Institut National Polytechnique de Lorraine *École doctorale IAE + M*

Contribution de l'approche Multi-Bandes à la reconnaissance automatique de la parole

THÈSE

présentée et soutenue publiquement le 20 septembre 1999

pour l'obtention du

Doctorat de l'Institut National Polytechnique de Lorraine
(Spécialité Informatique)

par

Christophe CERISARA

Composition du jury

Rapporteurs externes :	Henri Leich, Professeur, Université de Mons, Belgique
	Gérard Chollet, Directeur de Recherche CNRS, ENST, Paris
Rapporteur interne :	René Schott, Professeur, Université Henri Poincaré, Nancy I
Examinateurs :	Guy Pérennou, Professeur, Université Paul Sabatier, Toulouse
Directeurs de thèse :	Jean-Paul Haton, Professeur à l'Institut Universitaire de France
	Dominique Fohr, Chargé de Recherche CNRS, LORIA, Nancy

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Résumé

Le travail de recherche présenté dans ce mémoire réalise l'étude d'une nouvelle architecture pour les systèmes de reconnaissance automatique de la parole. Cette architecture est basée sur un découpage du spectre du signal en plusieurs zones fréquentielles, éventuellement recouvrantes, et sur un traitement indépendant de chacune de ces « bandes ». Celles-ci sont ensuite recombinées afin de fournir une réponse unique au problème de la reconnaissance. L'utilisation de cette méthode dite « Multi-Bandes » est à l'origine motivée par les travaux du psycho-acousticien H. Fletcher, travaux qui ont été récemment reconsidérés par J. B. Allen et qui aboutissent à un modèle de l'audition humaine proche de ce principe. L'application de celui-ci à une tâche de reconnaissance automatique de la parole est généralement motivée par sa robustesse aux bruits limités fréquentiellement. Nous montrons dans ce mémoire qu'il possède d'autres avantages, moins attendus, comme la résistance à tous les types de bruits stationnaires, et qu'il peut également surpasser le système de référence dans des environnements non bruités.

Ce mémoire commence par présenter les travaux existant dans le domaine du « Multi-Bandes », puis réalise une étude préliminaire du comportement de chacune des bandes. Ensuite, le problème du choix du module de recombinaison est posé, et plusieurs solutions sont proposées et testées. De même, nous montrons qu'il n'est pas possible d'utiliser l'algorithme classique de Viterbi en reconnaissance continue lorsque les bandes sont asynchrones, et différentes autres possibilités sont étudiées. Nous proposons ainsi deux algorithmes permettant de recombiner les bandes soit en fin de phrase, soit après des segments temporels associés à des unités de parole, comme les phonèmes.

Nous élaborons ensuite une architecture possible pour un système Multi-Bandes, architecture caractérisée notamment par l'utilisation conjointe du spectre complet et des sous-bandes, ainsi que par l'apprentissage du système réalisé grâce au critère de Minimisation de l'Erreur de Classification. Les tests de ce système sont réalisés en milieu plus ou moins bruité, avec différents types de bruit, et pour plusieurs tâches de traitement de la parole. Nous montrons ainsi que, selon l'environnement considéré, certains composants du système Multi-Bandes doivent être modifiés, comme par exemple l'importance accordée au spectre complet. Finalement, le dernier chapitre de ce mémoire est consacré à un nouvel algorithme d'apprentissage qui permet d'approcher l'optimum global du système. Cet algorithme a également la propriété de modifier les modèles dans chaque bande afin qu'ils soient réellement représentatifs de l'information contenue dans celles-ci, ce qui n'est pas le cas des phonèmes qui sont classiquement utilisés.

Mots-clefs : Reconnaissance Automatique de la Parole, Multi-Bandes, Reconnaissance Robuste de la Parole, Combinaison de Classifieurs, Modèles Stochastiques, Apprentissage.

Abstract

The subject of this dissertation concerns the study of a new architecture for automatic speech recognition systems. This architecture is based on a spectral splitting of the speech signal into several frequency bands. These bands are then treated by independent classifiers, which results are finally recombined, in order to obtain a unique answer to the recognition problem. This « Multi-Bands » method is originally motivated by the work of the psycho-acoustician H. Fletcher, which has been more recently reviewed by J. B. Allen. The main advantage of such a model is certainly its robustness to frequency-limited noise. However, we demonstrate in this thesis that it has several other benefits, such as the robustness to every kinds of stationary noise, or better performances than the full-band system even in clean speech.

This dissertation begins to describe the state-of-the art works about Multi-Bands speech recognition, and realizes a preliminary study of each band, taken individually. The problem of the choice of the recombination module is then considered, and several solutions are proposed and tested. It is also demonstrated that the Viterbi algorithm can not be used when the bands are not synchronous, and several other algorithms are studied to replace it. We have thus imagined two new methods to recombine the bands at the end of a sentence, and after shorter temporal segments which are associated with speech units, such as phones.

A Multi-Bands system is then built with respect to our theoretical results and our previous experiments. It is thus characterized, for example, by the simultaneous use of the full-band and the sub-bands, and by its training with the Minimum Classification Error (MCE) criterion. Several tests have been realized with this system, either in noisy or in clean conditions, with different kinds of noise and for different speech processing tasks. Finally, the last chapter is dedicated to a new training algorithm which allows the system to reach its global optimum. Moreover, it has the property to modify the HMM in the bands so that they model the real acoustic information which is present in the selected frequency region. This is usually not the case for the phonemes, which are classically used.

Keywords : Automatic Speech Recognition, Multi-Bands, Robust Speech Recognition, Classifiers Combination, Stochastic Models, Training.

Remerciements

Je voudrais tout d’abord remercier chaleureusement Jean-Paul Haton, sans qui cette thèse n'aurait jamais eu lieu. C'est effectivement grâce à lui que j'ai pu revenir en Lorraine et me consacrer à la recherche et à l'enseignement au LORIA. C'est encore grâce à ses conseils et aux nouvelles idées et orientations qu'il m'a souvent proposés au cours de nos réunions scientifiques, que j'ai pu mener à bien et dans les meilleures conditions mon travail. Ses remarques ont toujours été judicieuses et riches en enseignements, tout comme celles de Dominique Fohr, qui n'a jamais hésité à passer de longs après-midi avec moi pour revoir telle idée ou pour résoudre tel autre problème. Je suis donc extrêmement reconnaissant envers mes deux encadrants.

Je souhaite également remercier vivement Gérard Chollet, Directeur de recherche à l'ENST, Henri Leich, Professeur à l'Université de Mons, et René Schott, Professeur à l'Université Henri Poincaré, d'avoir accepté d'être les rapporteurs de cette thèse, et d'avoir consacré une partie de leur temps précieux à relire ce mémoire. Je remercie également Guy Pérennou, Professeur à l'Université Paul Sabatier à Toulouse, d'avoir accepté de siéger dans le jury et de s'être intéressé à mon travail de recherche.

Je tiens également à remercier Anne Bonneau et Irina Illina pour leurs judicieux conseils et pour le temps qu'elles ont consacré à lire ce document. De même, je remercie Yves Laprie, Frédéric Alexandre et Jean-François Mari pour leurs conseils et leur amitié. Tous mes remerciements vont également à Agnès, mon épouse, pour son aide dans la rédaction, son soutien sans failles, sa patience et son amour qui m'ont apporté le bonheur et qui m'ont permis de travailler dans les meilleures conditions possibles.

Finalement, je voudrais remercier tous mes collègues et amis, Arnaud, Jean, Imed, Laurent, Hervé, Nicolas et Yann, pour leur aide et leur soutien.

À Agnès.

Table des matières

Liste des figures....13
Liste des tableaux....15
Chapitre 1
Introduction....17
1.1. Objectifs de la thèse....17
1.2. Présentation pratique du mémoire....17
1.3. Présentation de la Reconnaissance Automatique de la Parole....18
Chapitre 2
Le « Multi-Bandes » en Reconnaissance Automatique de la Parole....21
2.1. Motivations....21
2.1.1. Motivations Psycho-acoustiques....21
2.1.2. Autres motivations....24
2.1.3. Les problèmes posés par le Multi-Bandes....25
2.2. Principe....27
2.3. État de l'art des domaines proches du Multi-Bandes....27
2.4. État de l'art du Multi-Bandes....29
2.4.1. Un nouveau problème : l'asynchronisme entre les bandes....29
2.4.2. Recombinaison trame par trame : état de l'art....30
2.4.3. Recombinaison après des segments de plusieurs trames : état de l'art....32
Chapitre 3
Contributions pour la reconnaissance Multi-Bandes en mode continu....36
3.1. Introduction....36
3.2. Impossibilité d'utiliser l'algorithme de Viterbi avec le Multi-Bandes....36
3.2.1. Présentation de l'algorithme de Viterbi....36
3.2.2. Problème posé par l'algorithme Viterbi....37
3.3. Un nouvel algorithme pour la recombinaison en fin de phrase....38
3.3.1. Présentation théorique....38
3.3.2. Expériences....40
3.4. Conception d'un algorithme pour la recombinaison après chaque phonème....43
3.4.1. Rappels sur les algorithmes de programmation dynamique....43
3.4.2. Description de notre algorithme....43
3.4.3. Complexité de l'algorithme....47
3.4.4. Comparaison avec les autres algorithmes de programmation dynamique....48
Chapitre 4
Étude des bandes de fréquence....50
4.1. Introduction....50
4.2. Présentation des reconnaisseurs....50
4.2.1. Définition des reconnaisseurs....50
4.2.2. Brève présentation des HMM du second ordre....51
4.3. Caractérisation des bandes de fréquence....53
4.3.1. État de l'art sur l'étude des limites des bandes....53
4.3.2. Définition des limites des bandes dans notre système....54
4.3.3. État de l'art sur l'étude des paramètres acoustiques....54

4.4. Étude des bandes de notre système....55
4.4.1. Introduction....55
4.4.2. Présentation des bandes de notre système....55
4.4.3. Étude phonétique....57
Chapitre 5
La recombinaison 63
5.1. État de l'art sur la fusion d'information....63
5.1.1. Présentation des différents types de recombinaison....63
5.1.2. Recombinaison sur les étiquettes des classes gagnantes....64
5.1.3. Recombinaison sur les rangs....66
5.1.4. Recombinaison sur les scores....67
5.1.5. Première conclusion....68
5.2. Notre étude sur la recombinaison linéaire....68
5.2.1. Définition....68
5.2.2. Étude théorique....70
5.2.3. Recombinaison empirique....72
5.2.4. L'algorithme de Minimisation de l'Erreur de Classification....73
5.3. Présentation de la recombinaison neuronale utilisée dans notre système....75
Chapitre 6
Expérimentations 77
6.1. Introduction....77
6.2. État de l'art des différentes méthodes utilisées en RAP robuste....78
6.3. Étude dans du bruit « artificiel »....79
6.3.1. Introduction, Résumé de notre système et conditions expérimentales....79
6.3.2. Étude dans le bruit limité fréquentiellement....83
6.3.3. Étude dans le bruit blanc....91
6.3.4. Conclusion....95
6.4. Étude avec du bruit naturel....96
6.4.1. Bruit d'automobile....96
6.4.2. Bruit de cantine....99
6.4.3. Conclusion....103
6.5. Apprentissage du module de recombinaison dans le bruit....103
6.5.1. Introduction....103
6.5.2. Expérimentations....104
6.5.3. Étude du perceptron après cet apprentissage....106
6.6. Étude dans les milieux extrêmement bruités....109
6.7. Identification du langage....110
6.7.1. Introduction 110
6.7.2. Adaptation du système à cette tâche....111
6.7.3. Étude des classifieurs individuels et comparaison avec d'autres résultats expérimentaux 112
6.7.4. Résultats des systèmes Multi-Bandes....113
6.8. Expériences en milieu non bruité....114
6.8.1. Introduction....114
6.8.2. Résultats des expériences....115
6.8.3. Commentaires....116
6.9. Passage au mode continu....118
6.9.1. Introduction....118

6.9.2. Utilisation d'un étage de pré-segmentation....119
6.9.3. Algorithme de programmation dynamique....120
6.9.4. Complexité des algorithmes....124
6.9.5. Comparaison avec les autres algorithmes de la littérature....125
6.10. Conclusions du chapitre....126
6.10.1. Robustesse du Multi-Bandes en milieu bruité : une explication possible....126
6.10.2. Comparaison entre les différents modules de recombinaison testés....127
6.10.3. Comparaison avec d'autres systèmes robustes....127
6.10.4. Expériences en milieu non bruité et passage au mode continu....128
Chapitre 7
Apprentissage global....129
7.1. Motivations....129
7.2. Choix du critère d'optimisation....131
7.3. Principe....131
7.3.1. Modification des HMM....131
7.3.2. Initialisation des paramètres....134
7.4. Résumé de l'algorithme....134
7.5. Résultats expérimentaux....136
7.5.1. Expériences dans un environnement non bruité....136
7.5.2. Expériences dans un environnement bruité....137
7.5.3. Apprentissage dans un environnement bruité....137
7.6. Étude des nouvelles classes phonétiques....139
7.6.1. Introduction....139
7.6.2. Principe....139
7.6.3. Résultats expérimentaux....143
7.7. Conclusion....145
Chapitre 8
Conclusions et Perspectives....146
8.1. Objectifs initiaux et conclusions de notre étude....146
8.1.1. Résumé des points principaux du mémoire....146
8.1.2. Originalité de notre travail....147
8.2. Perspectives....148
8.2.1. Perspectives à court terme....148
8.2.2. Perspectives à plus long terme....149
Bibliographie....150
Annexe 1
Résultats chiffrés de toutes les expériences présentées dans le mémoire....157
Annexe 2
Liste des phonèmes utilisés....161
Annexe 3
Étude individuelle du comportement phonétique des sous-bandes....163
Annexe 4
Étude individuelle de la modification des classes dans chaque bande pendant l'apprentissage global
....168

Liste des figures

Figure 2.1 : Principe général d'un système Multi-Bandes 27

Figure 2.2 : Exemple d'associations trames-états pour deux bandes. 30

Figure 2.3 : Exemple de HMM résultant de la méthode de recombinaison des HMM 34

Figure 3.2 : Segmentation des 4 sous-bandes pour la phrase « C'est ainsi que Jacques fut arrêté». De bas en haut : la réponse des bandes [0 ... 538 Hz], [461 ... 1000 Hz], [923 ... 2823 Hz] et [2374 ... 7983 Hz], la segmentation manuelle puis le signal de parole. 38

Figure 3.3 : Début du graphe associé à la figure 3.4 41

Figure 3.4 : Exemple d'alignement réalisé par l'algorithme sur les réponses 41

Figure 3.1 : Exemple du graphe utilisé par l'algorithme de Viterbi en mode continu. Une transition entre deux HMM est présentée. 37

Figure 4.1 : Topologie d'un HMM. 51

Figure 4.2 : Filtres utilisés pour l'analyse en cepstres et répartition de ces filtres dans les bandes 56

Figure 4.3 : Taux de reconnaissance comparés de la première sous-bande et du spectre complet pour chaque phonème 58

Figure 4.4 : Différences des taux de reconnaissance du spectre complet et de la première sous-bande pour chaque phonème. 60

Figure 5.1 : Représentation de la recombinaison linéaire 69

Figure 5.2 : Utilisation d'un perceptron en recombinaison 76

Figure 6.1 : Schéma de notre système Multi-Bandes 79

Figure 6.2 : Spectrogramme du bruit filtré aigu 84

Figure 6.3 : Spectrogramme du bruit filtré grave 84

Figure 6.4 : Taux de reconnaissance des systèmes linéaires sans apprentissage du module de recombinaison dans le bruit filtré aigu. 87

Figure 6.5 : Taux de reconnaissance des systèmes linéaires sans apprentissage du module de recombinaison dans le bruit filtré grave. 87

Figure 6.6 : Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans le bruit filtré aigu 89

Figure 6.7 : Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans le bruit filtré grave 90

Figure 6.8 : Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans du bruit blanc. 91

Figure 6.9 : Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans du bruit blanc. 92

Figure 6.10 : Spectrogramme du bruit de voiture 96

Figure 6.11 : Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans le bruit de voiture. 97

Figure 6.12 : Taux de reconnaissance, dans le bruit de voiture, des systèmes Multi-Bandes avec apprentissage du module de recombinaison 98

Figure 6.13 : Spectrogramme du bruit de cantine 100

Figure 6.14 : Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans le bruit de cantine. 101

Figure 6.15 : Taux de reconnaissance, dans le bruit de cantine, des systèmes Multi-Bandes avec apprentissage du module de recombinaison. 102

Figure 6.16 : Spectrogramme du bruit de sèche-cheveux 105

Figure 6.17 : Modification de l'influence de la première sous-bande due à l'apprentissage du module de recombinaison dans le bruit blanc 108

Figure 6.18 : Modification de l'influence du spectre complet due à l'apprentissage du module de recombinaison dans le bruit blanc 108

Figure 6.19 : Schéma général du système Multi-Bandes adapté à l'identification automatique de la langue 112

Figure 6.20 : Courbe du taux de reconnaissance pour un système Multi-Bandes en fonction de l'importance accordée au spectre complet 117

Figure 7.1 : Courbe extraite de la vraisemblance de sortie d'un HMM en fonction d'un de ses paramètres. 133

Figure 7.2 : Évolution du taux de reconnaissance des systèmes Multi-Bandes en fonction du nombre d'itérations de l'apprentissage global 136

Figure 7.3 : Exemple de comportement d'une classe dans le cas 1. 141

Figure 7.4 : Exemple de comportement d'une classe dans le cas 3. 142

Figure 7.5 : Étude comparée de la modification de la taille des classes et de leur taux de reconnaissance pour la première sous-bande 143

Liste des tableaux

Tableau 6.1 : Intervalles de confiance sur la partie « coretest » de TIMIT en fonction du taux de reconnaissance 82

Tableau 6.2 : Comportement des bandes dans le bruit filtré aigu et grave, à 20 dB 85

Tableau 6.3 : Taux de reconnaissance du Multi-Bandes, avec apprentissage du module de recombinaison dans le bruit blanc, dans différents environnements 105

Tableau 6.4 : Taux de reconnaissances de quelques systèmes dans le bruit blanc à 0 dB. 109

Tableau 6.5 : Taux d'erreur pour les classifieurs considérés individuellement en fonction du nombre de centroïdes par langue. 112

Tableau 6.6 : Taux d'erreur pour les systèmes Multi-Bandes sans apprentissage du module de recombinaison 113

Tableau 6.7 : Taux d'erreur pour les systèmes Multi-Bandes avec apprentissage de la recombinaison (8 centroïdes) 114

Tableau 6.8 : Taux de reconnaissance de tous le systèmes dans un environnement non bruité 116

Tableau 6.9 : Taux de reconnaissance du système Multi-Bandes 0,1*4+0,6 en mode continu basé sur la segmentation donnée par le système de référence 119

Tableau 6.10 : Taux de reconnaissance du système Multi-Bandes 0,05*4+0,8 en mode continu utilisant l'algorithme de programmation dynamique. 120

Tableau 6.11 : Taux de reconnaissance du système Multi-Bandes calculant la moyenne des cinq bandes en mode continu en présence d'un bruit de voiture à -10 dB. L'algorithme d'alignement utilisé est celui dérivé du two-level et du Level Building, auquel des contraintes de durées sur le silence ont été ajoutées. 121

Tableau 6.12 : Taux de reconnaissance, en mode continu et dans le bruit blanc, du système Multi-Bandes dont la recombinaison de segmentation est 0,8+0,05*4 et celle de sélection est 0,2*5 123

Tableau 7.1 : Taux de reconnaissance du système Multi-Bandes avec recombinaison par PMC dans plusieurs environnements bruités après quatre itérations d'apprentissage global. 137

Tableau 7.2 : Taux de reconnaissance du système Multi-Bandes avec recombinaison par PMC avant et après un apprentissage « classique » (i.e. indépendant) dans un milieu bruité. 138

Tableau 7.3 : Taux de reconnaissance du système Multi-Bandes avec recombinaison par PMC avant et après un apprentissage global dans un milieu bruité. 138

Chapitre 1
Introduction

1.1. Objectifs de la thèse

Le cœur de notre travail consiste à étudier et à évaluer certaines « contributions d'approches Multi-Bandes à la reconnaissance automatique de la parole ». Il s'agit donc avant tout de réfléchir au nouveau principe de modélisation acoustique de la parole qu'est le principe Multi-Bandes. Nous avons ainsi tenté de comprendre en profondeur les raisons qui ont fait émerger ce nouveau paradigme, mais aussi et surtout les conséquences que celui-ci a sur la conception de nouvelles méthodes d'extraction d'indices acoustiques pertinents du signal de parole. Ces conséquences sont en fait loin d'être aussi simples qu'il n'y paraît au premier abord, comme nous le verrons au cours de ce mémoire.

De plus, dans un domaine de recherche appliqué comme l'est celui de la reconnaissance de la parole, il ne suffit pas de comprendre, mais il faut aussi concevoir, puis appliquer ces principes et les mettre en œuvre dans des systèmes opérationnels. C'est pourquoi nous avons consacré près de la moitié de ce mémoire à l'implémentation de plusieurs algorithmes, ainsi qu'aux tests réalisés avec ceux-ci et à leurs analyses. L'objectif ultime est de réaliser un système complet, fonctionnel et efficace de reconnaissance de la parole basé sur le Multi-Bandes, mais il s'agit là d'un travail considérable, qui demanderait à lui seul de nombreux mois consacrés notamment à la conception d'algorithmes heuristiques et à l'optimisation du code-source du système. Nous n'espérons pas bien entendu atteindre cet objectif, mais nous avons tout de même abordé un certain nombre de problèmes liés à l'implémentation d'un tel système, et avons proposé plusieurs solutions à ces problèmes.

1.2. Présentation pratique du mémoire

Le paradigme Multi-Bandes qui est présenté dans ce mémoire est encore relativement nouveau. Il n'est en effet activement développé que depuis quatre ou cinq ans, et il n'existe donc pas encore de consensus sur l'architecture que doit posséder un système Multi-Bandes de référence. Le champ de recherche de ce domaine est très vaste et n'est encore qu'insuffisamment exploré, de sorte que chaque rapport ou compte-rendu de travaux étudiant ce principe aborde le problème sous un angle différent des autres. À travers ce mémoire, nous avons voulu apporter également de nouvelles considérations sur ce domaine, car ce foisonnement d'idées et de tentatives est très certainement profitable au développement dudit système de référence. C'est pourquoi nous avons essayé de mettre l'accent, dans les chapitres qui vont suivre, sur les points qui nous paraissent cruciaux, et de passer rapidement sur ceux qui ont déjà été traités par ailleurs. De nombreuses références permettent ainsi au lecteur intéressé de compléter cette présentation et éventuellement d'approfondir plus en détail certains points spécifiques.

Le mémoire est constitué de huit chapitres. Chacun d'entre eux traite un aspect directement ou indirectement lié au Multi-Bandes, et leur succession respecte un ordre que nous avons voulu progressif dans la difficulté. Les premiers chapitres sont consacrés à une présentation assez générale des domaines de recherche dans lesquels s'inscrit le Multi-Bandes, puis à une réflexion sur son principe qui mêle exposés théoriques et explications intuitives. Les derniers chapitres sont consacrés aux expériences réalisées avec notre système. Chacune de ces expériences est précédée d'une brève présentation expliquant notamment les objectifs visés, les résultats obtenus et un certain nombre de commentaires relatifs à ceux-ci.

Nous n'avons pas rédigé un unique chapitre où serait centralisé l'ensemble de l'état de l'art, car cela s'est avéré impossible à réaliser, vu la diversité des domaines de recherche concernés par le Multi-Bandes. Un tel chapitre aurait dû présenter de nombreuses notions qui ne sont pas a priori sémantiquement liées entre elles. Nous avons donc préféré, au début des chapitres 2 et 5, réaliser un état de l'art des domaines qui concernent directement la notion développée dans ces chapitres. Ce choix est risqué dans le sens où il est ainsi plus facile de confondre ce qui relève de l'état de l'art et de l'apport de notre travail. Nous avons donc tenté de rendre cette distinction la plus claire possible, aussi bien dans l'organisation des paragraphes que dans l'indication explicite des parties consacrées à l'état de l'art et de celles liées à notre travail. Dans le cas où la confusion serait tout de même possible, nous avons rappelé dans la conclusion l'originalité de notre travail par rapport aux travaux préexistants dans le domaine du Multi-Bandes.

1.3. Présentation de la Reconnaissance Automatique de la Parole

Ce mémoire traite essentiellement de la reconnaissance automatique de la parole, même si une incursion est faite en identification du langage dans la partie 6.7. Nous allons donc tout d'abord présenter brièvement ce qui se cache derrière ces termes et quels sont les autres domaines de recherche concernés.

La reconnaissance automatique de la parole (RAP) est une partie d'un domaine de recherche plus vaste que l'on nomme habituellement le traitement automatique du langage naturel (TALN). De nombreux autres domaines de recherche font également partie du TALN : nous pouvons ainsi citer la reconnaissance automatique du locuteur, l'identification du langage, la synthèse de la parole, le codage et la compression de la parole, mais aussi les systèmes de dialogue, l'indexation de documents, la modélisation sémantique des textes, etc.

La reconnaissance automatique de la parole emprunte aussi largement à un autre domaine de l'intelligence artificielle qui est la reconnaissance des formes. Ce vaste domaine comprend également la reconnaissance des caractères (manuscrits ou non) et l'analyse des images ou des scènes visuelles. De nombreuses techniques sont partagées par la reconnaissance de la parole et la reconnaissance des images, et nous citerons dans la suite plusieurs articles issus de ce domaine.

Tous ces domaines sont très actifs depuis de nombreuses années, mais les plus connus sont certainement la reconnaissance automatique de la parole, la reconnaissance du locuteur, la synthèse de la parole et la compression de la parole. Ceci s'explique aisément, car ce sont ces mêmes domaines qui offrent les applications les plus facilement perceptibles du grand public. Ainsi, la compression de la parole est abondamment utilisée dans les téléphones portables, dont le développement n'est plus aujourd'hui à remettre en cause, mais aussi dans les transmissions multimédias à travers les réseaux ou dans les disques optiques les plus récents (DVD). De même, la synthèse de la parole commence aujourd'hui à équiper tous les ordinateurs, mais aussi les téléphones portables, les voitures et de nombreux équipements électroménagers. Les applications de la reconnaissance du locuteur sont peut-être moins évidentes, car moins développées, mais une polémique commence notamment à s'élever concernant certaines d'entre elles. Ainsi en est-il des systèmes de sécurité qui reconnaissent un utilisateur référencé par sa voix, ou plus récemment de la validité juridique des « preuves » de la culpabilité d'un suspect obtenues par une telle analyse.

Quant à la reconnaissance automatique de la parole elle-même, ses applications sont nombreuses. Elles peuvent être grossièrement classées dans quatre domaines :

1. ***Les applications de dictée vocale*** : Ces produits sont très nombreux actuellement sur le marché. Nous pouvons par exemple citer ViaVoice d'IBM, Naturally Speaking de Dragon Systems ou encore WINSAPI de Microsoft. Toutefois, si de nouvelles versions de ces logiciels apparaissent très fréquemment, cela est sûrement dû à leur manque actuel de robustesse. Ils font encore malheureusement trop d'erreurs pour concurrencer sérieusement quelqu'un habitué à manipuler le clavier, et surtout leurs performances se dégradent beaucoup trop lorsque l'environnement est quelque peu bruité.

2. ***La télématique vocale*** : La plupart des grandes compagnies de télécommunications se tournent actuellement vers ce domaine afin de remplacer leurs opérateurs humains par des systèmes automatiques. Leurs rôles restent pour l'instant très simple, par exemple donner le numéro de téléphone d'un abonné. Ceci est cependant déjà très intéressant pour ces compagnies, car le coût horaire d'un opérateur humain est très élevé comparé à celui d'un logiciel, et des économies substantielles peuvent être réalisées de la sorte. De plus, la qualité du service de renseignement téléphonique peut être amélioré car le même logiciel peut facilement gérer plusieurs appels en même temps, et l'attente pour les usagers est donc moins grande.

3. ***Les commandes vocales*** : Ces systèmes, plus simples, sont beaucoup plus robustes et donc plus utiles d'une certaine manière, même si leur discrétion ne leur a pas valu la même renommée que les précédents. Ils sont ainsi particulièrement utiles, voire nécessaires, pour des applications dites « mains libres », i.e. pour lesquelles l'utilisateur ne peut utiliser ses mains afin de donner des ordres à un système. Ceci arrive, par exemple, lorsqu'il s'agit de conduire un véhicule, de réaliser une opération chirurgicale ou tout simplement lorsque l'utilisateur est dans l'incapacité d'utiliser ses mains à cause d'un handicap.

4. ***L'aide aux malentendants*** : Ces applications répondant à un réel besoin, de nombreux projets sont actuellement en cours ou sont déjà terminés pour réaliser de nouvelles prothèses auditives basées sur la reconnaissance de la parole, ou encore des systèmes d'apprentissage de la langue pour les malentendants.

La Reconnaissance Automatique de la Parole est donc très utile dans de nombreux domaines. C'est sans doute pour cette raison qu'un effort financier et humain très important a été consacré à cette recherche au cours de ces dernières années. Aujourd'hui, certains laboratoires s'orientent déjà vers la compréhension du langage naturel du point de vue sémantique, avec tous les problèmes de modélisation du monde que cela pose. Toutefois, nous pensons qu'il ne faut pas pour autant délaisser le niveau acoustico-phonétique en recherche, et ce pour plusieurs raisons : tout d'abord, il est évident que de nombreux progrès sont encore à réaliser dans ce domaine, notamment au vu des résultats actuellement trop médiocres des systèmes de reconnaissance en milieu bruité. De plus, nous avons encore beaucoup à apprendre sur le modèle auditif humain, et la plupart des études réalisées sur l'audition humaine montrent qu'elle est fondamentalement très différente de ce que nos modèles de reconnaissance proposent. C'est pourquoi la recherche doit encore expérimenter de nouveaux modèles acoustiques, comme nous essayons de le faire dans ce mémoire en nous inspirant autant que possible du modèle humain. Plusieurs références à des articles de psycho-acoustique sont d'ailleurs données afin de conserver un lien avec ce domaine.

Chapitre 2
Le « Multi-Bandes » en Reconnaissance Automatique de la Parole

Nous expliquons tout d'abord dans ce chapitre les raisons qui nous ont poussés à étudier la reconnaissance de la parole par la méthode dite Multi-Bandes, puis nous présentons les principes généraux de celle-ci. La partie 2.3 réalise un état de l'art des domaines proches du Multi-Bandes, c'est-à-dire des recherches en reconnaissance de la parole qui possèdent de nombreux points communs avec le Multi-Bandes, mais diffèrent également de celui-ci sur certains points. La partie 2.4 pose un certain nombre de problèmes généraux liés au Multi-Bandes, et expose les différentes solutions qui ont été élaborées dans les systèmes Multi-Bandes développés par d'autres laboratoires. Il s'agit donc en fait d'un état de l'art des travaux spécifiques à ces problèmes.

2.1. Motivations

Pourquoi étudier un nouveau modèle phonétique comme le Multi-Bandes alors que les systèmes actuels annoncent des taux de reconnaissances de l'ordre de 90 % et plus ? Comme nous l'avons déjà signalé en introduction, de tels taux de reconnaissance descendent très vite en dessous de 50 % lorsque les conditions d'utilisation ne sont pas idéales et de plus, ces systèmes utilisent des modèles phonétiques qui restent bien trop éloignés de l'audition humaine pour être pleinement satisfaisants. Le modèle que nous proposons ici est donc largement motivé par des considérations psycho-acoustiques que nous développons dans la partie 2.1.1. Les autres motivations sont exposées dans la partie 2.1.2.

2.1.1. Motivations Psycho-acoustiques

La principale motivation pour le paradigme Multi-Bandes est une motivation d'ordre psycho-acoustique. En effet, l'idée d'appliquer ce principe à la reconnaissance automatique de la parole provient initialement de l'article publié par Jont B. Allen en 1994. Cet article retrace les travaux que Harvey Fletcher a réalisés dans les années 50 concernant l'audition humaine. Le cœur de ces travaux, qui est repris dans l'article de Allen, concerne le traitement de l'information auditive par des canaux fréquentiels indépendants. Nous allons présenter dans ce paragraphe un résumé succinct des quelques points de l'article de Allen qui concernent directement notre travail. Les lecteurs intéressés par une présentation plus détaillée de cet article peuvent se reporter à [Mirghafori99] ou à l'article lui-même [Allen94].

Le modèle auditif humain développé par Fletcher est décomposé en cinq étages. Le premier de ces étages, qui correspond à la cochlée, décompose le signal en bandes de fréquences indépendantes, chacune d'entre elles codant le Rapport Signal sur Bruit (RSB) présent dans la bande. Le second calcule un taux de reconnaissance phonétique partiel tandis que les deux derniers forment les syllabes puis les mots. Pour ses expériences, Fletcher a distingué deux indices de mesure : l'articulation, qui mesure la probabilité d'identifier correctement des unités de parole dénuées de sens, et l'intelligibilité qui mesure la probabilité d'identifier correctement des unités de parole sémantiquement significatives, comme les mots. Il a ainsi clairement démontré l'importance du contexte pour la reconnaissance phonétique, ce qui l'a amené à réaliser une grande partie de ses expériences sur des phonèmes dénués de sens, afin d'éliminer l'influence de ce contexte. Ceci explique également pourquoi la plupart de nos expériences ont été réalisées sur des phonèmes hors-contextes, le paradigme Multi-Bandes développé dans la thèse étudiant précisément un modèle de reconnaissance phonétique qui se situe avant la prise en compte du contexte et des informations linguistiques. Ainsi, selon Fletcher, des indices de reconnaissance sont calculés indépendamment dans chaque bande fréquentielle dans le deuxième étage de son modèle. Ces indices sont ensuite recombinés de sorte à optimiser l'erreur phonétique finale. En effet, Fletcher a montré, en s'appuyant sur la définition de l'articulation et sur certaines de ses expériences, que le taux d'erreur après recombinaison est égal au produit des taux d'erreur dans chaque bande. Les modules de recombinaison que nous avons développés pour notre système sont bien entendu très loin d'être aussi performants. Mais nous ne pouvons actuellement qu'espérer comprendre comment une telle recombinaison est réalisée dans notre cerveau, et tenter de l'approcher au mieux.

D'autres travaux, qui ne sont pas directement liés à l'aspect Multi-Bandes ont également indirectement contribué à mettre en évidence l'intérêt de cette approche. Ainsi, Arai & al. ont montré la robustesse inhérente du système auditif humain lorsque différentes bandes de fréquences sont légèrement désynchronisées [Greenberg98a]. Ceci corrobore l'hypothèse selon laquelle chacune de ces bandes est traitée indépendamment dans notre cerveau. En effet, si un tel découpage n'avait pas lieu, toute l'information présente à un instant donné serait traitée comme une seule et même entité, entité qui serait alors composée de parties appartenant en réalité à des instants différents, et donc en fait à d'autres entités ! Il serait alors difficile de décider de la véritable nature du signal. Cette expérience montre qu'il est probable que le cerveau traite ces bandes fréquentielles séparément.

Steeve Greenberg a également suggéré à plusieurs reprises [Greenberg98b] que la robustesse du système auditif humain est en grande partie due au fait que celui-ci utilise l'information provenant d'une multitude de « canaux » parmi lesquels figurent les bandes fréquentielles. Ainsi, l'extrême redondance de l'information acoustique que présentent tous ces canaux se retrouve également dans les processus de décision qui leur sont associés. Lorsque le choix final doit être réalisé, il est alors facile de détecter les canaux dont les réponses sont altérées par du bruit. Cette explication est attrayante, mais nous ne pouvons malheureusement pas encore l'appliquer à la reconnaissance automatique de la parole, car nos systèmes ne disposent pas de toutes les sources d'information dont dispose notre cerveau. Et même si nous pouvions avoir accès à toutes ces informations, il faudrait un temps énorme à nos systèmes pour pouvoir exploiter autant d'information à la fois ! C'est pourquoi nous ne pouvons que nous limiter pour l'instant à l'information provenant de quelques bandes fréquentielles.

Cette idée de redondance apparaît également dans la thèse de Laurent Besacier [Besacier98a] qui a appliqué le principe du Multi-Bandes à la reconnaissance automatique du locuteur. Il a ainsi montré que les résultats de la recombinaison sont meilleurs lorsque les bandes fréquentielles se recouvrent que lorsqu'elles sont disjointes. Ceci conforte l'idée selon laquelle plus il y a de redondance entre les canaux, meilleure est la recombinaison.

D'autres travaux de psycho-acoustiques font intervenir une division du domaine fréquentiel. Ainsi, Oded Ghitza a montré que les traits acoustiques, comme la nasalisation[1] ou le voisement[2], sont perçus par notre cerveau dans différentes bandes de fréquences [Ghitza94]. L'expérience qu'il a menée consiste ainsi à diviser le signal de parole des syllabes CVC (consonne-voyelle-consonne) en quatre parties selon l'axe temporel et en trois parties selon l'axe fréquentiel. Il a ensuite choisi deux syllabes différant entre elles d'un unique trait acoustique, et a ensuite échangé certains de ces « pavés » entre les deux syllabes. Le but est donc de comprendre dans quelle zone temporelle et fréquentielle est perçu ce trait. Il a ainsi montré, par exemple, que la nasalisation est essentiellement perçue dans les basses fréquences.

L'idée selon laquelle les traits acoustiques sont perçus et décodés dans différentes bandes de fréquences apparaît également dans les travaux de Miller et Nicely [Miller55]. Cette idée est particulièrement intéressante car elle invalide dans une certaine mesure l'hypothèse qui a été classiquement utilisée dans les systèmes Multi-Bandes et qui postule que chaque bande fréquentielle doit reconnaître des phonèmes : il ne semble pas que ce soit le cas en réalité, et d'autres classes phonétiques, ou encore mieux des traits acoustiques, seraient certainement plus adaptés au paradigme Multi-Bandes. Nous en reparlerons au chapitre 7.

Les travaux de Miller et Nicely, de Lippmann [Lippmann97a] et de French et Steinberg [French47], ont également montré que l'audition humaine est particulièrement robuste au filtrage fréquentiel. Ainsi, même lorsque le signal est largement filtré dans les fréquences moyennes, ne laissant subsister que les très basses et les très hautes fréquences, l'intelligibilité du signal reste exceptionnellement bonne. Encore une fois, ceci montre que le cerveau humain sait tirer parti de toute l'information contenue dans une bande fréquentielle, même étroite. De plus, il ne traite pas l'information fréquentielle « en bloc », comme la plupart des systèmes de reconnaissance automatique actuels, mais il sait au contraire éliminer de son processus décisionnel les zones fréquentielles non pertinentes. Enfin, il semblerait qu'une information acoustique différente et complémentaire soit effectivement portée par les hautes et les basses fréquences, car si l'une de ces deux bandes vient à manquer, les résultats sont nettement moins bons.

Dans le même ordre d'idée, nous pouvons citer les travaux de Mokhtari et Clermont [Mokhtari96] qui ont montré que la classification automatique des voyelles de l'anglais, dans le cas multi-locuteurs, est meilleure lorsque le signal est filtré en dessous de 1700 Hz. Ils expliquent cela, en s'appuyant également sur plusieurs autres expériences, par le fait que l'information phonétique est codée plutôt dans la partie basse du spectre, tandis que l'information acoustique relative au locuteur est plutôt contenue dans la partie haute du spectre.

1 La nasalisation distingue les consonnes dans "ma" et "pa"
2 Le voisement distingue les premières consonnes dans "ville" et "file"

2.1.2. Autres motivations

Toutefois, les motivations du paradigme Multi-Bandes ne sont pas seulement d'ordre psycho-acoustiques. Celles-ci sont particulièrement importantes, car ce sont elles qui ont engendré l'intérêt actuel grandissant porté au Multi-Bandes, mais il en existe également de nombreuses autres. Nous donnons ci-dessous quelques-unes des ces motivations, puis nous considérons dans la partie 2.1.3. les inconvénients du Multi-Bandes.

1. Le Multi-Bandes peut être robuste au bruit limité fréquentiellement. Cette hypothèse est certainement la première qui apparaît lorsqu'on commence à réfléchir aux conséquences du Multi-Bandes. En effet, elle est très intuitive, dans le sens où lorsqu'un bruit n'affecte qu'une région limitée du spectre, les paramètres acoustiques qui sont classiquement utilisés pour la reconnaissance, à savoir les cepstres, sont tous affectés par ce bruit. Ce qui signifie que la totalité du vecteur acoustique est corrompu par le bruit. Il est bien entendu possible d'utiliser d'autres coefficients qui ne dépendent pas de l'ensemble du spectre, comme les niveaux d'énergie d'un banc de filtres par exemple, auquel cas l'argument précédent ne tient plus. Malheureusement, même dans ce cas, les densités de probabilités qui sont calculées dans les états des Modèles de Markov Cachés (HMM) tiennent compte de chaque coefficient du vecteur acoustique[3]. Ainsi, les erreurs occasionnées par le bruit dégraderont toujours les taux de reconnaissance du système. Or, dans un système Multi-Bandes, chaque bande n'utilisant qu'une zone limitée du cepstre, seules quelques bandes seront affectées par le bruit, tandis que les autres proposeront toujours la solution correspondant à un signal non bruité. Bien entendu, rien n'est dit ici sur les méthodes qui permettent d'exploiter cet avantage, ni si celui-ci aboutit à un accroissement réel des performances ou s'il ne s'agit que d'une simple conjecture. La confirmation expérimentale de cette hypothèse est réalisée dans la partie 6.3.2.

2. Une autre motivation très proche de la première ne concerne pas les bruits additifs, mais plutôt certains bruits convolutifs. Ainsi, la réverbération est un phénomène qui, a priori, affecte beaucoup plus les basses fréquences que les hautes fréquences. En effet, intuitivement, un signal haute fréquence a une dynamique beaucoup plus rapide que celle relative à la réverbération, et la séparation du signal primaire et du signal réverbéré est donc plus facile, ce qui n'est pas le cas pour un signal basse fréquence. Cette disparité de comportement entre les hautes et basses fréquences face à un signal réverbéré a été notamment mise en évidence par Steeve Greenberg dans [Greenberg96].

3 Les lecteurs intéressés par un rappel sur les coefficients cepstraux ou les HMM peuvent se référer à [Rabiner78] ou à [Rabiner93].

3. D'autres motivations plus théoriques peuvent également être considérées. En effet, les modèles que nous utilisons dans un système Multi-Bandes appartiennent chacun à une bande de fréquences beaucoup plus étroite que celle qui est utilisée dans les systèmes classiques. Il n'est donc pas nécessaire d'utiliser autant de coefficients. En fait, les modèles des sous-bandes possèdent beaucoup moins de coefficients que les modèles classiques, ce qui diminue d'autant le phénomène traditionnellement désigné par le terme de « *dimensionnality curse* ». Ce phénomène correspond à une propriété bien connue des modèles qui stipule que plus la dimension de l'espace des paramètres de ces modèles est grande, plus grandes sont également les erreurs de modélisation et plus difficile est la convergence de l'algorithme d'apprentissage. Réduire la dimension des vecteurs acoustiques permet donc d'améliorer la qualité des modèles. Bien entendu, cet avantage n'est exploitable que si le module de recombinaison n'« ajoute » pas les erreurs des bandes les unes aux autres, auquel cas la modélisation finale est tout aussi erronée, voire plus, que ne l'est la modélisation utilisant l'ensemble des coefficients spectraux. Ceci montre par ailleurs l'importance fondamentale du module de recombinaison dans un système Multi-Bandes.

4. La dernière motivation dont nous allons parler ici est plutôt en fait un défaut des modèles de Markov classiques. En effet, ces modèles, grâce à leurs transitions « en boucle » qui reviennent sur l'état d'où elles sont parties, permettent de modéliser la dynamique temporelle du signal. Ceci constitue en fait leur avantage principal et a largement contribué à leur succès actuel. Malheureusement, nous pouvons déplorer que la dynamique de la seconde dimension du signal, à savoir la dimension spectrale, ne soit quant à elle pas du tout modélisée. Ainsi, les HMM ne peuvent considérer que des événements spectraux rigoureusement simultanés. Or, le signal acoustique possède très certainement une dynamique variable selon la fréquence considérée, comme l'indiquent aujourd'hui plusieurs études [Ghitza94]. Il serait donc préférable de modéliser également la dynamique du signal dans la dimension spectrale. Nous reparlerons de ce point dans la partie 2.3.

Nous n'avons présenté ici qu'un petit nombre de motivations qui, soit nous paraissaient importantes, soit étaient directement liées au travail que nous avons réalisé. Bien entendu, d'autres motivations existent, comme par exemple celles relatives à la théorie de la recombinaison de classifieurs [Tumer99]. Les lecteurs intéressés peuvent se référer à la thèse de Nikki Mirghafori [Mirghafori99] qui cite d'autres motivations et d'autres pointeurs.

2.1.3. Les problèmes posés par le Multi-Bandes

Malgré tous les avantages exposés ci-dessus, le Multi-Bandes est loin d'être la solution miracle au problème actuel du manque de fiabilité de la reconnaissance automatique de la parole. Nous allons donner ici un bref aperçu des quelques problèmes spécifiques au Multi-Bandes, sachant que ces problèmes seront ensuite détaillés et traités plus en détail dans le mémoire.

Tout d'abord, ce qui constitue l'avantage même du Multi-Bandes peut se révéler un obstacle insurmontable à première vue. Ainsi, le fait de diviser le spectre en plusieurs parties réduit considérablement la quantité d'information disponible dans chacune de ces bandes. En conséquence, comme nous l'avons déjà remarqué, même si la modélisation est meilleure, les taux de reconnaissance des phonèmes dans chaque bande sont bien plus bas que si tout le spectre est utilisé. Ceci est exact, comme les expériences du chapitre 4 le démontrent. Toutefois, la première réponse à ce problème réside dans la méthode de comparaison utilisée : il est vain d'essayer de comparer les taux de reconnaissance des phonèmes dans tout le spectre et dans une bande, car les phonèmes ne constituent pas l'unité de base dans une sous-bande. En effet, il est intuitivement évident que l'information contenue dans une bande ne permet pas de discriminer entre tous les phonèmes. Par exemple, la différence entre un /f/ et un /s/ n'est pas visible dans les basses fréquences. Si nous devons comparer des taux de reconnaissance, il nous faut modéliser l'information réellement présente dans chaque bande et calculer les taux de reconnaissance sur ces modèles. Bien entendu, il ne s'agit pas là d'un problème simple ; nous en reparlerons dans la partie 7.6. La seconde réponse à ce problème revient à considérer le Multi-Bandes comme une procédure de division des tâches entre chaque bande et le module de recombinaison. En effet, en réalisant un système Multi-Bandes, nous spécialisons en quelque sorte un certain nombre de classifieurs à résoudre une sous-partie du problème global de reconnaissance. Quant au module de recombinaison, il doit tenir compte de l'ensemble des informations des bandes pour décider du phonème prononcé. Ainsi, le fait de diviser le spectre en bandes indépendantes réduit effectivement la quantité d'information dans chaque bande, mais ce n'est pas forcément un mal, car cela permet de faciliter la tâche du module de recombinaison, qui lui, a accès à l'ensemble des informations fournies par les classifieurs sur tout le spectre. Le tout est donc d'adapter la classification réalisée dans chaque bande à l'information qui y est réellement contenue. Il s'agit là d'un problème difficile mais fondamental qui, à mon avis, devrait prendre de l'importance au cours des années à venir.

Une autre critique assez fréquente vis-à-vis du Multi-Bandes consiste à remarquer que l'information mutuelle qui existe entre différentes régions fréquentielles disparaît suite au découpage fréquentiel. Par exemple, une occlusive est détectable grâce à l'énergie apparaissant simultanément sur toutes les fréquences. Ce phénomène n'est pas du tout géré par les classifieurs qui se trouvent dans les bandes, mais par le module de recombinaison. L'information mutuelle entre les bandes est ainsi considérée uniquement dans le dernier étage du processus de reconnaissance. Sachant que l'information retournée par les bandes n'est qu'une fraction de celle dont elles disposent en entrée (ce qui est normal, car leur rôle est justement de réduire la complexité et la quantité des données), il faut faire attention à ce que l'information qui n'est pas pertinente pour une bande seule mais qui l'est pour le module de recombinaison ne soit pas éliminée par les classifieurs. Il faut donc instaurer un compromis entre la suppression de l'information superflue par les classifieurs et la préservation de l'information nécessaire au dernier module du système, compromis qui justifie l'utilisation du troisième type de recombinaison défini au chapitre 5.

Enfin, et pour terminer cette première présentation des problèmes posés par le Multi-Bandes, nous devons considérer également les choix réalisés lors de la conception du système. Ceux-ci peuvent bien entendu affecter grandement les performances de celui-ci selon qu'ils sont avisés ou non. Ainsi en est-il du choix qui décide du nombre de bandes que doit comporter le système, de leurs limites fréquentielles ainsi que des paramètres acoustiques à utiliser dans les modèles de Markov. Ces trois problèmes sont détaillés dans la partie 4.3. De même, nous devons choisir une méthode de recombinaison, et nous avons déjà vu que ce choix est sans aucun doute le plus important de tous. C'est pourquoi nous lui consacrons tout le chapitre 5.

2.2. Principe

Le principe général d'un système Multi-Bandes peut être modélisé par la figure 2.1 .

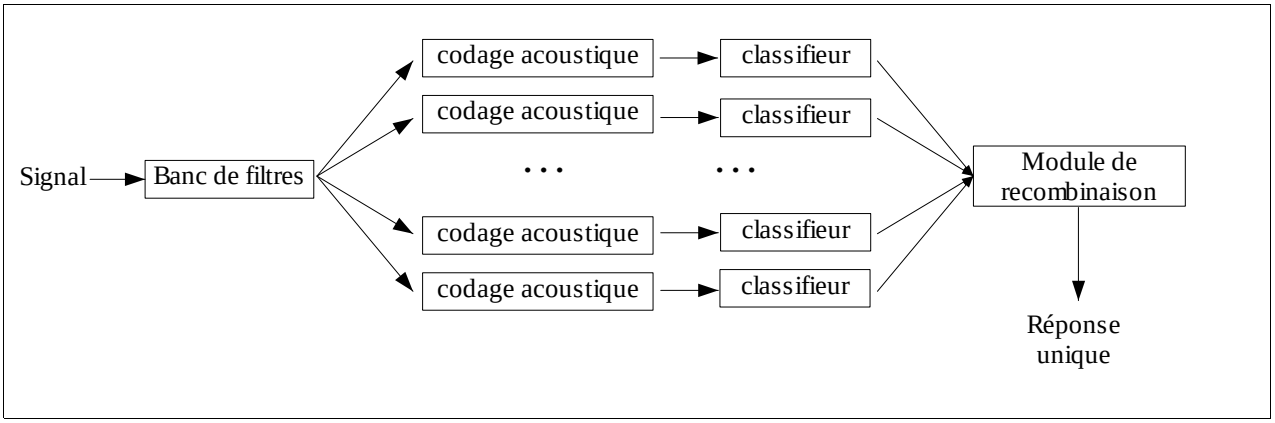

Figure 2.1 : Principe général d'un système Multi-Bandes

Le paradigme Multi-Bandes peut être décrit de la manière suivante : tout d'abord, le signal acoustique passe à travers un banc de filtres qui le décompose en plusieurs bandes fréquentielles. Puis, chacune de ces bandes est traitée indépendamment des autres grâce à un classifieur qui, le plus souvent, décide de l'appartenance d'un segment temporel à un phonème parmi ceux de la langue considérée. Enfin, les résultats de chacun de ces classifieurs sont passés à un module de recombinaison qui décide à quelle classe phonétique appartient le segment temporel initial.

Bien entendu, il s'agit là d'une présentation très générale du paradigme Multi-Bandes, et un grand nombre de problèmes se posent dès lors, comme par exemple le problème de la segmentation, du choix des classes phonétiques dans les bandes, de l'information à recombiner, etc. Tous ces problèmes seront traités au fur et à mesure dans le mémoire. La figure 2.1 ne fait que présenter le paradigme Multi-Bandes dans toute sa généralité. Le système Multi-Bandes que nous avons conçu diffère légèrement de ce système « canonique ». Il sera présenté en détails dans le chapitre 4 et chacun des choix d'implémentation que nous avons réalisés sera justifié. Mais avant de parler plus précisément de notre système, nous allons réaliser dans la prochaine partie un état de l'art des domaines de recherche proches du Multi-Bandes.

2.3. État de l'art des domaines proches du Multi-Bandes

Un certain nombre de recherches en parole sont en fait très proches du paradigme Multi-Bandes, mais ne revendiquent pas l'appartenance à ce courant, soit parce que leurs auteurs ne désirent pas s'associer à celui-ci, soit parce que les points communs entre ces systèmes et le Multi-Bandes ne sont pas assez nombreux pour justifier cette appellation. Il en va de même pour le Multi-Bandes que pour toutes les disciplines scientifiques : ses frontières sont loin d'être parfaitement définies. Toutefois, les travaux qui sont cités ici présentent un intérêt certain pour la recherche en Multi-Bandes et nous avons donc considéré qu'une présentation de ceux-ci devait être réalisée.

Dans le chapitre introductif, nous avons rappelé un des défauts principaux des modèles de Markov, à savoir qu'ils ne modélisent la dynamique que dans une seule dimension. Ceci est pénalisant car le signal est codé en deux dimensions, le temps et le cepstre par exemple. Il existe pourtant un cadre théorique permettant de modéliser le signal en tenant compte de ces deux dimensions : ce sont les Champs de Markov Cachés, ou *Markov Random Fields* (MRF). Malheureusement, ces MRFs sont beaucoup plus compliqués à manipuler que les Chaînes de Markov Cachées et surtout, il n'existe pas d'algorithme d'apprentissage efficace pour les MRF comme celui de Baum Welch pour les HMM. C'est essentiellement cette absence d'un tel algorithme qui a pendant si longtemps découragé les quelques chercheurs qui se sont penchés sur le sujet. Cependant, très récemment, Gravier & al. [Gravier98] ont tenté de régler l'épineux problème de l'apprentissage en « adaptant » un système Multi-Bandes de façon à autoriser les interactions entre les bandes fréquentielles. Ainsi, leur système commence par apprendre des HMM classiques dans chaque bande, puis un ajustement des paramètres des HMM est réalisé grâce à une heuristique modélisant les dépendances entre les bandes. Cet article réalise essentiellement une adaptation du cadre théorique des MRFs à la reconnaissance automatique de la parole, mais beaucoup de travail reste encore à réaliser afin d'élaborer des stratégies efficaces pour ce type d'approche. Les perspectives présentées dans l'article sont nombreuses et très intéressantes, notamment grâce à un nouvel algorithme d'apprentissage basé sur l'algorithme d'Estimation-Maximisation (EM) auquel est associée une descente de gradient.

Les deux autres études dont nous allons parler maintenant tentent également de pallier ce problème de l'absence de modélisation de la dynamique de la seconde dimension par les HMM classiques. La première a été réalisée par Ghahramani & al. dans [Ghahramani97]. Leur travail concerne le développement d'un nouveau formalisme qui remplace chaque état d'un HMM classique par un ensemble d'états associés. Nous retrouvons ainsi la structure d'un système Multi-Bandes, mais la différence réside essentiellement dans le fait qu'une unique suite d'observations est passée à ce HMM factoriel qui gère lui-même la distribution de ces observations dans l'espace des états. Un algorithme d'apprentissage exact est développé pour cette structure, mais il est malheureusement trop coûteux pour être réellement applicable. C'est pourquoi Ghahramani & al. ont développé dans leur article plusieurs algorithmes approchés qui rendent l'apprentissage possible. Un autre article écrit par D. Xu & al. [Xu96] se rapproche beaucoup plus du Multi-Bandes. Le système modélisé est appelé « HMM Multi-Canaux » où chaque état d'un HMM classique est également remplacé par plusieurs états. L'idée que D. Xu & al. cherchent à modéliser consiste à considérer chaque séquence unidimensionnelle à travers toutes les combinaisons possibles de canaux, chacune de ces séquences étant modélisée par un HMM. La moyenne est alors calculée en sommant les scores sur toutes les séquences. Ainsi, tous les sous-états correspondants au même état d'un HMM classique possèdent des transitions vers tous les sous-états correspondants à l'état suivant du HMM. De même, d'autres transitions bouclent vers les sous-états initiaux. L'information acoustique est alors divisée en autant de canaux. Dans ce modèle, les canaux ne sont pas indépendants entre eux et tous les états peuvent transporter tous les canaux. Quelques expériences de détection de mots montrent qu'il est possible d'obtenir de meilleures performances qu'avec un HMM discret, à condition de choisir correctement les canaux d'information utilisés.

Finalement, deux autres études menées par Lippmann & al. [Lippmann97b] et Cooke & al. [Cooke97] utilisent un découpage fréquentiel pour augmenter la robustesse des systèmes de RAP au bruit. Le principe consiste alors à calculer les densités de probabilités retournées par un HMM classique en supprimant ou régénérant les coefficients corrompus des vecteurs acoustiques. Ce calcul fait appel à la théorie des données manquantes qui permet d'obtenir une densité de probabilité sachant les coefficients du vecteur acoustique corrompus. Il faut donc, au cours d'une étape de pré-traitement du signal, utiliser un détecteur de rapport signal sur bruit afin de déterminer quels coefficients sont corrompus et lesquels ne le sont pas. De plus, Lippmann utilise des coefficients spectraux et non plus cepstraux, car dans ce dernier cas, un bruit limité fréquentiellement n'en affecte pas moins l'ensemble du cepstre. Sur ce point, une tentative très intéressante à été récemment réalisée afin d'adapter la théorie des données manquantes à l'utilisation de coefficients cepstraux [Siohan98]. Malgré ces restrictions, un système basé sur ce principe peut se montrer extrêmement robuste, à condition que l'étage de détection des coefficients corrompus effectue correctement son travail. Stéphane Dupont a également très récemment travaillé sur la reconstruction des données manquantes [Dupont98].

2.4. État de l'art du Multi-Bandes

2.4.1. Un nouveau problème : l'asynchronisme entre les bandes

Après cet état de l'art des travaux associés au Multi-Bandes, nous allons dans cette partie et dans les suivantes nous intéresser exclusivement au système Multi-Bandes lui-même. Le premier problème spécifique au Multi-Bandes que nous allons traiter est celui posé par l'absence de synchronisme entre les bandes pour la reconnaissance en mode continu. Nous expliquons de manière succincte ce problème dans le prochain paragraphe, puis nous le considérerons selon un autre point de vue dans le chapitre 3. Mais dans un premier temps, nous consacrons cette partie 2.4 à l'état de l'art des systèmes Multi-Bandes présentés dans la littérature, tout d'abord en les décrivant de manière générale, puis en les analysant essentiellement du point de vue de la résolution du problème évoqué ci-dessus. Les détails de ces systèmes concernant leurs autres aspects sont développés dans les parties concernées, c'est-à-dire essentiellement 4.3.1, 4.3.3 et 5.1. Les parties 3.3 et 3.4 présentent de nouveaux algorithmes que nous avons conçus pour résoudre ce problème. Les tests correspondants de reconnaissance de la parole Multi-Bandes en mode continu se trouvent dans la partie 6.9.

Avant toute chose, il nous faut expliquer en quoi consiste le problème d'asynchronisme entre les bandes. Nous avons vu qu'un système Multi-Bandes réalise une reconnaissance indépendante sur chaque bande, puis recombine les résultats de ces classifieurs. Le problème est donc essentiellement de savoir « quand » recombiner. Nous pouvons ainsi distinguer les deux « niveaux » de recombinaison suivants :

1. Recombinaison trame par trame : Cette recombinaison est équivalente à une recombinaison au niveau des états des HMM. Les sorties des classifieurs dans les bandes sont ainsi recombinées après chaque trame de signal. Les systèmes utilisant cette recombinaison sont présentés dans la partie 2.4.2.

2. Recombinaison après des segments temporels plus longs que la simple trame : Dans ce cas, les HMM dans les bandes sont indépendants et peuvent donc associer chaque trame à des états différents. Par exemple, la figure 2.2 montre une association différente des trames sur les états pour deux bandes. Nous rappelons que cette association est classiquement réalisée par l'algorithme de Viterbi, dont une description détaillée peut être consultée dans [Forney73].

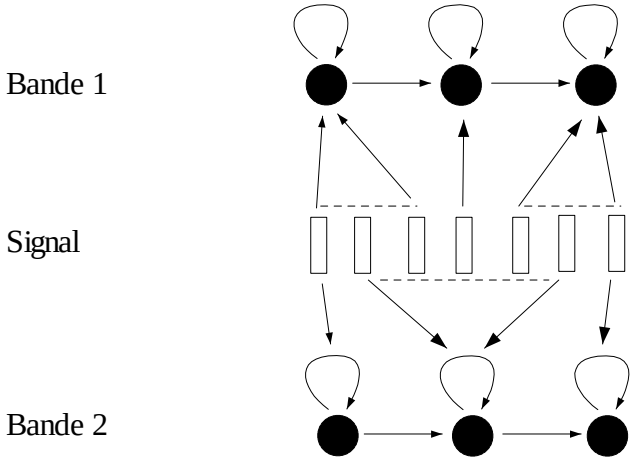

Figure 2.2 : Exemple d'associations trames-états pour deux bandes.

Ainsi, le cas extrême de cette recombinaison consiste à laisser les bandes totalement indépendantes sur toute la phrase et à ne recombiner qu'à la fin de la phrase prononcée. Le problème vient alors du fait que chaque HMM choisissant une séquence d'association trames-états différente, non seulement la segmentation mais aussi le nombre de phonèmes ou de mots dans chaque bande vont être différents. Comment donc recombiner de telles suites de phonèmes ? Un certain nombre de solutions ont été proposées dans la littérature : nous les étudions dans la partie 2.4.3. Dans le cadre de cette thèse, nous avons nous-mêmes conçu deux algorithmes qui sont présentés dans les parties 3.3 et 3.4.

2.4.2. Recombinaison trame par trame : état de l'art

Revenons à la recombinaison trame par trame. L'avantage de cette méthode est de résoudre d'emblée le problème exposé ci-dessus. En effet, puisque les bandes sont recombinées après chaque état, leur segmentation est exactement la même ainsi que le nombre d'unités reconnues dans la phrase. Cette solution, qui a également l'avantage d'être simple, a été adoptée dans un grand nombre de systèmes. Ainsi, le premier véritable système Multi-Bandes créé utilise cette méthode. Il s'agit du système que Duchnowsky a développé au cours de sa thèse, soutenue en septembre 1993. Ce système réalise une reconnaissance phonétique sur de la parole continue en mode indépendant du locuteur. Le but de ce travail était d'aider les malentendants à reconnaître les phrases prononcées. Duchnowsky a ainsi créé un système Multi-Bandes composé de quatre bandes dont les limites fréquentielles sont [100-700 Hz], [700-1500 Hz], [1500-3000 Hz] et [3000-4500 Hz]. Ces limites ont été choisies afin qu'un unique formant soit approximativement contenu dans une bande. La recombinaison qu'il a choisie est une recombinaison trame par trame, où seules les étiquettes des classes gagnantes dans chaque bande sont recombinées. L'appartenance à la classe q de chaque trame est alors décidée en maximisant la probabilité suivante :

$$Pr(c_1 c_2 c_3 c_4|q) = \prod_{i=1}^{4} Pr(c_i|q)$$

Avec

$$Pr(c_i|q) = \frac{x_{c_i}^{q}}{N_q}$$

où $x_{c_i}^{q}$ représente le nombre de trames classifiées c_i par le reconnaisseur de la bande *i* lorsque *q* est prononcé dans la base d'apprentissage, et N_q représente le nombre total de trames correspondant à la classe *q* prononcée.

De plus, puisque les $x_{c_i}^{q}$ ne sont pas assez nombreux dans la base d'apprentissage, Duchnowsky utilise une méthode d'interpolation linéaire afin d'estimer ces nombres d'occurrences. Le dernier étage de son système est composé d'un autre HMM dont le rôle est de corriger les trop nombreux changements d'étiquettes des trames finales. Il a également testé plusieurs paramètres acoustiques, dont les coefficients de prédiction linéaire (LPC), les cepstres et les coefficients d'autocorrélation. Les cepstres se sont révélés les meilleurs paramètres, et les expériences qu'il a réalisées sur la base de données TIMIT ont donné un taux de reconnaissance phonétique de 58,5 % en utilisant 39 phonèmes. Ce résultat correspondait à ceux obtenus par les systèmes de référence de l'époque [Lee89].

Un autre système, développé par Tibrewala et Hermansky [Tibrewala97], utilise également une recombinaison au niveau des trames. En fait, plusieurs expériences ont été réalisées sur le corpus américain Switchboard afin de comparer les différents niveaux de recombinaison, mais les résultats étant quasiment identiques, que la recombinaison soit réalisée après chaque trame ou après chaque phonème, Tibrewala et Hermansky ont décidé de conserver dans leur système une recombinaison trame par trame. À la différence du système de Duchnowsky, le module de recombinaison utilise cette fois les probabilités retournées par les HMM et associe ainsi un score final à chaque phonème, soit en réalisant une somme pondérée de ces probabilités, soit par l'intermédiaire d'un perceptron multi-couches (PMC). Ces différents types de recombinaison sont détaillés dans le chapitre 5. Tibrewala et Hermansky montrent clairement que la meilleure recombinaison possible est celle utilisant un PMC.

De même, Okawa & al. [Okawa98] ont réalisé un système recombinant soit les densités de probabilités des états des HMM, soit les coefficients acoustiques eux-mêmes après chaque trame. Ils ont essentiellement testé plusieurs types de recombinaisons linéaires, en considérant des poids dépendant du Rapport Signal-Bruit dans chaque bande, ou dépendant de l'entropie conditionnelle de chaque bande. Les expériences qu'ils ont réalisées sur le corpus ATIS d'ARPA montrent que presque tous les systèmes Multi-Bandes surpassent le système de référence, et que les meilleurs scores sont obtenus avec des poids choisis empiriquement.

Enfin, je terminerai cette première approche des autres systèmes Multi-Bandes en parlant de celui conçu par Naghmeh Nikki Mirghafori [Mirghafori99], qui utilise également une recombinaison trame par trame s'appuyant sur les probabilités retournées par les HMM, probabilités qui sont passées à un perceptron dont le rôle est d'associer un score final à chaque phonème.

Ainsi, un grand nombre de systèmes ont choisi une recombinaison après chaque trame. Ceci peut aisément se comprendre au vu de la difficulté à concevoir un algorithme efficace permettant de recombiner après des segments temporels plus longs. Néanmoins, avec une recombinaison trame par trame, les bandes sont obligatoirement synchrones. Ceci est pénalisant dans le cadre d'un système Multi-Bandes, car l'une des motivations principales de ce paradigme est justement de modéliser des dynamiques temporelles différentes dans chaque bande.

2.4.3. Recombinaison après des segments de plusieurs trames : état de l'art

✓ *Justification de la méthodologie utilisée pour les tests*

La majeure partie de nos expériences sont réalisées en utilisant une recombinaison après des segments de plusieurs trames, lorsque la segmentation est connue. Ceci s'explique essentiellement par les deux raisons suivantes :

Tout d'abord, il est facile de recombiner des segments de plusieurs trames lorsque les limites de ces segments sont connues, et cette solution est certainement la plus simple pour étudier le comportement d'un système Multi-Bandes en l'absence de synchronisme entre les bandes. C'est pourquoi l'essentiel de ce mémoire utilise ce mode de reconnaissance, ce qui nous a permis de comparer aisément différentes méthodes de recombinaison et différents algorithmes propres à l'apprentissage du système tout en conservant cette hypothèse d'indépendance temporelle entre les bandes qui est fondamentalement liée au paradigme Multi-Bandes à nos yeux. La solution que nous avons choisie pour notre étude, c'est-à-dire une reconnaissance en mode « isolé » avec un alignement[4] indépendant de chaque bande, s'oppose à une autre solution qui a été la plus souvent choisie dans la littérature, c'est-à-dire une reconnaissance en mode continu avec une recombinaison trame par trame. Nous n'avons pas choisi celle-ci car, d'une part, nous voulions explorer de nouvelles voies de recherche, mais surtout, nous considérons que cet asynchronisme entre les bandes fait intimement partie du paradigme Multi-Bandes et qu'il est dommageable de s'en séparer. Ce qui ne signifie pas du tout que nous abandonnons la reconnaissance en mode continu, mais nous considérons plutôt que c'est un problème du Multi-Bandes qu'il nous faut résoudre par des moyens autres que synchroniser les bandes, comme nous le verrons dans la suite.

Enfin, si le mode isolé est trop contraignant pour certaines applications en reconnaissance de la parole, il ne l'est pas pour d'autres tâches de traitement de la parole. Par exemple, Laurent Besacier a développé au cours de sa thèse un système de reconnaissance du locuteur utilisant le principe Multi-Bandes [Besacier98a]. Cette tâche pouvant être réalisée sans que la segmentation ne soit prise en compte, il est alors possible de recombiner les réponses des classifieurs à la fin de la phrase. De même pour une tâche d'identification du langage, la recombinaison peut être facilement réalisée à la fin de la phrase, comme nous le verrons dans la partie 6.7. Enfin, les applications utilisant des commandes vocales s'appuient également sur une reconnaissance en mode isolé.

4 Nous rappelons qu'un alignement est l'association de chaque trame de parole avec un état d'un HMM. Un alignement doit bien entendu respecter les contraintes de transitions entre les états des HMM et entre les HMM eux-mêmes. Nous ne considérons donc pas les alignements qui sont impossibles au sens de ces transitions.

Néanmoins, une conséquence de ce choix est que les taux de reconnaissance calculés ne sont pas les mêmes que ceux qui apparaissent généralement dans la littérature et qui concernent le mode continu. Ainsi, le taux de reconnaissance correspondant au terme anglophone d'« *accuracy* » fait intervenir un algorithme de programmation dynamique qui minimise la distance entre la véritable succession de phonèmes prononcés et celle reconnue par le système. Cette distance s'exprime de la manière suivante :

$$\mathrm{Acc} = \frac{nPhon - nIns - nSub - nSup}{nPhon}$$

où *nPhon* représente le nombre de phonèmes réellement prononcés, *nIns* le nombre de phonèmes insérés, *nSub* le nombre de phonèmes substitués, i.e. confondus par le reconnaisseur, et *nSup* le nombre de phonèmes supprimés, i.e. oubliés. Chaque alignement entre les séquences de phonèmes prononcés et reconnus constitue un chemin dans le graphe de tous les alignements possibles. Le score de chacun de ces chemins se calcule par la formule :

$$\mathrm{Score} = 300 \cdot nIns + 300 \cdot nSup + 400 \cdot nSub$$

et le chemin de score minimal est conservé.

Nous utilisons ce taux pour nos expériences en mode continu dans la partie 6.9. Par contre, le taux de reconnaissance que nous employons lorsque la segmentation est connue correspond en fait au terme anglophone de « *phoneme classification error* », comme l'indique De Mori dans son livre [DeMori98] à la page 269. Il se calcule par la formule suivante :

$$\frac{N_{rec}}{N_{pron}}$$

où N_{rec} désigne le nombre de phonèmes correctement reconnus et N_{pron} le nombre total de phonèmes prononcés.

✓ *Reconnaissance en mode continu : état de l'art*

Néanmoins, pour en revenir à la reconnaissance de la parole, il est indispensable de pouvoir utiliser un système en mode continu. Deux algorithmes ont été cités dans la littérature sur le Multi-Bandes afin de réaliser cette tâche :

Méthode basée sur la combinaison de HMM :

Le premier d'entre eux réalise une combinaison de plusieurs HMM, selon une méthode dérivée d'un principe issu des travaux de M. J. F. Gales et S. J. Young [Gales98]. Ces travaux présentent un formalisme permettant de modéliser conjointement la parole et le bruit présents dans le signal. Ce principe est généralement connu sous le nom de « Combinaison Parallèle de Modèles ». Il est donc possible de l'appliquer au Multi-Bandes en transformant l'ensemble des HMM des bandes en un unique HMM, dont la topologie est cependant beaucoup plus complexe. Par exemple, si nous considérons deux HMM à trois états, le HMM unique correspondant est représenté sur la figure 2.3. Le principe pour construire ce nouveau HMM est simple, dans la mesure où chaque état du HMM résultant correspond à deux états des HMM initiaux qui sont associés à la même trame du signal.

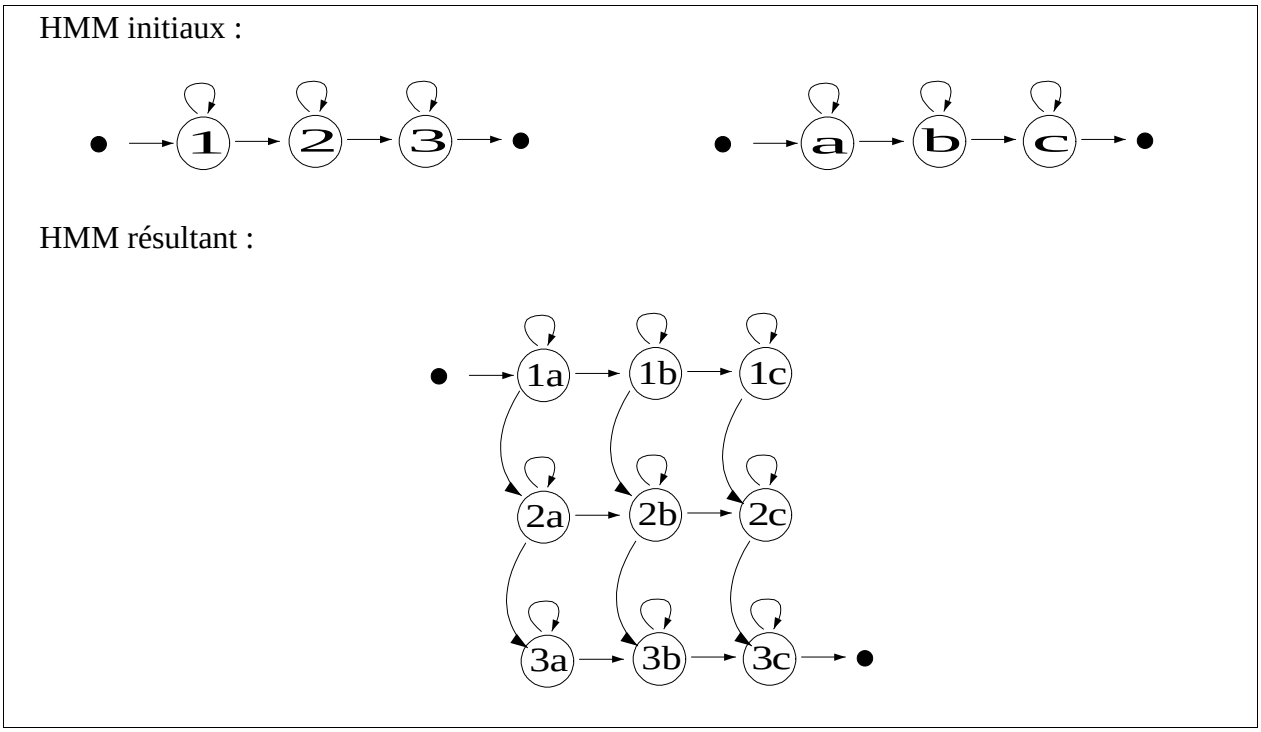

Figure 2.3 : Exemple de HMM résultant de la méthode de recombinaison des HMM

Une fois cet HMM unique obtenu, il suffit alors d'appliquer l'algorithme de Viterbi afin de connaître la meilleure séquence de modèles possible.

Le problème de cette méthode vient du fait que le HMM résultant est considérablement plus complexe que les HMM initiaux. Ce qui signifie que la reconnaissance d'une phrase complète est tout simplement impossible, car il faudrait appliquer cette méthode à la suite complète des états, après que toutes les séquences de modèles possibles aient été construites. Une solution à ce problème consiste à contraindre les alignements dans les bandes à n'être décalés que d'un nombre maximum d'états, auquel cas la taille du HMM résultat ne dépend plus de la taille de la phrase, mais n'en reste pas moins très grand.

Méthode basée sur le « two-level dynamic programming » :

La seconde méthode proposée dans la littérature consiste à utiliser l'algorithme de programmation dynamique à deux niveaux [Rabiner93], qui a été utilisé avec succès en reconnaissance de la parole au début des années 80. Cet algorithme est caractérisé par le fait qu'il se décompose en deux étapes : la première consiste à calculer les taux de reconnaissance de tous les modèles isolés sur des portions arbitraires de la phrase à reconnaître, tandis que la seconde construit le graphe de tous les alignements possibles en concaténant ces segments. La recherche du meilleur chemin dans ce graphe est ensuite réalisée.

L'intérêt de cet algorithme, qui a été proposé par Sakoe en 1979, est qu'il améliore la complexité de l'algorithme de recherche « brut » de tous les alignements possibles. En effet, le fait de calculer les scores de tous les segments possibles au cours de la première étape de l'algorithme à deux niveaux permet de ne plus recalculer ces alignements lorsque tous les chemins sur la phrase complète sont considérés pendant la deuxième étape.

N. N. Mirghafori a testé dans sa thèse ces deux méthodes [Mirghafori99]. Pour les deux algorithmes, les résultats montrent une légère dégradation par rapport au système avec une recombinaison trame par trame. Nous pourrions donc en déduire que l'absence de synchronisme entre les bandes n'est pas bénéfique pour le système. Cependant, H. Bourlard et S. Dupont ont obtenu des résultats différents avec leur système [Bourlard96]. Celui-ci est constitué par des HMM modélisant des phonèmes, mais les auteurs ont introduit un état supplémentaire ne générant pas de trames de signal dont le rôle est de recombiner et de resynchroniser les différentes bandes. Cette architecture nécessite néanmoins toujours l'utilisation de l'un des deux algorithmes cités ci-dessus. Pour compléter la description de ce système, il faut noter que les densités de probabilités dans chaque état du HMM sont calculées par un perceptron multi-couches. Deux modules de recombinaison ont été testés : un linéaire, dont les poids correspondent soit aux taux de reconnaissance des modèles individuels, soit au rapport signal-bruit présent dans la bande, et un non linéaire, constitué d'un perceptron multi-couches. Les meilleurs résultats ont été observés pour le perceptron et pour une recombinaison au niveau de la syllabe, qui était l'unité phonétique de recombinaison la plus longue utilisée. Ceci tendrait donc à confirmer l'importance de l'hypothèse d'asynchronisme entre les bandes.

L'explication qui me paraît la plus probable quant aux mauvais résultats obtenus par Mirghafori relève plutôt de la complexité accrue des algorithmes utilisés, ce qui a malheureusement pour effet d'augmenter le nombre d'erreurs et d'imprécisions dans ceux-ci. Ceci semble confirmé par le fait que, dans les expériences de Bourlard, la reconnaissance est réalisée sur des mots isolés. L'accroissement en complexité du HMM résultant dans la méthode de combinaison des HMM est donc moindre que pour une recombinaison sur la phrase.

Chapitre 3
Contributions pour la reconnaissance Multi-Bandes en mode continu

3.1. Introduction

Nous venons de voir dans le chapitre 2 deux algorithmes proposés dans la littérature et destinés à réaliser une reconnaissance de la parole Multi-Bandes en mode continu. Classiquement, l'algorithme utilisé pour réaliser ceci est l'algorithme de Viterbi. Or, celui-ci ne figure pas parmi les deux algorithmes sus-cités. Ceci vient du fait qu'il n'est pas possible d'utiliser l'algorithme de Viterbi avec le Multi-Bandes. Dans une première partie 3.2, nous allons expliquer pourquoi, puis nous proposons deux nouveaux algorithmes accomplissant cette tâche. Dans la partie 3.3, chaque bande est laissée totalement indépendante sur toute la phrase, puis la recombinaison est réalisée à la fin de la phrase. Dans la partie 3.4, les bandes sont indépendantes à l'intérieur de chaque modèle, puis sont recombinées après chaque modèle. Seuls les algorithmes théoriques sont présentés dans ce chapitre, et les résultats expérimentaux correspondants sont présentés dans le chapitre 6.

3.2. Impossibilité d'utiliser l'algorithme de Viterbi avec le Multi-Bandes

3.2.1. Présentation de l'algorithme de Viterbi

L'algorithme de base utilisé pour aligner les trames de parole avec les états des HMM est l'algorithme de Viterbi [Forney73]. Dans le but de mieux comprendre l'algorithme qui est présenté dans la suite, nous devons tout d'abord brièvement rappeler quel est le principe général de l'algorithme de Viterbi, algorithme de programmation dynamique stochastique. En fait, il s'agit d'une recherche d'un meilleur chemin dans le graphe de tous les alignements possibles. Toutefois, cette recherche d'un meilleur chemin, à la différence des algorithmes classiques tels que l'algorithme de Dijsktra, utilise les propriétés du graphe en question, à savoir qu'il est connexe, orienté et acyclique. L'algorithme de Viterbi permet donc de trouver ce meilleur chemin en un temps $N\,T$, où N est le nombre de prédécesseurs de chaque état, et T est le nombre de trames.

Le principe est le suivant : les trames de parole étant considérées de la première à la dernière, notons *t* la trame courante. Tous les états de tous les HMM sont alors considérés l'un après l'autre, dans un ordre quelconque ; soit *e* l'état courant. Le meilleur chemin permettant d'associer la trame *t* à l'état *e* est calculé en choisissant le chemin entrant dans le nœud (*e*,*t*) qui fournit le score maximum dans ce nœud. La figure 3.1 montre une partie d'un tel graphe pour une transition entre deux modèles.

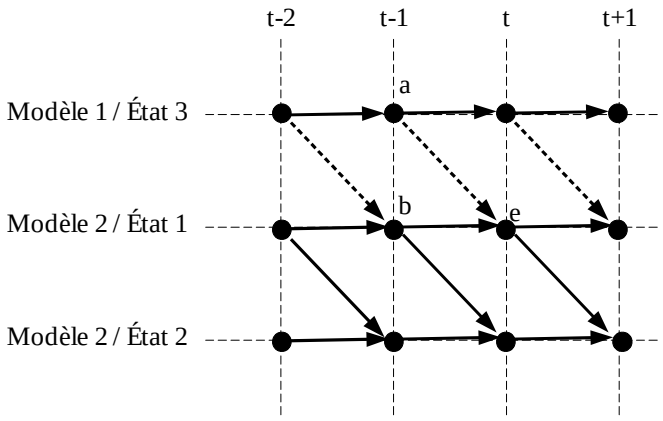

Figure 3.1 : Exemple du graphe utilisé par l'algorithme de Viterbi en mode continu. Une transition entre deux HMM est présentée.

Dans cet exemple, le score obtenu au nœud (*e*,*t*) est, soit le score du nœud (*a*,*t*-1) auquel est ajoutée la probabilité de transition entre *a* et *e* puis la probabilité d'émission de *t* par *e*, soit le score du nœud (*b*,*t*-1) auquel est ajoutée la probabilité de transition entre *b* et *e* puis la même probabilité d'émission de *t* par *e*. Le score maximum parmi ces deux scores indique le meilleur chemin recherché.

Une fois que tous ces scores ont été calculés jusqu'à la dernière trame et le dernier état, une étape classique de « *back-tracking* » permet, en partant du dernier état du modèle du silence par exemple, de remonter jusqu'au premier état du premier phonème et de mettre ainsi en évidence le meilleur chemin.

3.2.2. Problème posé par l'algorithme Viterbi

L'avantage principal de l'algorithme de Viterbi réside dans le fait que la segmentation du signal en phonèmes n'est pas définie avant que l'algorithme n'ait parcouru tout le signal et n'ait réalisé l'étape de « retour-arrière ». En fait, il est impossible de connaître la meilleure segmentation avant d'être arrivé à la fin du signal, et c'est justement cette faculté de pouvoir tester toutes les segmentations possibles en un seul parcours du signal qui fait la force de l'algorithme de Viterbi. Malheureusement, ceci n'est plus possible si nous voulons utiliser le principe Multi-Bandes avec recombinaison à la fin des phonèmes. En effet, la recombinaison s'effectuant seulement à la fin de chaque segment, il faut obligatoirement connaître cet instant afin de pouvoir appliquer la recombinaison. Le problème vient donc du fait qu'avec le Multi-Bandes, il est impossible d'associer un score à un chemin lorsque le dernier phonème de celui-ci n'est pas complet. La recombinaison impose qu'un nombre entier de phonèmes soient considérés pour que le score puisse être calculé, ce qui empêche l'utilisation de l'algorithme de Viterbi.

Les solutions que nous avons déjà présentées (cf. partie 2.4) pour résoudre ce problème sont les suivantes :

1. Recombiner le signal trame par trame ;
2. Combiner les modèles HMM entre les bandes ;
3. Utiliser l'algorithme de programmation dynamique à deux niveaux ;

Nous en avons conçu trois autres que nous présentons dans cette partie. La première consiste à utiliser un étage de pré-segmentation du signal, comme celui décrit dans [Husson98], puis à appliquer l'algorithme Multi-Bandes en mode isolé sur cette segmentation. Cette solution est très simple et nous verrons dans la partie 6.9.2 comment nous pouvons l'appliquer en pratique. La deuxième, qui est présentée dans la partie 3.3, laisse les bandes totalement indépendantes sur toute la phrase, et la troisième, qui est présentée dans la partie 3.4, utilise un algorithme de programmation dynamique et lui adjoint des méthodes de « recherche en faisceaux », ou de « *Beam Search* ».

3.3. Un nouvel algorithme pour la recombinaison en fin de phrase

3.3.1. Présentation théorique

Aucune étude n'a été réalisée, à notre connaissance, avec un niveau de recombinaison en fin de phrase dans le cas de la parole continue. Nous avons cependant réfléchi au problème, car il nous a semblé qu'il était possible de tirer parti de cette totale indépendance des bandes, aussi bien dans les modèles que du point de vue temporel, comme nous allons le voir.

Dans le cas où les bandes sont laissées totalement libres tout au long de la phrase, comme nous l'avons déjà remarqué, la segmentation dans chacune d'elle ainsi que le nombre de phonèmes reconnus diffèrent. La figure 3.2 nous donne un aperçu des réponses possibles pour chaque bande pour la phrase « *C'est ainsi que Jacques fut arrêté* ».

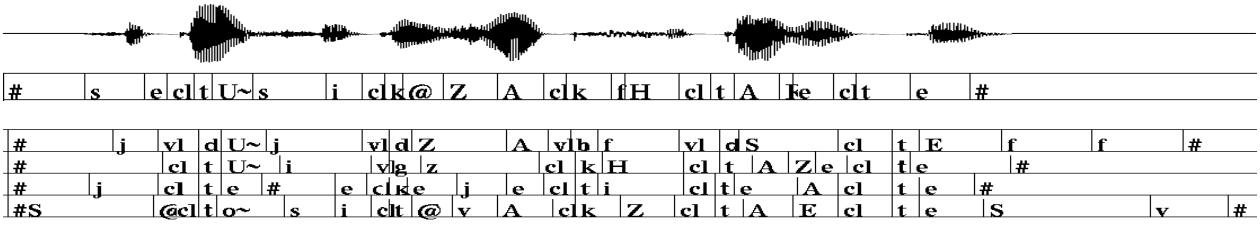

Figure 3.2 : Segmentation des 4 sous-bandes pour la phrase « C'est ainsi que Jacques fut arrêté». De bas en haut : la réponse des bandes [0 ... 538 Hz], [461 ... 1000 Hz], [923 ... 2823 Hz] et [2374 ... 7983 Hz], la segmentation manuelle puis le signal de parole.

Dans une approche statistique, un système de reconnaissance automatique de la parole recherche la suite de phonèmes $\widehat{\Omega}$ telle que

$$P(X|\widehat{\Omega}) = \underset{\Omega}{argmax}\, P(X|\Omega)$$

où *X* représente la suite des vecteurs acoustiques prononcés.

Dans le cas d'un système Multi-Bandes sans synchronisme entre des bandes considérées comme indépendantes, cette probabilité est définie par :

$$P(X|\Omega)=\prod_{i=1}^{B} P(X|\Omega_i, \text{bande } i)$$

où Ω_i représente la suite de phonèmes reconnue dans la bande *i*, qui peut être différente de la suite de phonèmes finale que le système reconnaît. Il nous faut donc trouver une relation liant Ω aux Ω_i. Notons $\Re(\Omega_1,\ldots,\Omega_B,\Omega)$ cette relation, *B* représentant le nombre de bandes.

Il nous faut émettre ici des hypothèses simplificatrices afin de pouvoir résoudre le problème. Dans un premier temps, puisque chaque Ω représente une suite de phonèmes, nous avons décidé que la relation $\Re$ serait en fait définie comme la concaténation de relations $r(\omega_1,\ldots,\omega_B,\omega)$ mettant en jeu un phonème au plus de chaque bande. Ceci signifie qu'il ne sera pas possible de définir de relation entre un phonème d'une bande et plusieurs phonèmes d'une autre bande. Ce choix est très certainement discutable et nous sommes intéressés par des développements n'utilisant pas cette restriction.

Comment définir la relation *r* ? Le but étant de retrouver la phrase prononcée, *r* doit permettre d'associer à un ensemble de phonèmes $(\gamma(1),\ldots,\gamma(B))$ proposés par les bandes, un unique phonème ω_0 ayant été prononcé. *r* peut donc correspondre à une mesure de similarité entre phonèmes. Toujours pour des raisons de simplicité, nous avons choisi *r* comme étant la probabilité que les bandes aient reconnu $(\gamma(1),\ldots,\gamma(B))$ en supposant que ω est le phonème effectivement prononcé. Ceci nous permet de déduire ω_0 en maximisant $r(\gamma(1),\ldots,\gamma(B),\omega)$ sur tous les ω possibles. Nous avons donc :

$$r(\gamma(1),\ldots,\gamma(B),\omega)=P(\gamma(1),\ldots,\gamma(B)|\omega)P(\omega)$$

Soit :

$$r(\gamma(1),\ldots,\gamma(B),\omega)=\prod_{i=1}^{B} P(\gamma(i)|\omega, \text{bande } i)P(\omega) \qquad \text{(Eq-1)}$$

Nous pouvons remarquer que *r* définit ici une relation de recombinaison entre les étiquettes des classes gagnantes de chaque bande.

Maintenant que *r* est définie, il nous reste à décider quels phonèmes de chaque bande nous allons inclure dans le calcul de la même relation. Ce choix va être réalisé en trois étapes :

1. Une séquence arbitraire Ω est générée ;
2. Le graphe de toutes les associations de phonèmes possibles entre les bandes et Ω est créé ;
3. Le chemin maximisant le produit des probabilités retournées par *r* est calculé.

Ceci se traduit par :

$$\Re(\Omega(1),\ldots,\Omega(B),\Omega)=\max_{\Gamma\in\Lambda}\prod_{\gamma\in\Gamma} r(\gamma(1),\ldots,\gamma(B),\gamma(0)) \qquad \text{(Eq-2)}$$

où $\gamma(0)$ désigne un phonème de Ω, Γ une association de phonèmes possibles, i.e. le n-uplet $(\gamma(1),\ldots,\gamma(B),\gamma(0))$, et Λ l'ensemble de toutes les associations Γ possibles.

Par exemple, si deux bandes ont reconnu respectivement [a,b] et [c,d], alors :

$$\Lambda=\{((\text{a},\text{c}),(\text{b},\text{d})),((\text{a},\text{c}),(\text{b},\square),(\square,\text{d})),\ldots\}$$

Un Γ possible est alors $\Gamma=((\text{a},\text{c}),(\text{b},\text{d}))$, et pour ce Γ, une association possible est $\gamma(0)=\text{a}$ et $\gamma(1)=\text{c}$.

Finalement, la séquence finale des phonèmes prononcés peut être calculée de la manière suivante :

$$\widehat{\Omega}=\underset{\Omega}{\mathit{argmax}}\,\Re(\Omega_1,\ldots,\Omega_B,\Omega) \qquad \text{(Eq-3)}$$

3.3.2. Expériences

Nous avons vu que la méthode exposée ci-dessus réalise en fait une recombinaison utilisant les étiquettes des classes gagnantes dans chaque bande. Ces classes sont recombinées grâce l'équation Eq-1. Cette formule nous oblige donc à connaître $P(\omega_i|\omega_{j,}\text{ bande }k)$ pour tous les triplets (*i*,*j*,*k*) possibles. Ces probabilités sont pratiquement calculées à partir de la matrice de covariance des bandes. Eq-2 et Eq-3 sont réalisées en utilisant un algorithme classique de recherche d'un meilleur chemin dans un graphe. Ce graphe peut être défini de la manière suivante :

Chaque nœud contient un ensemble de *B* pointeurs indiquant la position du phonème courant dans la suite des phonèmes de chaque bande. Une transition d'un nœud à l'autre est alors réalisée en incrémentant un ou plusieurs de ces pointeurs, l'ensemble des phonèmes ainsi considérés constituant une association Γ possible. Par exemple, avec quatre bandes, chaque nœud possède 16 transitions possibles. Le nœud final du graphe est celui pour lequel tous ses pointeurs ont atteint le dernier phonème proposé dans chaque bande.

Nous définissons la probabilité de chaque chemin par l'équation :

$$P(\Gamma)=\prod_{\gamma\in\Gamma}\left\{\max_{\omega}\left\langle\begin{matrix}P(\text{bande } i \text{ a trouvé } \gamma(i)|\omega)\\ P(\text{bande } i \text{ a oublié } \omega)\end{matrix}\right\rangle\cdot P(\omega)\right\}$$

Comme cette équation l'indique, le cas des bandes vides a été traité en calculant la probabilité qu'une bande ait oublié un phonème, probabilité qui se calcule également à partir de la matrice de confusion d'un classifieur.

Le début d'un tel graphe est présenté en exemple sur la figure 3.3. Ce graphe correspond à la phrase « *C'est ainsi que Jacques fut arrêté* », et dont nous pouvons voir la transcription phonétique sur la figure 3.4. Un exemple d'alignement réalisé par l'algorithme sur cette phrase est également présenté sur la figure 3.4.

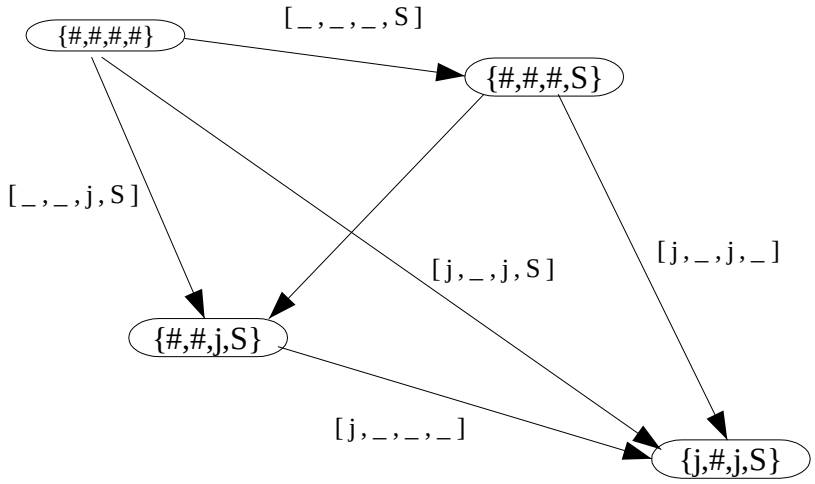

Figure 3.3 : Début du graphe associé à la figure 3.4

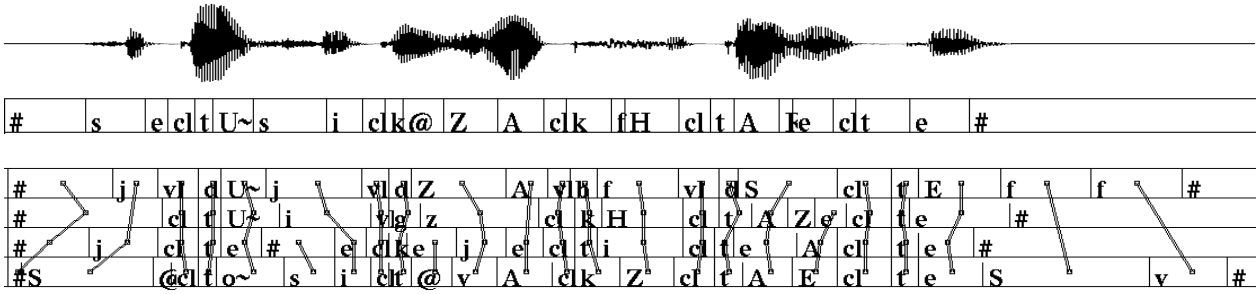

Figure 3.4 : Exemple d'alignement réalisé par l'algorithme sur les réponses de 4 sous-bandes fréquentielles

Comme nous pouvons le voir, l'alignement semble correct, dans la mesure où les segments groupés par l'algorithme sont effectivement très proches d'un point de vue phonétique et où la plupart de ces groupes correspondent à un phonème réellement prononcé. Toutefois, il est bien évident que cette analyse sur un seul exemple ne permet pas de valider l'algorithme, mais qu'une étude beaucoup plus approfondie des résultats est nécessaire. Nous n'avons pas réalisé cette étude, car les premiers tests de reconnaissance de cet algorithme ne fournissent pas de résultats exploitables. L'analyse des erreurs réalisées nous a montré que ces mauvais résultats étaient dus à la recombinaison qui n'utilise que les étiquettes des classes gagnantes dans chaque bande. Or, nous n'avons pas réussi à obtenir des taux de reconnaissance corrects avec une telle recombinaison, comme nous le verrons dans la partie 6.2. La seule recombinaison qui donne de bons résultats utilise l'ensemble des densités de probabilités retournées par les HMM pour tous les modèles de toutes les bandes. Il est très difficile d'utiliser une recombinaison de ce type avec l'algorithme exposé ci-dessus, car les densités de probabilités que nous pouvons calculer avec un tel algorithme appartiennent à des segments temporels dont les limites varient, ce qui induit une trop grande variabilité de ces mêmes densités de probabilités pour qu'elles soient apprises par exemple par un perceptron.

Ceci signifie que la recombinaison utilisant les scores des classifieurs nécessite une segmentation commune à tous les classifieurs. C'est pour cette raison que nous n'avons pas poursuivi plus loin l'étude de cet algorithme, tant qu'une recombinaison valable sur les étiquettes des classes gagnantes n'a pas été trouvée. Cependant, nous avons quand même présenté cet algorithme, car nous pensons que l'idée de recombiner les bandes au niveau de la phrase est très intéressante, or aucun article à notre connaissance ne traite de ce sujet. De plus, il n'est pas impossible de trouver dans un futur proche une bonne recombinaison n'utilisant que les étiquettes des classes gagnantes, comme cela est suggéré dans la partie 6.2.

Nous allons voir dans la partie suivante un autre algorithme permettant d'utiliser cette fois une recombinaison sur les scores tout en réalisant une reconnaissance en mode continu avec toutefois une contrainte supplémentaire par rapport à la partie 3.3 : la segmentation de chaque bande est la même.

3.4. Conception d'un algorithme pour la recombinaison après chaque phonème

3.4.1. Rappels sur les algorithmes de programmation dynamique

Comme nous l'avons vu dans la partie 3.2, il est impossible d'utiliser l'algorithme de Viterbi avec un système de reconnaissance automatique de la parole Multi-Bandes. Il nous faut donc nous tourner vers d'autres algorithmes de programmation dynamique qui nous permettent de calculer la meilleure segmentation possible de la phrase en phonèmes. Un certain nombre de tels algorithmes étaient utilisés dans les années 80 en reconnaissance de la parole, avant que l'algorithme de Viterbi ne fasse son apparition. Nous en avons déjà présenté deux : la recherche « brute » du meilleur alignement, qui considère tous le alignements possibles et détermine le meilleur d'entre eux, et l'algorithme de « two-level dynamic programming », qui améliore ce premier algorithme en calculant les alignements sur tous les segments possibles au cours d'une première étape. Nous pouvons remarquer que si ce deuxième algorithme a souvent été cité dans la littérature du Multi-Bandes comme une alternative à l'algorithme de Viterbi, quasiment aucun résultat n'a été publié avec un tel algorithme.

De plus, un autre algorithme de programmation dynamique, le « Level Building » [Myers81], n'a quant à lui jamais été cité à notre connaissance dans la littérature du Multi-Bandes, bien qu'il soit également applicable dans un tel cas. Nous allons donc dans un premier temps décrire cet algorithme puis proposer une adaptation des différents principes mis en œuvre dans celui-ci et dans le *two-level* afin de tester notre système Multi-Bandes en mode continu.

Le principe de base du *Level Building* est de construire les meilleurs chemins « niveau par niveau », où un niveau est caractérisé par le nombre de phonèmes inclus dans ce chemin. Ainsi, le *Level Building* ajoute un nouveau phonème à toutes les trames de fin des chemins du niveau précédent pour construire le niveau suivant. L'avantage est que les alignements ne sont plus construits que pour un seul niveau à la fois, alors que dans le cas du *two-level*, ils sont construits sur toute la phrase. Le *Level Building*, une fois qu'il a construit tous les alignements d'un niveau, ajoute ceux-ci aux meilleurs chemins du niveau précédent, puis recommence pour le niveau suivant : il n'y a donc plus besoin de calculer les alignements pour tous les segments possibles au cours d'une étape préalable, car les alignements déjà construits des niveaux précédents ne sont plus jamais reconsidérés.

Une description complète de ces différents algorithmes peut être trouvée dans [Rabiner93].

3.4.2. Description de notre algorithme

L'algorithme que nous avons utilisé est un mélange entre le *two-level dynamic programming* et le *Level Building*. En effet, nous progressons trame par trame, comme dans le cas du *two-level*, mais nous ne construisons les alignements que sur un seul niveau à la fois, sans reconsidérer les alignements des phonèmes précédents dans la phrase, comme le fait le *Level Building*. Cet algorithme est le suivant :

1. Initialement, un chemin ne contenant aucun phonème est créé et s'arrête, par définition, sur la trame –1.
2. Pour chaque trame *t* du signal,
   3. Pour chaque chemin *c* s'arrêtant sur la trame *t*-1 et qui a été retenu,
      4. Pour chaque modèle *m*,
         5. Pour chaque durée *d* possible de *m*,
            6. Un nouveau chemin *c*' est construit en concaténant le modèle *m* de durée *d* avec le chemin *c*. Ce chemin *c*' s'arrête donc sur la trame *t*-1+*d* incluse.
            7. La vraisemblance *s* du modèle *m* sur le segment [*t* , *t*-1+*d*] est alors calculée par le système Multi-Bandes. Ce calcul est possible, car la segmentation est connue.
            8. Le score associé à *c*', noté *S*(*c*'), est alors défini par :

$$S(c') = S(c) + Trans(c,m) + s$$

               où *S*(*c*) est le score final du chemin *c*, et *Trans*(*c*,*m*) représente le logarithme de la probabilité de transition entre *c* et *m*. Nous avons utilisé un bi-gramme pour calculer cette probabilité.
            9. Le chemin *c*' est alors inséré dans la liste ordonnée des *Nbest* meilleurs chemins calculés jusqu'à l'étape courante et se terminant à la trame *t*-1+*d*. Cette liste est ordonnée en fonction du score final de chaque chemin. Il y a une liste par trame du signal. Si *S*(*c*') est inférieur au dernier élément de la liste, le chemin *c*' est supprimé.
10. Lorsque la dernière trame est considérée, l'algorithme s'arrête et choisit le premier chemin de la liste correspondant à cette dernière trame.

Nous pouvons faire plusieurs remarques concernant cet algorithme :

- Dans le cas où *Nbest*=1 et *Trans*(*c*,*m*)=0 (aucune grammaire n'est utilisée et la succession de deux modèles est équiprobable), le chemin optimal est obtenu. Nous pouvons démontrer ceci en considérant le meilleur chemin *c* se terminant sur la trame *t*. Ce chemin est composé d'un dernier modèle *m* débutant sur la trame *t*°, et d'un chemin *a* qui précédé *m* . Le score de *c* est alors :

$$S(c) = S(a) + s(m)$$

  Nous pouvons alors démontrer que *a* est le meilleur chemin se terminant sur *t*°-1, car si ce n'était pas le cas, alors il existerait un chemin *b* tel que $S(b) > S(a)$, et donc $S(b) + s(m) > S(a) + s(m)$.

  Puisque le meilleur chemin *c* se terminant sur *t* se construit en ajoutant un modèle à un meilleur chemin se terminant sur *t*° < *t*, nous sommes alors certain d'avoir construit *c* grâce à l'algorithme décrit ci-dessus. Un raisonnement par récurrence trivial permet de conclure quant à l'optimalité du chemin obtenu sur la dernière trame du signal.

- L'optimalité est obtenue sous réserve qu'un phonème n'ait pas une durée dépassant les bornes autorisées. Si ce cas se produit, le chemin final obtenu sera alors sous-optimal. Il est possible d'assurer l'optimalité globale en choisissant une longueur autorisée qui couvre toute la phrase, mais ceci n'est jamais réalisé en pratique, car la complexité de l'algorithme devient alors exponentielle. Les limites de longueur pour chaque modèle sont calculées sur le corpus d'apprentissage de la manière suivante :

$$Dmin(m) = min_{u \in A(m)}(D(u))$$
$$Dmax(m) = max_{u \in A(m)}(D(u))$$

  où m représente un phonème, $A(m)$ l'ensemble de toutes les occurrences de m sur le corpus d'apprentissage et u un exemple de ce corpus. $D(u)$ représente la durée de u.

  Théoriquement, à cause de cette contrainte, nous ne pouvons assurer que la solution fournie par l'algorithme est optimale, et puisque l'arbre de recherche n'est pas parcouru entièrement, nous devrions plutôt classer la méthode dans la catégorie des algorithmes de recherche en faisceaux. Toutefois, la probabilité qu'un phonème (différent du silence) dépasse la longueur maximale qu'il a sur le corpus d'apprentissage est quasiment nulle en pratique. Et même si cela arrive, alors la solution la plus probablement fournie par l'algorithme sera une succession de deux phonèmes identiques, et ce cas peut être facilement géré en ajoutant un étage supplémentaire qui concatène en un seul segment deux phonèmes identiques successifs. De plus, la même contrainte est utilisée dans la méthode dite de « *Level Building* », et la communauté scientifique considère également que cet algorithme fournit une solution optimale [Rabiner93].

- Dans le cas où *Nbest = Nmod*, et lorsque un bi-gramme entre les phonèmes est utilisé, le chemin optimal est obtenu. En effet, la seule différence par rapport au cas précédent vient du fait que le score d'un chemin dépend de son dernier modèle, du chemin précédent et de la transition entre les deux. Ceci implique que, pour obtenir le chemin optimal, il faut conserver les meilleurs chemins qui se terminent par chacun des modèles possibles. Le raisonnement pour démontrer l'optimalité est alors identique à celui présenté dans le point précédent.

- Néanmoins, dans le cas où un bi-gramme de phonèmes est utilisé, le fait de conserver pour chaque trame *Nmod* chemins possibles augmente considérablement la complexité de l'algorithme. C'est pourquoi nous avons décidé de ne conserver que *Nbest* < *Nmod* chemins pour chaque trame. Ceci revient à élaguer l'arbre de recherche et relève ainsi des méthodes dites de *recherche en faisceaux*.

  Celles-ci sont généralement utilisées pour résoudre l'explosion combinatoire occasionnée par les grammaires modélisant la succession des mots d'une phrase dans les systèmes de reconnaissance grand vocabulaire [Ravishankar96]. Nous retrouvons également cette idée dans le système de reconnaissance de la parole basé sur les modèles de trajectoire développé par Yifan Gong [Gong97].

  L'algorithme considéré élimine certaines solutions avant d'avoir pu calculer le score final de celles-ci. La solution obtenue n'est donc plus optimale. En fait, l'hypothèse à la base de cette méthode est la suivante : il est inutile de poursuivre le calcul des chemins les moins probables, car ceux-ci ont extrêmement peu de chance d'être parmi les meilleures solutions finales. Cette heuristique s'avère presque toujours vraie en pratique, et a déjà été utilisée à plusieurs reprises avec succès [Haton74].

- Un inconvénient de cette méthode vient du fait qu'elle n'autorise qu'une recombinaison linéaire entre les bandes, et n'est pas valable avec une recombinaison par un perceptron. En effet, celui-ci retourne des probabilités *a posteriori*, et s'il est possible de calculer le meilleur alignement possible en maximisant la vraisemblance de chaque chemin, ceci n'est plus vrai lorsque les probabilités *a posteriori* sont considérées. Ainsi, P(*M* | *X*) représente la probabilité que le modèle *M*, plutôt qu'un autre, ait généré les observations *X*, sachant que le segment *X* correspond effectivement à un des modèles en compétition. Mais P(*M* | *X*) ne donne aucune indication quant à la probabilité que le segment *X* ait été généré par un de ces modèles. Le chemin le plus probable est donc composé d'un seul phonème, quelle que soit la longueur de la phrase, car au-delà de deux phonèmes, les probabilités de chaque segment se multiplient entre elles et la probabilité totale baisse.

Ceci est un problème classique et commun à tous les systèmes de reconnaissance de la parole utilisant des probabilités *a posteriori*. Reconsidérons la définition d'une tâche de reconnaissance de la parole en mode continu : si *X* désigne une suite de vecteurs acoustiques observés et si *A* désigne une suite de phonèmes, alors la tâche de reconnaissance consiste à trouver la suite $\hat{A}$ telle que :

$$\hat{A} = \underset{A}{argmax}(P(A|X))$$

où *A* décrit l'ensemble des suites de phonèmes possibles. Définissons une segmentation S_i de la suite *A* comme la suite des indices des trames de début de chaque phonème. Si nous considérons alors toutes les segmentations possibles de *A*, nous avons :

$$\hat{A} = \underset{A}{argmax}(\sum_{S_i \in S} P(AS_i|X))$$

où *S* est l'ensemble de toutes les segmentations possibles de *A*.

Généralement, le terme $P(A \mid X)$ est approché par :

$$P(A|X) = \underset{S}{argmax}\, P(AS|X)$$

car ceci simplifie grandement les calculs. Nous avons donc, dans le cas où le système retourne des probabilités *a posteriori* :

$$\hat{A} = \underset{AS}{argmax}(P(AS|X)) = \underset{AS}{argmax}(P(A|S,X)P(S|X))$$

Nous voyons ainsi que la segmentation a besoin d'être explicitement modélisée à travers le terme $P(S \mid X)$. Nous verrons dans la partie 6.9.3 comment cela peut être réalisé en pratique.

3.4.3. Complexité de l'algorithme

Pour simplifier le calcul de la complexité, nous supposons que tous les phonèmes ont la même durée maximale D_{max} et une durée minimale égale à zéro. Dans ce cas, pour chaque trame et pour chaque chemin se terminant sur cette trame, $Nmod \cdot D_{max}$ nouveaux chemins sont générés. Cependant, l'algorithme de Viterbi n'est appliqué en pratique qu'une seule fois pour chaque modèle, sur une longueur de D_{max} trames : en effet, au cours de ce calcul, il fournit les scores de tous les segments de taille inférieure à D_{max} trames. De plus, ce calcul est appelé pour chaque bande (le coût de la recombinaison est constant). Pour chaque chemin se terminant sur une trame, le coût de la génération des nouveaux chemins est donc $B \cdot Nmod \cdot 2D_{max}$ (cf. la complexité de l'algorithme de Viterbi, page 36). B représente le nombre de bandes.

En comptant également le coût de l'insertion de chacun de ces nouveaux chemins dans la liste ordonnée des chemins déjà calculés, la complexité dans le pire cas de la création des nouveaux chemins devient $B \cdot Nbest \cdot Nmod \cdot 2D_{max}$.

Sachant qu'il faut réaliser ceci pour chaque chemin se terminant sur chaque trame, la complexité totale dans le pire cas est alors :

$$B \cdot Nbest^2 \cdot Nmod \cdot 2D_{max} \cdot T$$

où T représente le nombre de trames du signal.

La complexité de l'algorithme est donc linéaire en fonction du temps, tout comme celle de l'algorithme de Viterbi. Certes, la constante multiplicative devant le T est beaucoup plus grande pour notre algorithme, toutefois le comportement asymptotique de l'algorithme de Viterbi et du nôtre est exactement le même. À moyen et long terme, le coût en pratique de cet algorithme devrait donc devenir négligeable, au vu de la montée en puissance des nouveaux ordinateurs, même si à l'heure actuelle, il reste trop élevé pour envisager une application temps-réel.

3.4.4. Comparaison avec les autres algorithmes de programmation dynamique

✓ *Comparaisons générales avec les algorithmes de programmation dynamique*

- Le principe de récurrence que nous avons utilisé dans notre algorithme n'est pas un principe nouveau en programmation dynamique. En effet, la plupart des algorithmes de programmation dynamique utilisés en reconnaissance automatique de la parole utilisent un principe de récurrence similaire qui consiste à calculer les meilleurs chemins au rang *n* à partir des meilleurs chemins au rang *n*-1.
- Les algorithmes de programmation dynamique en reconnaissance de la parole ont été initialement conçus pour fonctionner sans grammaire, ou avec une grammaire déterministe et finie (*Finite State Network*). C'est pourquoi seule la meilleure solution sur chaque trame est conservée. Notre algorithme utilise quant à lui une grammaire statistique modélisant la probabilité de transition entre les mots (ou phonèmes). Nous avons alors dû conserver sur chaque trame *Nbest* meilleurs chemins, et non plus seulement un seul, afin d'obtenir une complexité raisonnable.

✓ *Comparaison avec le Level Building*

Le *Level Building* est un algorithme asynchrone. En effet, il revient fréquemment en arrière au cours de son exécution, car les trames de début d'un niveau sont souvent antérieures aux trames de fin du niveau précédent. Notre algorithme est quant à lui synchrone, car il progresse trame par trame sans jamais revenir en arrière.

✓ *Comparaison avec le Two-Level Dynamic Programming*

- Le *two-level* construit chaque alignement qu'il considère sur toute la phrase, comme l'algorithme de recherche « brut ». Notre algorithme ne construit les alignements que sur un seul niveau à la fois tout comme le *Level Building*.
- Le *two-level* calcule le score pour tous les segments possibles sur toute la phrase, alors que nous avons utilisé dans notre algorithme une contrainte sur la durée des segments, qui est comprise entre une durée minimale et une durée maximale, comme c'est le cas pour le *Level Building.*

✓ *Comparaison avec l'algorithme de Viterbi*

La différence principale entre notre algorithme et l'algorithme de Viterbi réside dans le fait que le second calcule le meilleur chemin trame par trame, tandis que le premier calcule le meilleur chemin segment par segment. Ainsi, pour Viterbi, une trame est ajoutée à chaque chemin, et le nouveau score est calculé, tandis que pour notre algorithme, un phonème complet est ajouté à chaque chemin, puis son score est calculé.

Ceci a pour inconvénient de devoir tester plusieurs longueurs possibles pour chaque segment ajouté, ce que n'a pas à faire l'algorithme de Viterbi, car la longueur des trames est fixée. C'est pourquoi notre algorithme est plus coûteux que celui de Viterbi, mais ce surcoût peut néanmoins être maîtrisé en introduisant une longueur maximale autorisée pour les phonèmes.

De plus, pour Viterbi, la décision quant au choix du meilleur chemin se terminant sur la trame courante se fait immédiatement après que ces chemins aient été calculés, alors que dans notre algorithme, le calcul des scores des chemins « précède » le choix du meilleur chemin de plusieurs trames (autant que la longueur du dernier phonème du chemin considéré). Ceci impose de conserver le score de ces chemins pendant plusieurs trames, ce qui n'est pas nécessaire avec l'algorithme de Viterbi; Il y a donc également un léger surcoût en espace mémoire, qui peut également être réduit grâce à l'introduction d'une recherche en faisceau (cf. ci-dessus).

Ces inconvénients sont cependant mineurs devant l'avantage obtenu, i.e. la possibilité de réaliser une reconnaissance Multi-Bandes en mode continu, dans laquelle les bandes sont indépendantes à l'intérieur de chaque modèle et utilisant une recombinaison sur les scores.

Chapitre 4
Étude des bandes de fréquence

4.1. Introduction

Nous avons introduit dans le chapitre deux le principe général du traitement de la parole Multi-Bandes. Un certain nombre de problèmes liés à ce principe ont ainsi été exposés et des solutions ont été proposées. Nous allons parler dans ce chapitre de problèmes plus « précis », dans le sens où ces problèmes sont liés à l'implémentation de l'algorithme et non à la méthode sous-jacente. En fait, dans tout système, il faut faire des choix d'implémentation, et nous allons décrire les nôtres dans les prochaines parties. Ce chapitre est ainsi consacré à l'étude des bandes de fréquence que l'on peut considérer comme des « experts » dont il faut recombiner les avis. L'étude du module de recombinaison lui-même est réalisée dans le chapitre 5.

Nous allons débuter ce chapitre en présentant les classifieurs phonétiques que nous avons utilisés dans chaque bande. Puis, nous allons définir les bandes elles-mêmes et étudier leur comportement en réalisant quelques expériences simples sur celles-ci. Ces tests sont présentés ici et non pas au chapitre 6, car celui-ci est dédié aux expériences du système global. Or, il n'est pas question du système final dans ce chapitre, et seul le comportement des bandes est étudié. De plus, les tests réalisés sur les classifieurs utilisés dans les bandes ne sont pas exhaustifs, car ce sont des reconnaisseurs classiques et leur comportement est bien connu et détaillé dans un grand nombre d'ouvrages [Calliope89] [Rabiner93]. Nous avons donc voulu présenter ici essentiellement en quoi le comportement de ces classifieurs diffère du cas général lorsque seule une bande de fréquences est considérée.

4.2. Présentation des reconnaisseurs

4.2.1. Définition des reconnaisseurs

Les reconnaisseurs que nous avons utilisés sont des Modèles de Markov Cachées (HMM) du second ordre. Nous avons utilisé ceux-ci plutôt que des HMM du premier ordre, qui sont beaucoup plus répandus, parce qu'ils ont déjà été soigneusement étudiés dans notre laboratoire, notamment par Jean-François Mari qui a développé la plus grande partie du code source les concernant [Mari97]. De plus, leurs performances sont supérieures à celles des HMM du premier ordre, car ils modélisent mieux, entre autre, la durée des phonèmes. C'est pourquoi, dans la suite de ce mémoire, nous désignons sous le terme « système de référence » un HMM du second ordre utilisant l'ensemble du spectre.

Une description complète de ces modèles peut être trouvée dans [Mari97], mais nous allons présenter brièvement les points essentiels qui différencient ces modèles des modèles classiques. Il est à noter que dans la majorité des systèmes Multi-Bandes développés dans les autres laboratoires [Berthommier98] [Bourlard97] [Mirghafori99], les reconnaisseurs utilisés sont des modèles hybrides composés d'un HMM du premier ordre qui décide de l'alignement temporel, mais dont les densités de probabilités exprimées dans chaque état sont calculées grâce à un perceptron. Ce modèle est plus performant que le HMM classique, en grande partie car il permet de modéliser des densités de probabilités quelconques et non pas seulement gaussiennes ou multi-gaussiennes. Mais il n'est pas judicieux pour étudier un nouveau paradigme comme celui du Multi-Bandes que tous les laboratoires utilisent des systèmes similaires, et notre choix contribue en conséquence à la variété des systèmes Multi-Bandes existants.

Les reconnaisseurs que nous avons utilisés sont donc des HMM du second ordre composés de 3 états, dont la topologie respecte celle présentée dans la figure 4.1. Les densités de probabilités dans chaque état sont calculées par un mélange de gaussiennes, dont le nombre varie d'un état à l'autre et est déterminé au cours de l'apprentissage, comme cela est expliqué dans [Mari96].

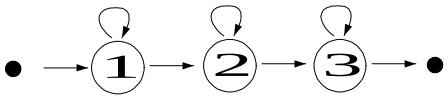

Figure 4.1 : Topologie d'un HMM.

4.2.2. Brève présentation des HMM du second ordre

En fait, les algorithmes qui sont utilisés dans les HMM du second ordre sont très peu différents de ceux utilisés classiquement, à savoir l'algorithme de Viterbi pour le décodage et l'algorithme de Baum-Welch pour l'apprentissage. L'algorithme de Viterbi pour les HMM du premier ordre calcule itérativement la probabilité du meilleur alignement entre les instants 1 et *t* et s'achevant sur l'état s_j par :

$$\delta_t(j) = \max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(O_t) \qquad 1 \leqslant j \leqslant N \qquad 2 \leqslant t \leqslant T$$

où *N* est le nombre d'états total, *T* la durée totale de la phrase, a_{ij} la probabilité de transition entre les deux états *i* et *j*, et $b_j(O_t)$ la densité de probabilité que le vecteur d'observation au temps *t* ait été émis par l'état *j*.

L'adaptation de cette formule itérative pour un HMM du second ordre donne :

$$\delta_t(j,k) = \max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i,j) \cdot a_{ijk}] \cdot b_k(O_t) \qquad 1 \leqslant j \leqslant N \qquad 1 \leqslant k \leqslant N \qquad 3 \leqslant t \leqslant T$$

a_{ijk} représente la probabilité d'emprunter la transition (*j*,*k*) sachant l'état précédent *i*.

Comme nous le voyons, ces deux formules sont très similaires, mise à part l'introduction d'une dépendance supplémentaire envers un autre état pour δ, qui représente maintenant la probabilité du meilleur alignement entre les instants 1 et *t* et s'achevant sur s_j en *t*-1 et sur s_k en *t*.

De même, l'algorithme *forward-backward* réalisant l'apprentissage des HMM s'écrit pour le second ordre :

$$\alpha_t(j,k)=\sum_{i=1}^{N}\alpha_{t-1}(i,j)\cdot a_{ijk}\cdot b_k(O_{t+1}) \qquad 2\leqslant t\leqslant T-1 \qquad 1\leqslant j\leqslant N \qquad 1\leqslant k\leqslant N$$

$\alpha_t(j,k)$ représente la probabilité d'avoir simultanément la suite d'observations partielle $O_1,\ldots,O_t$ et la transition $s_j \rightarrow s_k$ entre les instants *t*-1 et *t* d'une part et *t* et *t*+1 d'autre part.

Au cours de l'apprentissage, ce premier calcul *forward* est suivi d'un calcul *backward* qui s'exprime de la même manière :

$$\beta_t(i,j)=\sum_{k=1}^{N}\beta_{t+1}(j,k)\cdot a_{ijk}\cdot b_k(O_{t+1}) \qquad 2\leqslant t\leqslant T-1 \qquad 1\leqslant j\leqslant N \qquad 1\leqslant k\leqslant N$$

$\beta_t(i,j)$ représente la probabilité de la suite d'observations depuis *t*+1 à *T*, connaissant le modèle et la transition $s_i \rightarrow s_j$ entre les instants *t*-1 et *t*.

Finalement, on définit :

$$\eta_t(i,j,k)=\frac{\alpha_t(i,j)\,a_{ijk}\,b_k(O_{t+1})\,\beta_{t+1}(j,k)}{P(O|M)} \qquad 2\leqslant t\leqslant T-1$$

qui représente la probabilité de la transition $s_i \rightarrow s_j \rightarrow s_k$ entre les instants *t*-1 et *t*+1 pendant l'émission de la suite d'observations.

De même, nous définissions :

$$\xi_t(i,j)=\sum_{k=1}^{N}\eta_t(i,j,k)$$

et

$$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$$

L'apprentissage se termine alors en réestimant les paramètres grâce aux formules suivantes :

$$\overline{a_{ijk}}=\frac{\sum_t \eta_t(i,j,k)}{\sum_{k,t}\eta_t(i,j,k)}$$

$$\overline{\mu_i}=\frac{\sum_t \gamma_t(i)O_t}{\sum_t \gamma_t(i)}$$

$$\overline{\sum_i}=\frac{\sum_t \gamma_t(i)(O_t-\mu_i)(O_t-\mu_i)^t}{\sum_t \gamma_t(i)}$$

Des explications plus détaillées concernant ces algorithmes, ainsi que ceux ayant été utilisés pour répartir les données d'apprentissage dans plusieurs gaussiennes par état, peuvent être trouvées dans [Mari96].

4.3. Caractérisation des bandes de fréquence

La conception d'un système Multi-Bandes dépend de la réponse que l'on apporte aux deux questions suivantes : combien de bandes faut-il utiliser et quelles limites fréquentielles doivent-elles posséder ? Nous allons considérer ces deux questions dans cette partie 4.3, ainsi que le problème du choix des coefficients acoustiques à utiliser dans les bandes.

4.3.1. État de l'art sur l'étude des limites des bandes

Plusieurs travaux se sont déjà penchés sur l'étude des bandes. C'est pourquoi nous n'avons pas réalisé à nouveau cette étude, mais nous nous sommes plutôt appuyés sur celles existantes pour construire notre système.

Le système Multi-Bandes de Duchnowsky utilise quatre bandes dont les limites sont [100-700 Hz], [700-1500 Hz], [1500-3000 Hz] et [3000-4500 Hz]. Ces bandes ont été choisies afin que chacune d'entre elles englobe grossièrement un formant des voyelles de l'anglais. Une autre solution a été proposée par Bourlard et Dupont [Bourlard96] qui ont testé leur système avec 3, 4 et 6 bandes. Les meilleurs résultats qu'ils ont obtenus correspondent à l'utilisation de 4 bandes, dont les limites sont [0-901 Hz], [797-1661 Hz], [1493-2547 Hz] et [2298-4000 Hz]. Ces limites sont calculées à partir des limites des bandes critiques. De même, Tibrewala et Hermansky ont étudié 2, 4 et 7 bandes également composées de plusieurs bandes critiques, leurs meilleurs résultats ayant été obtenus cette fois avec deux bandes seulement, dont les limites sont [0-1140 Hz] et [1046-4000 Hz]. Des résultats presque similaires ont été obtenus par Okawa et al. qui ont testé 2, 3, 4 et 6 bandes et qui ont obtenu les meilleurs résultats avec 2 bandes dont les limites correspondent à une partition égale de l'échelle Mel, à savoir [0-1850 Hz] et [1691-8000 Hz].

Quelques études ont également été menées avec un grand nombre de bandes, malgré les difficultés de conception que cela pose. Ainsi, Hermansky et al. ont montré très récemment [Hermansky98] que la conjonction de 15 bandes très étroites et d'une bande codant le spectre entier permettait d'obtenir de bons résultats. De même, Besacier a testé dans sa thèse, pour une tâche de reconnaissance du locuteur [Besacier98a], 21 bandes, chacune d'entre elle étant composée de 3 des 24 filtres utilisés pendant l'analyse acoustique du signal, avec recouvrement partiel entre elles. Ceci constitue d'ailleurs un des résultats intéressants des travaux de Besacier, dont les expériences semblent montrer que plus grand est le recouvrement entre les bandes, meilleurs sont les résultats.

4.3.2. Définition des limites des bandes dans notre système

Beaucoup de ces résultats semblent contradictoires, dans la mesure où certains préconisent un faible nombre de bandes tandis que d'autres s'appuient sur les travaux de psycho-acoustiques pour utiliser un grand nombre de bandes. Chaque système possédant ses propres particularités, il semble donc difficile de comparer toutes ces études entre elles. Toutefois, un compromis autour d'un système à quatre bandes semble s'être instauré dans la communauté de recherche, et nous avons décidé d'utiliser également dans notre propre système quatre sous-bandes, auxquelles nous avons ajouté une cinquième bande composée du spectre complet[5]. Le choix de quatre sous-bandes est motivé en partie par les raisons que nous avons déjà invoquées ci-dessus, mais également parce que nous pensons qu'un plus petit nombre de bandes n'aurait pas permis d'utiliser pleinement les possibilités du paradigme Multi-Bandes, alors qu'un plus grand nombre de bandes aurait rendu le système trop complexe et aurait surtout allongé considérablement les temps d'apprentissage, voire de tests.

En ce qui concerne les limites fréquentielles de ces quatre sous-bandes, nous avons dans un premier temps réalisé un système utilisant les mêmes limites fréquentielles que Bourlard. Ensuite, essentiellement pour des raisons pratiques, nous avons modifié ces limites de sorte à ce que chaque bande contienne le même nombre de coefficients. Nous avons donc équitablement réparti les différents filtres utilisés au cours de l'analyse acoustique et qui respectent cette échelle dans les différentes bandes, comme le montre la figure 4.2. Ainsi, nos sous-bandes possèdent les limites suivantes : [0-538 Hz], [461-1000 Hz], [923-2823 Hz] et [2374-7983 Hz].

Nous avons comparé les premières limites testées avec celles-ci, mais nous n'avons pu déceler aucune différence significative dans les résultats. Aussi nous sommes-nous ensuite contentés d'utiliser ces nouvelles limites.

4.3.3. État de l'art sur l'étude des paramètres acoustiques

Classiquement, les HMM utilisent des coefficients acoustiques qui correspondent à des coefficients cepstraux répartis sur une échelle logarithmique, dite « échelle Mel ». Cette échelle correspond aux bandes critiques de notre oreille interne et est donc adaptée à l'audition. Ces coefficients sont désignés sous le terme de MFCC (*Mel Features Cepstral Coefficients*). Une description plus détaillée de ceux-ci ainsi que de la plupart des autres coefficients utilisés en reconnaissance de la parole peut être trouvée dans [Rabiner78]. Ces MFCCs sont utilisés dans la plupart des systèmes actuels, car ils se sont révélés être les meilleurs pour modéliser l'information acoustique contenue dans un signal de parole. Mais la question de l'efficacité de tels coefficients se pose lorsque seulement une partie du spectre est utilisée, ce qui est le cas dans une seule bande.

5 La justification de l'utilisation conjointe de quatre sous-bandes et du spectre complet est donnée dans la partie 6.3.1 qui définit notre système, page 80.

Si nous reprenons les différents systèmes Multi-Bandes existants que nous avons déjà cités, nous pouvons tout d'abord nous référer à celui de Duchnowsky qui a ainsi étudié plusieurs types de paramètres acoustiques, dont l'énergie dans un banc de filtre, les paramètres LPC, les coefficients cepstraux et les paramètres d'autocorrélation du signal. Il désapprouve également l'utilisation d'une échelle logarithmique pour les filtres d'analyse à l'intérieur d'une même bande, car, selon lui, cette échelle est totalement inutile au vu de la « faible » largeur des bandes. Après avoir comparé les résultats obtenus avec ces différents paramètres, il conclut que les meilleurs paramètres sont les coefficients cepstraux. De même, une comparaison réalisée par Tibrewala et Hermansky entre des coefficients calculés à partir de l'énergie dans les bandes critiques et les coefficients cepstraux montre que ces derniers donnent les meilleurs résultats. Enfin, Bourlard et Dupont ont également comparé les performances de l'énergie dans les bandes critiques et des coefficients lpc-cepstraux et ont conclu en faveur des coefficients cepstraux. Pour sa part, Mirghafori a opté pour l'utilisation de coefficients RASTA-PLP [Hermansky94] qui sont théoriquement plus robustes au bruit que les autres coefficients.

4.4. Étude des bandes de notre système

4.4.1. Introduction

Nous allons dans cette partie étudier plus précisément le comportement de chacune de ces bandes. Tous les tests qui sont réalisés ici consistent en des tests de reconnaissance des phonèmes de l'anglais sur TIMIT lorsque la segmentation est connue[6]. Nous avons choisi ce type d'expériences car elles sont très simples, et elles permettent d'isoler le phénomène que l'on veut observer, i.e. essentiellement le type d'information phonétique apparaissant dans les zones fréquentielles. En effet, une expérience de reconnaissance en mode continu aurait introduit des effets indésirables liés aux erreurs de segmentation et à l'interaction entre les phonèmes.

Toutefois, bien que le corpus TIMIT soit l'un des mieux adaptés pour réaliser ce genre d'études, car la segmentation est relativement bien faite et la prononciation des locuteurs est souvent irréprochable, il existe tout de même un grand nombre d'imprécisions, liées notamment aux erreurs de segmentations dues à un étiquetage manuel imparfait, ou encore aux phénomènes de coarticulation qui peuvent modifier totalement les propriétés acoustiques d'un phonème. Il faut donc tenir compte de ces imprécisions dans l'interprétation des résultats.

4.4.2. Présentation des bandes de notre système

- Notre système Multi-Bandes est composé de cinq bandes, les quatre premières étant en réalité des sous-bandes dont les limites sont [0-538 Hz], [461-1000 Hz], [923-2823 Hz] et [2374-7983 Hz], comme cela a été expliqué ci-dessus. La cinquième bande est composée du spectre complet[7].

6 cf. partie 2.4.3.
7 cf. partie 6.3.1.

- Les exemples trouvés dans la littérature donnent presque toujours la préférence aux coefficients cepstraux. Nous avons nous-mêmes décidé d'utiliser ces coefficients. Nous calculons ceux-ci en décomposant tout d'abord le signal par une série de filtres triangulaires, puis nous répartissons équitablement ces filtres dans les différentes bandes, comme cela est montré sur la figure 4.2. Les coefficients cepstraux sont alors calculés en appliquant la transformation discrète en cosinus (DCT) aux coefficients énergétiques de chaque filtre. Puisque les filtres originaux respectent l'échelle Mel, les coefficients calculés sont effectivement des MFCCs et non simplement des coefficients cepstraux. Nous avons fait ce choix car, malgré le fait que l'échelle Mel puisse être inutile à l'intérieur même d'une bande comme l'a fait remarquer Duchnowsky, elle ne nuit en rien aux résultats et le calcul des coefficients reste ainsi assez simple. Le nombre de coefficients de chaque bande est défini dans la partie 6.3.1.

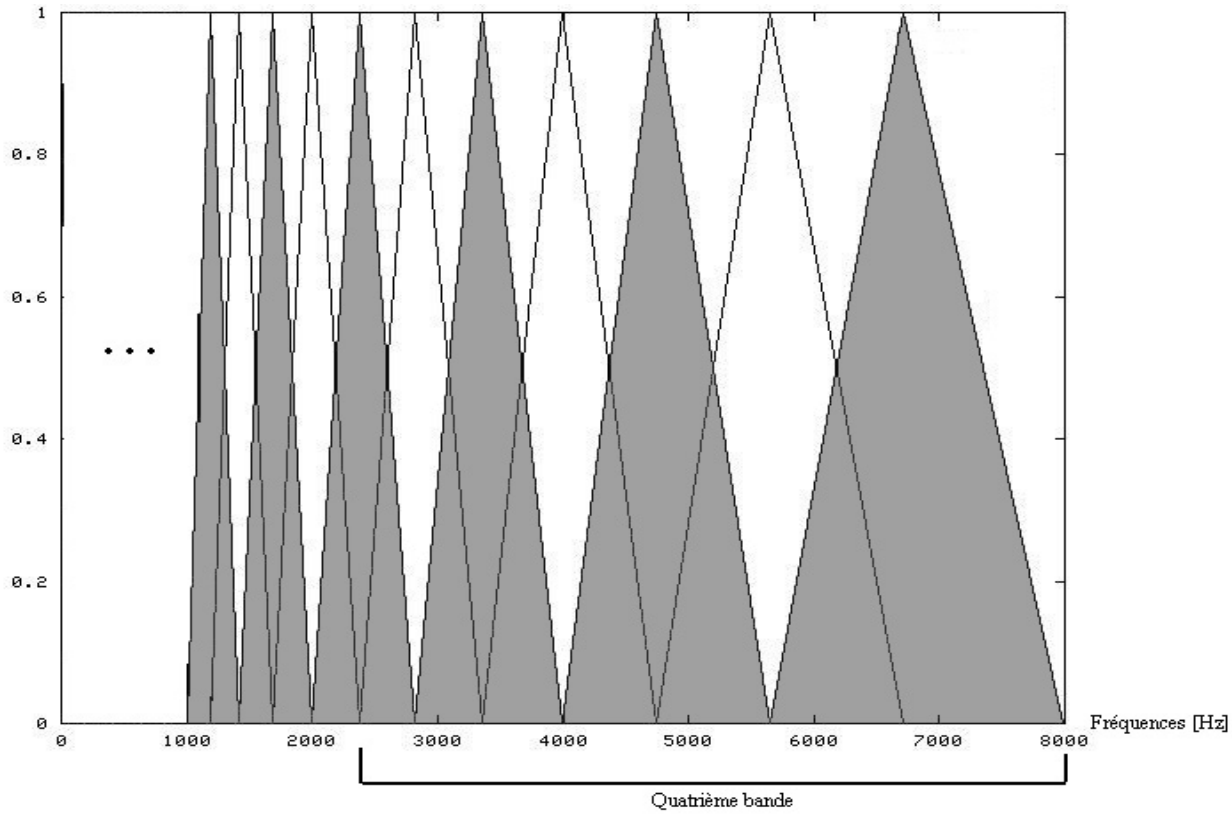

Figure 4.2 : Filtres utilisés pour l'analyse en cepstres et répartition de ces filtres dans les bandes

4.4.3. Étude phonétique

✓ ***Pourquoi cette analyse ?***

Comme nous l'avons déjà signalé, toutes les sous-bandes ne peuvent pas différencier tous les phonèmes[8] aussi clairement que ne le fait le système de référence. En effet, chaque phonème est caractérisé par son spectre fréquentiel à court terme et par l'évolution temporelle de ce dernier, et lorsque le signal est filtré, une partie des traits caractéristiques du phonème peuvent également disparaître. Néanmoins, un certain nombre de ces indices, comme le formant d'un son voisé, affectent une région très limitée fréquentiellement et qui peut rester entièrement contenue dans la sous-bande considérée. Ces formants correspondant physiologiquement aux fréquences de résonance du conduit vocal, nous comprenons qu'elles restent relativement stables d'une prononciation à l'autre, au moins pour la même personne, et sont très importantes dans la reconnaissance de tels sons. Ces indices ne sont toutefois pas les seuls à être utiles pour l'identification d'un phonème. Le contexte, par exemple, est également très important et a une grande influence sur la réalisation acoustique d'un son. Nous avons cependant choisi de ne pas considérer celui-ci dans l'état actuel de notre système, car son intégration relève plutôt de l'amélioration des modèles phonétiques, et n'est donc pas utile pour la validation d'un nouveau principe comme le Multi-Bandes. Ainsi, dans le cadre de la modélisation hors-contexte des phonèmes, les caractéristiques principales de ceux-ci s'apparentent effectivement à des indices spectraux et à leur évolution dans le temps.

Le paradigme Multi-Bandes étant essentiellement fondé sur un découpage fréquentiel du spectre, nous nous intéressons ici uniquement à l'aspect statique de ces indices, leur aspect dynamique étant modélisé par les HMM. Or, comme nous venons de le voir, le filtrage a pour effet d'isoler certains de ces indices acoustiques. Ceci a vraisemblablement pour conséquence de faciliter la reconnaissance de ces phonèmes dans les bandes concernées, et inversement de rendre cette reconnaissance très difficile dans les autres bandes. C'est pourquoi nous pouvons supposer que chaque bande est adaptée à la reconnaissance d'un type spécifique de phonèmes.

Nous voulons donc tout d'abord vérifier expérimentalement cette hypothèse, c'est-à-dire savoir s'il est vrai que certaines bandes sont meilleures que d'autres pour distinguer un phonème particulier. Ensuite, nous voulons connaître quels sont les phonèmes les plus facilement identifiables dans chaque bande, et à l'inverse lesquels sont les moins faciles à détecter. Une réponse intuitive peut être donnée à cette question, basée sur un raisonnement qui met en jeu les propriétés de chaque phonème, mais il n'est pas toujours évident qu'un système réel se comporte de la sorte, notamment à cause de multiples facteurs qui ne sont pas d'ordre phonétiques mais relèvent plutôt des stratégies de modélisation et d'implémentation. Nous voulons donc vérifier si les bandes se conforment bien à nos prédictions et si les aspects phonétiques ne sont pas marginaux par rapport aux autres caractéristiques de notre système. Enfin, cette étude constitue une étude préliminaire qui est poursuivie plus en profondeur dans le chapitre 7, où nous tenterons d'utiliser tout ce que nous avons compris du comportement des bandes afin d'améliorer notre système. De plus, nous étudierons alors quels sont les véritables fondements et implications du paradigme Multi-Bandes, qui est trop souvent vu simplement comme une architecture parallélisant plusieurs HMM. Nous pensons au contraire que ce paradigme est bien plus que cela, et qu'il peut aller jusqu'à modifier en profondeur le principe de reconnaissance des systèmes automatiques actuels.

8 Une liste des phonèmes utilisés est donnée en annexe 2

✓ ***Étude de la première sous-bande : [0 ... 538 Hz]***

La figure 4.3 montre les taux de reconnaissance par phonème de la première sous-bande, c'est-à-dire celle dont les limites sont comprises entre 0 et 538 Hz, et les taux de reconnaissance par phonème du système de référence. Ces taux de reconnaissance ont été calculés sur le corpus d'apprentissage, afin que les différences soient significatives. En effet, certains phonèmes n'apparaissent pas assez souvent dans le corpus de test pour que l'on puisse interpréter les résultats obtenus.

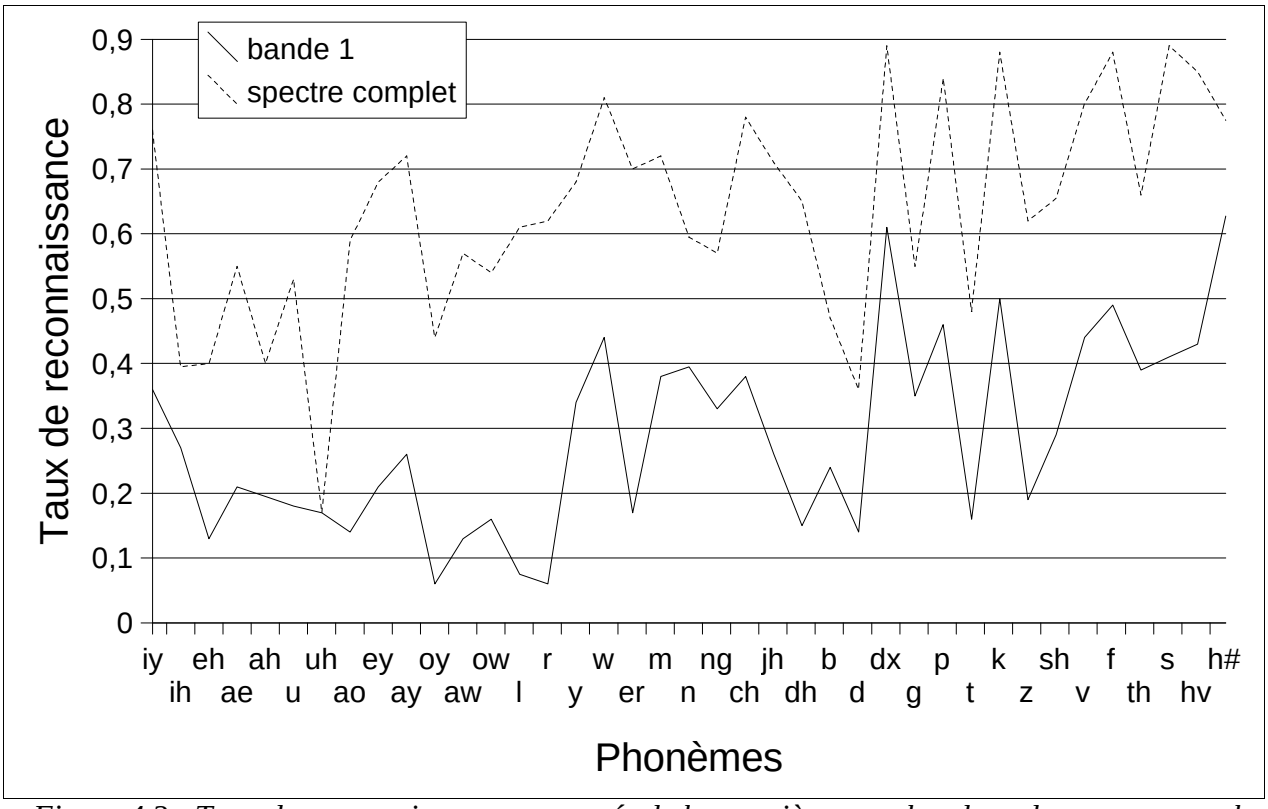

Figure 4.3 : Taux de reconnaissance comparés de la première sous-bande et du spectre complet pour chaque phonème

L'axe des abscisses sur cette figure représente tous les phonèmes de l'anglais utilisés, après regroupement de ceux-ci en 39 phonèmes (cf. annexe 2). Ce regroupement est celui classiquement utilisé et qui a été proposé par K. F. Lee dans sa thèse. Regrouper ces phonèmes permet d'une part de ne pas surcharger le graphique en réduisant le nombre de données à comparer, et d'autre part augmente le nombre d'exemples utilisés par phonème pour calculer ces taux de reconnaissance, donc leur représentativité.

On remarque que, quel que soit le phonème observé, le taux de reconnaissance du système de référence est presque toujours supérieur au taux de reconnaissance de la première sous-bande. Ceci nous amène, d'une part à constater la supériorité très nette du système de référence sur les sous-bandes en milieu non bruité, et d'autre part à considérer que les taux de reconnaissance ne sont pas parfaitement adaptés à une analyse des caractéristiques des sous-bandes. En effet, le taux de reconnaissance fait intervenir tous les modèles, en en opposant un seul par rapport à tous les autres. Dans ces conditions, il est normal que le système de référence obtienne de meilleurs résultats que les sous-bandes, car il possède beaucoup plus d'information qu'une sous-bande pour distinguer un phonème de tous les autres.

La deuxième remarque que nous pouvons faire est que la forme des courbes est grossièrement la même pour le système de référence et la première bande. Ceci signifie que, pour notre système, les phonèmes qui sont difficiles à reconnaître lorsque le spectre complet est considéré le sont également lorsque seule une partie du spectre est utilisée. Ceci, à notre avis, est dû à deux raisons : la première est liée au fait que la topologie des classifieurs utilisés sur le spectre complet et sur une sous-bande est la même ; ce qui signifie que si le classifieur utilisé ne parvient pas à modéliser un indice acoustique caractéristique d'un phonème à cause de sa topologie, alors il ne le pourra pas quelle que soit la zone du spectre qui lui est attribuée. La deuxième raison provient du fait que certains phonèmes sont intrinsèquement plus difficiles à reconnaître que les autres, soit par exemple parce qu'il y a peu de données d'apprentissage concernant ces phonèmes, soit parce qu'ils peuvent être facilement confondus avec d'autres qui leur sont proches.

Nous voyons donc que l'on peut classer les difficultés des classifieurs en plusieurs catégories :

1. Les difficultés liées au ***type d'information*** caractéristique d'un phonème : si cette information n'est pas reconnaissable par la topologie du classifieur utilisé, alors le paradigme Multi-Bandes ne peut rien faire pour augmenter les performances du système.

2. Les difficultés liées à la ***base d'apprentissage*** : si notre système dispose de peu d'exemples concernant un phonème, sa modélisation sera grossière et les taux de reconnaissance assez faibles, que l'on utilise le paradigme Multi-Bandes ou pas.

3. Les difficultés liées à la ***confusion naturelle*** qui peut exister entre certains phonèmes : dans ce cas, il est peu probable que le Multi-Bandes puisse améliorer les résultats pour ces phonèmes, mais nous pensons néanmoins qu'en divisant la tâche et les indices de reconnaissance entre plusieurs classifieurs, il est peut-être possible de gagner quelques pourcentages de reconnaissance sur ce type d'erreurs.

4. Les difficultés liées à la trop ***grande quantité d'information*** présente dans le signal. Les classifieurs ont parfois toutes les données pour identifier un indice permettant de reconnaître un phonème, mais ne le peuvent pas en pratique, car ils ont en fait trop de données à leur disposition, et ne parviennent pas à en extraire l'information pertinente.

C'est cette quatrième catégorie d'erreurs qui nous intéresse tout particulièrement, car nous pensons que le Multi-Bandes permet, en réduisant la quantité d'information mise à la disposition des classifieurs, de faciliter la tâche de ceux-ci qui ont alors moins de chance d'omettre un indice important. Nous pouvons remarquer d'ailleurs que le paradigme Multi-Bandes, dans sa version la plus élaborée telle qu'elle est présentée au chapitre 7, peut également avoir une certaine influence sur les erreurs dues à la troisième catégorie, mais nous en reparlerons dans le chapitre 7.

Comme il est difficile d'analyser le comportement détaillé de la première sous-bande et du spectre complet sur la figure 4.3, nous avons représenté sur la figure 4.4 la différence des taux de reconnaissance du système de référence et de la première sous-bande. L'intérêt d'utiliser la différence entre les deux courbes est essentiellement d'éliminer les trois premières causes de variation du taux de reconnaissance citées ci-dessus. Intuitivement, cette courbe peut s'analyser ainsi : lorsque la valeur portée par la courbe est haute, alors le phonème noté en abscisses est mal reconnu par la première sous-bande. A l'inverse, lorsque la valeur portée par la courbe est basse, alors le phonème correspondant est bien reconnu par la première sous-bande.

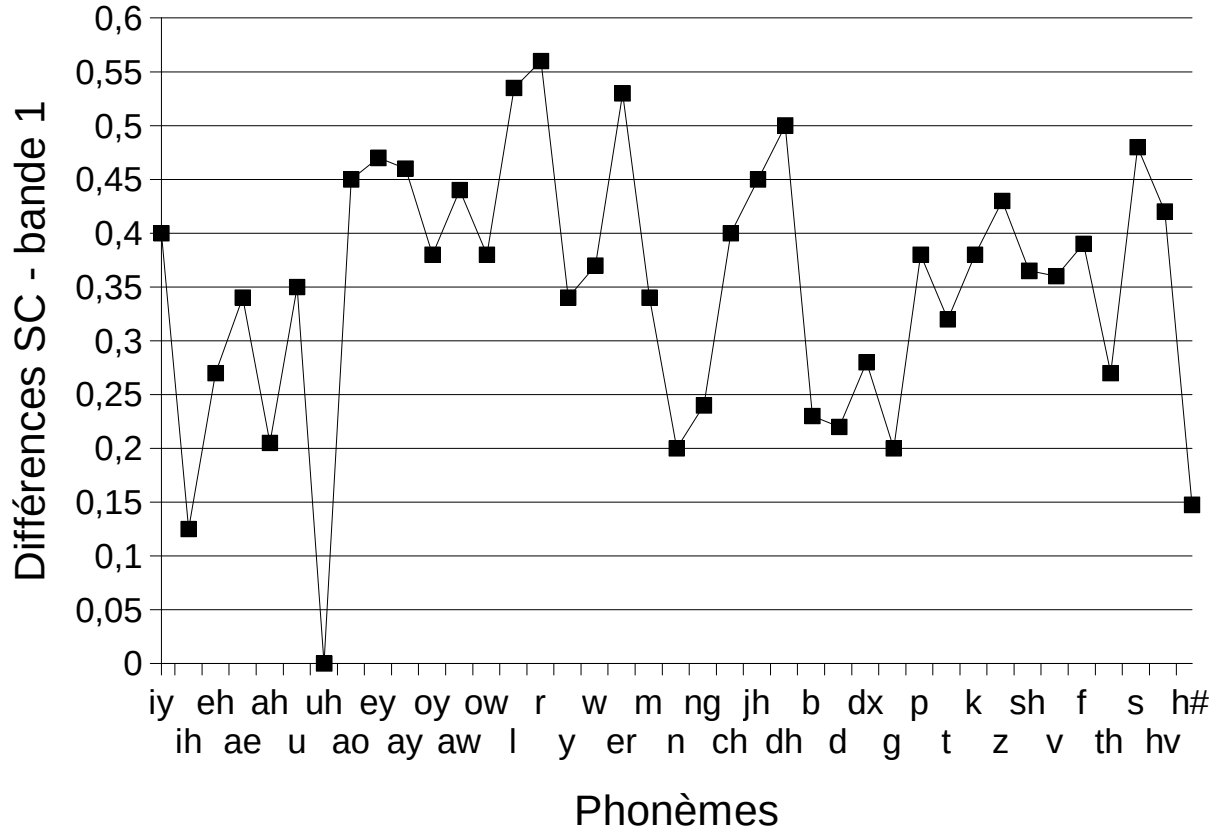

Figure 4.4 : Différences des taux de reconnaissance du spectre complet et de la première sous-bande pour chaque phonème.

Une première analyse nous permet de distinguer plusieurs phonèmes assez bien reconnus par la première sous-bande :

- La voyelle /ih/ de "b**i**t" ;
- La voyelle /ah/ de "b**u**t" ;
- La voyelle /uh/ de "b**oo**k" ;

- La consonne /n/ de "**n**oon" ;
- La consonne /ng/ de "si**ng**" ;
- La consonne /b/ de "**b**eet" ;
- La consonne /d/ de "**d**ay" ;
- La consonne /g/ de "**g**ay" ;
- Le silence.

Ces résultats correspondent relativement bien aux caractéristiques connues des phonèmes. Ainsi, les deux voyelles /uh/ et /ih/ possèdent chacune un formant F1 situé dans les très basses fréquences. De même, nous trouvons dans cette liste toutes les occlusives nasales voisées /n/, /ng/, /b/, /d/ et /g/ (à l'exception du /m/), qui possèdent également un formant dans les très basses fréquences. L'interprétation est toutefois moins claire en ce qui concerne le silence. Peut-être ce résultat vient-il du fait que le bruit affecte peu les basses fréquences et donc que celles-ci sont privilégiées pour détecter le silence.

Au contraire, les phonèmes suivants sont très mal reconnus par la première bande :

- Le diphtongue /ey/ de "b**ai**t" ;
- Le diphtongue /ay/ de "b**i**te" ;
- La consonne /l/ de "**l**ay" ;
- La consonne /r/ de "**r**ay" ;
- La consonne /er/ de "bi**r**d" ;
- La consonne /th/ de "**th**en" ;
- La consonne /s/ de "**s**ea" ;

Nous voyons que la première sous-bande a beaucoup de mal à reconnaître particulièrement les diphtongues et les liquides, les premières étant en règle générale caractérisées par un « glissement » des indices fréquentiels d'une fréquence aiguë vers une fréquence grave, ou inversement, et les secondes ayant des indices relativement aigus. Nous pouvons donc concevoir que les basses fréquences, seules, ne soient pas adaptées à détecter ce type de phénomènes. L'interprétation est plus facile pour les son /th/ et /s/ qui se situent essentiellement dans les hautes fréquences.

Il est ainsi possible d'exhiber assez facilement des différences de comportement entre les bandes, mais leur interprétation n'est pas toujours simple, car les phénomènes qui interviennent sont eux-mêmes complexes.

Cette étude est poursuivie en annexe 3 pour les autres sous-bandes. Nous avons préféré séparer cette partie du corps de la thèse car elle est assez longue, et nous ne souhaitons pas qu'elle fasse perdre au lecteur la vision globale qu'il peut avoir du document.

✓ *Conclusion*

En réalisant cette étude, nous nous étions fixés quatre objectifs :

- Mettre en évidence les différences de comportement des classifieurs en fonction des bandes et des phonèmes ;

- Essayer d'établir un critère phonétique qui nous permette de décider à quel phonème correspond un segment de signal de parole donné. Ce critère doit prendre sa décision en se basant sur les réponses des différentes bandes du système et en s'appuyant sur la connaissance que nous avons pu extraire de cette étude quant au comportement de chacune d'entre elles ;
- Mieux comprendre comment les indices acoustiques de la parole se répartissent d'une bande à l'autre ;
- Valider une hypothèse que nous avons formulée dans les motivations du Multi-Bandes, à savoir que des unités différentes des phonèmes peuvent mieux modéliser l'information acoustique présente dans les bandes de fréquences.

Avons-nous atteint ces objectifs ?

En ce qui concerne le premier, oui, car cette étude nous a effectivement permis de mettre en évidence des comportements différents des classifieurs dans chaque bande en fonction de chaque phonème. Ce résultat ne fait cependant que confirmer ce à quoi nous nous attendions, et n'est donc pas très surprenant, même s'il est au moins rassurant quant au fondement du paradigme Multi-Bandes !

En ce qui concerne le deuxième objectif, nous sommes obligés de constater que nous ne l'avons pas atteint : en fait, nous nous sommes rendu compte que les différences de comportement existent, mais qu'un grand nombre d'autres facteurs, qui ne sont pas d'ordre phonétiques, entrent en compte. Ainsi, nous pouvons déduire des conclusions générales de l'étude qui précède, par exemple le fait que la première sous-bande soit bien adaptée à la reconnaissance des occlusives nasales voisées, ou que les sifflantes soient bien reconnues par la quatrième sous-bande, mais l'analyse précise de ces différences en termes phonétiques est très difficile, voire impossible. Il est donc trop ambitieux de vouloir construire un module de recombinaison basé uniquement sur des critères phonétiques. Nous devons plutôt réduire nos objectifs et considérer que des connaissances de ce type ne peuvent être utiles que en tant que compléments d'un module de recombinaison plus classique, c'est-à-dire statistique.

Le troisième objectif n'est de la même manière que partiellement atteint, car les phénomènes mis en œuvres sont trop complexes pour être appréhendés par une étude aussi sommaire. Une réflexion beaucoup plus longue et entièrement consacrée à cet aspect phonétique du Multi-Bandes serait nécessaire. En revanche, le quatrième objectif est beaucoup plus satisfaisant, car les expériences précédentes nous ont montré que certaines catégories de phonèmes ne sont pas correctement reconnues dans certaines bandes, et qu'une autre unité de base, adaptée à chaque bande, permettrait de résoudre ce problème. Ceci ne signifie pas qu'il est impossible de réaliser un système Multi-Bandes avec des phonèmes, comme nous le montrons dans la suite, mais seulement que le potentiel du Multi-Bandes n'est pas exploité pleinement avec de telles unités de base. Cette étude constitue en fait la première étape d'une démarche réflexive concernant les classes phonétiques dans les sous-bandes qui se poursuit au chapitre 7.

Chapitre 5
La recombinaison

Ce chapitre est consacré à l'étude du module de recombinaison. Nous présentons les différentes recombinaisons applicables à notre système, uniquement d'un point de vue théorique. Tous les résultats expérimentaux qui permettent de comparer les différentes recombinaisons que nous avons jugées intéressantes sont donnés au chapitre 6.

5.1. État de l'art sur la fusion d'information

5.1.1. Présentation des différents types de recombinaison

Dans le paradigme Multi-Bandes, le terme de ***recombinaison*** correspond en fait à ce qui est appelé dans d'autres domaines la ***fusion d'informations*** ou encore la ***combinaison de classifieurs***. En effet, le module de recombinaison doit fusionner les réponses de différents classifieurs afin d'en déduire une réponse unique. Ce problème de la fusion d'information constitue un vaste domaine de recherche très actif en reconnaissance des formes, et notamment en vision. Les réseaux de neurones constituent également des acteurs privilégiés dans le domaine de la recombinaison de classifieurs. Nous allons donc largement nous appuyer sur les travaux de ces différents champs d'applications pour essayer ensuite d'adapter ces techniques au paradigme Multi-Bandes.

Nous présentons dans cette première partie 5.1 un état de l'art des recherches dans ce domaine. Cet état de l'art est loin d'être exhaustif, au vu du nombre important de publications existantes, mais il est plutôt sélectif et ne présente que les algorithmes pouvant finalement être appliqués à la reconnaissance de la parole. Les parties suivantes, à partir de la partie 5.2, concernent exclusivement notre travail.

La méthode la plus répandue permettant de recombiner des classifieurs est certainement la simple moyenne des sorties correspondantes des différents classifieurs. Pondérer ces sorties par des poids permet de généraliser cette méthode et constitue une première approche que nous qualifierons de *recombinaison linéaire*. Ces recombinaisons linéaires ont déjà été mathématiquement analysées pour des problèmes de régression, mais pas en classification. Plusieurs autres recombinaisons, non linéaires, ont également été étudiées, et peuvent être classées de la manière suivante :

1. Les recombinaisons de type « vote majoritaire » qui n'utilisent que l'étiquette de la classe gagnante de chaque classifieur.
2. Les recombinaisons utilisant une information basée sur le rang des classes.
3. Les recombinaisons de type neuronales qui utilisent les scores retournés par tous les classifieurs pour chaque classe.

L'avantage principal des recombinaisons de type 1 est qu'elles peuvent être utilisées avec à peu près n'importe quels classifieurs. Les autres catégories de recombinaison (2 et 3) ajoutent des contraintes supplémentaires sur ceux-ci.

Nous allons présenter successivement dans les parties 5.1.2 à 5.1.4 les recombinaisons de type 1, 2 et 3. La recombinaison linéaire, qui fait partie du type 3, sera détaillée spécifiquement dans la partie 5.2, de même que la recombinaison neuronale qui est traitée dans la partie 5.3.

5.1.2. Recombinaison sur les étiquettes des classes gagnantes

Le système Multi-Bandes idéal est un système composé de classifieurs dans chaque bande qui réduisent au maximum l'information dont ils disposent, de façon à ce que la tâche du module de recombinaison soit la plus aisée possible. L'information minimale que de tels classifieurs peuvent fournir est tout simplement l'étiquette de la classe qu'ils considèrent comme la meilleure. C'est pourquoi ces recombinaisons sont très intéressantes pour notre système. Malheureusement, le problème qu'elles posent est justement lié à cette réduction drastique de l'information, qui au final peut ne plus être suffisante pour permettre au module de recombinaison de réaliser correctement sa tâche.

✓ *Méthode des votes majoritaires*

La méthode de ce type la plus utilisée dans la littérature est celle dite des « ***votes majoritaires*** ». Celle-ci est très simple, car elle choisit comme classe finale la classe qui est le plus souvent proposée par les différents classifieurs. En cas d'égalité, elle choisit une des classes au hasard. Bien entendu, différentes améliorations sont souvent utilisées avec cette méthode de base, comme par exemple le fait de choisir en cas d'égalité la classe la plus probable, ou encore de pondérer les votes par des fonctions de probabilités des classes. Mais le principe fondamental de cette méthode reste le même, et si un problème ne se prête pas à une telle analyse, aucune de ces variantes ne pourra améliorer véritablement les taux de reconnaissance. Une description détaillée de ces méthodes de recombinaison peut être trouvée dans [Dasarathy94].

Cette méthode se révèle particulièrement utile lorsqu'un grand nombre de classifieurs sont utilisés. Dans notre cas, nous n'utilisons que cinq classifieurs, ce qui n'est pas assez pour employer cet algorithme. Nous l'avons tout de même testé, car cette solution est souvent associée au Multi-Bandes, et nous avons obtenu un taux de reconnaissance phonétique de 63,1 %, qui est à comparer avec le système de référence dont le taux de reconnaissance est de 73,3 %. Ce faible résultat peut être expliqué par le peu de bandes mis en jeu, mais nous ne pensons pas qu'accroître le nombre de bandes permettrait réellement d'augmenter les scores. En effet, cette méthode prend une décision en utilisant uniquement l'information disponible dans une bande. Or, à notre avis, l'utilisation simultanée de toutes les bandes devrait permettre d'obtenir de meilleurs scores.

Les partisans de cette recombinaison peuvent trouver cette analyse un peu rapide, et ils n'auraient pas tort. En effet, comme le suggère Laurent Besacier dans sa thèse [Besacier98a], si nous n'utilisons pas des bandes étroites, mais plutôt des bandes largement recouvrantes, les résultats peuvent devenir nettement meilleurs. Nous n'avons cependant pas poursuivi l'étude de cette recombinaison car nos résultats préliminaires nous ont incités à nous intéresser à d'autres méthodes, mais nous ne rejetons pas totalement celle-ci et pensons qu'elle pourrait avoir une certaine importance dans l'avenir.

✓ *Méthodes par choix dynamique*

Une autre recombinaison n'utilisant que les étiquettes des classes gagnantes est également largement étudiée : il s'agit de la recombinaison par « ***choix dynamique*** ». Elle consiste à sélectionner au cours du décodage un des classifieurs, et à retourner la réponse qu'il fournit. Le critère de choix est bien entendu l'élément le plus important de cette recombinaison, et c'est sur lui que repose la validité de la méthode.

L'avantage de celle-ci est qu'il est très simple de connaître la borne supérieure du taux de reconnaissance atteignable, en utilisant la recombinaison dite de l' « ***Oracle*** ». Celle-ci considère que la classe correcte a été effectivement donnée si elle fait partie des réponses des différents classifieurs. Il s'agit donc bien de la borne supérieure atteignable (mais rarement atteinte !) par ce type de méthode. À l'inverse du vote majoritaire, le taux de reconnaissance fourni par l'Oracle est d'autant plus représentatif des taux de reconnaissance atteignables que peu de classifieurs sont utilisés. En effet, plus le nombre de classifieurs est grand et plus il y a de chances pour qu'un de ces classifieurs fournisse la bonne réponse par hasard ! Notre système respectant cette condition, nous avons ainsi voulu savoir si la recombinaison par choix dynamique pouvait être utilisée en Multi-Bandes. La réponse est clairement oui, car la borne supérieure du taux de reconnaissance est alors de 83 %, qui est très supérieur au taux de référence de 73 %.

Une première catégorie de méthodes utilisant le choix dynamique s'appuie sur le formalisme bayésien. Elles apparaissent notamment dans le domaine de la reconnaissance des caractères manuscrits [Xu96]. Le principe est le suivant : étant donnée la matrice de confusion $A=\left[n_{ij} \right]$ d'un classifieur, nous pouvons exprimer la probabilité qu'un exemple appartienne à la classe C_i par la formule :

$$P\left(x\in C_i | r(x)=j\right)=\frac{n_{ij}}{n_{.j}}=\frac{n_{ij}}{\sum_i n_{ij}}$$

où n_{ij} compte le nombre de fois où le phonème *j* est reconnu alors que le phonème *i* est prononcé, et $n_{.j}$ représente le nombre d'exemples total assimilés à la classe C_j par le classifieur, dont *r(x)* symbolise la réponse. La probabilité $P\left(x\in C_j | r(x)=j\right)$ représente la confiance que nous avons dans le choix d'un classifieur ; il est ainsi possible de sélectionner le classifieur qui a la confiance la plus élevée.

Les premiers tests que nous avons réalisés avec cette méthode donnent des résultats très mauvais, semblables au choix aléatoire d'une bande. Néanmoins, de nombreuses améliorations ont été proposées dans la littérature, le plus souvent en reconnaissance des caractères manuscrits. En ce qui concerne la reconnaissance de la parole, nous avons connaissance du travail de Duchnowsky qui a utilisé une variante de cette méthode, ou plus récemment de Besacier qui présente également dans sa thèse une extension basée sur ce principe bayésien.

Un autre algorithme de recombinaison faisant intervenir la notion de confiance dans un classifieur utilise la théorie de Dempster-Shafer, dont une bonne introduction peut être trouvée dans [Haton91]. Cette théorie attribue de la confiance à des sous-ensembles aussi bien qu'à des éléments de l'ensemble des hypothèses d'apparition des événements et peut également composer ces mesures de confiance lorsqu'elles proviennent de plusieurs classifieurs indépendants.

Enfin, une autre possibilité de recombinaison par choix dynamique consiste à calculer le rapport signal-bruit dans chaque bande de fréquence et à choisir la bande possédant le rapport signal-bruit le plus élevé. Nous-mêmes n'avons réalisé aucune expérience semblable, mais nous pouvons faire un parallèle entre une telle méthode et le formalisme que nous avons déjà présenté sous le nom de « Théorie des Données Manquantes » [Lippmann97b]. En effet, les systèmes concernés proposent d'éliminer les zones fréquentielles dont le rapport signal-bruit est trop bas, ce qui est très semblable à un choix dynamique, tout en étant plus général. Certains travaux récents [Dupont98] étudient cette possibilité qui semble prometteuse.

✓ ***Généralisation du choix dynamique***

Plutôt que d'éliminer les bandes corrompues du signal, il est possible de les considérer toutes, mais en leur accordant une importance dépendant du niveau de bruit. Un certain nombre de travaux récents se sont penchés sur ce problème. Nous pouvons ainsi citer [Glotin98a] qui a fait précéder le système de reconnaissance Multi-Bandes d'un module d'analyse de scènes auditives, dont le rôle est d'estimer le rapport signal-bruit dans chaque bande. L'estimation de ce rapport est réalisée à partir du contenu harmonique des bandes et de l'entropie des reconnaisseurs [Glotin98b]. Cette même idée a ensuite évoluée vers un nouveau type de recombinaison, dite « *full-combination* » [Hagen99], dans laquelle toutes les possibilités de combinaison des bandes sont considérées, chacune de ces possibilités étant pondérée par un facteur dépendant de la probabilité de bruitage des bandes inclues dans la combinaison. Selon ce schéma, une « combinaison de bandes » est en fait une sous-partie de l'ensemble des bandes, celles qui ne sont pas inclues dans ce sous-ensemble étant considérées comme bruitées. Plusieurs méthodes de calcul des facteurs pondérants ont été proposées : la première utilise l'indice d'harmonicité d'un signal monophonique [Berthommier99] tandis que la seconde utilise un indice de localisation extrait d'un signal stéréophonique [Glotin99].

5.1.3. Recombinaison sur les rangs

Pour présenter les différentes méthodes exposées ici, nous nous appuyons sur les travaux de Tim K. Ho en reconnaissance des formes [Ho94].

La première méthode permettant de recombiner les rangs des classes retournés par les classifieurs consiste à calculer une ***intersection*** d'ensembles de classes. Une fois que les rangs des classes ont été obtenus pour tous les exemples du corpus d'apprentissage, le rang maximum (i.e. le moins bon) retourné par chaque classifieur pour chaque classe effectivement prononcée est conservé. Ces rangs maximum constituent alors un seuil au-dessus duquel, en phase de test, la classe est éliminée des réponses possibles. Un autre processus de décision permet alors de sélectionner une classe unique parmi les classes restantes.

La deuxième méthode consiste à calculer une ***union*** d'ensembles de classes. Le meilleur rang obtenu pour chaque exemple du corpus d'apprentissage par la classe à laquelle appartient cet exemple, est calculé pour chaque classifieur. Le maximum de ces rangs (i.e. le moins bon) est alors utilisé comme un seuil pour chaque classifieur et chaque classe. Pendant le test, seules les classes ayant un rang inférieur (i.e. meilleur) à ces seuils sont conservées.

La troisième est une généralisation du vote majoritaire appelée « ***compte de Borda*** ». Le compte de Borda pour une classe est la somme du nombre de classes dont le rang est inférieur (i.e. meilleur) pour tous les classifieurs. Les classes sont alors ordonnées par ordre décroissant de leur compte de Borda.

Les taux de reconnaissance que nous avons obtenus avec ces trois méthodes sont nettement inférieurs au taux de référence, mais ce sont des algorithmes classiques de fusion d'informations que nous devions évaluer.

5.1.4. Recombinaison sur les scores

Ce dernier type de recombinaison utilise presque toute l'information qu'il est raisonnablement possible d'obtenir de la part des classifieurs. L'avantage est donc de minimiser le risque de perdre une information essentielle à la reconnaissance, mais d'un autre côté nous compliquons grandement la tâche du module de recombinaison.

Les méthodes principales utilisant ce type de recombinaison sont essentiellement la recombinaison linéaire et les réseaux de neurones. Celles-ci étant étudiées en détail dans les parties 5.2 et 5.3, nous n'en parlons pas ici. Nous présentons plutôt brièvement les méthodes moins connues et que nous ne développons pas par ailleurs, tout en donnant au lecteur intéressé plusieurs pointeurs permettant d'approfondir l'une quelconque d'entre elle.

La première de ces méthodes est une généralisation de la recombinaison linéaire et a été proposée par Kagan Tumer [Tumer99] sous le nom de « recombinaison par statistiques d'ordre » (*Order Statistics Recombination*). Étant donnée une suite ordonnée par ordre croissant (X_1, ..., X_n) de valuations d'une variable aléatoire X, la statistique d'ordre i de X est définie comme étant X_i. Tumer a appliqué ce principe avec N classifieurs en ordonnant de la même manière leur réponse pour une même classe i. Il définit alors trois modules de recombinaison qui associent à la classe i un score égal respectivement à celui du classifieur de statistique d'ordre 1, d'ordre médian et d'ordre N. Il étudie alors pratiquement et démontre théoriquement l'intérêt d'une telle combinaison de classifieurs, notamment lorsque les observations sont bruitées.

Une autre méthode, dite de « régression par piles », a été introduite par Wolpert en 1992 [Wolpert92] puis reconsidérée par Breinman en 1996 [Breinman96]. Elle commence également à être utilisée pour des tâches de recombinaison de classifieurs. L'idée de base est de minimiser le taux d'erreur des classifieurs en calculant les biais de ceux-ci sur un ensemble d'apprentissage. Ces biais sont estimés en réalisant une classification dans un nouvel espace dont les entrées sont les réponses des classifieurs lorsque ceux-ci sont entraînés sur une partie seulement du corpus d'apprentissage et testés sur la partie restante, et dont les sorties sont les véritables classes des exemples. Ce principe s'apparente à la méthode dite de validation croisée, mais utilise une stratégie plus complexe que celle du « tout-ou-rien » traditionnellement associée à celle-ci.

Enfin, nous terminons ce rapide tour d'horizon de quelques méthodes de fusion d'information en présentant le travail de Patrick Verlinde [Verlinde99], qui a conçu un système de vérification du locuteur qui généralise le principe Multi-Bandes en recombinant différentes sources d'informations, notamment visuelles et acoustiques. Il a lui aussi choisi d'utiliser les scores retournés par les classifieurs, et a testé plusieurs méthodes pour recombiner ces scores, dont les suivantes :

1. La théorie bayésienne : sous l'action de certaines hypothèses, celle-ci se voit transformée en une régression logistique et associe à chaque classe une vraisemblance qui est le produit des vraisemblances fournies par les classifieurs partiels ;
2. Le Perceptron Multi-Couches (cf. partie 5.3) ;
3. La méthode dite des « k plus proches voisins » [Duda73] ;
4. Un arbre de décision [Mitchell97].

Les résultats mettent tous en avant une amélioration des taux de vérification lorsque plusieurs sources d'information sont combinées, le choix de la meilleure méthode de recombinaison étant cependant fonction du corpus disponible et du type d'application visé.

5.1.5. Première conclusion

En testant les différentes méthodes présentées aussi bien dans les parties 5.1.2 que 5.1.3, nous n'avons pas pu obtenir des taux de reconnaissance comparables à ceux du système de référence. Ceci peut provenir du fait que la quantité d'information transmise par les classifieurs, i.e. seulement le rang des classes, n'est pas suffisante pour pouvoir retrouver la véritable classe prononcée. C'est pourquoi nous n'avons pas poursuivi l'étude de ces recombinaisons, mais nous nous sommes plutôt intéressés à d'autres qui transmettent une plus grande quantité d'information au dernier étage du système. Celles-ci sont détaillées dans la suite. En ce qui concerne les recombinaisons présentées dans cette première partie, nous pensons qu'elles ont un véritable potentiel mais qui ne peut pas être exploité avec une architecture semblable à celle de notre système. Ainsi, la méthode de l'Oracle a montré qu'un potentiel existait. De plus, les travaux de Duchnowsky, ou plus récemment ceux ayant trait à la théorie des données manquantes et qui s'appuient sur de telles recombinaisons, ont également exhibé des taux de reconnaissance tout à fait honorables, voire très bons. Il nous paraît donc prématuré de tirer un trait sur ces méthodes, même si nous ne les avons pas choisies comme composantes dans notre système. Nous avons essentiellement testé notre système avec les modules de recombinaison qui sont étudiés dans la suite.

5.2. Notre étude sur la recombinaison linéaire

5.2.1. Définition

La recombinaison linéaire consiste à associer à chaque phonème un score qui est défini comme la somme pondérée des densités de probabilités retournées par les HMM modélisant ce phonème dans chacune des bandes. Cette structure peut être représentée schématiquement par la figure 5.1.

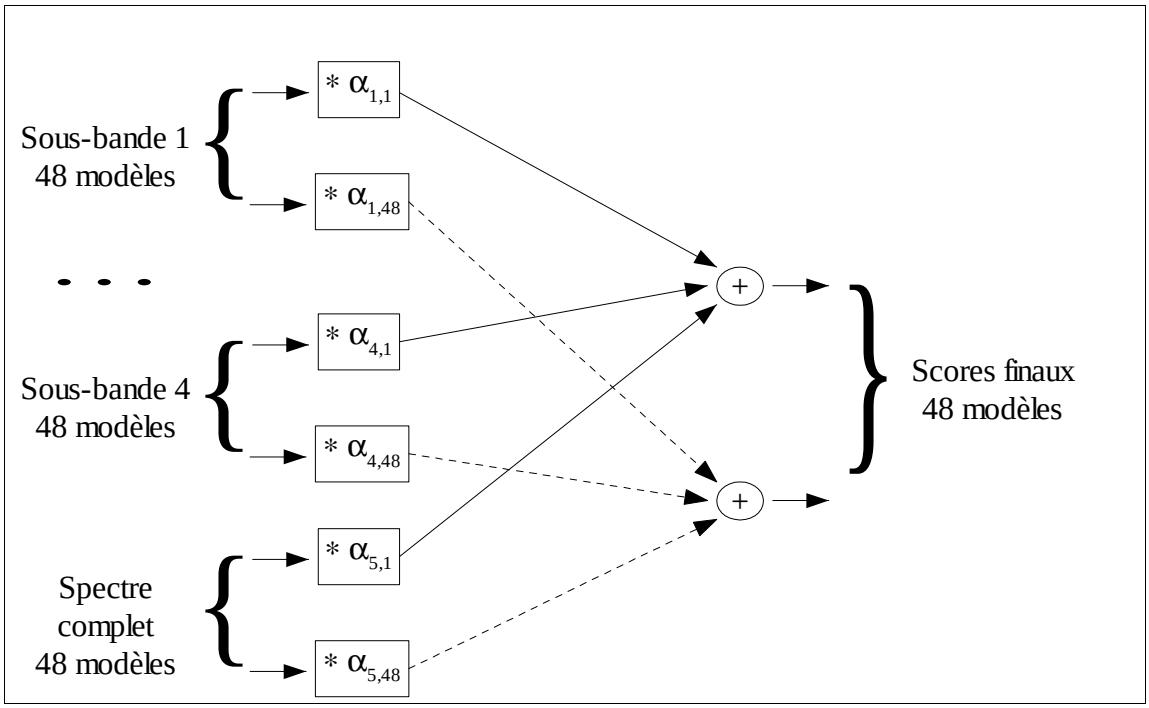

Figure 5.1 : Représentation de la recombinaison linéaire

Elle peut être également définie formellement :

$$S(x,M)=\sum_{b=1}^{B} \alpha_{b,M} P(x|M,b)$$

S (*x,M*) représente le score du modèle *M* pour l'exemple *x*, *B* indique le nombre total de bandes, et P(*x* | *M,b*) représente la densité de probabilité retournée par le HMM modélisant *M* dans la bande *b*. Nous disposons donc finalement d'autant de scores qu'il y a de phonèmes. Il est alors facile de choisir comme solution unique le phonème dont le score final est maximal.

Nous pouvons toutefois remarquer que cette structure linéaire n'autorise pas les interactions entre deux modèles différents : le score final d'un modèle *M* ne dépend que des mêmes modèles *M* des bandes, et aucunement des autres modèles *M'*. Ceci signifie que si les coefficients de la recombinaison sont ajustés grâce à un algorithme d'apprentissage, cet apprentissage ne sera pas discriminant, c'est-à-dire qu'il tendra à faire coïncider au maximum le modèle global de chaque phonème avec les données d'apprentissage, mais qu'il n'essayera pas de « séparer » au mieux les classes entre elles. Si nous voulons un apprentissage discriminant, nous devons choisir une autre fonction linéaire, par exemple :

$$S(x,M)=\sum_{b=1}^{B} \sum_{M'} \alpha_{b,M,M'} P(x|M',b)$$

où *M'* parcourt tous les modèles possibles.

Dans ce cas, le score final de chaque phonème dépend de tous les modèles dans toutes les bandes. Cette solution est certainement préférable à la première, mais elle augmente le nombre de coefficients à ajuster. Lorsque nous avons commencé à travailler avec les recombinaisons linéaires, nous avions préalablement décidé de réaliser tous les tests avec la première recombinaison, plus simple et donc plus facile à entraîner, puis, lorsque la méthode serait validée, de tester la seconde recombinaison. Toutefois, après avoir effectivement implémenté la première recombinaison linéaire, nous nous sommes aperçus que celle-ci était beaucoup moins intéressante que la recombinaison neuronale, et nous n'avons donc pas poursuivi plus loin son développement.

5.2.2. Étude théorique

Lorsqu'un nouveau système doit être optimisé, il faut en premier lieu l'étudier de manière formelle afin de vérifier si son optimum n'est pas calculable. Si c'est le cas, il suffit alors d'appliquer la formule obtenue sur des données réelles puis de vérifier expérimentalement la validité de la méthode. Sinon, et c'est ce qui se produit la plupart du temps, il faut expliquer et montrer pourquoi cette solution n'est pas calculable, puis proposer un algorithme d'apprentissage qui converge vers cet optimum. Nous allons suivre la même démarche dans cet exposé, en présentant tout d'abord dans cette partie notre tentative de résolution formelle du problème d'optimisation posé par la recombinaison linéaire. La fonction objectif que nous considérons est le taux de reconnaissance phonétique, que nous cherchons à maximiser. L'hypothèse de base qui va nous permettre de développer nos calculs est la suivante :

> « *Le corpus d'apprentissage que nous utilisons pour entraîner les coefficients de la recombinaison linéaire est suffisamment grand pour être représentatif de l'ensemble des locutions possibles* ».

Cette hypothèse nous autorise à maximiser la fonction objective sur le corpus d'apprentissage, et à considérer que cette fonction ainsi optimisée est également la fonction optimale sur le corpus de test. Elle n'est certes pas vraie en pratique, mais elle est indispensable en traitement de la parole, car le corpus d'apprentissage est le seul dont nous disposions pour optimiser nos paramètres. Cette hypothèse est généralement implicitement admise dans tous les travaux actuels sur le traitement de la parole.

Le but recherché est donc de calculer la recombinaison linéaire optimale pour le système Multi-Bandes et le corpus d'apprentissage considérés. Il faut ainsi minimiser le nombre d'erreurs de reconnaissance du système sur le corpus d'apprentissage, les paramètres à ajuster au cours de cette minimisation étant les coefficients de la recombinaison.

Soit *Nmod* le nombre de classes phonétiques. Il s'agit en fait du nombre de phonèmes utilisés ou, ce qui revient au même, du nombre de modèles de Markov par bande. Par exemple, en anglais et sur le corpus TIMIT, nous utilisons *Nmod* = 48. Soit x un segment du signal de parole correspondant au phonème $M(x)$. Une erreur de reconnaissance sur x se produit lorsque le score final d'un modèle M différent de $M(x)$ est supérieur au score final du modèle $M(x)$, ce qui se traduit par l'équation :

$$S(x,M) > S(x,M(x))$$

où $S(x,M)$ est la fonction qui retourne le score final du segment x pour le modèle M, à savoir :

$$S(x,M)=\sum_{b=1}^{B}\alpha_{b,M}P(x|M,b)$$

Pour obtenir la recombinaison optimale, il faut donc minimiser la fonction objectif suivante :

$$f_{\text{obj}}=\sum_{i=1}^{i=Nex}\left\{\begin{array}{ll}1 & \text{si } \exists M\neq M(i) \text{ tq } S(i,M)>S(i,M(i))\\ 0 & \text{sinon}\end{array}\right\}$$

où N_{ex} est le nombre d'exemples du corpus d'apprentissage.

Il est impossible de minimiser théoriquement cette fonction, car elle n'est pas dérivable. Il faut donc tout d'abord choisir un sous-ensemble *E* d'exemples du corpus d'apprentissage, puis vérifier si les *Nmod - 1* inéquations

$$S(x,M)<S(x,M(i)) \qquad \forall M\in\{1,...,M(i)-1,M(i)+1,...,Nmod\}$$

sont vérifiées pour chaque élément *x* de *E*.

Écrite de manière plus détaillée, chacune de ces équations est équivalente à :

$$\sum_{b=1}^{B}\alpha_{b,M}P(x|M,b)<\sum_{b=1}^{B}\alpha_{b,M(i)}P(x|M(i),b)$$

soit

$$\sum_{b=1}^{B}\left(\alpha_{b,M}P(x|M,b)-\alpha_{b,M(i)}P(x|M(i),b)\right)<0$$

Ceci revient donc à résoudre un système de $(Nmod-1)\times\text{card}[E]$ inéquations linéaires à 240 inconnues, i.e. les coefficients $\alpha_{b,M}$. Typiquement, puisque Nmod = 48 et le corpus d'apprentissage possède environ 130000 exemples, nous devons résoudre un système de 6 240 000 inéquations à 240 inconnues !

Il est bien entendu hors de question de résoudre un tel système par des algorithmes classiques de résolution, comme celui du Simplex. Il est toutefois peut-être possible de le résoudre par des méthodes géométriques. En effet, nous pouvons remarquer que chacune de ces inéquations coupe l'espace des paramètres en deux hyper-espaces. De plus, les hyper-plans séparateurs passent tous par l'origine. Nous sommes donc confrontés à un cône polyédrique convexe [Berger90] qui délimite une zone de l'espace dans laquelle doit se trouver la solution, si l'on veut que tous les exemples de *E* soient correctement reconnus par le système. Un certain nombre de propriétés de ces systèmes peuvent alors être utilisées. Par exemple, nous savons que si une solution existe, alors une infinité de solutions existe. De même, si nous connaissons deux solutions, c'est-à-dire deux points dans l'espace des paramètres, alors le segment formé par ces points ne contient que des solutions du système.

Nous avons essayé de résoudre de tels systèmes géométriquement, mais nous n'avons pas trouvé de solution parfaite, et les algorithmes géométriques que nous avons testés n'ont pas réussi à converger, à cause du manque de précision des calculs réels sur notre ordinateur. De plus, même dans le cas où nous arrivions à résoudre un tel système, il faudrait alors recommencer la résolution pour tous les sous-ensembles *E* du corpus d'apprentissage, c'est-à-dire $2^{\mathrm{card}[E]}$ fois ! Il est certes possible de réduire ce nombre, par exemple en commençant par les plus grands sous-ensembles possibles, mais la complexité d'une telle résolution rend le système pratiquement insoluble.

C'est pourquoi nous nous sommes tournés vers des méthodes approchées, qui, bien qu'imparfaites, permettent d'obtenir un résultat en temps raisonnable. Les différentes méthodes que nous avons testées pour la recombinaison linéaire sont présentées dans les deux prochaines parties.

5.2.3. Recombinaison empirique

Reconsidérons l'équation de base de la recombinaison linéaire, à savoir :

$$S(x,M)=\sum_{b=1}^{B}\alpha_{b,M}P(x|M,b)$$

Il faut donner des valeurs aux coefficients $\alpha_{b,M}$ de sorte à maximiser le taux de reconnaissance final. La manière la plus simple, et qui n'est pas forcément la moins efficace comme les expériences le montreront, consiste à ajuster empiriquement ces coefficients. Dans notre système, nous avons testé trois ensembles de valeurs possibles donnés aux coefficients.

1. Le premier définit la moyenne des bandes, i.e. $\alpha_{b,M}=\frac{1}{B} \quad \forall\, 1\leqslant b\leqslant B \quad \forall\, M$.

2. Le deuxième définit la moyenne des sous-bandes lorsque la bande représentant le spectre complet n'est plus considéré du tout dans le système. Nous avons alors :

   $\alpha_{b,M}=\frac{1}{B-1} \quad \forall\, 1\leqslant b\leqslant B-1 \quad \forall\, M$ et
   $\alpha_{B,M}=0 \quad \forall\, M$

3. Le troisième a été choisi de sorte à donner plus d'importance à la bande représentant tout le spectre, car celle-ci possède un taux de reconnaissance plus élevé que les autres bandes, du moins dans un environnement non bruité. Nous avons par exemple :

   $\alpha_{b,M}=0{,}1 \quad \forall\, 1\leqslant b\leqslant B-1 \quad \forall\, M$
   $\alpha_{B,M}=0{,}6 \quad \forall\, M$

Dans la suite du mémoire nous considérons que la bande représentant l'ensemble du spectre est la dernière bande d'indice *B*, tandis que les sous-bandes n'utilisant qu'une partie du spectre sont indicées de 1 à *B*-1, la bande d'indice 1 correspondant aux basses fréquences et celle d'indice *B*-1 aux hautes fréquences.

5.2.4. L'algorithme de Minimisation de l'Erreur de Classification

La meilleure méthode pour obtenir une recombinaison linéaire efficace consiste à rechercher le point minimum selon le taux de reconnaissance final dans l'espace des coefficients de la recombinaison. Une telle recherche est possible grâce à l'utilisation de l'algorithme de Minimisation de l'Erreur de Classification (MCE) développé par Juang et al. [Juang97][Juang92]. Nous allons dans un premier temps présenter cet algorithme pour des classifieurs quelconques, puis nous l'appliquerons plus spécifiquement à la recombinaison linéaire.

Le principe de base de cet algorithme est d'utiliser une descente de gradient de l'erreur de classification. Malheureusement, cette erreur n'étant pas dérivable, elle doit être approchée par la fonction sigmoïde suivante :

$$l_{C(x)}(x)=\frac{1}{1+\exp(-\gamma\, d_{C(x)}(x))}$$

x est un exemple du corpus d'apprentissage, C(x) est la véritable classe de x, et γ est une constante évaluée expérimentalement.

La fonction $d_{C(x)}$ est une fonction qui « quantifie » l'erreur de classification et a la forme suivante :

$$d_{C(x)}(x)=-g_{C(x)}(x)+\left(\frac{1}{N-1}\sum_{j\neq C(x)} g_j^{\eta}(x)\right)^{1/\eta}$$

Dans cette expression, $g_j(x)$ est le score final retourné par le système pour la classe j lorsque l'exemple x est présenté. N est le nombre de classes et η est une constante ajustée empiriquement. Lorsque η tend vers $+\infty$, le terme $\left(\frac{1}{N-1}\sum_{j\neq C(x)} g_j^{\eta}(x)\right)^{1/\eta}$ tend vers $g_{\overline{C(x)}}(x)$, $\overline{C(x)}$ étant la « meilleure mauvaise classe de x », c'est-à-dire la classe différente de $C(x)$ possédant le score le plus élevé. Dans nos expériences, nous avons utilisé cette approximation afin de simplifier les calculs. L'équation de $d_{C(x)}$ devient alors :

$$d_{C(x)}=-g_{C(x)}(x)+g_{\overline{C(x)}}(x)$$

Si nous considérons maintenant que λ est un paramètre du système, alors la méthode de descente du gradient de l'erreur appliquée à l'approximation de l'erreur de classification définie ci-dessus donne :

$$\lambda(t+1)=\lambda(t)-\epsilon\frac{\partial l_{C(x)}(x)}{\partial\lambda}$$

où ϵ est une valeur positive proche de 0 définissant la « vitesse » de convergence de la méthode. Cette dernière équation doit alors être appliquée à tous les paramètres jusqu'à convergence.

Il nous reste donc à calculer la dérivée de la fonction sigmoïde approchant l'erreur de classification. Cette dérivée est la suivante :

$$\frac{\partial l_{C(x)}(x)}{\partial \lambda} = \frac{-1}{(1+\exp(-\gamma\, d_{C(x)}(x)))^2}\left(-\gamma \frac{\partial d_{C(x)}(x)}{\partial \lambda}\right)\exp(-\gamma\, d_{C(x)}(x))$$

soit :

$$\frac{\partial l_{C(x)}(x)}{\partial \lambda} = \gamma \frac{\partial d_{C(x)}(x)}{\partial \lambda} \times \frac{1+\exp(-\gamma\, d_{C(x)}(x))-1}{(1+\exp(-\gamma\, d_{C(x)}(x)))^2}$$

ou encore :

$$\frac{\partial l_{C(x)}(x)}{\partial \lambda} = \gamma \frac{\partial d_{C(x)}(x)}{\partial \lambda}(l_{C(x)}(x) - l^2_{C(x)}(x))$$

Lorsque λ est un paramètre du modèle $C(x)$, cette dernière équation devient :

$$\frac{\partial l_{C(x)}(x)}{\partial \lambda} = -\gamma \frac{\partial g_{C(x)}(x)}{\partial \lambda}(l_{C(x)}(x) - l^2_{C(x)}(x))$$

et si λ est un paramètre du modèle $\overline{C(x)}$, elle devient :

$$\frac{\partial l_{C(x)}(x)}{\partial \lambda} = \gamma \frac{\partial g_{\overline{C(x)}}(x)}{\partial \lambda}(l_{C(x)}(x) - l^2_{C(x)}(x))$$

Appliquons maintenant ces équations à la recombinaison linéaire. Dans ce cadre bien précis, les paramètres λ sont en fait les coefficients $\alpha_{b,M}$. De plus, nous avons :

$$g_j(x) = \sum_{b=1}^{B} \alpha_{b,M} P(x|M,b)$$

Donc :

$$\frac{\partial g_{C(x)}(x)}{\partial \alpha_{b,C(x)}} = P(x|C(x),b)$$

et

$$\frac{\partial g_{\overline{C(x)}}(x)}{\partial \alpha_{b,\overline{C(x)}}} = P(x|\overline{C(x)},b)$$

Les modifications que l'on doit appliquer aux coefficients sont donc :

$$\boxed{\alpha_{b,C(x)}(t+1) = \alpha_{b,C(x)}(t) + \epsilon\, \gamma\, (l_{C(x)}(x) - l^2_{C(x)}(x))\, P(x|C(x),b)}$$

et

$$\alpha_{b,\overline{C(x)}}(t+1)=\alpha_{b,\overline{C(x)}}(t)-\epsilon\gamma(l_{C(x)}(x)-l^2_{C(x)}(x))P(x|\overline{C(x)},b)$$

Nous allons terminer la présentation de cet algorithme par quelques remarques :

1. Cet algorithme est très proche de la célèbre rétro-propagation du gradient de l'erreur habituellement utilisée dans les réseaux de neurones. Il possède notamment les mêmes inconvénients, à savoir la difficulté de choisir correctement ϵ. Ainsi, si ϵ est choisi trop grand, l'algorithme risque de « passer au-dessus » de l'optimum global, et s'il est choisi trop petit, la convergence risque d'être extrêmement lente. Nous avons expérimenté ces problèmes dans le choix de ϵ et plusieurs essais ont été nécessaires à chaque utilisation de cet algorithme afin de trouver une valeur de ϵ qui convienne à la tâche demandée.

2. Tout algorithme d'apprentissage optimise une certaine fonction qui est définie selon un critère d'apprentissage. Par exemple, l'algorithme classiquement utilisé pour les HMM utilise le critère MLE, qui maximise l'estimation de la vraisemblance. De même, l'algorithme MCE utilise le critère, qui se note également MCE, et qui minimise une approximation de l'erreur de classification. Ce critère est particulièrement intéressant, car il est discriminant, c'est-à-dire qu'au cours de l'apprentissage, tous les modèles sont modifiés pour chaque exemple, de façon à éloigner le plus possible les classes concurrentes les unes des autres. C'est en grande partie cette considération qui nous a amenés à utiliser un tel critère. De plus, nous n'avons pas choisi d'autres critères discriminants, comme le MAP qui maximise la probabilité a posteriori $P(M|x)$, ou le MMIE qui maximise l'information mutuelle, car ces deux derniers rendent l'apprentissage beaucoup trop complexe. Enfin, puisque nous avons également testé un module de recombinaison neuronal, il est intéressant d'utiliser le même critère d'apprentissage, et donc d'avoir un formalisme unique, pour nos deux systèmes. Ceci a grandement facilité la mise au point des algorithmes que nous développons au chapitre 7.

3. Nous avons été les premiers à utiliser cet algorithme de minimisation de l'erreur de classification dans le cadre du Multi-Bandes [Cerisara98a], mais il semble aujourd'hui que d'autres tentatives dans ce sens commencent à être développées [McMahon98] [Chu98].

5.3. Présentation de la recombinaison neuronale utilisée dans notre système

La recombinaison non linéaire par excellence est celle qui permet d'apprendre à peu près n'importe quel type de fonction de recombinaison. C'est dans ce but que nous avons envisagé une recombinaison par réseau de neurones. Nous avons testé le réseau de neurones le plus connu, à savoir le Perceptron Multi-Couches (PMC). L'inconvénient de cette recombinaison est que, puisqu'elle peut apprendre à peu près n'importe quelle fonction, l'espace des valeurs des paramètres est beaucoup plus complexe et vaste qu'il ne l'est dans le cas de la recombinaison linéaire. Ce qui signifie également que l'apprentissage d'un tel système est beaucoup plus difficile et délicat à réaliser. Cependant, nous pouvons espérer que le point optimum de cet espace est également meilleur.

Le perceptron que nous avons utilisé a la forme suivante :

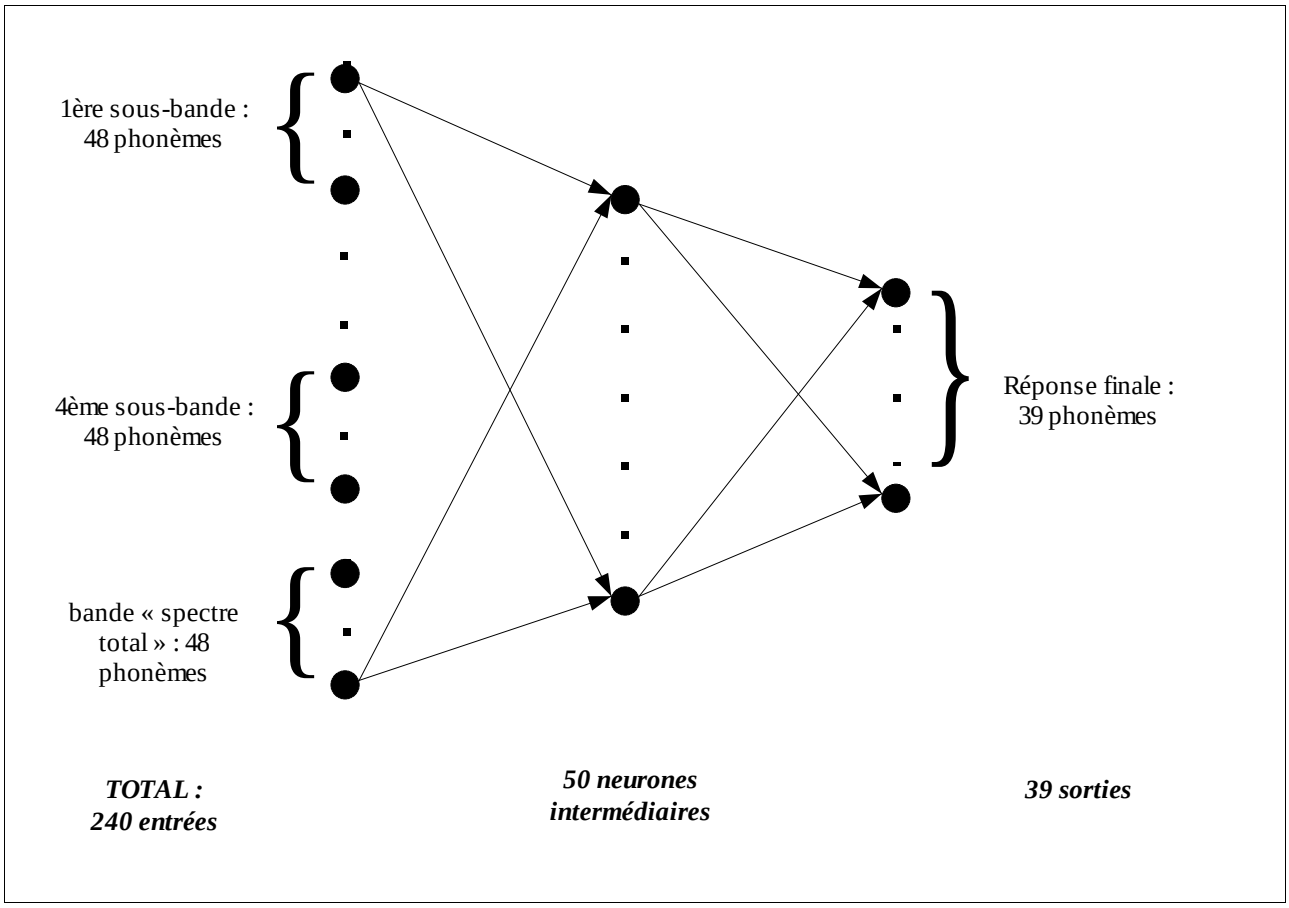

Figure 5.2 : Utilisation d'un perceptron en recombinaison

Pour nos tests, nous n'avons utilisé qu'un réseau de neurones classique, sans essayer l'un des nombreux algorithmes qui sont souvent proposés afin d'accroître ses performances. Ainsi, l'apprentissage de ce perceptron a été réalisé grâce à la seule rétro-propagation du gradient de l'erreur directement dérivée du critère MCE. De même, plusieurs autres types de réseaux de neurones peuvent paraître plus adaptés à la reconnaissance de la parole, comme les réseaux de neurones à fonction radiales (RBF) ou les réseaux récurrents [Haton95]. Toutefois, nous n'avons pas voulu compliquer inutilement le système sans avoir au préalable obtenu le maximum de renseignements possibles sur ce qu'on peut considérer comme un « système minimal ». Il serait néanmoins très intéressant d'essayer d'autres architectures plus adaptées et nous pouvons ainsi espérer améliorer à l'avenir les performances du système que nous décrivons ici. Nous rappelons également que toutes les expériences menées dans la littérature du Multi-Bandes montrent que cette recombinaison avec un perceptron est la meilleure possible, mais aucune expérience n'a été réalisée à notre connaissance avec des réseaux de neurones autres que le perceptron.

Chapitre 6
Expérimentations

6.1. Introduction

De nombreux systèmes de reconnaissance de la parole existent actuellement sur le marché. Malheureusement, la plupart d'entre eux réagissent très mal lorsque les conditions d'utilisation se dégradent, par exemple lorsque du bruit ambiant vient perturber le signal ou lorsque le signal passe à travers le téléphone. Or, les grandes sociétés de télécommunications comme AT&T aux États-Unis ou France Télécom, parmi d'autres utilisateurs potentiels, sont très demandeurs de tels systèmes pour des applications de télématique vocale. Il faut donc impérativement proposer de nouveaux modèles et de nouvelles méthodes afin de rendre la reconnaissance automatique de la parole vraiment robuste.

L'une des motivations du paradigme Multi-Bandes étant justement la robustesse au bruit limité fréquentiellement, il est intéressant d'étudier le Multi-Bandes de ce point de vue, ce qui est fait dans ce chapitre. Celui-ci se décompose en plusieurs parties : tout d'abord, nous présentons un bref état de l'art des méthodes classiquement utilisées en reconnaissance robuste de la parole, puis nous résumons les caractéristiques de notre système en nous plaçant d'un point de vue pratique. Nous testons ensuite ce système, tout d'abord avec du bruit artificiel, puis avec du bruit naturel qui est ajouté au signal de parole. Enfin, nous présentons quelques expériences de reconnaissance après avoir réalisé un apprentissage du module de recombinaison dans le bruit, et nous terminons cette première partie du chapitre en testant notre système sur un corpus composé de parole téléphonique très bruitée. La seconde partie du chapitre complète cette étude expérimentale dans un milieu non bruité. Finalement, les algorithmes permettant d'adapter le Multi-Bandes à la reconnaissance en mode continu sont abordés.

Tous ces tests sont clairement motivés, et nous introduisons chacun d'entre eux en expliquant les objectifs recherchés. En fait, au cours de la thèse, un très grand nombre d'expériences ont été réalisées. Nous ne les présentons pas toutes dans ce mémoire, mais nous avons sélectionné celles qui nous semblent les plus représentatives du comportement du système et les plus intéressantes du point de vue de l'interprétation. Il est donc peu probable que les résultats qui sont présentés ici soient des résultats marginaux correspondants à des cas particuliers du système, car chacun d'entre eux a été choisi parmi plusieurs autres expériences similaires.

6.2. État de l'art des différentes méthodes utilisées en RAP robuste

Le combat contre le bruit se réalise à tous les niveaux, aussi bien dans l'analyse acoustique du signal que dans les méthodes de décodage phonétique, et de très nombreuses méthodes dites « robustes » existent donc dans la littérature. Un état de l'art de toutes ces méthodes peut être trouvé dans [Junqua96]. Nous nous contentons de présenter rapidement ici celles qui sont les plus répandues ou qui ont un lien direct avec notre travail.

Une première catégorie de méthodes robustes tente d'éliminer le bruit du signal avant de l'envoyer à un reconnaisseur automatique. Ces méthodes sont dites de « débruitage ». En fait, le débruitage n'est pas seulement intéressant pour la reconnaissance de la parole, mais aussi pour l'intelligibilité humaine : ainsi, un certain nombre de recherches se consacrent exclusivement à enlever le bruit du signal de façon à le rendre plus compréhensible et « agréable » à l'oreille. Ceci peut être appliqué par exemple dans un réceptacle téléphonique ou dans les implants auditifs qui aident les malentendants à percevoir la parole. La plus connue des méthodes de débruitage est celle dite de « soustraction spectrale » [Boll79] qui estime le spectre du bruit dans les zones de silence puis qui soustrait ce spectre à celui du signal bruité. Cette méthode de base a connu de nombreuses améliorations, comme par exemple celle proposée par Ephraïm et al. [Ephraïm95]. Le principe de base consiste à utiliser un autre espace de projection que le spectre, ce nouvel espace étant défini par les vecteurs propres de la matrice de covariance du signal. Néanmoins, si ces méthodes permettent généralement d'améliorer l'intelligibilité du signal, elles ne sont que rarement bénéfiques pour les taux de reconnaissance des systèmes de RAP.

Une autre catégorie de méthodes définit de nouveaux paramètres qui sont moins influencés par le bruit que les paramètres classiques comme les MFCC. Les plus connus de ces paramètres sont les RASTA-PLP [Hermansky94], mais d'autres paramètres utilisant des filtres spécifiques, comme les filtres de Wiener [Junqua96] sont également utilisés. Nous pouvons considérer que les méthodes s'appuyant sur la théorie des données manquantes [Lippmann97a] se classent dans cette catégorie.

Une troisième catégorie de méthodes acceptent au contraire la présence de bruit dans le signal et tentent de créer des modèles ou des algorithmes de décodage qui utilisent essentiellement l'information du signal et non du bruit. C'est typiquement le cas des systèmes Multi-Bandes ou des systèmes hybrides combinant les réseaux de neurones et les HMM [Bourlard94]. Les systèmes de cette catégorie n'ont pas pour but de modéliser le bruit, mais essayent plutôt de s'en accommoder.

Enfin, une dernière catégorie de méthodes tentent au contraire de modéliser le bruit en même temps que la parole. M. J. Gales [Gales98] a ainsi conçu un algorithme dit de « Combinaison Parallèle de Modèles » (CPM), qui modélise la parole non bruitée par un HMM et le bruit par un autre HMM. Ces deux HMM sont ensuite combinés afin de décoder le signal en autorisant les deux modèles à cohabiter. Cet algorithme a connu beaucoup de succès, et il est notamment à la base de l'algorithme de composition des HMM utilisé en Multi-Bandes et présenté dans la partie 2.4. Malheureusement, cette méthode traite essentiellement les bruits stationnaires, i.e. qui ne varient pas « beaucoup » avec le temps, et, comme pour la soustraction spectrale, elle nécessite la connaissance *a priori* du bruit ou au moins d'une partie du signal ne contenant que du bruit pour pouvoir le modéliser.

6.3. Étude dans du bruit « artificiel »

6.3.1. Introduction, Résumé de notre système et conditions expérimentales

✓ *Introduction*

Nous entendons par bruit artificiel du bruit ayant été généré et synthétisé par des méthodes numériques. Par exemple, les bruits tels que le bruit blanc ou le bruit rose ainsi que tout filtrage de ces bruits sont considérés comme des bruits artificiels. Nous étudions dans un premier temps notre système dans de tels bruits, car leurs caractéristiques étant parfaitement connues, il est beaucoup plus facile d'analyser le comportement du système que dans le cas d'un bruit naturel.

✓ *Résumé de notre système et conditions expérimentales*

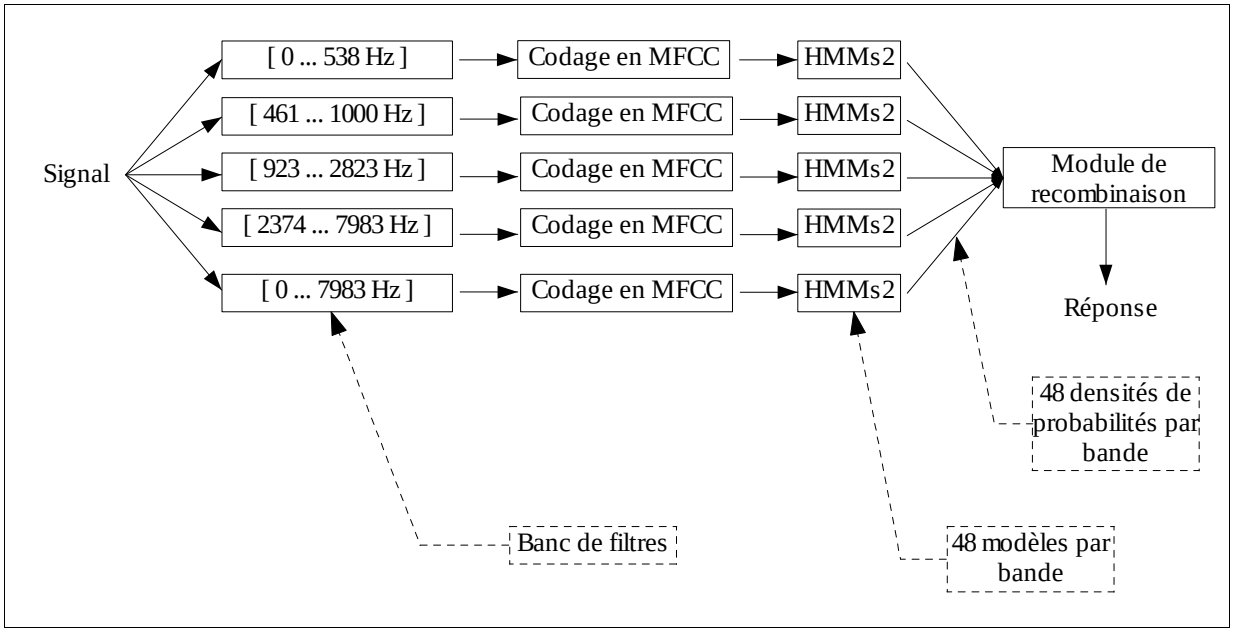

Figure 6.1 : Schéma de notre système Multi-Bandes

Toutes les expériences qui vont suivre sont réalisées avec notre système Multi-Bandes, qui est schématisé sur la figure 6.1 et dont les caractéristiques sont les suivantes :

1. **Analyse acoustique :** Le signal est filtré par 24 filtres triangulaires respectant l'échelle Mel, comme le montre la figure 4.2, et l'énergie du signal dans chacun de ces filtres est calculée. Chaque bande est constituée par 6 de ces filtres, ce qui définit 4 sous-bandes fréquentielles. Dans chaque bande est alors appliquée une transformée discrète en cosinus (DCT) aux indices représentants l'énergie dans chaque filtre, ce qui permet de calculer 3 coefficients par bande. Nous calculons alors les dérivées premières et secondes de ces coefficients, ce qui amène le nombre de coefficients total à 9, puis nous supprimons le premier coefficient qui représente l'énergie totale présente dans la bande. Le nombre de coefficients final du vecteur acoustique dans une sous-bande est donc 8. La même opération est appliquée à l'ensemble des 24 filtres pour le système de référence qui possède donc des vecteurs acoustiques de 35 coefficients.

2. **Nombre de bandes et limites des bandes :** Aux quatre sous-bandes définies ci-dessus, nous ajoutons le système de référence (utilisant l'ensemble du spectre) qui constitue une cinquième bande : notre système utilise donc cinq bandes au total, dont les limites sont respectivement [0 ... 538 Hz], [461 ... 1000 Hz], [923 ... 2823 Hz], [2374 ... 7983 Hz] et [0 ... 7983 Hz]. Nous avons utilisé conjointement les sous-bandes et l'ensemble du spectre, car notre but n'est pas de remplacer le modèle classique par un nouveau modèle n'utilisant plus que les sous-bandes, mais plutôt d'améliorer le système de référence. Or, l'information acoustique contenue dans le spectre ***n'est pas*** la somme des informations acoustiques contenues dans chaque sous-bande, comme le montre [Allen94]. De plus, certaines de nos expériences nous ont amenés à penser que le filtrage du signal détruit une information nécessaire à la modélisation de la parole. Intuitivement, il s'agit de l'intéraction qui existe entre les bandes à un instant donné. Nous avons donc voulu ré-introduire cette information en ajoutant cette bande supplémentaire qui contient le spectre complet. Ainsi, nous voulons faire jouer un rôle différent à cette bande supplémentaire, dont le but est de modéliser l'intéraction entre les bandes, et aux sous-bandes, dont le rôle est d'améliorer la modélisation des indices phonétiques dont elles disposent. Cette amélioration est rendue possible par le fait que les vecteurs acoustiques considérés dans les sous-bandes ont une dimension moins grande que dans le spectre complet, ce qui signifie également que le phénomène connu sous le nom de « dimensionality curse » est réduit, et donc que la modélisation est meilleure, comme cela est expliqué page 25. Cette hypothèse théorique est expérimentalement prouvée dans la partie 6.8 : en effet, les taux de reconnaissance du système global excèdent ceux du système de référence, ce qui ne serait pas le cas si l'information des sous-bandes était contenue dans l'ensemble du spectre. Il est donc dommageable de rejeter cette cinquième bande, car elle-même contient une grande quantité d'information, tout particulièrement dans le cas de la parole non bruitée. Ce point particulier est développé dans les commentaires des expériences pages 88 et 93, et surtout dans la partie 6.8.

   Nous avons été parmi les premiers à utiliser conjointement les sous-bandes et l'ensemble du spectre [Cerisara98a] et plusieurs autres travaux l'ont ensuite également proposé [Mirghafori98] [Hermansky98].

3. **Reconnaisseurs dans chaque bande :** Les reconnaisseurs utilisés dans chaque bande sont des Modèles de Markov Cachés d'ordre 2 composés de 3 états. Les densités de probabilités associées aux observations dans chaque état sont générées par un mélange de gaussiennes :

$$f(X)=\sum_{i=1}^{N}\frac{\alpha_i}{(2\pi)^{k/2}\sqrt{\det\ \Lambda_i}}\cdot\exp\left(-\frac{1}{2}(X-M_i)^t\Lambda_i^{-1}(X-M_i)\right)$$

   X représente le vecteur acoustique observé, M_i le vecteur-moyenne d'une gaussienne, Λ_i la matrice de covariance d'une gaussienne, k la dimension des vecteurs, α_i un coefficient mesurant l'influence de la gaussienne i et N le nombre de gaussiennes considérées dans cet état.

4. **Module de recombinaison :** Le module de recombinaison utilise les scores retournés pour chaque phonème par tous les reconnaisseurs. Avec les 48 phonèmes de l'anglais américain utilisés, le module de recombinaison possède 240 entrées. Nous avons testé deux types de modules de recombinaison : plusieurs modules de recombinaison linéaires, qui associent à ces 48 phonèmes un score final qui est la somme pondérée des scores correspondants dans les bandes, et un module de recombinaison neuronal, composé d'un perceptron à 3 couches, possédant 240 entrées, 50 neurones intermédiaires et 39 neurones en sortie. Chaque neurone de sortie correspond à une classe phonétique, après regroupement de certains phonèmes proches comme le préconise K. F. Lee [Lee88]. La liste des phonèmes regroupés peut être consultée en annexe 2.

5. **Calcul des taux de reconnaissance :** Les taux de reconnaissance sont calculés dans ce chapitre en supposant que la segmentation manuelle des phonèmes est connue[9]. 39 phonèmes sont utilisés pour calculer ces taux, ce qui signifie que dans le cas linéaire, le regroupement des phonèmes est réalisé juste après le module de recombinaison.

6. **Corpus utilisé :** Le corpus que nous avons utilisé pour ces expériences est le corpus TIMIT [Garofolo93] qui contient des phrases en anglais lues par des américains natifs de plusieurs régions des États-Unis. Il s'agit d'un corpus multi-locuteurs, étiqueté phonétiquement, l'étiquetage ayant été réalisé manuellement. Le corpus étant préalablement découpé en plusieurs parties, nous avons respecté cette partition. Notre système a ainsi été entraîné sur la partie « *train* » (apprentissage) du corpus et nous l'avons testé sur la partie « *coretest* ». Sachant que nous avons toujours utilisé le même corpus de test, tous les résultats ont donc des intervalles de confiance qui respectent la même échelle. Plutôt que de répéter ces intervalles de confiance pour chaque résultat, nous avons récapitulé la valeur de l'intervalle de confiance en fonction du taux de reconnaissance dans le tableau 6.1.

9 cf. partie 2.4.3 pour la méthode de calcul et une justification de ce choix.

Taux de reconnaissance	***Intervalle de confiance***
10 %	± 0,6 %
20 %	± 0,8 %
40 %	± 1,0 %
60 %	± 1,0 %
80 %	± 0,8 %
90 %	± 0,6 %

Tableau 6.1 : Intervalles de confiance sur la partie « coretest » de TIMIT en fonction du taux de reconnaissance

Ces intervalles de confiance ont été calculés grâce au raisonnement suivant : estimer un taux de reconnaissance est équivalent à identifier la proportion *p* d'une population qui vérifie une certaine propriété. Soit f l'estimation de *p* sur une population donnée de taille *n*. Sachant que nf suit une loi binomiale, nous pouvons en déduire un intervalle de confiance de l'estimation de *p* lorsque *n* est grand ([Saporta90], page 308). Cet intervalle est alors :

$$f - u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} < p < f + u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$$

$u_{\alpha/2}$ prend la valeur 1,96 pour un intervalle de confiance à 95 %.

7. **Apprentissage :** L'apprentissage de notre système a été réalisé en deux étapes : la première ajuste les paramètres des HMM dans chaque bande, en isolant ceux-ci du reste du système. Ce premier apprentissage utilise la partie « *train* » du corpus réservée à cet usage. La seconde étape, lorsqu'elle a lieu, utilise les HMM issus de la première comme des classifieurs qui ne sont plus modifiés. Elle ajuste ensuite les paramètres du module de recombinaison en fonction des scores retournés par les HMM sur le même corpus d'apprentissage. Ce choix peut ne pas sembler judicieux, dans la mesure où les scores des reconnaisseurs sont issus de leur propre corpus d'apprentissage. C'est pourquoi nous avons dans un premier temps testé une autre méthode, connue sous le nom de *validation croisée*, qui permet de réaliser l'apprentissage et le test d'un système sur un unique corpus, sans réduire sensiblement la taille du corpus d'apprentissage et en évitant le problème que nous venons de mentionner.

   Nous avons adapté cette méthode à notre cas en réalisant l'apprentissage du module de recombinaison sur les neuf dixièmes du corpus de test, puis en testant le système final sur le dixième restant. Nous pouvons alors recommencer cette opération en choisissant à chaque fois un autre dixième du corpus pour le test, la moyenne des taux de reconnaissance obtenus sur les dix tests effectués représentant le taux de reconnaissance global. Les résultats obtenus étaient cependant identiques à ceux correspondant à la première méthode décrite ci-dessus. Ceci peut s'expliquer tout d'abord par le fait que le corpus d'apprentissage est suffisamment représentatif des données, mais également parce que l'apprentissage du module de recombinaison n'entraîne pas une dépendance aussi grande vis à vis du corpus qu'elle ne l'est pour un HMM par exemple. Nous reconsidérons d'ailleurs ce point dans la partie 6.5. Ainsi ces expériences nous ont amenés à utiliser le même corpus d'apprentissage pour les HMM et pour le module de recombinaison, ce qui nous a permis d'éviter la validation croisée, qui reste tout de même assez lente et coûteuse à employer.

6.3.2. Étude dans le bruit limité fréquentiellement

L'une des principales motivations pour le Multi-Bandes est sa robustesse supposée aux bruits n'affectant qu'une partie du spectre. Nous avons voulu dans cette partie tester directement cette hypothèse en générant un bruit blanc que nous filtrons ensuite pour qu'il affecte une zone bien précise du spectre. Nous multiplions ensuite l'amplitude de ce bruit par un coefficient que nous pouvons faire varier et qui permet de tester le système pour différents niveaux de bruit. Celui-ci est ensuite ajouté au signal original de la parole, et le rapport signal-bruit du signal obtenu est calculé.

✓ *Calcul du rapport signal-bruit*

Le rapport signal-bruit est défini par la formule suivante :

$$\text{RSB [dB]} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{bruit}}} \right)$$

où P_{signal} représente la puissance instantanée du signal et P_{bruit} la puissance instantanée du bruit.

Théoriquement, cette formule est définie pour les puissances instantanées, et non pas leur moyenne, car le rapport signal-bruit est alors plus représentatif de l'intelligibilité du signal que dans le second cas. En effet, si le locuteur ne dit rien pendant un « long » moment au cours de la phrase, la puissance moyenne du signal est beaucoup plus faible que s'il parle tout le temps. Le rapport signal-bruit est alors également plus faible, malgré le fait que l'intelligibilité reste la même ! Or, ce rapport signal-bruit nous est utile uniquement comme une mesure de l'intelligibilité. Toutefois, nous ne pouvons pas raisonnablement indiquer sur nos courbes le rapport signal-bruit à chaque instant, et il faut donc utiliser sa moyenne. En utilisant la segmentation parole / silence fournie avec le corpus TIMIT, nous avons calculé un rapport signal-bruit pour chaque segment de parole (et uniquement pour ceux-ci), puis nous avons calculé leur moyenne sur tout le corpus.

Ces moyennes graduent généralement l'axe des abscisses de nos courbes. Cependant, elles ne doivent pas être considérées comme une échelle absolue de l'intelligibilité. En effet, celle-ci varie beaucoup selon le type de bruit considéré, et le rapport signal-bruit ne peut pas modéliser ce phénomène [Hunt99]. C'est pourquoi il faut s'attacher bien plus à la forme des courbes et à leur comparaison relative qu'aux valeurs absolues données par les rapports signal-bruit.

Nous présentons dans un premier temps les résultats que nous avons obtenus avec deux types de bruits limités fréquentiellement : le premier, que nous appelons « bruit filtré aigu », est obtenu à partir d'un bruit blanc que nous avons filtré entre 2900 Hz et 5300 Hz. Le second, appelé « bruit filtré grave », est également obtenu à partir d'un bruit blanc qui est cette fois filtré entre 500 Hz et 1200 Hz. Leurs spectrogrammes respectifs peuvent être consultés sur les figures 6.2 et 6.3.

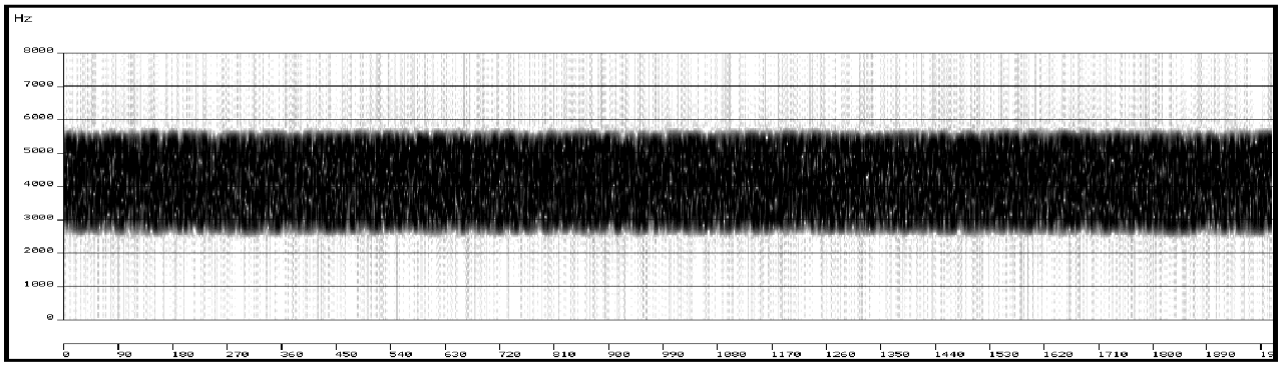

Figure 6.2 : Spectrogramme du bruit filtré aigu

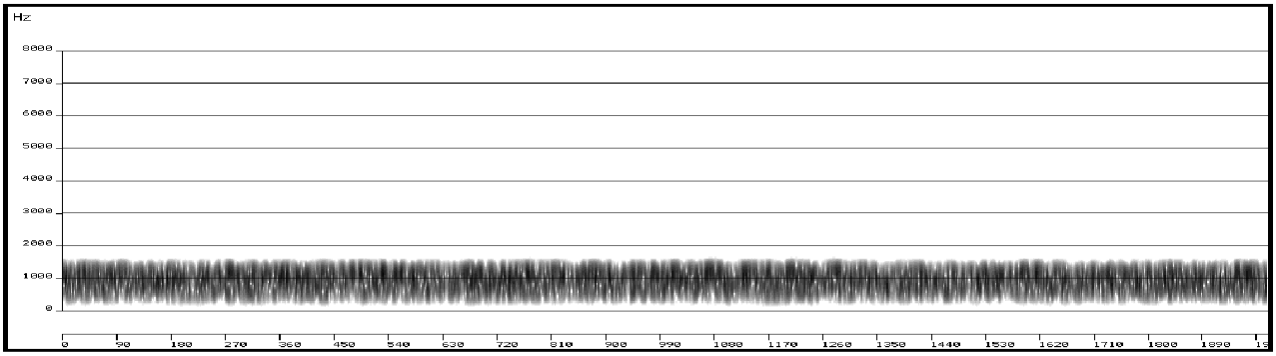

Figure 6.3 : Spectrogramme du bruit filtré grave

Le premier bruit est donc un bruit affectant une région relativement grande du spectre, mais qui ne contient pas *a priori* beaucoup d'information phonétique. Même si cette affirmation peut être discutée, il est en tout cas communément admis que la plus grande partie de l'information phonétique se trouve en dessous de 4000 Hz. Inversement, le deuxième bruit est plus restreint mais il affecte directement cette zone fréquentielle sensée contenir beaucoup d'information phonétique. Nous avons testé ces deux bruits car, comme nous le montrons dans la suite, ils peuvent avoir un effet différent sur le système.

✓ *Étude des reconnaisseurs partiels*

Observons tout d'abord le comportement des reconnaisseurs partiels face à ces deux types de bruit, tel que résumé dans le tableau 6.2.

	[0 ... 538 Hz]	**[461 ... 1000 Hz]**	**[923 ... 2823 Hz]**	**[2374 ... 7983 Hz]**	***Référence (spectre complet)***
aigu	38,0 %	34,0 %	47,0 %	6,0 %	44,8 %
grave	19,1 %	6,0 %	8,4 %	34,0 %	19,8 %
Aucun bruit	39,4 %	36,8 %	48,2 %	40,4 %	73,3 %

Tableau 6.2 : Comportement des bandes dans le bruit filtré aigu et grave, à 20 dB

Nous pouvons faire les remarques suivantes sur ce tableau :

- Le bruit blanc filtré entre 500 Hz et 1200 Hz affecte beaucoup plus les taux de reconnaissance que le bruit blanc filtré entre 2900 Hz et 3500 Hz, ce qui semblerait confirmer le fait que la zone [500 ... 1200 Hz] considérée isolément contient plus d'information phonétique que la zone [2900 ... 5300 Hz] également considérée isolément.

- Pour le bruit filtré aigu, seule la quatrième sous-bande, i.e. celle des hautes fréquences, est affectée par le bruit et voit ses taux de reconnaissance chuter. Les trois autres sous-bandes ne sont presque pas affectées par le bruit. De même, le bruit filtré grave affecte théoriquement les trois premières sous-bandes, ce qui se traduit par une nette chute des taux de reconnaissance pour ces trois premières sous-bandes. Le bruit utilisé est donc particulièrement destructeur même pour le système Multi-Bandes. Seule la quatrième sous-bande n'est presque pas affectée par celui-ci, même si son taux de reconnaissance chute légèrement, ce qui peut s'expliquer par les imprécisions du filtre employé. Ces observations étaient donc parfaitement prévisibles, et elles ne font que confirmer expérimentalement l'indépendance théorique d'une bande par rapport aux autres.

- Une remarque beaucoup plus intéressante vient du fait que, pour les deux bruits utilisés ci-dessus, le système de référence utilisant l'ensemble du spectre voit ses taux de reconnaissance décroître considérablement. Cette baisse de performances est d'autant plus spectaculaire que, dans chaque cas, une des sous-bandes considérée seule a de meilleurs taux de reconnaissance que le système utilisant l'ensemble du spectre. Ainsi en est-il de la troisième sous-bande pour le bruit filtré aigu (47 % vs. 45 %) et surtout de la quatrième sous-bande pour le bruit filtré grave (34 % vs. 20 %). Ce phénomène montre expérimentalement l'intérêt qu'il peut y avoir à utiliser un système Multi-Bandes en milieu bruité.

Nous avons donc déjà pu constater avec ces expériences préliminaires qu'il est possible de dépasser les performances du système de référence simplement en filtrant le bruit de façon à isoler une zone non bruitée du spectre, observation qui est notamment à l'origine des systèmes utilisant la théorie des données manquantes. L'objectif du paradigme Multi-Bandes est donc maintenant double :

1. Créer un système qui tire profit des bonnes performances des bandes non bruitées, mais sans connaître les zones fréquentielles affectées par le bruit !

2. Obtenir de bonnes performances même en milieu non bruité, afin que le système soit réellement robuste au bruit, et non pas seulement adapté à un environnement bruité.

✓ *Étude des systèmes Multi-Bandes linéaires sans apprentissage du module de recombinaison*

Observons maintenant le comportement des systèmes Multi-Bandes linéaires les plus simples, c'est-à-dire sans apprentissage du module de recombinaison. Les figures[10] 6.4 et 6.5 donnent un aperçu des taux de reconnaissance correspondants.

10 Les résultats chiffrés correspondants à ces figures et aux suivantes peuvent être trouvés en Annexe 1.

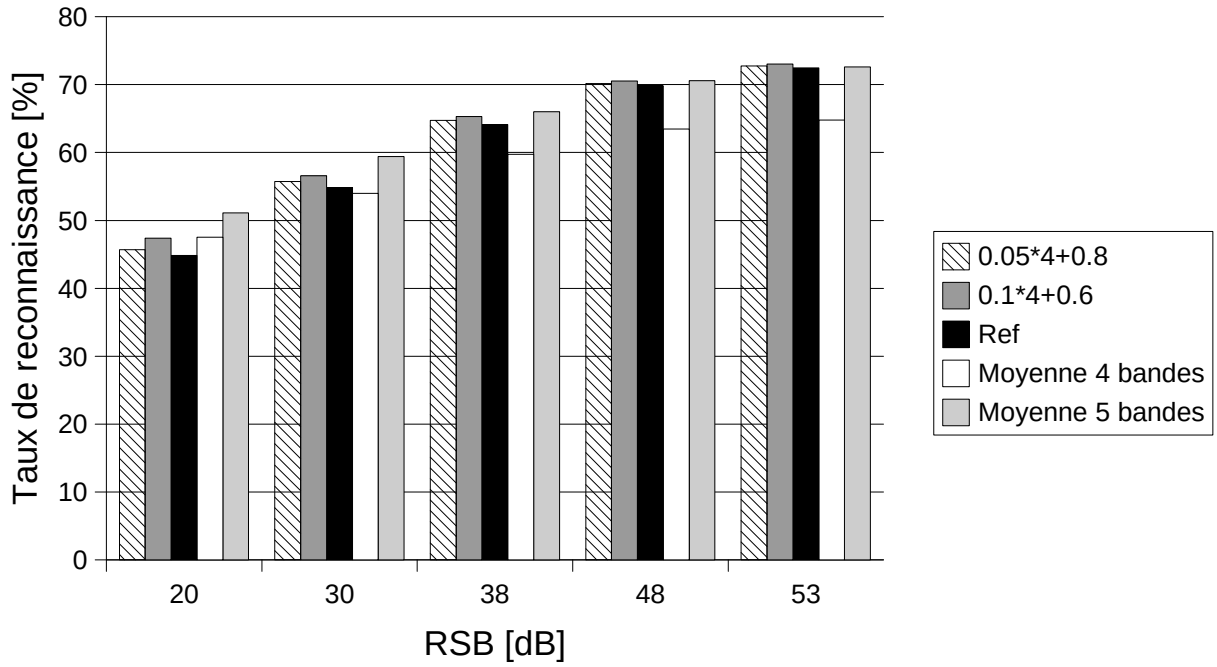

Figure 6.4 : Taux de reconnaissance des systèmes linéaires sans apprentissage du module de recombinaison dans le bruit filtré aigu.

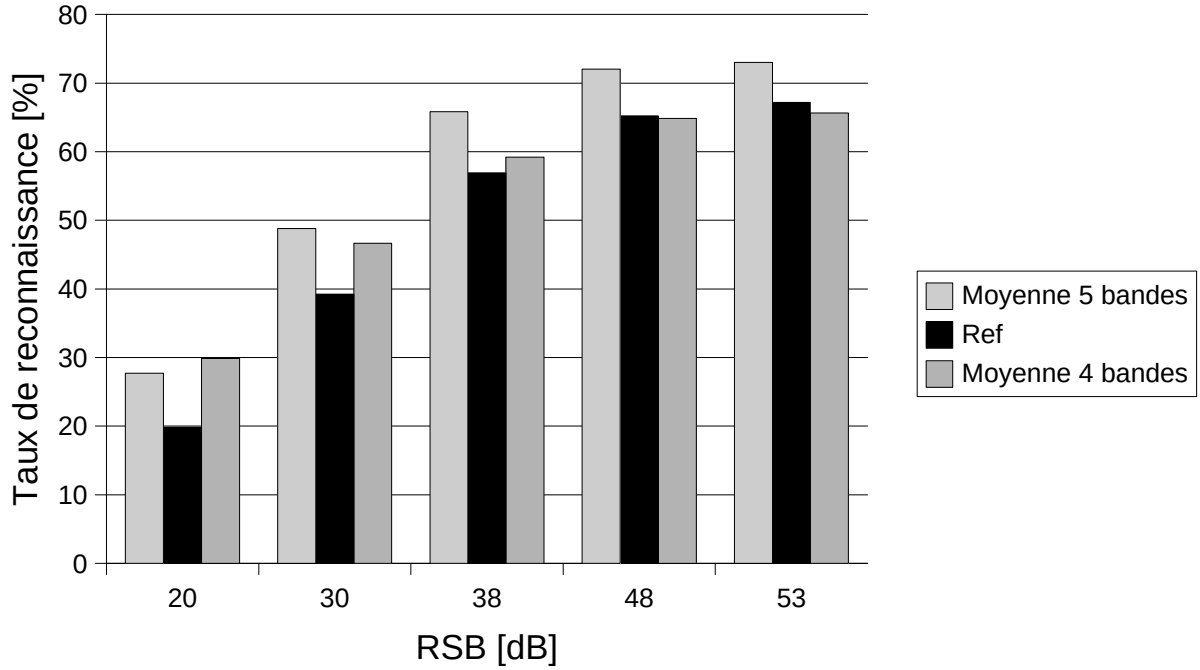

Figure 6.5 : Taux de reconnaissance des systèmes linéaires sans apprentissage du module de recombinaison dans le bruit filtré grave.

✓ ***Description :***

- Nous avons considéré sur la figure 6.4, de gauche à droite, le système ayant des poids fixés à 0,05 pour les quatre sous-bandes et à 0,8 pour la bande totale (en hachuré), le système ayant des poids à 0,1 pour les quatre sous-bandes et à 0,6 pour la bande totale (en gris foncé), le système de référence (en noir), la moyenne seulement des quatre sous-bandes (en blanc) et la moyenne des cinq bandes (en gris clair). Nous avons considéré tous ces modèles afin d'avoir une vue d'ensemble de toutes les possibilités de recombinaison linéaire sans apprentissage.

- Sur la figure 6.5, nous n'avons plus considéré que la moyenne des cinq bandes (en gris clair), le système de référence (en noir) et la moyenne des quatre sous-bandes (en gris foncé). En fait, nous avons omis les deux autres systèmes car leur comportement est identique à celui qui est le leur sur la première figure, et qui n'est pas le plus intéressant. Nous avons donc voulu isoler certains résultats afin de les mettre en valeur.

✓ ***Commentaires :***

- La première remarque à faire sur ces courbes est le bon comportement du système Multi-Bandes dans le bruit, comparé au système de référence. En fait, tous ces systèmes linéaires simples, excepté la moyenne des quatre sous-bandes, ont de meilleurs scores que le système de référence quel que soit le bruit et le niveau de bruit.

- Cette moyenne n'utilisant que les quatre sous-bandes a cependant un comportement très intéressant, dans le sens où elle réalise un score nettement inférieur au système de référence lorsqu'il y a peu de bruit et lorsqu'il détruit peu l'information phonétique du signal. Ceci montre que le spectre pris dans son ensemble contient une grande quantité d'information acoustique qu'il est extrêmement difficile de retrouver dans les sous-bandes prises séparément. Il ne faut donc pas utiliser uniquement les sous-bandes, car le système serait alors privé d'une grande partie de l'information qui peut lui être utile. Ceci explique pourquoi nous avons presque toujours considéré conjointement les sous-bandes et le spectre total dans nos expériences. À l'opposé, cette même moyenne sur les quatre sous-bandes est étonnamment performante lorsque le niveau de bruit s'élève et lorsqu'il affecte directement l'information pertinente de la parole, car dans ce cas, elle se révèle être de loin le meilleur système ! Ceci est un point très intéressant dont nous reparlerons dans la section 6.3.3.

- En ce qui concerne les autres systèmes linéaires étudiés ici, nous pouvons constater que le meilleur système lorsque le niveau de bruit est faible est le système affectant un poids 0,6 au spectre complet et un poids 0,1 aux sous-bandes. Il est également intéressant de remarquer qu'accorder plus d'importance au spectre complet permet d'augmenter les taux de reconnaissance, mais qu'en accorder « trop » les fait décroître, comme le montre le système affectant un poids de 0,8 au spectre complet et de 0,05 aux sous-bandes. Il semblerait donc que le meilleur système réalise un compromis judicieux entre le spectre complet, dont l'importance est capitale, et les sous-bandes qui peuvent lui apporter un certain nombre d'indices supplémentaires.

- Si nous nous intéressons maintenant à l'ensemble des résultats présentés ici, nous nous apercevons que le système qui a le meilleur comportement global est la simple moyenne des cinq bandes. Enfin, d'une manière plus générale, les systèmes Multi-Bandes ont sensiblement le même comportement, que le bruit soit aigu ou grave. La seule différence entre les deux courbes est un décalage vers le bas des résultats de reconnaissance dans le bruit filtré grave, ce qui est parfaitement compréhensible étant donné que l'essentiel de l'information phonétique est dans ce cas perdue. Ceci montre que la robustesse du Multi-Bandes n'est pas spécifique à un bruit n'affectant que certaines fréquences particulières, mais qu'elle semble apparaître quelle que soit la zone du spectre atteinte, et donc quels que soient les phonèmes touchés (par exemple les voyelles pour les fréquences basses et les fricatives pour les fréquences élevées). Ceci est d'autant plus vrai que dans le cas du bruit filtré grave, trois sous-bandes sur quatre sont affectées, ce qui n'empêche pas le système Multi-Bandes de se comporter beaucoup mieux que le système de référence. Il semble donc qu'il suffise qu'une seule bande ne soit pas affectée par le bruit pour que le Multi-Bandes donne de bons résultats. Nous verrons dans la section 6.3.3 ce qu'il advient lorsque toutes les sous-bandes sont touchées par le bruit.

✓ ***Étude des systèmes Multi-Bandes plus complexes***

Étudions maintenant les résultats des systèmes Multi-Bandes avec apprentissage du module de recombinaison pour les mêmes bruits. Dans ces expériences, nous étudions toujours la robustesse inhérente au système, ce qui signifie que s'il y a apprentissage du module de recombinaison, celui-ci est réalisé dans un milieu exempt de tout bruit. Notre système n'a donc aucune connaissance a priori du bruit qui va être utilisé, ni même s'il va y avoir du bruit. Les deux systèmes que nous étudions sont le système Multi-Bandes linéaire avec apprentissage des coefficients de recombinaison par l'algorithme MCE décrit auparavant, et le système Multi-Bandes avec une recombinaison par perceptron. Nous rappelons également les résultats du système de référence et d'un système linéaire simple à des fins comparatives.

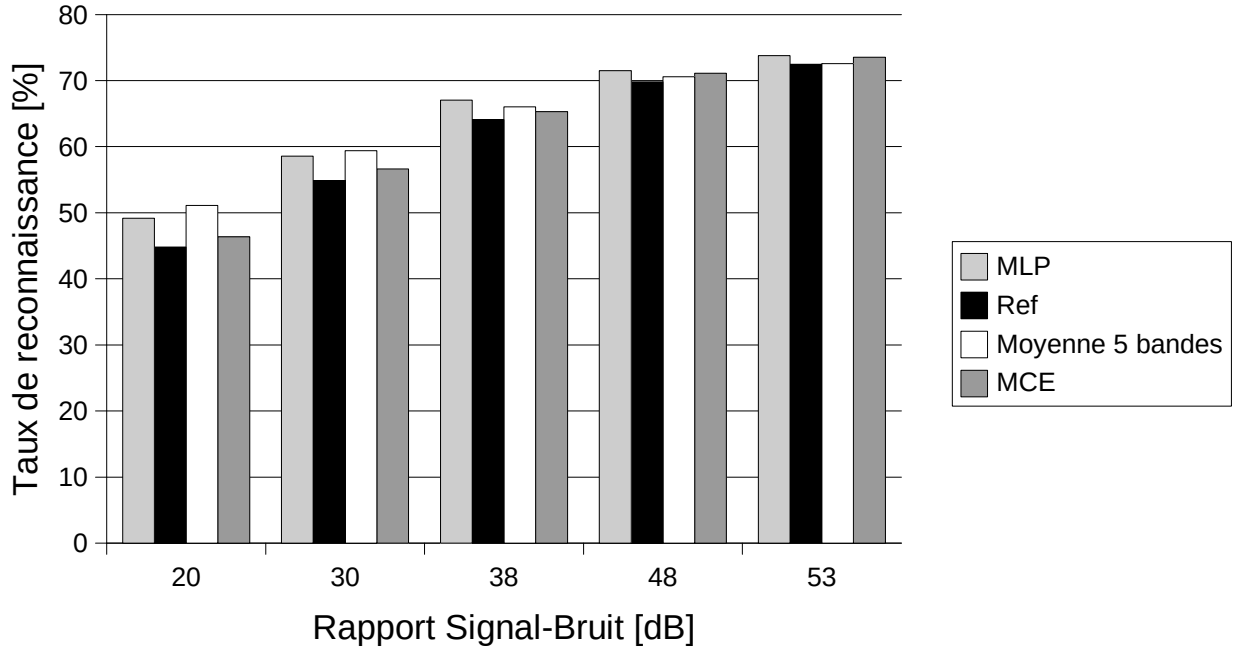

Figure 6.6 : Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans le bruit filtré aigu

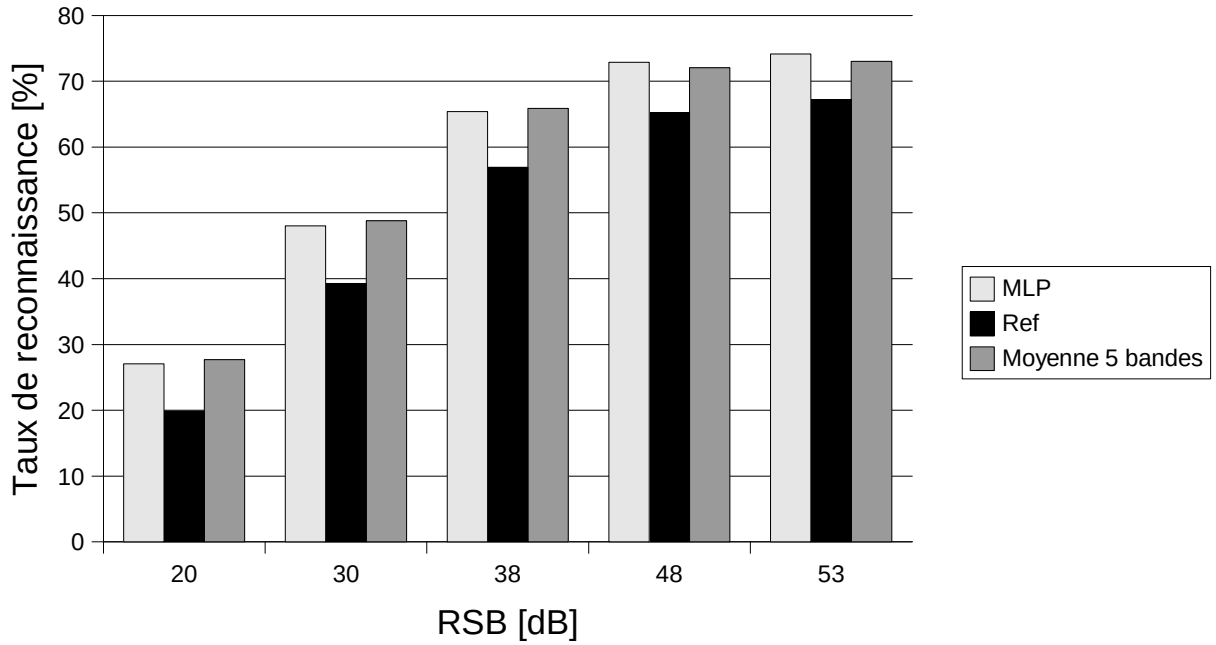

Figure 6.7 : Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans le bruit filtré grave

✓ ***Commentaires des courbes***

- Encore une fois, tous les systèmes Multi-Bandes présentés sont meilleurs que le système de référence, et ce quel que soit le niveau de bruit. Les différences sont plus flagrantes lorsque le bruit est fort, mais elles existent également lorsque le bruit est faible. Ce cas particulier est étudié plus en détail dans la partie 6.8.

- D'une manière générale, le meilleur système est le système utilisant une recombinaison par un perceptron : en effet, ce système réalise les meilleures performances, quel que soit le niveau de bruit, excepté lorsque celui-ci devient très important. Dans ce cas, la moyenne des cinq bandes dépasse légèrement le système avec recombinaison par un perceptron. Ceci peut s'expliquer aisément, car avec un apprentissage en milieu non bruité, ce dernier système acquiert une dépendance supplémentaire vis-à-vis des conditions d'apprentissage, dépendance qui n'apparaît pas pour le système moyennant les cinq bandes. Ce phénomène est encore plus net pour le système linéaire avec apprentissage MCE, qui se comporte relativement bien lorsque le bruit est faible, mais qui est cependant beaucoup moins robuste dans un environnement très bruité. En fait, ce système présente une courbe de reconnaissance qui, tout en restant légèrement supérieure à celle du système de référence, semble « reproduire » la forme de celui-ci. Nous reparlons de ce phénomène dans la partie 6.3.3. Nous n'avons pas reproduit le système MCE sur la figure 6.7, car son comportement est le même que sur la figure 6.6.

✓ ***Conclusion***

Nous avons démontré expérimentalement dans cette partie l'hypothèse de base qui est la robustesse du système Multi-Bandes aux bruits limités fréquentiellement. Nous avons de plus montré l'importance qu'il y a à utiliser conjointement les sous-bandes et le spectre complet dans le système. Nous avons également commencé à analyser certains résultats concernant le comportement des sous-bandes dans le bruit, analyse que nous poursuivons dans la section suivante. En conclusion de cette première étude, nous remarquons que le meilleur système Multi-Bandes est certainement le système avec une recombinaison par un perceptron, car c'est un système qui se montre excellent en milieu peu bruité, et qui reste très bon en milieu très bruité. Toutefois, comme nous l'avons remarqué en présentant d'un point de vue théorique cette recombinaison, l'apprentissage du perceptron est assez délicat à réaliser à cause de l'espace de recherche qui est beaucoup plus complexe que dans le cas de la recombinaison linéaire. De plus, ce système ne doit pas occulter les surprenants résultats obtenus par la simple moyenne des cinq bandes, notamment en milieu très bruité, qui est alors clairement le système le plus robuste parmi tous ceux testés.

6.3.3. Étude dans le bruit blanc

Nous venons de démontrer que le Multi-Bandes est robuste au bruit limité fréquentiellement. Nous allons maintenant observer le comportement des systèmes Multi-Bandes dans du bruit affectant l'ensemble des fréquences du spectre, à savoir le bruit blanc. Les résultats des systèmes avec et sans apprentissage du module de recombinaison sont cette fois présentés ensembles.

✓ ***Résultats des systèmes Multi-Bandes dans le bruit blanc***

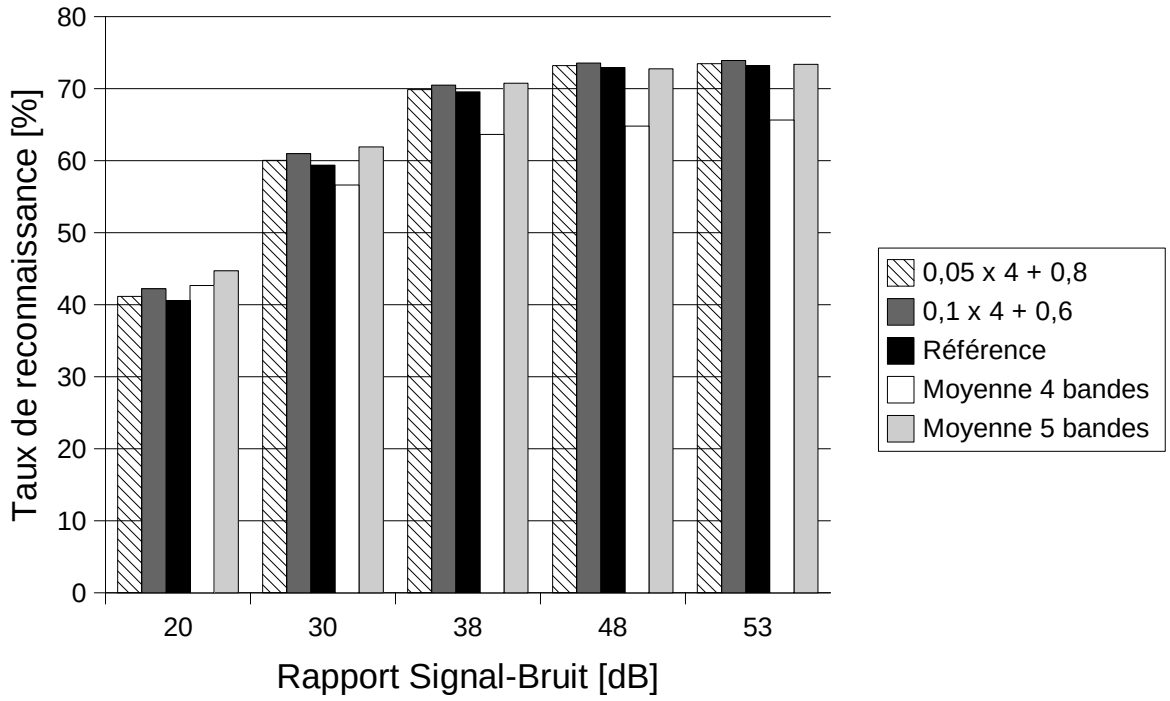

Figure 6.8 : Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans du bruit blanc.

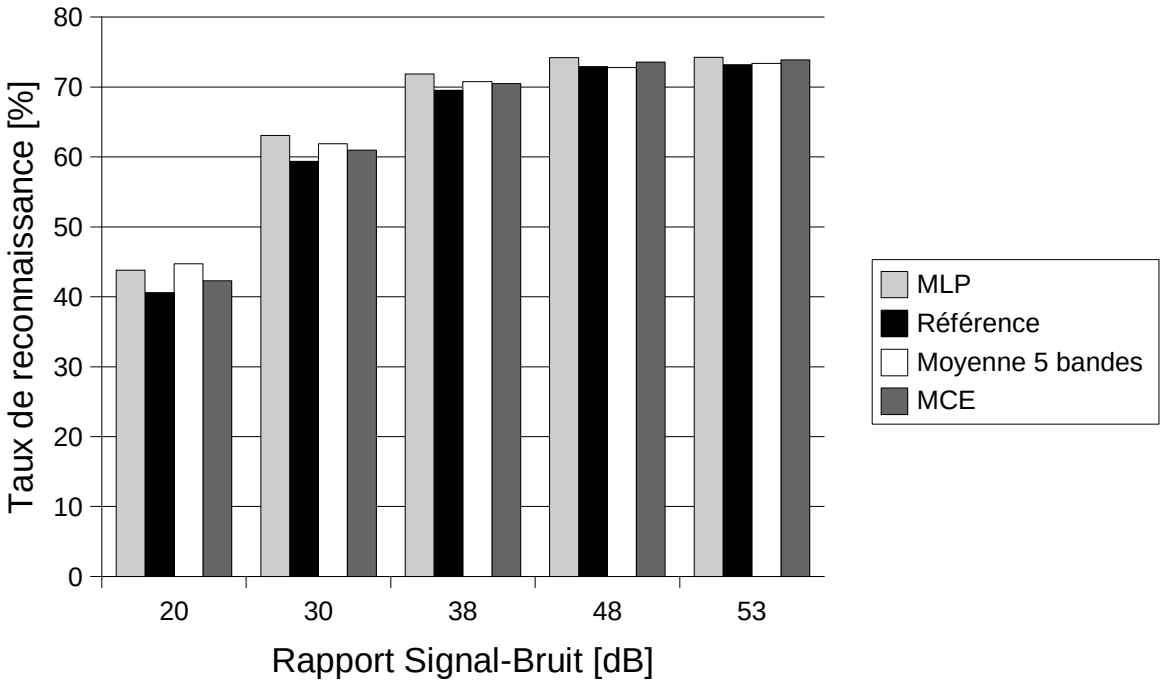

Figure 6.9 : Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans du bruit blanc.

✓ ***Commentaires***

Plusieurs remarques peuvent être faites en ce qui concerne ces résultats :

a) Remarques concernant le système MCE

Comme nous l'avons déjà constaté dans les expériences de la section 6.2, le système linéaire avec apprentissage des poids de la recombinaison par l'algorithme MCE est certainement le système qui semble le plus « proche » du système de référence, au moins de par la forme de sa courbe représentant le taux de reconnaissance. En fait, en étudiant plus attentivement son comportement, nous voyons qu'il n'est pas *exactement* semblable au système de référence sur au moins deux aspects : tout d'abord, il est toujours légèrement meilleur que le système de référence quel que soit le niveau de bruit ajouté, et ensuite l'écart entre le taux de reconnaissance du système MCE et du système de référence, bien que restant assez faible, varie néanmoins légèrement selon le type et le niveau de bruit considérés. Malgré quelques différences de comportement assez subtiles semblables à cet exemple, la ressemblance entre ces deux systèmes est tout de même frappante. Nous concevons deux explications possibles à celle-ci : la première consiste à supposer que les coefficients de la recombinaison du système MCE sont devenus très proches, au cours de leur apprentissage, du « tout cinquième bande », c'est-à-dire proches de 1 pour la cinquième bande et de 0 pour les quatre autres sous-bandes. Cette hypothèse s'est révélée fausse lorsque nous avons observé ces coefficients. Initialement, nous les avions fixés à 0,6 pour la cinquième bande et à 0,1 pour les quatre premières sous-bandes, valeurs que nous supposons être proches des valeurs « optimales », afin d'accélérer l'apprentissage du module de recombinaison. Nous avons ensuite réalisé l'apprentissage MCE, et l'observation des coefficients obtenus montre que ceux-ci sont restés très proches des valeurs que nous leur avions données initialement, mais qu'ils se sont en quelque sorte « spécialisés » autour de ces valeurs selon les phonèmes concernés. Cette observation, ajoutée à la figure 6.20 que nous étudions dans la suite, montre que cette première hypothèse est fausse.

La seconde explication possible à la ressemblance entre le système MCE et le système de référence est tout simplement l'adaptation du premier aux conditions d'apprentissage, à savoir un environnement non bruité. Cette hypothèse permet d'expliquer la différence de comportement entre le système MCE et le système dans lesquels les coefficients de la recombinaison sont fixés à 0,1 pour les sous-bandes et à 0,6 pour le spectre complet. Mais un autre problème se pose alors : le système basant sa recombinaison sur un perceptron devrait avoir le même comportement. C'est le cas d'une certaine manière, car ce dernier obtient des résultats moins bons que les simples moyennes des bandes lorsque le bruit devient très important. Toutefois, il semblerait que le perceptron ait mieux réussi son apprentissage, dans le sens où l'initialisation aléatoire de ses paramètres lui a permis de mieux « doser » l'importance de la cinquième bande, directement modèle par modèle, et pas globalement comme c'est le cas pour le système MCE. Ceci a d'ailleurs été facilité par le fait que le perceptron a été entraîné avec des pas d'apprentissage décroissants, ce qui n'est pas le cas du MCE, pour lequel nous avons fixé un pas constant assez faible : ce dernier a ainsi pu augmenter ses taux de reconnaissance, ce qui n'aurait pas été possible avec un pas plus grand, vu la complexité de l'espace d'apprentissage, mais il n'a néanmoins pas pu s'échapper des minima locaux présents au voisinage des coefficients initiaux. Toutefois, il aurait peut-être été possible d'améliorer le système MCE en utilisant un apprentissage plus long et se basant sur des poids initiaux aléatoires. Nous n'avons malheureusement pas réalisé cette expérience car le temps nous a manqué. Pour conclure cette discussion, nous pouvons dire que le système MCE n'est en fait, à cause de son initialisation, qu'un système de référence amélioré, ce qui montre par ailleurs que le paradigme Multi-Bandes ne se limite pas à ajuster le poids du système de référence, mais se base beaucoup plus sur le comportement individuel de chaque modèle dans chaque bande.

b) Remarques concernant la moyenne des quatre sous-bandes

Dans le cas du bruit blanc, le système Multi-Bandes qui réalise simplement la moyenne des quatre sous-bandes est certes bien meilleur que le système de référence lorsque le bruit devient très élevé, mais la différence n'est plus aussi nette que précédemment, dans les expériences réalisées avec le bruit filtré grave. De plus, il n'est plus le meilleur système à 20 dB, mais se trouve juste en dessous du système Multi-Bandes moyennant les cinq bandes. Ceci est intéressant, car nous pouvons, en comparant la moyenne des quatre sous-bandes et celle des cinq bandes, catégoriser les bruits en deux classes en fonction de leur effet sur le spectre total :

1. Les bruits affectant des zones très informatives du spectre et laissant les autres fréquences à peu près saines. Il s'agit notamment ici du bruit filtré grave. Ce type de bruit est très pénalisant pour le système utilisant le spectre complet, en comparaison des systèmes n'utilisant qu'une petite partie du spectre, et les systèmes Multi-Bandes qui accordent beaucoup d'importance au spectre complet obtiennent donc également de mauvais résultats. C'est par exemple le cas de la moyenne des 5 bandes.

2. Les bruits affectant équitablement toutes les zones les plus informatives du spectre. Ils peuvent donc affecter plus fortement une zone donnée du spectre, mais à condition que celle-ci ne soit pas une zone très informative. Nous retrouvons dans cette catégorie le bruit blanc et le bruit filtré aigu. Dans ce cas, les systèmes Multi-Bandes qui utilisent de manière non négligeable le spectre total obtiendront de meilleurs résultats que ceux qui ne l'utilisent pas. Il faut toutefois faire attention à cette remarque, car elle ne signifie pas du tout que les performances des systèmes Multi-Bandes augmentent conjointement avec l'importance accordée au spectre complet : comme nous le verrons, ce n'est pas du tout le cas, bien au contraire.

Nous voyons donc qu'il faut prendre soin, lorsque l'on parle de la robustesse du Multi-Bandes dans le bruit, de toujours considérer ces deux phénomènes simultanément, car ils ont chacun une influence sur le système Multi-Bandes, influence qui dépend du bruit présent. Ceci est particulièrement vrai pour des bruits « naturels » qui peuvent rapidement changer de forme et combiner les deux types de comportement que nous venons d'évoquer.

c) Remarques concernant les performances du Multi-Bandes dans le bruit blanc

Nous pouvons remarquer que le comportement des différents systèmes Multi-Bandes est globalement le même dans le bruit blanc que dans le bruit n'affectant qu'une partie du spectre, à savoir :

- Domination très nette des systèmes Multi-Bandes sur le système de référence, mis à part la simple moyenne des quatre sous-bandes lorsque le bruit est faible.

- Très bon comportement du système Multi-Bandes utilisant un perceptron pour la recombinaison, qui s'avère être le meilleur système dans presque tous les cas.

- Mais aussi, lorsque le bruit est très important, très bonnes performances des systèmes Multi-Bandes calculant la simple moyenne des bandes.

Ces résultats sont surprenants dans la mesure où l'hypothèse la plus communément admise expliquant la robustesse du Multi-Bandes est le fait que certaines zones de fréquences laissées intactes par le signal peuvent être utilisées lorsqu'une autre partie du spectre est affectée par le bruit. Nous voyons bien que cette explication n'est plus du tout suffisante dans notre cas. Elle reste vraie, comme certaines expériences que nous avons réalisées ou apparaissant dans la littérature le montrent [Lippman97a], mais elle n'explique qu'une partie de la robustesse du Multi-Bandes. Suite aux expériences que nous venons de présenter, ainsi qu'aux études menées sur les sous-bandes considérées individuellement dans le bruit, nous suggérons une autre explication à la robustesse du Multi-Bandes, qui est la suivante :

> *Une zone filtrée du spectre est individuellement plus robuste au bruit que le spectre considéré dans son ensemble, même lorsque ce bruit affecte également toutes les fréquences du spectre.*

Nous expliquons ce phénomène en partie grâce au fait que les indices acoustiques qui caractérisent un phonème correspondent souvent à une énergie très forte, concentrée sur une zone fréquentielle très étroite, comme le sont les formants. En conséquence, si nous considérons une sous-bande fréquentielle contenant ce formant, l'énergie du bruit blanc est beaucoup moins importante dans cette bande qu'elle ne l'est dans le spectre complet, alors que l'énergie du formant reste la même dans cette bande que dans le spectre complet. La différence d'énergie entre l'information « utile », i.e. celle du formant, et l'information « inutile », i.e. celle du bruit est donc beaucoup plus importante dans le cas de la sous-bande que du spectre complet, ce qui explique qu'il soit plus « facile » pour le reconnaisseur d'identifier la voyelle dans la bande que dans le spectre complet. Nous avons imagé cette explication en prenant le cas d'une voyelle, car c'est sans doute le cas le plus facile à comprendre, mais ce phénomène peut être étendu à des indices fréquentiels quelconques. Enfin, nous verrons dans la partie 6.8 que cette explication n'est pas la seule qui permette de comprendre l'importance d'une sous-bande dans le bruit, et nous la compléterons à ce moment-là.

6.3.4. Conclusion

Nous avons montré dans cette partie que, en plus d'être robuste aux bruits limités fréquentiellement, le Multi-Bandes est également robuste aux bruits affectant toutes les fréquences du spectre. Nous avons de plus émis une nouvelle explication concernant cette robustesse, qui proviendrait en partie de la sélection d'une sous-bande non bruitée du spectre, mais aussi de la robustesse des sous-bandes elles-mêmes. Nous avons ainsi proposé un début d'explication quant à cette robustesse intrinsèque des sous-bandes qui serait une conséquence directe de la forme des indices phonétiques dans le spectre. Les commentaires et remarques concernant ces explications sont poursuivis dans la partie 6.8.

6.4. Étude avec du bruit naturel

Nous allons maintenant étudier le comportement du paradigme Multi-Bandes dans du bruit « naturel », c'est-à-dire qui est issu de situations réelles. Nous avons utilisé dans ce but le corpus de bruit NOISE-ROM-0 [NOISE-ROM90] qui contient essentiellement des bruits militaires, mais également quelques bruits à usage plus « civil ». Une infinité de situations de la vie courante mettent en jeu un grand nombre de type de bruits, et nous avons donc été obligé de faire des choix quant aux bruits à utiliser. Chacune des parties suivantes est consacrée à un bruit particulier et expose et analyse les résultats obtenus avec ce type de bruit. Tous ces bruits sont additifs, c'est-à-dire qu'ils sont ajoutés au signal initialement issu du corpus de parole. Nous n'avons pas traité les bruits convolutifs, c'est-à-dire ceux qui proviennent d'une déformation du signal. Un exemple de tel bruit convolutif, cité dans les motivations du paradigme Multi-Bandes, est la réverbération. Le Multi-Bandes s'est montré également robuste dans ce type de bruit, comme l'a montré Mirghafori [Mirghafori99], et nous invitons le lecteur intéressé à se reporter à ces travaux.

6.4.1. Bruit d'automobile

Un grand nombre d'applications de reconnaissance de la parole sont actuellement destinées à être embarquées dans les automobiles. Ceci permet notamment de libérer l'attention visuelle du conducteur lorsque celui-ci désire réaliser des tâches « secondaires », comme par exemple choisir une station de radiophonie ou contrôler un système GPS. Ces applications sont donc très importantes du point de vue de la sécurité et leur enjeu économique n'est pas à négliger non plus. Le bruit présent dans les habitacles des véhicules est donc très étudié et nous confrontons dans ce chapitre notre propre système à ce type de bruit.

Le bruit de voiture est très particulier car il affecte essentiellement et presque exclusivement les très basses fréquences. Ce qui a pour conséquence de laisser pratiquement intacte la zone phonétiquement informative du spectre. L'intelligibilité de la parole dans ce bruit, pour un auditeur humain, reste donc très bonne. De la même manière, les systèmes automatiques réalisent des taux de reconnaissance étonnamment bons dans ce type de bruit. Afin de pouvoir mettre tout de même en évidence les différences de comportement des différents systèmes dans le bruit de voiture, nous avons dû utiliser des niveaux de bruit très élevés.

Le bruit est un bruit de voiture Volvo, enregistré sur autoroute à une vitesse stable. Il est extrait de la base NOISE-ROM-0. Un spectrogramme du bruit seul est représenté sur la figure 6.10. Les taux de reconnaissance des différents systèmes dans ce type de bruit sont présentés dans la suite.

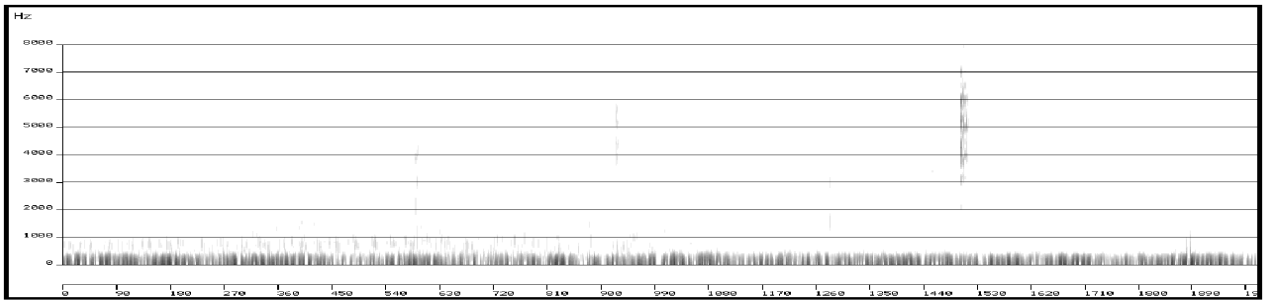

Figure 6.10 : Spectrogramme du bruit de voiture

✓ ***Étude des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans le bruit de voiture***

Les résultats de ces systèmes sont présentés sur la figure 6.11.

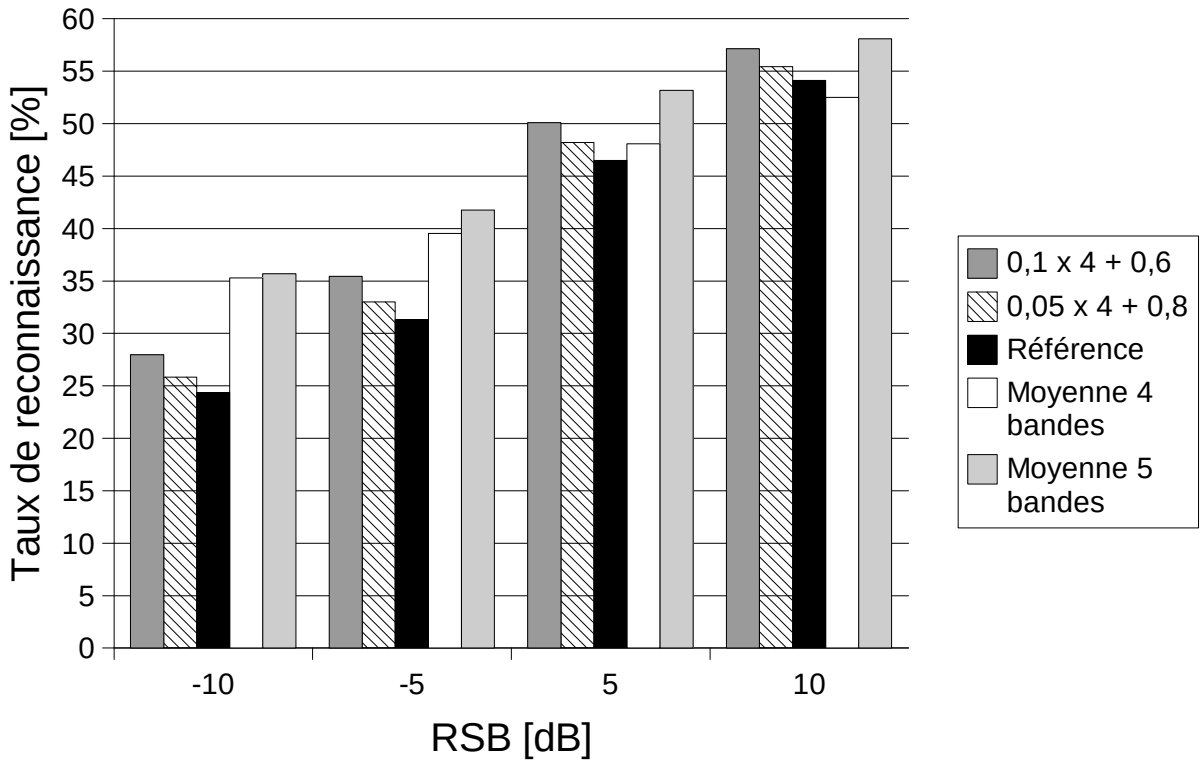

Figure 6.11 : Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans le bruit de voiture.

Les différents systèmes ont globalement le même comportement dans ce bruit que dans ceux considérés précédemment. Notamment, les résultats obtenus par les systèmes Multi-Bandes sont nettement meilleurs que ceux correspondants au système de référence : par exemple, à –10 dB, la différence entre le système de référence et la moyenne des cinq bandes dépasse les 10 %, et à 10 dB, elle est de l'ordre de 5 %. Cette moyenne des cinq bandes est clairement le meilleur système, même si les autres systèmes Multi-Bandes réalisent également de bonnes performances comparées au système de référence, sauf encore une fois pour la moyenne des quatre sous-bandes lorsque le bruit est relativement faible. Les remarques déduites des précédentes expériences sont donc encore valables dans ce bruit de voiture. En fait, celui-ci n'affecte qu'une zone fréquentielle réduite qui ne transporte pas la majorité de l'information acoustique : nous pouvons donc classer ce bruit dans la même catégorie que le bruit filtré aigu, comme nous l'avons expliqué un peu plus haut. C'est pourquoi nous retrouvons un comportement assez similaire entre les courbes des résultats concernant ces deux bruits.

✓ *Étude, dans le bruit de voiture, des systèmes Multi-Bandes avec apprentissage du module de recombinaison*

Observons maintenant les résultats des systèmes avec apprentissage du module de recombinaison dans le bruit de voiture.

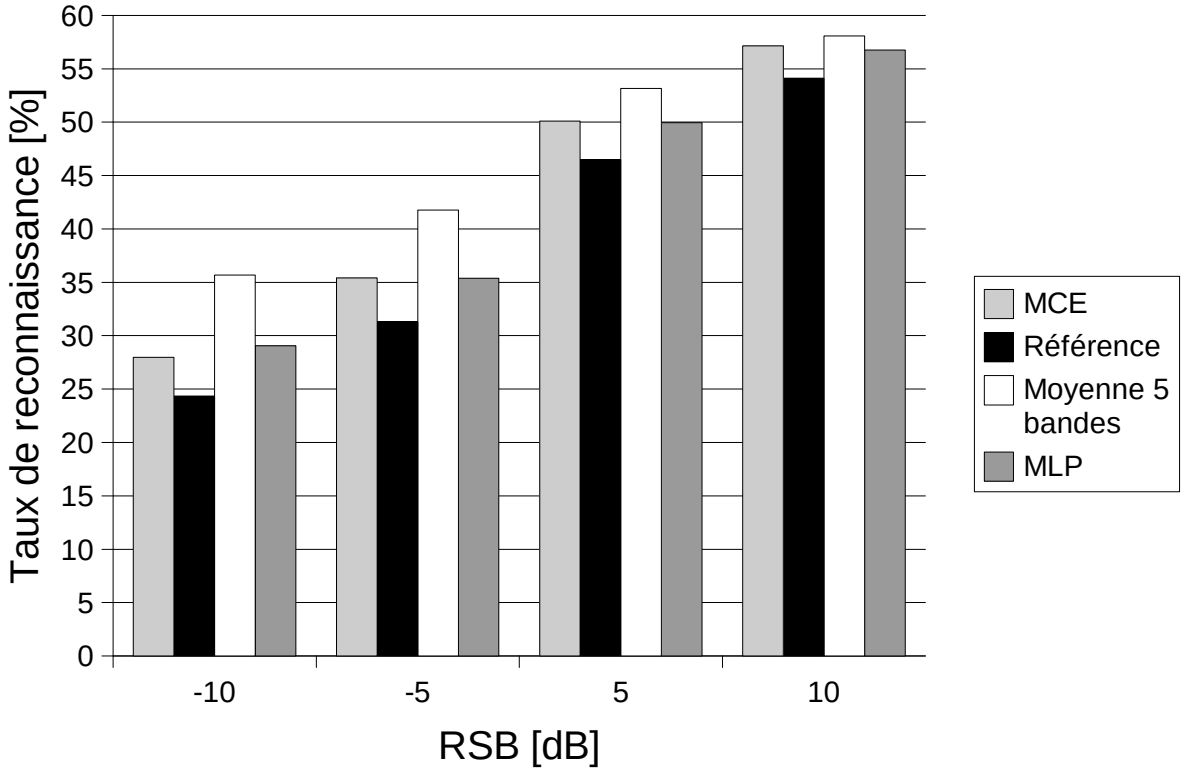

Figure 6.12 : Taux de reconnaissance, dans le bruit de voiture, des systèmes Multi-Bandes avec apprentissage du module de recombinaison

Comparons cette courbe avec celle obtenue dans le bruit filtré aigu : les deux courbes sont très similaires. La première remarque que nous pouvons faire est qu'ici, le système utilisant un perceptron ne semble jamais être le meilleur système, mais que la moyenne des 5 bandes le dépasse toujours. En fait, ce phénomène se retrouve avec le bruit filtré aigu pour des taux de reconnaissance comparables : par exemple, dans le bruit de voiture à 10 dB, le PMC atteint 56,8 % de reconnaissance et la moyenne des 5 bandes 58,1 %. De la même manière, dans le bruit filtré aigu à 30 dB, le PMC a un taux de reconnaissance de 58,6 % et la moyenne des 5 bandes de 59,4 %. Nous voyons donc que, à condition d'observer des plages de valeurs similaires entre les courbes des systèmes dans les deux types de bruit, le comportement de ceux-ci est très semblable.

Dans le cas du bruit de voiture, nous n'avons pas testé les systèmes pour des valeurs très faibles du niveau de bruit, comme cela a été le cas avec le bruit filtré aigu. Toutefois, au vu de ces conclusions ou même simplement en extrapolant les courbes présentées sur la figure 6.12, nous pouvons conclure que la différence entre la moyenne des 5 bandes et le PMC va en s'amenuisant au fur et à mesure que le niveau de bruit baisse. De plus, nous pouvons aussi prédire que les taux de reconnaissance du système Multi-Bandes avec recombinaison par perceptron finissent sûrement par dépasser ceux du système linéaire calculant la moyenne des 5 bandes, pour des taux de reconnaissance de l'ordre de 67 %.

Nous avons vu avec ce premier bruit « naturel » que les conclusions issues des expériences sur du bruit artificiel ne sont pas démenties, loin de là. Nous pouvons également noter que le bruit de voiture n'est pas très éloigné de certains bruits artificiels que nous avons utilisés jusqu'à présent : en effet, c'est un bruit très stable et dont le spectre ne varie que très peu dans le temps. Nous allons donc maintenant appliquer notre système sur un bruit qui varie beaucoup plus au cours du temps.

6.4.2. Bruit de cantine

Les bruits dynamiques par excellence sont ceux qui apparaissent à un instant donné pour disparaître ensuite rapidement et laisser éventuellement la place à un autre bruit. Par exemple, cela se produit dans la vie courante lorsqu'une porte claque, puis des bruits de pas apparaissent, puis un raclement de chaise sur le sol, etc. Mais, s'il est possible de reproduire effectivement ce type de bruit en enregistrant une scène de la vie courante, l'étude des taux de reconnaissance des systèmes automatiques dans ceux-ci est beaucoup plus difficile. En effet, les bruits « explosifs » comme ceux du claquement d'une porte affectent très fortement toutes les fréquences du spectre pendant un temps très court, et il est quasiment impossible, même pour un être humain, de comprendre ce qui a été prononcé juste à ce moment. En fait, le cerveau compense cette perte d'information en utilisant l'information acoustique dont il dispose juste avant et juste après cette brève interruption, information qui, associée au contexte pragmatique du discours, permet de reconstruire les données perdues. Or, nous ne nous intéressons pas au traitement sémantique de la parole, mais seulement à la manière d'extraire le maximum d'information du signal, même en présence de bruit. Ainsi, si nous enregistrions le claquement de la porte, puis que nous le repassions en boucle, de sorte à n'avoir aucun silence entre les instants de bruit, alors dans ces conditions, même un homme aurait beaucoup de mal à discerner ce que dit son interlocuteur. Néanmoins, il comprendrait mieux que la plupart des systèmes automatiques actuels de reconnaissance de la parole, même sans faire intervenir de niveaux sémantiques. Or, ce que nous recherchons avec le modèle Multi-Bandes, c'est justement de pouvoir améliorer ces performances en présence de bruit, uniquement en utilisant l'information acoustique que l'on peut extraire du signal. C'est pourquoi l'essentiel de notre étude porte sur des bruits ne variant pas beaucoup au cours du temps.

Nous avons néanmoins voulu tester notre système dans un bruit totalement différent de ceux que nous avons testés jusqu'à présent. Le bruit que nous avons choisi est généralement connu sous son appellation anglaise de « *cocktail party noise* ». Il s'agit en fait d'un enregistrement du bruit occasionné par une ou plusieurs conversations simultanées à celle qui nous intéresse et que nous pouvons appeler « conversation principale ». La tâche est donc extrêmement difficile, car il faut isoler un discours parmi tous ceux qui sont perçus. Le cerveau réalise remarquablement bien ceci, probablement en utilisant toute une multitude d'informations simultanément, comme par exemple la localisation spatiale de la source du signal grâce aux deux oreilles, ou encore l'identification du timbre d'une voix pour isoler celle-ci des autres, ou même des informations visuelles permettant d'associer un son aux mouvements des lèvres du locuteur. Ceci n'est toutefois pas possible pour les systèmes automatiques, pour qui aucun des ces trois indices n'est généralement utilisable. Ce qui ne nous empêche pas de tester notre système dans un tel bruit et d'essayer d'améliorer les résultats des HMM classiques.

Nous avons utilisé le bruit appelé « *babble noise* » extrait de la base NOISE-ROM-0, qui est un enregistrement réalisé dans un réfectoire occupé par plusieurs dizaines de personnes qui parlent bruyamment ensemble ou par petits groupes. Le fait même qu'il y ait de nombreuses personnes parlant en même temps facilite la tâche de reconnaissance, car il se crée une sorte de « brouhaha » duquel il est plus facile d'isoler la conversation principale. Ainsi, cette tâche aurait été beaucoup plus difficile si nous avions simplement enregistré une ou deux autres conversations rapprochées, elles-mêmes isolées, et que nous avions ajouté celles-ci à la conversation principale. Nous nous sommes contentés cependant pour le moment de tester notre système avec ce bruit de « cantine » qui était à notre disposition. Les résultats des différents systèmes dans ce bruit sont présentés dans les paragraphes suivants. Un spectrogramme de ce bruit est disponible sur la figure 6.13.

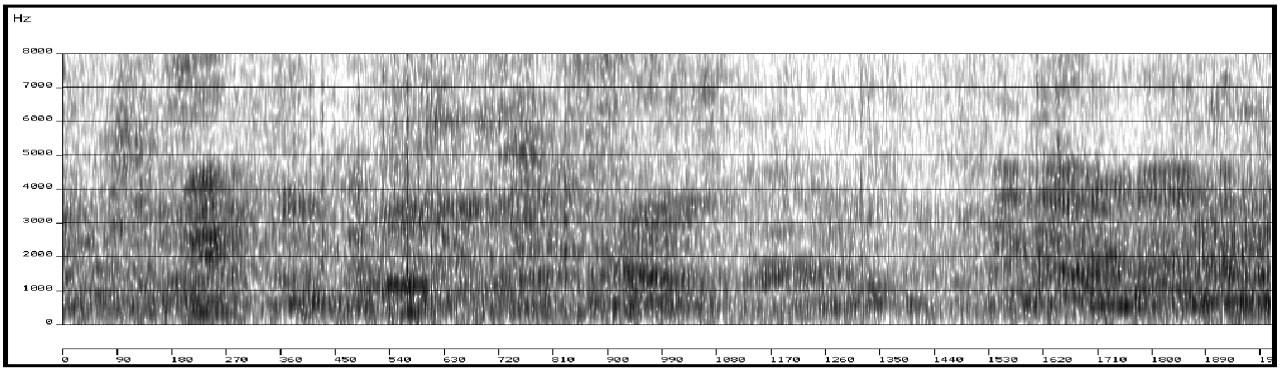

Figure 6.13 : Spectrogramme du bruit de cantine

✓ *Étude, dans le bruit de « cantine », des systèmes Multi-Bandes sans apprentissage du module de recombinaison*

Les résultats de cette étude sont donnés sur la figure 6.14.

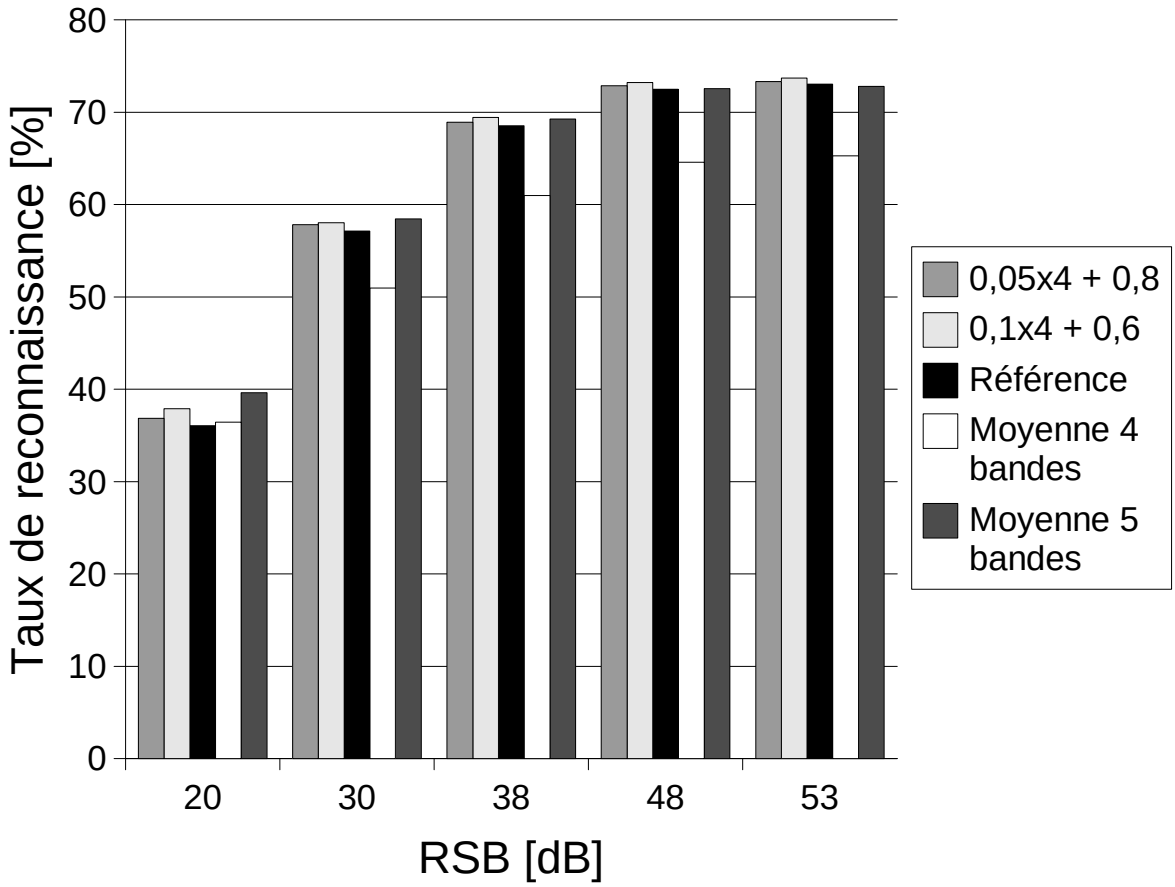

Figure 6.14 : Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans le bruit de cantine.

✓ ***Étude, dans le bruit de « cantine », des systèmes Multi-Bandes avec apprentissage du module de recombinaison***

Les résultats de cette étude sont donnés dans la figure 6.15.

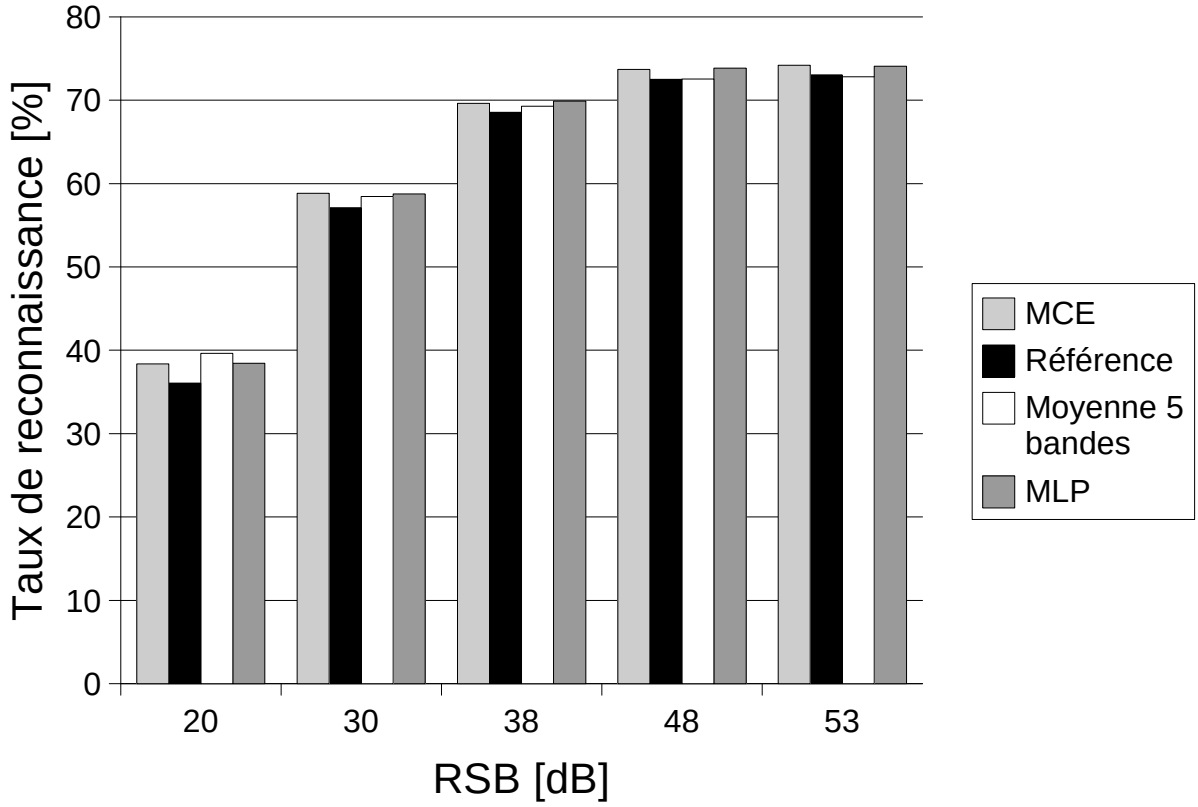

Figure 6.15 : Taux de reconnaissance, dans le bruit de cantine, des systèmes Multi-Bandes avec apprentissage du module de recombinaison.

✓ ***Commentaires***

Encore une fois, nous retrouvons exactement le même comportement des systèmes Multi-Bandes dans ce bruit que dans ceux que nous avons étudiés précédemment, malgré le fait que le type de bruit considéré ici est tout de même assez difficile par rapport aux autres. Si nous devions classer ce bruit dans l'une des deux catégories que nous avons définies ci-dessus, puisque la moyenne des quatre bandes obtient des résultats nettement inférieurs à ceux obtenus par la moyenne des cinq bandes, il semble probable que ce bruit appartienne à la catégorie du bruit blanc et du bruit filtré aigu, c'est-à-dire qu'il affecte de la même manière toutes les fréquences « utiles » du spectre de parole. Le fait que ces fréquences soient atteintes n'est guère étonnant, vu la définition même du bruit : nous ne devons pas oublier en effet qu'il s'agit avant tout de parole ! Toutefois, le fait qu'elles soient toutes équitablement affectées par le bruit corrobore ce que nous avions préalablement remarqué concernant la multitude des voix enregistrées : il en aurait peut-être été totalement différemment si seulement une ou deux autres conversations isolées avaient été ajoutées à la conversation principale, car dans ce cas, seules des parties réduites du spectre de la parole auraient été affectées par le bruit à un instant donné, et celui-ci aurait probablement appartenu à la deuxième catégorie de bruits, à savoir ceux qui n'affectent qu'une portion du spectre à un instant donné.

6.4.3. Conclusion

Les expériences que nous avons menées dans le bruit « naturel » ont clairement montré que la robustesse du Multi-Bandes est valable quel que soit le type de bruit additif concerné. De plus, les explications de cette robustesse issues des expériences avec du bruit artificiel se sont révélées toujours valables dans le bruit naturel et nous ont même permis de « prévoir » les résultats que nous avons obtenus en fonction du type de bruit utilisé. La conclusion de cette partie est donc globalement positive quant à l'utilité du Multi-Bandes dans le bruit. Toutefois, comme nous l'avons fait remarquer au début de celle-ci, il existe une multitude de bruits de la vie courante que nous n'avons pas pu tester avec notre système, même si nous avons considéré le plus grand nombre possible de types de bruits différents.

6.5. Apprentissage du module de recombinaison dans le bruit

6.5.1. Introduction

Après avoir vérifié l'hypothèse de robustesse du Multi-Bandes, nous allons maintenant essayer d'accroître celle-ci en utilisant un moyen très simple, l'apprentissage du module de recombinaison dans le bruit. Le but de cette expérience est donc d'augmenter les performances du Multi-Bandes dans le bruit. Intuitivement, cette démarche n'est pas justifiable, et les raisons suivantes peuvent être invoquées à l'encontre de celle-ci :

- Un système qui est entraîné dans le bruit est effectivement performant dans ce bruit, mais il devient dépendant de celui-ci et n'offre plus que de piètres résultats lorsque l'environnement sonore change.

- Le système Multi-Bandes serait bien plus robuste à un bruit particulier si les HMM eux-mêmes étaient entraînés dans ce bruit ! Pourquoi donc se restreindre au module de recombinaison ?

En ce qui concerne cette dernière remarque, il faut rappeler que notre but ici n'est pas de rendre le système Multi-Bandes robuste à un type de bruit particulier, auquel cas il faudrait effectivement entraîner également les HMM dans ce bruit, mais bien de le rendre robuste au plus grand nombre d'environnements possibles. Quant à la première remarque, nous avons pensé, avant de réaliser ces expériences, qu'entraîner le module de recombinaison dans le bruit permet de rendre le système Multi-Bandes plus robuste sans qu'il n'en devienne pour autant dépendant du bruit d'apprentissage, ou tout au moins qu'il ne le devienne que très peu. Cette hypothèse, peu intuitive au premier abord, mérite quelques explications. En fait, elle repose sur le rôle que nous accordons au module de recombinaison : celui-ci n'a pas pour but de mesurer une « distance » entre le signal et un modèle, mais plutôt d'accorder plus ou moins d'importance à telle ou telle bande, pour tel ou tel modèle. C'est le rôle des HMM de mesurer cette « distance » entre le signal et un modèle de référence. C'est pour cela que nous n'avons pas entraîné les HMM dans le bruit, car dans ce cas, les modèles de référence seraient en fait des modèles de phonèmes bruités, ce qui expliquerait alors la dépendance des HMM au type de bruit utilisé au cours de l'apprentissage. Néanmoins, comme nous l'avons signalé, le module de recombinaison ne construit pas de tels modèles, mais pondère plutôt les réponses des HMM. En entraînant le module de recombinaison dans le bruit, nous voulons donc lui faire apprendre quels sont les modèles de quelles bandes qui sont effectivement robustes et lesquels ne le sont pas, afin que les poids qu'il attribue à ceux-ci soient pertinents du point de vue de la robustesse. Or, si nous continuons à entraîner le module de recombinaison dans un environnement non bruité, il ne *pourra pas* savoir quels modèles sont robustes. Intuitivement, nous-mêmes savons parfaitement qu'il est vain de tenter de distinguer une fricative dans un bruit blanc en ne regardant que les hautes fréquences, et dans l'hypothèse où nous devions quand même décider si une fricative est présente ou pas, nous nous focaliserions certainement sur des indices situés dans les basses fréquences. C'est ce type de comportement que nous voulons faire acquérir au module de recombinaison.

Nous avons choisi d'entraîner le module de recombinaison dans du bruit blanc. En effet, il faut que le bruit ajouté au corpus d'apprentissage affecte équitablement toutes les fréquences afin que la robustesse réelle d'une bande pour un modèle soit apprise et qu'un biais ne soit pas introduit dès l'apprentissage. De plus, nous voulions utiliser un bruit dont les caractéristiques sont les plus « générales » possibles, afin que le système ne perde pas sa capacité d'abstraction.

6.5.2. Expérimentations

Rappelons que les HMM ont été entraînés dans un environnement non bruité et que seul le module de recombinaison, c'est-à-dire un perceptron dans les expériences qui suivent, a été entraîné dans un bruit blanc avec un rapport signal-bruit de 25 dB. Comme nous voulons essentiellement tester la robustesse de ce système pour le plus grand nombre d'environnements sonores possibles, nous ne présentons les résultats que pour un seul rapport signal-bruit. Ceux-ci apparaissent, pour tous les environnements considérés, dans le tableau 6.3. Trois systèmes sont testés dans ce tableau :

- Le système de référence ;
- Le système Multi-Bandes avec une recombinaison par un perceptron qui est entraîné dans un environnement non bruité : ce système est appelé « PMC » ;
- Le même système mais après un petit apprentissage (moins de cinq itérations) dans un environnement corrompu par le bruit blanc décrit ci-dessus. Ce dernier système est appelé « PMC bruité ».

Par rapport aux expériences précédentes, nous avons ajouté un nouveau bruit, appelé « sèche-cheveux », et qui est l'enregistrement du bruit d'un sèche-cheveux. Nous avons voulu modéliser, par l'intermédiaire de celui-ci, toutes les nuisances sonores qui apparaissent très fréquemment dans la vie quotidienne et qui sont en grande partie dues à des petits moteurs d'appareils électroménagers. Dans cet appareil intervient également un bruit de ventilateur et un « sifflement » provoqué par l'air s'écoulant à travers un orifice réduit. C'est donc un bruit relativement complexe, mais également assez proche du bruit blanc que nous avons étudié. Un spectrogramme de ce bruit est donné sur la figure 6.16.

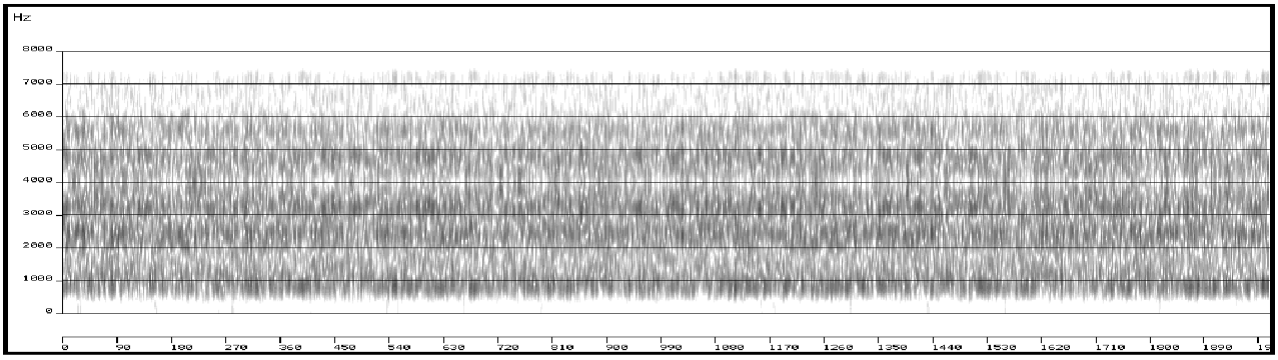

Figure 6.16 : Spectrogramme du bruit de sèche-cheveux

	PMC bruité	***Référence***	***PMC***
Environnement non bruité	72,6 %	73,3 %	74,4 %
Bruit blanc RSB = 20dB	56,8 %	40,6 %	43,8 %
Bruit filtré aigu RSB=20 dB	53,2 %	44,8 %	49,2 %
Bruit filtré grave RSB = 20 dB	28,8 %	19,9 %	27,1 %
Sèche-cheveux RSB=5 dB	34,4 %	29,5 %	32,2 %
Bruit de cantine RSB=20 dB	37,4 %	36,1 %	38,5 %
Bruit d'automobile RSB = -15 dB	31,3 %	24,4 %	29,1 %

Tableau 6.3 : Taux de reconnaissance du Multi-Bandes, avec apprentissage du module de recombinaison dans le bruit blanc, dans différents environnements

Commentaires :

- Tout d'abord, nous pouvons remarquer que lorsque le perceptron est entraîné dans le bruit blanc, le taux de reconnaissance du système Multi-Bandes s'améliore beaucoup dans ce bruit (+13 % absolus), ce qui est tout à fait normal, même si le rapport signal-bruit est moins élevé pendant le test qu'au cours de l'apprentissage. Néanmoins, nous pouvons remarquer que l'amélioration est très grande au regard de la brièveté de l'apprentissage du perceptron dans le bruit, ce qui pourrait bien suggérer qu'il y a une profonde différence entre le comportement que devrait avoir le perceptron dans un environnement bruité et celui qu'il a dans un milieu non bruité.

- Ensuite, nous pouvons relever plusieurs environnements, autres que le bruit blanc, pour lesquels l'apprentissage du module de recombinaison a été bénéfique. Il s'agit notamment du bruit filtré aigu, du bruit filtré grave, du bruit de sèche-cheveux et du bruit d'automobile. La majorité des environnements étudiés sont ainsi mieux traités par notre système lorsque le perceptron est entraîné dans le bruit blanc. Ceci montre donc que l'hypothèse qui est à la base de ces expériences est correcte, ou du moins qu'elle contient une part de vérité, même si d'autres phénomènes interviennent.

- En effet, deux environnements ne sont pas aussi favorables au nouveau système : il s'agit de l'environnement non bruité et du bruit de cantine. Pour le premier, nous pouvons facilement le comprendre : le perceptron étant initialement entraîné dans un environnement non bruité, il était alors parfaitement adapté à cet environnement. Si nous l'entraînons ensuite dans du bruit blanc, il va modifier cette configuration et s'adapter au nouvel environnement bruité. Il est donc normal que ses taux de reconnaissance dans l'environnement initial baissent légèrement. Cependant, la chute des performances du nouveau système dans le bruit de cantine, même si elle reste faible, montre un des points faibles de l'hypothèse qui est à la base de ces expériences. Ainsi, lorsque nous entraînons le perceptron dans le bruit blanc, le nouveau système n'est plus robuste à tous les autres types de bruit, mais il en existe au moins une catégorie pour laquelle ses résultats se dégradent. En fait, la principale différence entre le bruit de cantine et les autres types de bruits étudiés ici ne se situe pas dans la dimension spectrale, mais plutôt dans la dimension temporelle. En effet, le premier est dynamique, c'est-à-dire qu'il se modifie constamment au cours du temps, alors que les autres sont plutôt statiques. Il est donc fort possible que, en ayant entraîné le perceptron dans le bruit blanc, nous lui ayons fait apprendre quelle est la robustesse de chaque bande et de chaque modèle dans le bruit, comme nous le voulions au départ, mais il semble également que cette amélioration de la sélectivité dans le domaine spectral s'accompagne d'une plus grande sensibilité dans le domaine temporel. Cette dualité pourrait être également intéressante à considérer.

6.5.3. Étude du perceptron après cet apprentissage

En plus des expériences qui ont été décrites ci-dessus, nous avons voulu comprendre un peu mieux ce qui se passe lorsque le perceptron est entraîné dans un environnement bruité. En effet, si au cours de cet apprentissage le perceptron réalise bien ce que nous souhaitions, à savoir accroître l'importance qu'il accorde aux bandes et aux modèles robustes, alors nous avons voulu savoir quelles étaient effectivement les bandes et les modèles robustes. Or, nous avons montré expérimentalement dans la section 6.3.3 que les sous-bandes sont plus robustes que le spectre complet, et nous devons donc nous attendre à ce que l'importance accordée par le perceptron à ces sous-bandes augmente au dépend de la cinquième bande. Les deux raisons principales qui motivent cette étude du perceptron sont donc :

1. Vérifier grâce à cette expérience l'hypothèse de plus grande robustesse des sous-bandes vis-à-vis du spectre complet.
2. Vérifier dans le même temps que le perceptron a réellement appris à donner plus d'importance aux bandes robustes au cours de son apprentissage dans le bruit.

✓ ***Comment mesurer l'importance accordée à un modèle ou à une bande par le perceptron ?***

Pour réaliser cette étude, nous avons fourni au perceptron des entrées artificielles, pour lesquelles toutes les valeurs sauf une sont positionnées à 0. Nous disposons alors en sortie du perceptron d'un bon indicateur de l'importance qu'accorde celui-ci à une entrée. Deux remarques peuvent être faites en ce qui concerne ces indicateurs :

1. Pour chaque entrée considérée, il y a autant d'indicateurs qu'il y a de sorties du perceptron, c'est-à-dire de classes. En effet, chaque entrée a une influence différente selon la classe en sortie que nous considérons. De plus, cette méthode nous permet également d'analyser l'influence d'une entrée spécifique sur une classe donnée, ce qui fait intervenir le pouvoir discriminant du perceptron.
2. Ces sorties ne sont que des indicateurs, et ils ne mesurent pas exactement l'importance qu'accorde le perceptron à une entrée. En effet, en fonction de la définition que l'on donne à cette « importance », il faut selon les cas considérer ou pas les biais qui existent dans chaque neurone du perceptron. Toutefois, dans notre cas, ces biais n'ont pas une grande importance car nous allons dans la suite soustraire ces indicateurs entre eux, ce qui a pour effet de réduire considérablement l'influence des biais.

Pour chaque modèle *i* d'une bande, nous activons donc l'unique entrée correspondante du perceptron et mesurons dans A(*i*) et B(*i*) la valeur obtenue à la sortie *i* du perceptron, respectivement avant et après son apprentissage dans le bruit blanc. Nous ne considérons donc pas ici l'influence d'un modèle *i* sur un autre modèle *j*, mais uniquement l'influence d'un modèle sur lui-même, afin de ne pas compliquer inutilement l'étude. La figure 6.17 reporte les différences B(*i*) - A(*i*) pour les modèles de la première sous-bande et la figure 6.18 reporte les mêmes différences pour les modèles du spectre complet.

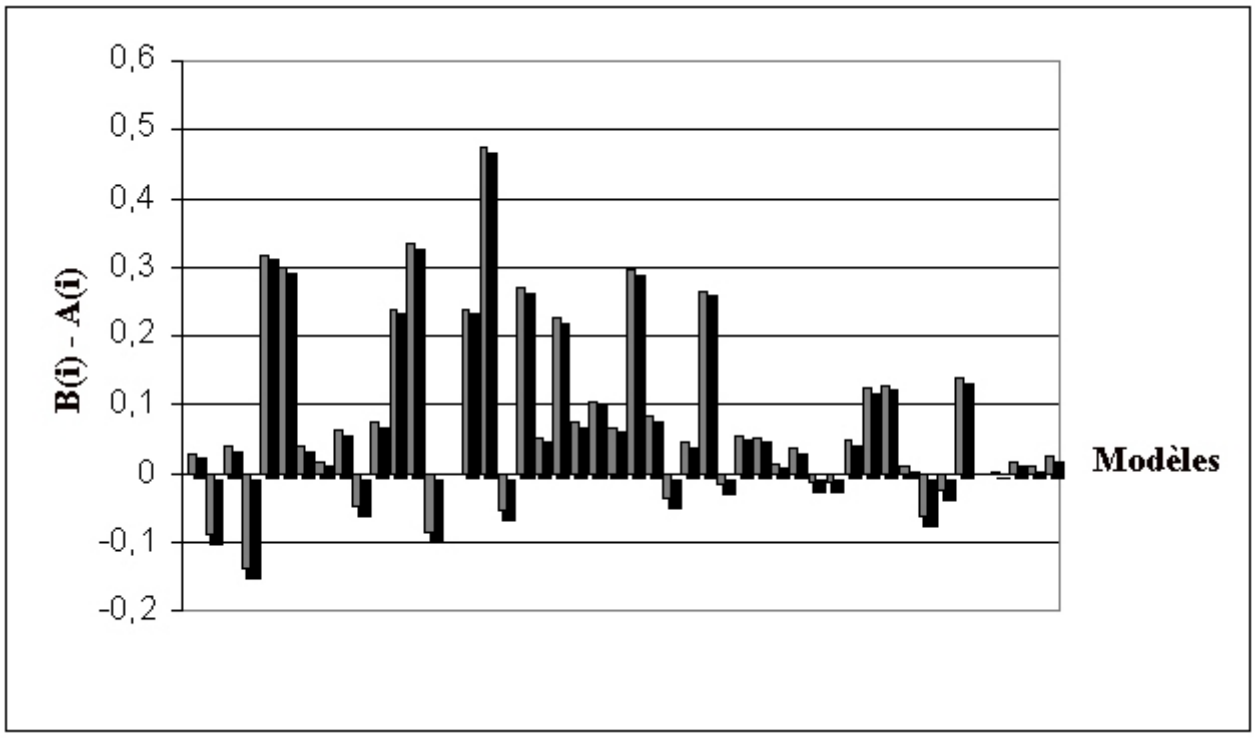

Figure 6.17 : Modification de l'influence de la première sous-bande due à l'apprentissage du module de recombinaison dans le bruit blanc

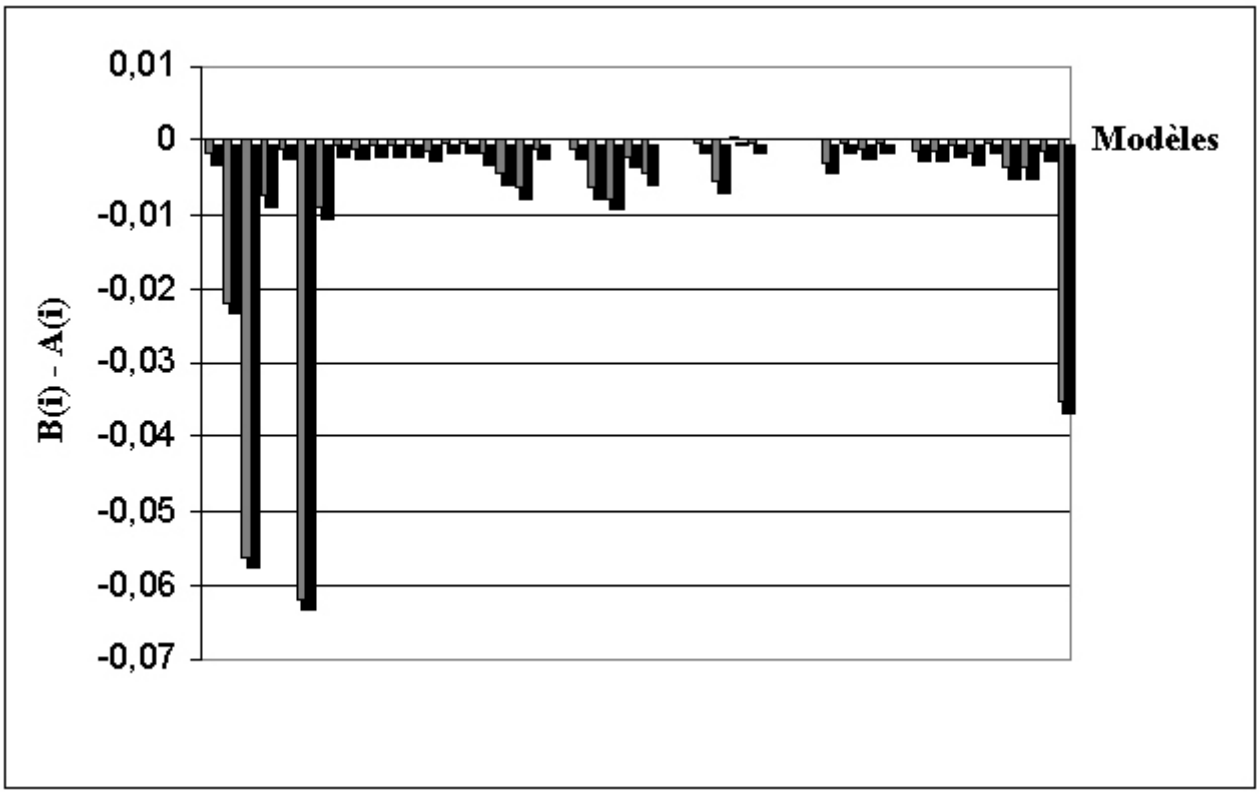

Figure 6.18 : Modification de l'influence du spectre complet due à l'apprentissage du module de recombinaison dans le bruit blanc

✓ ***Commentaires***

L'opposition entre les deux courbes est très claire, car presque toutes les différences calculées pour la première sous-bande sont positives, tandis que toutes celles correspondant à la cinquième bande sont négatives. Ceci signifie qu'après son apprentissage dans le bruit blanc, le perceptron a « choisi » d'accorder plus d'importance à la première sous-bande, pour quasiment tous les modèles, et moins d'importance au spectre complet, également pour tous les modèles. Nous ne présentons pas les résultats des trois autres sous-bandes qui sont identiques. Nous avons ainsi clairement vérifié les deux hypothèses exposées ci-dessus, c'est-à-dire d'une part que les sous-bandes sont plus robustes au bruit que le spectre complet, même dans le bruit blanc, et d'autre part que le perceptron a appris à tirer parti de ce phénomène grâce à son apprentissage en milieu bruité. Nous ne poursuivons pas ici plus en détail l'analyse de l'influence des modèles un par un dans le bruit blanc, mais il est certain qu'une telle analyse, bien que demandant beaucoup de temps, nous permettrait peut-être de mieux comprendre quelles sont les caractéristiques qui permettent ou au contraire empêchent certains phonèmes d'être robustes au bruit.

6.6. Étude dans les milieux extrêmement bruités

Dans presque toutes les expériences présentées dans ce chapitre, le niveau du bruit ajouté reste relativement faible, dans la mesure où les rapports signal-bruit étudiés sont rarement inférieurs à 20 dB. En fait, nous avons choisi le niveau de bruit à ajouter au signal non pas en fonction du rapport signal-bruit obtenu, mais plutôt en fonction des taux de reconnaissance que les systèmes parvenaient à réaliser dans de tels bruits. En effet, d'une part, comme nous l'avons déjà signalé, le calcul du rapport signal-bruit n'est qu'approximatif, mais surtout, lorsque les taux de reconnaissance des systèmes passent sous le seuil des 20 %, l'interprétation des résultats devient impossible.

Nous pouvons illustrer ce phénomène en donnant les taux de reconnaissance de plusieurs systèmes dans le bruit blanc à 0 dB :

Système :	**Référence**	**PMC**	**Moyenne 4 bandes**	**Moyenne 5 bandes**
Taux de reconnaissance :	20,2 %	20,4 %	20,4 %	20,4 %

Tableau 6.4 : Taux de reconnaissances de quelques systèmes dans le bruit blanc à 0 dB.

En fait, dans un tel niveau de bruit, tous les systèmes associent une réponse unique aux segments de parole qui leur sont présentés. Dans le cas du bruit blanc, ils répondent tous que le modèle qui a généré le signal est celui correspondant à l'absence de parole (noté h# dans l'annexe 2). Ceci explique pourquoi ils ont tous le même taux de reconnaissance, qui est en fait le pourcentage d'apparition de h# dans le corpus de test.

Pourquoi donnent-ils tous cette réponse ? Tout simplement parce que le bruit est si fort qu'il couvre le signal de parole et les systèmes automatiques n'arrivent plus à extraire une information pertinente de ce signal. Ils donnent donc pour réponse le modèle qui est le plus proche, pour eux, du bruit ajouté. Dans le cas du bruit blanc, nous comprenons facilement qu'il s'agit du modèle h#. Toutefois, dans le cas d'un autre bruit n'affectant que certaines fréquences, il peut s'agir d'un tout autre modèle, par exemple une voyelle dont un formant correspondrait au bruit ajouté, et les taux de reconnaissance seraient alors tout autres.

Quoi qu'il en soit, ceci prouve que tous les systèmes qui ont été considérés jusqu'à présent sont incapables de gérer un tel niveau de bruit. Or, il se trouve que l'oreille humaine parvient encore, certes avec difficulté mais tout de même assez clairement, à distinguer certains phonèmes entre eux sans faire intervenir d'information syntaxiques ou sémantiques. Nos modèles acoustiques ont donc encore de nombreux progrès à réaliser avant de parvenir à gérer ces niveaux de bruit !

Nous ne reparlerons plus dans ce mémoire de bruits de cette intensité, tout simplement parce qu'ils ne peuvent être gérés tels quels, ni par nos systèmes, ni par le système de référence que nous utilisons. Il faudrait donc pré-traiter le signal grâce à un étage de débruitage, mais ce n'est pas le but de notre travail.

6.7. Identification du langage

6.7.1. Introduction

Nous n'avons pas seulement étudié le paradigme Multi-Bandes dans des tâches de reconnaissance de la parole, mais également pour une tâche d'identification du langage. Nous avons fait ceci pour les raisons suivantes :

1. Le paradigme Multi-Bandes est un principe général de modélisation de la parole, ce qui signifie qu'il peut être appliqué à plusieurs tâches de Traitement Automatique du Langage Naturel (TALN), et pas seulement à la Reconnaissance Automatique de la Parole (RAP). Ainsi, il nous a paru parfaitement légitime de tester ce principe sur plusieurs corpus, et même sur plusieurs tâches de TALN, à savoir la RAP et l'identification du langage qui est présentée ici. Il faut toutefois remarquer que le cœur de notre travail concerne la RAP car il s'agit de l'application privilégiée de notre système, mais nous sommes parfaitement disposés à étendre ce principe à d'autres applications et c'est ce que nous montrons ici.

2. Nous voulions tester le plus grand nombre de corpus possible afin de valider correctement le principe Multi-Bandes, et nous avons ainsi utilisé dans cette partie un nouveau corpus, le corpus OGI-ML [Muthusamy92b]. Ce corpus est un corpus multilingue, composé de conversations spontanées enregistrées à travers le téléphone dans onze langues différentes.

3. Le paradigme Multi-Bandes est robuste au bruit additif, comme nous l'avons montré dans les parties précédentes, mais nous avons également voulu le tester dans du bruit convolutif. Or, le signal de parole qui compose OGI-ML a été enregistré au travers de lignes téléphoniques qui, comme nous le savons, introduisent des bruits convolutifs très importants.

4. Enfin, ce corpus est un corpus *extrêmement* difficile, car il est composé de parole *spontanée* et enregistrée par *téléphone*. De plus, la qualité de ces enregistrements est très *hétérogène*, car elle dépend directement de la qualité des lignes téléphoniques dans les différents pays concernés. Tous ces facteurs rendent l'utilisation de ce corpus particulièrement délicate et nous avons donc voulu tester ce que devient le Multi-Bandes dans des conditions aussi difficiles. Ceci explique également pourquoi nous avons placé cette partie dans la partie concernant les expériences en milieu bruité.

5. L'identification du langage est une tâche qui a une utilité certaine dans un grand nombre d'applications de reconnaissance de la parole. Ainsi, elle devra très certainement être utilisée par les systèmes automatiques qui équipent petit à petit les centres d'appels téléphoniques, comme par exemple les services après-vente en ligne ou les services téléphoniques de réservation d'hôtels, pour qui il est nécessaire de savoir dans quelle langue parle leur correspondant. De même, afin que les applications de dictée vocale soient performantes, il faut qu'elles puissent reconnaître que certains passages du texte qui leur est dicté appartiennent à une langue étrangère de façon à les orthographier correctement.

6.7.2. Adaptation du système à cette tâche

La tâche d'identification du langage que nous avons choisie a pour but d'identifier la langue dans laquelle est parlé un signal donné de parole. Le corpus OGI-ML étant particulièrement bruité, nous avons utilisé un module d'élimination des segments de bruit, car ceux-ci ne permettent pas de discrimination entre les langues.

Ce module de discrimination bruit/parole utilise deux classes modélisées par des HMM, l'une représentant la parole et l'autre le bruit. Ces modèles ont été entraînés sur une petite partie du corpus d'apprentissage dont un étiquetage en classes phonétiques a été réalisé manuellement. Nous n'utilisons de cet étiquetage que l'information de segmentation permettant de différencier les segments du signal correspondant au silence/bruit de ceux correspondant à de la parole. Nous n'utilisons donc pas l'information provenant de l'étiquetage des segments de parole en différentes unités phonétiques. En fait, pour cet étage de pré-traitement du signal, une seule classe représente tous les segments de parole confondus. C'est une classification très grossière, mais qui a le mérite d'être simple et dont le seul but est d'isoler les segments utiles du signal de ceux qui ne le sont pas. Ces deux modèles sont ensuite utilisés pour segmenter tout le corpus d'apprentissage et de test en segments de parole et de bruit, et seuls les segments de parole sont passés au classifieur principal.

La classification des langues est réalisée grâce à un simple dictionnaire (*codebook*) obtenu par quantification vectorielle de tous les vecteurs de parole et constitué de 8, 16, 32 ou 64 centroïdes par langue. L'apprentissage de ces modèles est réalisé par l'algorithme LBG, dont une description est donnée dans [Siohan98]. Cette modélisation d'une langue est très rudimentaire, et de nombreux systèmes plus perfectionnés existent [Reyes94][Zissman96]. Toutefois, le but que nous nous étions fixé étant de valider l'utilisation du modèle multi-bandes pour une tâche d'identification du langage, nous avons délibérément choisi un modèle qui soit facilement manipulable de par sa simplicité. Il est néanmoins bien entendu possible, et même souhaitable, de tester le Multi-Bandes en identification des langues avec des modèles plus complexes.

Les classifieurs décident finalement de la langue parlée après avoir traité une phrase de 10 secondes. Ils ont alors calculé un score pour chaque langue, score qui correspond à la somme des distances entre chaque vecteur acoustique et le centroïde le plus proche appartenant au modèle de la langue considérée. Ces scores sont ensuite passés au module de recombinaison. Nous avons testé le système sur deux langues, l'anglais et le japonais, le corpus d'apprentissage étant composé de 939 phrases de parole spontanée, et le corpus de test de 252 phrases. Les vecteurs acoustiques sont composés simplement de 12 coefficients MFCCs. Le système complet est représenté sur la figure 6.19.

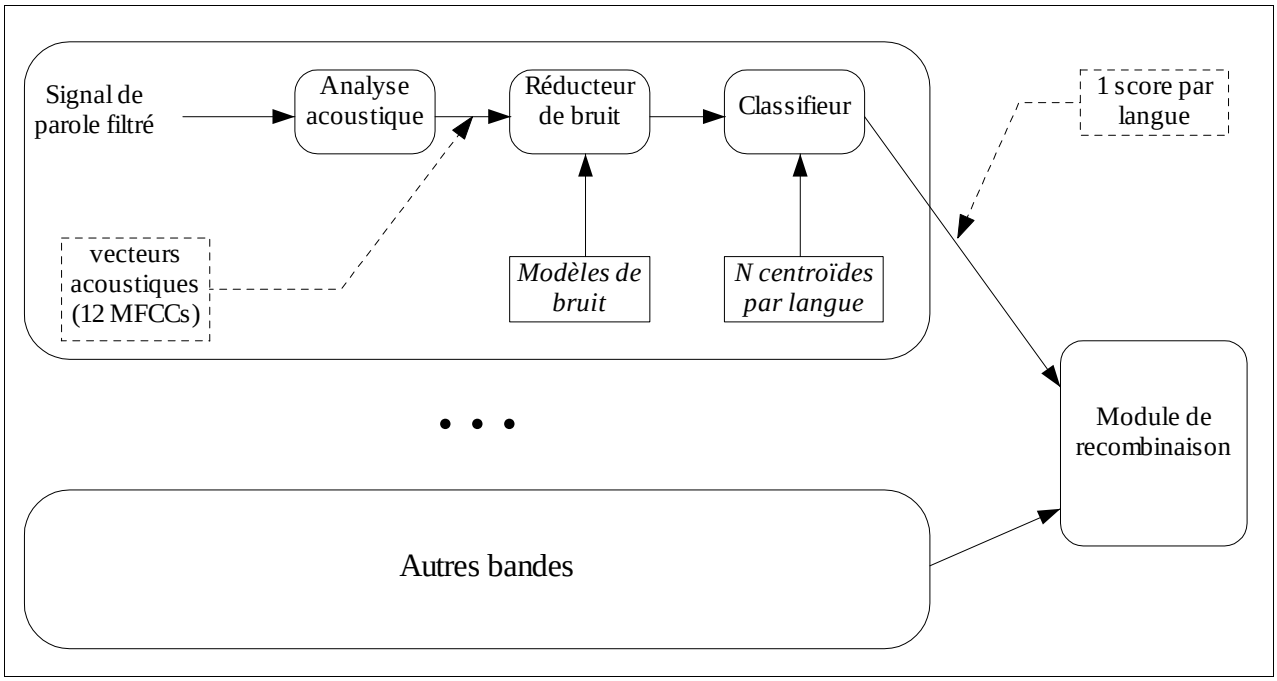

Figure 6.19 : Schéma général du système Multi-Bandes adapté à l'identification automatique de la langue

6.7.3. Étude des classifieurs individuels et comparaison avec d'autres résultats expérimentaux

Pour cette tâche, le système de référence est le système composé d'un dictionnaire de *N* centroïdes par langue, appliqué à l'ensemble des fréquences du spectre. Chaque test consiste à identifier la langue parlée parmi les deux suivantes : l'anglais et le japonais. Les résultats de ce système et des classifieurs partiels considérés un par un sont présentés dans le tableau 6.5. L'intervalle de confiance de ces résultats est seulement de +/- 4,4 %. C'est peu, mais il est très difficile de faire mieux sur seulement deux langues, car la taille du corpus de test disponible dans OGI n'est pas suffisante. Les bandes que nous avons utilisées ont les limites fréquentielles suivantes : [0 ... 900 Hz], [800 ... 1600 Hz], [1500 ... 2500 Hz] et [2300 ... 4000 Hz]. Comme précédemment, le spectre total est ajouté en tant que cinquième bande au système Multi-Bandes.

	8 centroïdes	***16 centroïdes***	***32 centroïdes***	***64 centroïdes***
référence	29,4 %	27,0 %	26,6 %	23,4 %
sous-bande 1	27,8 %	33,3 %	29,8 %	32,9 %
sous-bande 2	40,4 %	33,7 %	36,1 %	33,7 %
sous-bande 3	38,1 %	39,7 %	34,1 %	34,5 %
sous-bande 4	39,3 %	34,9 %	36,5 %	35,30%

Tableau 6.5 : Taux d'erreur pour les classifieurs considérés individuellement en fonction du nombre de centroïdes par langue.

La première remarque que nous pouvons faire est que ces taux d'erreur sont très élevés, ce qui, d'une part prouve la difficulté du corpus, et d'autre part peut s'expliquer par le fait que les modèles sont assez simples. De plus, dans un certain nombre de cas, les taux d'erreur du spectre total et des sous-bandes, notamment pour la première bande, sont du même ordre de grandeur, ce qui est sûrement dû en partie au bruit convolutif présent dans le signal. Enfin, nous pouvons remarquer qu'accroître le nombre de centroïdes n'augmente pas obligatoirement les performances du système, surtout pour les sous-bandes, ce qui montre la difficulté que le système a pour partitionner les vecteurs en classes distinctes et représentatives d'une langue.

✓ ***Comparaison avec les résultats obtenus par d'autres systèmes***

Certains systèmes apparaissant dans la littérature parviennent néanmoins à obtenir des taux de reconnaissance très corrects sur cette tâche et ce corpus, mais c'est généralement au prix d'un modèle beaucoup plus complexe de la parole et d'un étage plus efficace de débruitage. Par exemple, pour la même tâche et le même corpus, Zissman obtient 16 % d'erreur avec des mélanges de gaussiennes et seulement 8 % d'erreur avec un modèle noté PPR qui parallélise plusieurs reconnaisseurs phonétiques [Zissman96].

6.7.4. Résultats des systèmes Multi-Bandes

✓ ***Étude des systèmes Multi-Bandes sans apprentissage du module de recombinaison***

De la même manière que pour la tâche de reconnaissance de la parole, nous avons testé plusieurs systèmes, selon que le module de recombinaison est entraîné ou pas. Dans le cas où il n'y a pas d'apprentissage du module de recombinaison, nous avons testé deux systèmes : la moyenne des cinq bandes et le système accordant un poids de 0,6 au spectre complet et un poids de 0,1 aux quatre sous-bandes. Les résultats des tests pour ces systèmes sont présentés dans le tableau 6.6.

	8 centroïdes	***16 centroïdes***	***32 centroïdes***	***64 centroïdes***
référence	29,4 %	27,0 %	26,6 %	23,4 %
moyenne 5 bandes	29,4 %	27,3 %	26,2 %	25,4 %
0,1 x 4 + 0,6	31,3 %	26,9 %	25,8 %	23,4 %

Tableau 6.6 : Taux d'erreur pour les systèmes Multi-Bandes sans apprentissage du module de recombinaison

Globalement, aucune amélioration n'est apportée par les modèles linéaires simples.

✓ ***Étude des systèmes Multi-Bandes avec apprentissage du module de recombinaison***

Nous avons donc réalisé un apprentissage sur les coefficients de la recombinaison linéaire par l'algorithme MCE, et utilisé également un perceptron pour réaliser la recombinaison, ce qui a permis d'améliorer les résultats qui sont présentés dans le tableau 6.7.

Classifieurs	***Taux d'erreur***
Référence (spectre complet)	29,4 %
MCE	21,0 %
PMC	25,4 %

Tableau 6.7 : Taux d'erreur pour les systèmes Multi-Bandes avec apprentissage de la recombinaison (8 centroïdes)

Seuls les modèles à 8 centroïdes ont été testés, car le nombre de coefficients à estimer étant directement proportionnel au nombre de centroïdes, les algorithmes d'apprentissage comme le MCE ou le PMC ne convergeaient pas lorsque le nombre de paramètres est trop grand.

✓ ***Conclusion***

Ces expériences sont satisfaisantes dans la mesure où leur rôle qui consistait à appliquer le paradigme Multi-Bandes à une nouvelle tâche de parole a été rempli. Elles le sont moins si on considère les taux d'erreur obtenus qui restent tout de même relativement élevés. Ceci peut s'expliquer par la simplicité des modèles mis en jeu, qui ne peuvent modéliser correctement un corpus aussi difficile que celui-ci. Ces résultats ne remettent pas en cause le principe du Multi-Bandes, qui a tout de même réussi à améliorer sensiblement les performances assez mauvaises des modèles initiaux, mais des tests s'appuyant sur une modélisation plus poussée et plus adaptée au corpus sont nécessaires afin de pouvoir concevoir un système réellement efficace. Nous pouvons donc dire que ces expériences sont intéressantes d'un point de vue théorique mais qu'elles sont insuffisantes pour mettre en œuvre un système destiné à une application réelle, dans lequel des étages initiaux de débruitage du signal doivent être inclus.

6.8. Expériences en milieu non bruité

6.8.1. Introduction

Le Multi-Bandes a tout d'abord été considéré comme « utile seulement dans le bruit » par une majorité de chercheurs, et les premières expériences réalisées dans la littérature montraient qu'effectivement, dans un environnement non bruité, le Multi-Bandes n'améliorait pas les résultats du système de référence. Depuis peu, l'utilisation conjointe des sous-bandes, qui sont limitées fréquentiellement, et du spectre complet, dément cette affirmation et permet d 'exhiber des résultats surpassant ceux du système de référence dans un environnement non bruité. Nous avons été parmi les premiers à publier cette idée [Cerisara98a], qui trouve son origine dans les premières expériences que nous avons réalisées et qui montrent que le Multi-Bandes donne de meilleurs résultats que le système de référence lorsque du bruit blanc est ajouté au signal de parole.

Comme nous l'avons déjà vu précédemment, il est possible d'expliquer ces résultats grâce à une plus grande robustesse au bruit des sous-bandes que du spectre complet. Toutefois cette expérience, jointe aux considérations psycho-acoustiques qui ont été présentées dans le chapitre introductif, nous a amené à émettre l'hypothèse qui est à la base de cette partie 6.8, et que l'on peut formuler ainsi :

L'information fournie par l'ensemble des sous-bandes est différente de celle disponible dans le spectre complet.

Ceci peut paraître étonnant dans la mesure où nous utilisons les mêmes filtres lors de l'analyse acoustique, ce qui signifie que les énergies à partir desquelles sont calculées les coefficients cepstraux sont les mêmes pour le spectre complet et pour les sous-bandes. Ce point est incontestable, et l'information disponible à ce niveau est effectivement la même pour les deux systèmes. Mais si on considère les HMM qui sont utilisés ensuite comme des systèmes permettant de réduire la quantité d'information disponible, alors ce processus de réduction est différent selon que le HMM est appliqué au spectre complet ou à une sous-bande. Dans ce dernier cas, la quantité d'information mise à disposition du HMM est plus petite que dans le premier cas, ce qui implique que l'information disponible en sortie des HMM est effectivement différente pour les deux systèmes.

Différente, mais pas forcément moindre, car la réduction n'est pas la même dans les deux cas. Il est donc tout à fait possible de retrouver en sortie des HMM des sous-bandes de l'information qui aurait été perdue pendant le décodage du HMM travaillant sur le spectre complet. Auquel cas l'utilisation conjointe des sous-bandes et du spectre complet permettrait d'augmenter les taux de reconnaissance du système final, dans la mesure où le module de recombinaison sait tirer parti de cette information supplémentaire fournie par les sous-bandes. Ceci est loin d'être évident, car il est probable que la quantité d'information supplémentaire fournie par les sous-bandes est « petite » comparée à l'information fournie par le spectre complet. Nous avons tout de même réalisé un certain nombre d'expériences afin de mettre en évidence ce phénomène, expériences que nous présentons dans la suite.

6.8.2. Résultats des expériences

Tous les systèmes qui ont été étudiés jusqu'à présent sont maintenant testés directement sur le signal issu de TIMIT, qui peut aisément être assimilé à de la parole enregistrée dans un environnement non bruité. Ces résultats sont reportés dans le tableau 6.8.

Système	***Taux de reconnaissance***
[0 ... 538 Hz]	39,4 %
[461 ... 1000 Hz]	36,8 %
[923 ... 2823 Hz]	48,2 %
[2374 ... 7983 Hz]	40,4 %
[0 ... 7983 Hz] = Système de Référence	73,3 %
Moyenne 4 sous-bandes	65,5 %
Moyenne 5 bandes	73,4 %
0,1*4 + 0,6	73,9 %
0,05*4 + 0,8	73,5 %
MCE	74,0 %
PMC	74,4 %

Tableau 6.8 : Taux de reconnaissance de tous le systèmes dans un environnement non bruité

6.8.3. Commentaires

Plusieurs remarques peuvent être effectuées concernant ces résultats :

1. Contrairement au cas bruité, la différence entre les taux de reconnaissance du système de référence utilisant tout le spectre et des sous-bandes qui n'utilisent qu'une partie du spectre est très nette, de l'ordre de 30 % absolu en faveur du spectre complet. Ceci montre l'importance de celui-ci pour la reconnaissance en environnement non bruité. C'est pourquoi il nous faut non seulement obligatoirement considérer le spectre complet lui-même dans le système Multi-Bandes si nous ne voulons pas perdre une grande quantité d'information, mais également accorder une grande influence à celui-ci dans le processus de décision final. C'est ce qui est réalisé par exemple lorsque nous fixons les coefficients de la recombinaison linéaire à 0,1 pour les sous-bandes et à 0,6 pour le spectre total.

2. Néanmoins, nous voyons également qu'il ne faut pas considérer uniquement le spectre complet, ni même lui donner trop d'importance, comme le montre le système linéaire dont les coefficients de la recombinaison sont fixés à 0,05 pour les sous-bandes et à 0,8 pour la cinquième bande. En effet, les taux de reconnaissance sont plus faibles pour ce système que pour le système qui accorde un peu moins d'importance au spectre complet. Or, ceci n'est pas un phénomène marginal n'apparaissant que pour certaines valeurs des coefficients. Nous avons en effet testé les taux de reconnaissance pour d'autres valeurs de ces coefficients, en accordant de plus en plus d'importance au spectre complet. Ces tests sont présentés sur la figure 6.20, et nous voyons clairement que la décroissance s'amorce dès 0,5 et continue jusqu'à 1 (l'axe des abscisses représente le coefficient accordé à la cinquième bande, ceux accordés aux sous-bandes étant la différence entre l'unité et la valeur portée en abscisse, multipliée par le nombre de bandes).

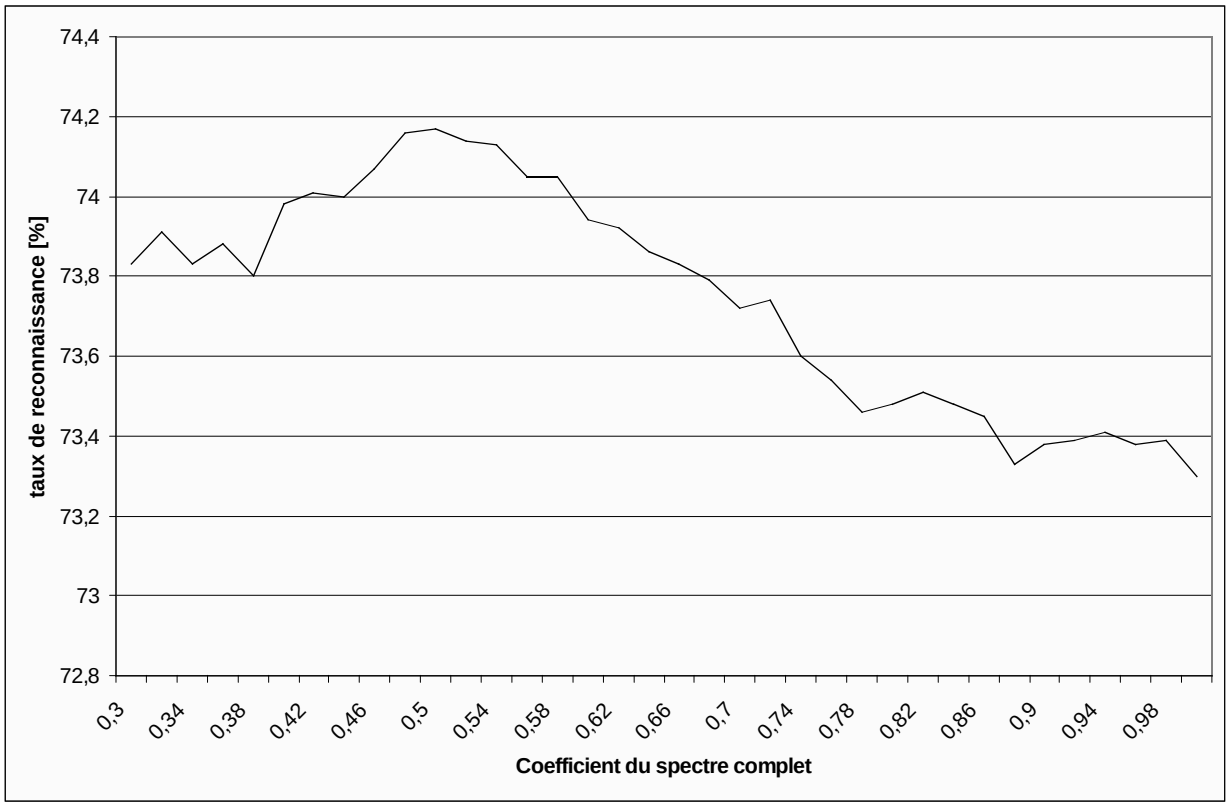

Figure 6.20 : Courbe du taux de reconnaissance pour un système Multi-Bandes en fonction de l'importance accordée au spectre complet

3. Tous les systèmes Multi-Bandes présentés (sauf la moyenne des 4 sous-bandes) ont des taux de reconnaissance supérieurs au système de référence. De plus, contrairement à ce que nous avions observé en milieu bruité, la hiérarchie entre les systèmes Multi-Bandes avec ou sans apprentissage du module de recombinaison est cette fois très nettement en faveur des systèmes avec apprentissage. Le meilleur système est celui qui utilise une recombinaison neuronale. Il faut toutefois noter que la différence entre ce système et celui de référence n'est pas significative. Nous avons donc testé à nouveau ces deux systèmes sur l'ensemble du corpus de test de TIMIT, qui est dix fois plus volumineux que la partie « coretest » utilisée pour les autres systèmes. Nous trouvons alors approximativement les mêmes résultats (73,4 % contre 74,3 %), la différence étant cette fois significative.

4. Ces bons résultats du Multi-Bandes dans un environnement non bruité n'étaient pas attendus. En effet, les motivations principales du Multi-Bandes concernaient essentiellement sa robustesse au bruit limité fréquentiellement, mais il ne semblait pas évident de prime abord que le fait de simplement découper le spectre en bandes puisse permettre d'améliorer les taux de reconnaissance, même lorsque aucun bruit n'est présent dans le signal. Il est difficile d'expliquer intuitivement ce phénomène simplement en observant ces dernières expériences, mais si nous considérons celles-ci en même temps que celles que nous avons présentées auparavant et qui concernent la robustesse du Multi-Bandes dans du bruit blanc, alors nous pouvons analyser plus globalement le phénomène et émettre une explication plausible.

   Celle que nous proposons met en jeu une information supplémentaire obtenue dans les sous-bandes qui n'apparaît pas dans le spectre complet, comme nous l'avons suggéré dans l'introduction de cette partie. Nous allons néanmoins essayer d'expliquer à nouveau ici ce phénomène, mais d'un point de vue légèrement différent. Tout d'abord, nous devons nous rappeler que les systèmes de reconnaissance n'utilisent pas toute l'information utile contenue dans le signal, mais tentent plutôt d'extraire de celui-ci le maximum d'information utile, et seulement de l'information utile. En effet, le signal enregistré possède en lui une quantité telle d'information qu'il est impossible de la traiter globalement. Au contraire, le but des reconnaisseurs est d'éliminer la plus grande partie de cette information afin de n'en conserver qu'une petite partie utilisable. Or, les reconnaisseurs automatiques n'étant pas parfaits, cette phase d'élimination de l'information redondante supprime également obligatoirement un peu de l'information utile et conserve de même une petite part d'information inutile. *Le fait de filtrer le spectre peut alors être vu comme un moyen de réduire la quantité d'information initiale à traiter*. Ainsi, les reconnaisseurs disposent d'une quantité d'information beaucoup moins grande, et il est alors probable qu'ils éliminent moins d'information utile.

   L'étage suivant du système de reconnaissance, c'est-à-dire le module de recombinaison, considère alors l'information « minimale » issue de tous ces reconnaisseurs, et la réduit encore pour ne plus avoir qu'une seule réponse finale. C'est ainsi que nous expliquons l'existence d'une information dans les sous-bandes qui n'est pas disponible dans le spectre complet.

6.9. Passage au mode continu

6.9.1. Introduction

Toutes les expériences ayant été réalisées jusqu'à présent ont utilisé la segmentation manuelle des phrases en phonèmes, en la supposant connue[11]. Cette partie 6.9 s'intéresse à la reconnaissance en mode continu, c'est-à-dire lorsque la segmentation de la phrase est inconnue.

Nous avons présenté dans les chapitres 2 et 3 plusieurs algorithmes permettant de réaliser une reconnaissance en mode continu. Rappelons-les ici :

1. Recombinaison trame par trame ;
2. Combinaison de HMM ;
3. Programmation dynamique à deux niveaux ;
4. Utilisation d'un étage de pré-segmentation ;
5. Algorithme de programmation dynamique dérivé du *two-level* et du *Level Building*.

11 cf. 2.4.3 pour une justification.

La première méthode a été abondamment testée dans la littérature concernant le Multi-Bandes [Mirghafori99][Okawa98]. La seconde méthode est difficilement applicable en pratique à cause de sa complexité. De plus, un nombre exponentiel de chemins possibles peuvent être parcourus dans les HMM générés, ce qui a pour effet de multiplier d'autant les erreurs et imprécisions obtenues au cours de l'algorithme de Viterbi, et les résultats ont donc de fortes chances de ne pas être très bons. La troisième méthode est assez ancienne et était auparavant utilisée en reconnaissance de la parole, mais elle a rapidement été remplacée par de nouveaux algorithmes plus performants. Nous avons choisi de réaliser quelques expériences basées sur les méthodes 4 et 5, essentiellement parce que ce sont des méthodes qui restent assez simples et intuitives, mais aussi parce qu'elles n'ont pas été proposées dans la littérature sur le Multi-Bandes jusqu'à présent.

6.9.2. Utilisation d'un étage de pré-segmentation

Il existe de nombreuses manières de segmenter le signal. Toutefois, celle qui a donné jusqu'à présent les meilleurs résultats est fondée sur l'utilisation d'un HMM et de l'algorithme de Viterbi. Il en existe d'autres plus récentes [Husson98], mais nous avons voulu conserver notre système simple et nous avons donc dans un premier temps utilisé le système de référence afin de segmenter le signal, puis nous avons appliqué notre système Multi-Bandes au signal segmenté. Les résultats sont reportés dans le tableau 6.9. La recombinaison utilisée est une recombinaison linéaire empirique affectant un coefficient de 0,6 à la cinquième bande est de 0,1 aux quatre sous-bandes.

Systèmes	***Taux de reconnaissance***
Référence	68,9 %
Multi-Bandes basé sur la segmentation donnée par le système de référence	69,3 %

*Tableau 6.9 : Taux de reconnaissance du système Multi-Bandes 0,1*4+0,6 en mode continu basé sur la segmentation donnée par le système de référence*

L'inconvénient de cette méthode est bien entendu de ne pas corriger les erreurs dues à la segmentation du système de référence. Les seules erreurs qui peuvent être corrigées sont les erreurs de substitution du système de référence. Sachant que, en comparant celui-ci avec le même système Multi-Bandes en mode isolé, l'amélioration apportée par le système Multi-Bandes est de 0,6 %. Nous voyons que nous arrivons, avec une différence de 0,3 %, à récupérer tout de même une partie de cette amélioration. Néanmoins, le problème principal de cette méthode de pré-segmentation basée sur le système de référence est que les résultats du système Multi-Bandes en milieu bruité risquent d'être particulièrement mauvais. En effet, nous avons vu que le système de référence est très sensible au bruit, nous devons donc nous attendre à ce que la segmentation qu'il propose en milieu bruité soit très erronée, auquel cas, le système Multi-Bandes qui dépend de cette segmentation verra ses scores chuter de la même manière. La solution serait alors d'utiliser un autre module de segmentation, une autre bande par exemple, ou alors la bande possédant le rapport signal-bruit le plus élevé. Toutefois, nous ne nous satisfaisons pas de cette solution qui relève plus alors de l'ingénierie de développement que de la recherche. C'est pourquoi nous avons également proposé et testé une deuxième solution, dont les résultats sont présentés dans la suite.

6.9.3. Algorithme de programmation dynamique

Nous avons vu dans la partie 3.4 que cet algorithme ne peut être utilisé tel quel lorsque la recombinaison du système Multi-Bandes est neuronale. En effet, nous sommes alors en présence de probabilités *a posteriori* et nous devons donc modéliser explicitement la probabilité de la segmentation à travers le terme $P(S \mid X)$. Nous avons donc pour le moment testé notre algorithme seulement avec des recombinaisons linéaires, qui décident elles-mêmes de leur segmentation.

✓ ***Expériences en mode continu avec une recombinaison linéaire***

Dans un premier temps, nous avons testé l'algorithme présenté ci-dessus avec la recombinaison linéaire empirique qui associe les coefficients 0,05 aux sous-bandes et 0,8 au spectre complet. Le résultat de ce test, réalisé en milieu non bruité, est présenté dans le tableau 6.10.

Systèmes	***Taux de reconnaissance***
Référence	68,9 %
Multi-Bandes basé sur la programmation dynamique	69,4 %

*Tableau 6.10 : Taux de reconnaissance du système Multi-Bandes 0,05*4+0,8 en mode continu utilisant l'algorithme de programmation dynamique.*

Ceci montre, d'une part qu'il est possible d'utiliser l'algorithme développé afin d'obtenir une reconnaissance de la parole Multi-Bandes en mode continu sans synchronisme complet entre les bandes, et d'autre part qu'un tel système améliore les performances du système de référence. Cette amélioration reste cependant assez faible, mais la recombinaison utilisée ici est rudimentaire. Nous pouvons espérer de meilleurs résultats avec une recombinaison neuronale, mais il nous faut auparavant résoudre un certain nombre de problèmes, comme nous le verrons dans la suite.

Testons maintenant le système en présence de bruits, afin de vérifier si les bonnes performances du Multi-Bandes, observées lorsque la segmentation était connue, apparaissent également en mode continu. Nous avons choisi la simple moyenne des cinq bandes, car celle-ci s'est avérée être dans les expériences précédentes une des meilleures recombinaisons possibles en milieu bruité. Le bruit de voiture a été ajouté au corpus de test, pour un RSB voisinant les -10 dB.

Les premiers résultats obtenus étaient clairement en-dessous du modèle de référence dans ces conditions. Après avoir analysé quelques réponses du système Multi-Bandes, nous nous sommes rendus compte que celui-ci proposait un nombre de phonèmes nettement inférieur à celui effectivement prononcé, et que la plupart des segments proposés étaient composés de longs silences. Ceci nous a amené aux conclusions suivantes concernant le Multi-Bandes :

Lorsque la segmentation est connue et correspond effectivement aux limites des phonèmes prononcés, nous avons vu que le Multi-Bandes présente des résultats très bons, notamment en milieu bruité. De même, lorsque la segmentation est inconnue, mais que nous sommes dans un environnement non bruité, alors les segments obtenus par l'algorithme d'alignement sont assez proches des limites réelles des phonèmes prononcés, et le Multi-Bandes a également des performances très correctes. Par contre, lorsque le milieu est bruité et que les limites des segments proposées par l'algorithme d'alignement sont très différentes des réelles, alors le Multi-Bandes n'offre que de piètres résultats.

Nous en avons déduit que ceci provient certainement du fait qu'il y a deux types d'erreurs possibles en reconnaissance continue de la parole : les erreurs liées à la segmentation, et celles liées au choix d'un phonème pour chaque segment. Or, ***si le Multi-Bandes est beaucoup plus robuste que le système de référence pour les erreurs du deuxième type, il l'est certainement moins pour les erreurs du premier type***.

Il faut donc essayer de réduire par d'autres moyens les erreurs de segmentation afin d'obtenir de bons résultats en utilisant le Multi-Bandes en mode continu. Nous avons testé deux méthodes dans ce but :

- Utilisation d'un contrainte de durée sur le silence

Si, en milieu bruité, la segmentation proposée par l'algorithme d'alignement est très erronée, c'est en partie à cause des longs silences qui sont souvent utilisés pour modéliser des zones où le système de reconnaissance est indécis. Nous avons dans un premier temps choisi d'imposer une contrainte de durée sur le modèle associé au silence, en réduisant la durée *d* autorisée pour le silence à 15 trames (c.f. page 44). Bien entendu, il faut alors également modifier la grammaire, afin d'autoriser plusieurs silences successifs à apparaître au cours du décodage, car il est tout-à-fait possible d'imaginer un silence supérieur à 15 trames au cours d'une discussion normale. En fait, nous avons rendu équiprobables les probabilités de transition entre le silence et tout autre phonème dans le bi-gramme. De plus, nous avons ajouté un étage final qui concatène les silences successifs en un seul phonème, afin de ne pas accroître le nombre d'insertions dans le calcul du taux de reconnaissance.

Nous avons modifié le système Multi-Bandes linéaire réalisant la moyenne des cinq bandes de la sorte. Ses résultats en mode continu dans un bruit de voiture à -10 dB sont présentés dans le tableau 6.11.

Systèmes	***Taux de reconnaissance***
Référence	27,6 %
Multi-Bandes basé sur l'algorithme de programmation dynamique	32,8 %

Tableau 6.11 : Taux de reconnaissance du système Multi-Bandes calculant la moyenne des cinq bandes en mode continu en présence d'un bruit de voiture à -10 dB. L'algorithme d'alignement utilisé est celui dérivé du two-level et du Level Building, auquel des contraintes de durées sur le silence ont été ajoutées.

Nous voyons que ce système permet d'obtenir des résultats très corrects dans un bruit de voiture, mais malheureusement, ceux-ci ne sont pas aussi bons dans du bruit blanc. C'est pourquoi nous avons imaginé une autre modification de l'algorithme d'alignement.

- Modélisation séparée de la segmentation et de la reconnaissance des segments

Puisque le Multi-Bandes se comporte moins bien que le système de référence en ce qui concerne les erreurs de segmentation, il peut être profitable d'utiliser un algorithme d'alignement utilisant deux recombinaisons différentes : une pour le choix d'un alignement, et l'autre pour le choix des phonèmes associés à chaque segment. Le nouvel algorithme que nous proposons est le suivant :

11. Initialement, un chemin ne contenant aucun phonème est créé et s'arrête, par définition, sur la trame –1.
12. Pour chaque trame *t* du signal,
    13. Pour chaque chemin *c* s'arrêtant sur la trame *t*-1 et qui a été retenu,
        14. Pour chaque modèle *m*,
            15. Pour chaque durée *d* possible de *m*,
                16. Un nouveau chemin *c*' est construit en concaténant le modèle *m* de durée *d* avec le chemin *c*. Ce chemin *c*' s'arrête donc sur la trame *t*-1+*d* incluse.
                17. La vraisemblance *s* du modèle *m* sur le segment [*t* , *t*-1+*d*] est alors calculée par le système Multi-Bandes en utilisant un recombinaison linéaire R_1 favorisant les sous-bandes.
                18. Le score associé à *c*', noté *S*(*c*'), est alors défini par :

                $$S(c') = S(c) + Trans(c,m) + s$$

                où *S*(*c*) est le score final du chemin *c*, et *Trans*(*c*,*m*) représente le logarithme de la probabilité de transition entre *c* et *m*. Nous avons utilisé un bi-gramme pour calculer cette probabilité.
                19. Le modèle $m'(d) = \underset{m}{argmax}(S(c'))$ qui rend le score en *t*-1+*d* maximum est conservé.
        20. Pour chaque durée *d* possible de *m*'(*d*),
            21. Le score du chemin *c*, auquel le modèle *m*'(*d*) est ajouté, est recalculé en utilisant cette fois une seconde recombinaison linéaire R_2 qui favorise le spectre complet. Nous obtenons donc un nouveau score *S*(*c*') pour le chemin *c*' = *c* + *m*'(*d*).
            22. Le chemin *c*' est alors inséré dans la liste ordonnée des *Nbest* meilleurs chemins calculés jusqu'à l'étape courante et se terminant à la trame *t*-1+*d*. Cette liste est ordonnée en fonction du score final de chaque chemin. Il y a une liste par trame du signal. Si *S*(*c*') est inférieur au dernier élément de la liste, le chemin *c*' est supprimé.
23. Lorsque la dernière trame est considérée, l'algorithme s'arrête et choisit le premier chemin de la liste correspondant à cette dernière trame.

Intuitivement, cet algorithme décide du modèle à ajouter au chemin courant en utilisant une recombinaison qui favorise les sous-bandes (nous avons utilisé la moyenne des cinq bandes), puis associe à ce chemin un score qui utilise une recombinaison favorisant le spectre complet (nous avons utilisé les coefficients 0,05 pour les sous-bandes et les coefficients 0,8 pour le spectre complet). Ce score est ensuite stocké dans la liste ordonnée des chemins se terminant sur chaque trame.

De plus, nous avons ajouté un nouveau terme au calcul de la vraisemblance d'un modèle sur un segment, terme qui représente la probabilité de durée de ce phonème. Nous avons introduit cette probabilité en vue des expériences utilisant un réseau de neurones pour la recombinaison. Il existe deux catégories de méthodes pour modéliser la probabilité de durée d'un phonème : les méthodes paramétriques [Levinson86] qui font appel à une loi de Poisson ou une loi Gamma pour modéliser la durée d'un modèle, et les méthodes non paramétriques [Ostendorf92]. Nous avons choisi ces dernières, car elles sont beaucoup plus simples à implémenter que les premières. De plus, il a été démontré [Ostendorf96] que, sur des segments de durée aussi courtes que les phonèmes, les différences entre ces deux modèles sont insignifiantes. Nous avons donc simplement calculé les fréquences relatives de chaque phonème sur le corpus d'apprentissage, et utilisé celles-ci comme probabilités de durée.

Les résultats que nous avons obtenus avec cet algorithme dans le bruit blanc sont présentés dans le tableau 6.12.

	Référence	***Multi-Bandes***
Bruit blanc à 20 dB	35,6 %	36,1 %

*Tableau 6.12 : Taux de reconnaissance, en mode continu et dans le bruit blanc, du système Multi-Bandes dont la recombinaison de segmentation est 0,8+0,05*4 et celle de sélection est 0,2*5*

Cette méthode, qui utilise deux recombinaisons différentes pour la segmentation et la sélection du phonème, n'est pas plus coûteuse que dans le cas d'une unique recombinaison. En effet, si le calcul de la vraisemblance de chaque modèle dans chaque bande est la partie la plus coûteuse de tout le processus de reconnaissance, le calcul d'une recombinaison ne consiste qu'en une simple somme pondérée. Réaliser celui-ci deux fois n'a donc aucun effet visible sur la complexité globale du système.

6.9.4. Complexité des algorithmes

En ce qui concerne la possibilité d'implémenter efficacement ce principe dans des systèmes de reconnaissance réels, il faut encore considérer le problème de la complexité des algorithmes qui ont été présentés pour le mode continu. Considérons tout d'abord la complexité du système Multi-Bandes en mode isolé, c'est-à-dire sans avoir ajouté un des algorithmes présentés dans la partie 3.4 permettant la reconnaissance de la parole en mode continu. Nous pouvons quantifier cette complexité en fonction de celle du reconnaisseur de base sur lequel s'appuie le Multi-Bandes. Par exemple, sachant que notre système utilise 5 HMM en parallèles et donc 5 algorithmes de Viterbi s'exécutant simultanément, nous pouvons dire que la complexité du système Multi-Bandes est inférieure à 5 fois celle d'un HMM (inférieure, car la dimension des vecteurs acoustiques est plus petite dans les sous-bandes que pour le spectre complet). Mais cet intervalle n'est pas suffisamment précis et nous pouvons encore considérer le nombre d'opérations réellement exécutées au cours de ces algorithmes. En fait, nous avons choisi les vecteurs dans les sous-bandes de façon à ce que la somme des dimensions des vecteurs dans toutes les sous-bandes soit égale à la dimension des vecteurs dans le spectre complet. Sachant que notre système Multi-Bandes utilise simultanément les sous-bandes et le spectre complet, si nous ne considérons pas le coût des transitions d'un état à l'autre, alors le coût du Multi-Bandes serait donc exactement égal à 2 fois le coût d'un HMM. Mais ces coûts dus aux transitions entre les états des HMM n'étant pas nuls, il est plus probable que le coût *réel* du Multi-Bandes soit égal à α fois le coût d'un HMM, avec $2 < \alpha < 5$. Le coût du module de recombinaison, qui vient s'additionner au coût des reconnaisseurs, est négligeable dans la mesure où il s'agit simplement d'une somme pondérée d'un nombre constant de valeurs. Nous voyons donc que ce coût est très raisonnable pour les machines actuelles.

Si nous voulons maintenant calculer le coût du Multi-Bandes en mode continu, il nous faut ajouter au coût préalablement cité celui dû à l'un des algorithmes dont nous avons parlé ci-dessus. Essayons d'évaluer ces coûts selon l'algorithme utilisé :

1. Dans le cas de la recombinaison trame par trame, aucun coût n'est ajouté. En fait, puisque les transitions ne sont empruntées qu'une seule fois, le coût global est en fait maintenant strictement égal à 2 fois le coût d'un HMM.

2. Dans le cas de l'utilisation d'un module de pré-segmentation du signal, le coût à ajouter est simplement le coût de ce module. Par exemple, avec un algorithme de segmentation composé d'un simple HMM, le coût final est (α + 1) fois celui d'un HMM, ce qui reste tout aussi raisonnable.

3. Dans le cas de la combinaison de HMM, nous avons déjà vu que le HMM résultant de la combinaison des HMM dans les différentes bandes a un nombre exponentiel d'états : l'algorithme est donc impossible à utiliser tel quel, et des heuristiques doivent être utilisées.

4. Dans le cas de la programmation dynamique à deux niveaux, la complexité en temps est plus difficile à calculer, mais des heuristiques sont presque toujours utilisées afin de la rendre raisonnable.

5. Dans le cas de l'algorithme de programmation dynamique, nous avons déjà vu que la complexité de l'algorithme Multi-Bandes final est asymptotiquement la même que celle d'un simple HMM, à savoir linéaire selon le nombre de trames de la phrase. Cet algorithme est donc très intéressant et mérite à notre avis d'être développé plus en profondeur dans la suite de ce travail.

6.9.5. Comparaison avec les autres algorithmes de la littérature

Une fois que nous avons obtenu des résultats avec notre système pour la reconnaissance de phonèmes en mode continu sur TIMIT, nous pouvons comparer ces résultats à ceux qui ont été préalablement publiés dans la littérature sur la même base de donnée. Nous avons regroupé les principaux systèmes dans le tableau ci-dessous, et affiché leur taux d'erreur sur TIMIT, les tests étant réalisés en milieux non bruité :

Auteurs	***Méthode***	***tout TIMIT***	***coretest***
Lee et Hon, 1989	Continuous density HMM	33,9 %	
Digalakis, Ostendorf et Rohlicek, 1992	Stochastic Segment Model	36,0 %	
Lamel et Gauvain	Continuous denisty HMM		30,9 %
Goldenthal, 1994	Trajectory model		30,5 %
Robinson, 1994	Recurrent ANN	25,0 %	26,1 %
Glass, Chang et McCandless, 1996	Feature based recognition		30,5 %
Ström, 1997	Sparse, recurrent, time-delay ANN	27,0 %	27,8 %
Haton, Fohr et Cerisara, 1999	Système Multi-Bandes linéaire basé sur des HMM du second ordre.		**30,6 %**

Nous voyons que notre système obtient des résultats tout-à-fait comparables aux résultats des systèmes basés sur des modèles de Markov cachés. En fait, il obtient même des résultats très légèrement meilleurs (mais pas significativement !) que le système de Lamel et de Gauvain, qui est le meilleur système que nous connaissons basé sur des HMM. Nous pouvons également remarquer que les meilleurs systèmes de reconnaissance sur TIMIT ne sont pas ceux basés sur des HMM, mais sur des réseaux de neurones, notamment le système de Robinson. Ceci nous incite à tenter de remplacer les HMM qui modélisent actuellement les phonèmes dans les bandes par des réseaux de neurones récurrents, mais il ne s'agit pour l'instant que de perspectives.

6.10. Conclusions du chapitre

6.10.1. Robustesse du Multi-Bandes en milieu bruité : une explication possible

Nous avons dans la première partie du chapitre présenté un certain nombre d'expériences menées avec notre système Multi-Bandes en milieu bruité. Dans un premier temps, nous avons vérifié expérimentalement les hypothèses qui étaient présentées initialement comme des motivations du paradigme Multi-Bandes. Ainsi en est-il de l'hypothèse selon laquelle le Multi-Bandes est robuste aux bruits limités fréquentiellement. Nous avons donc testé celle-ci et avons démontré expérimentalement sa véracité. Toutefois, il est apparu au cours de cette étude que même lorsque presque toutes les bandes sont affectées par le bruit, les systèmes Multi-Bandes sont néanmoins plus performants que le système de référence. Nous avons alors eu l'idée d'étudier la robustesse du Multi-Bandes lorsque le bruit affecte l'ensemble du spectre, comme c'est le cas pour le bruit blanc. Ces expériences ayant également été positives, nous en avons déduit une nouvelle explication au phénomène de robustesse du Multi-Bandes, à savoir la plus grande robustesse des bandes elles-mêmes. Cette hypothèse est plus forte que celle qui avait été avancée dans les motivations du Multi-Bandes : en effet, la communauté scientifique explique habituellement la robustesse du Multi-Bandes grâce au comportement efficace du module de recombinaison qui sélectionne les bandes qui ne sont pas affectées par le bruit. Nous avons montré ici que cette explication est vraie, mais qu'elle n'est pas suffisante, et que les sous-bandes elles-mêmes sont plus robustes que le spectre total. Ces résultats ont été confirmés par deux autres expériences : la première provient de l'étude de la « dynamique » des courbes de reconnaissance des bandes prises isolément en fonction du niveau de bruit. Cette expérience est expliquée au chapitre 4. La seconde montre que le perceptron, dont le rôle est d'accorder plus ou moins d'importance aux bandes dans le choix final, modifie cette répartition, originellement apprise dans un environnement non bruité, en augmentant l'importance accordée aux sous-bandes et en diminuant celle accordée au spectre total lorsqu'il est entraîné dans le bruit blanc.

Nous avons ensuite voulu tester la robustesse du Multi-Bandes dans du bruit « naturel », i.e. ayant été enregistré et ajouté au signal. Les bons résultats du Multi-Bandes n'ayant pas été désavoués par ces tests, nous avons tenté d'améliorer les performances de notre système en entraînant le module de recombinaison dans le bruit. Les résultats obtenus sont assez bons, dans le sens où une amélioration de la robustesse dans la plupart des environnements testés apparaît, même si ces conclusions ne se vérifient pas lorsque le bruit utilisé pendant le test est totalement différent de celui utilisé pendant l'apprentissage.

Enfin, nous avons appliqué le principe Multi-Bandes à une nouvelle tâche de TALN qui est l'identification des langues. Au cours de ces expériences, le paradigme Multi-Bandes lui-même s'est assez bien comporté mais les résultats ne sont pas concluants car les classifieurs utilisés dans les bandes sont trop sensibles au bruit (très important) présent dans le corpus utilisé et ne permettaient pas au Multi-Bandes d'assumer pleinement son rôle. D'autres tests sont donc encore nécessaires dans ce domaine, mais ceux que nous avons réalisés ont néanmoins mis en évidence une caractéristique fondamentale du paradigme Multi-Bandes, qui n'est pas un classifieur en soi mais dont la seule utilité est de fédérer un ensemble de reconnaisseurs préexistants. Il est donc particulièrement dépendant de ces reconnaisseurs, qui doivent déjà permettre une reconnaissance correcte si nous voulons pouvoir améliorer leurs résultats.

6.10.2. Comparaison entre les différents modules de recombinaison testés

Nous avons testé tout au long de ce chapitre plusieurs modules de recombinaison, et nous avons ainsi montré que le meilleur d'entre eux est certainement le perceptron. En effet, ce module de recombinaison possède plus de paramètres que le système linéaire, et son espace de recherche, non linéaire, est plus grand. Il est donc potentiellement le meilleur module de recombinaison possible.

Toutefois, cette conclusion est remise en cause lorsque les conditions de test sont très différentes des conditions d'apprentissage. Ce cas se produit par exemple lorsque le niveau de bruit ajouté pendant le test est élevé, ou lorsque ce bruit « déséquilibre » l'importance relative des bandes les unes par rapport aux autres, par exemple en affectant une zone restreinte du spectre de parole. Il vaut mieux alors opter pour une recombinaison linéaire, plus simple et surtout n'ayant pas subi d'apprentissage dans un milieu non bruité, et n'ayant donc pas développé la même dépendance vis-à-vis de cet environnement que le perceptron. De surcroît, il faut diminuer d'autant plus l'influence du spectre total que le niveau de bruit est élevé, comme cela a été suggéré ci-dessus. Ceci permet donc à la simple moyenne des quatre sous-bandes d'être le modèle le plus performant dans certaines conditions.

6.10.3. Comparaison avec d'autres systèmes robustes

Il peut nous être reproché de ne pas avoir réalisé de comparaison entre notre système Multi-Bandes et d'autres méthodes robustes au bruit. Toutefois, cette comparaison est extrêmement difficile dans notre cas. En effet, le problème principal vient du fait que *le Multi-Bandes n'est pas un système de reconnaissance en soi*, mais est plutôt assimilable à une architecture permettant d'utiliser plusieurs systèmes de reconnaissance ensemble. Ceci signifie que si nous utilisons un système de base qui est lui-même robuste au bruit, il suffit alors d'utiliser ce même système comme reconnaisseur dans chaque bande pour immédiatement bénéficier de l'avantage de celui-ci et obtenir en conséquence un système Multi-Bandes encore plus robuste au bruit. Ainsi, comme le montrent les résultats que nous avons obtenus dans un environnement non bruité, ce nouveau système Multi-Bandes dépasserait presque sûrement les performances du système sur lequel il s'appuie !

Ce problème remet en cause le concept même de comparaison dans notre cas, car la principale utilité d'une comparaison entre deux systèmes est de pouvoir sélectionner le plus performant, ce qui n'est donc pas possible ici. De plus, pour réaliser une telle comparaison, il faut nous assurer que nous comparons des systèmes *comparables*, c'est-à-dire réalisant bien la même fonction. Or, comme nous l'avons vu en introduction de ce chapitre, les systèmes robustes aux bruits ne manquent pas, mais ils sont très différents les uns des autres tant par leur action que par leur fonction. Ainsi, nous ne pouvons comparer le Multi-Bandes avec des systèmes de débruitage ou des méthodes calculant de nouveaux coefficients pour les vecteurs acoustiques, car le principe robuste de ces systèmes se situe à un autre niveau. De même, il est difficile de comparer le Multi-Bandes avec des systèmes comme les perceptrons, car, bien qu'agissant au même niveau, leurs conditions d'utilisation ne sont pas les mêmes : les perceptrons supposent le bruit connu et tentent de le modéliser, ce qui est très différent des hypothèses de base du Multi-Bandes pour lequel le bruit est totalement inconnu et qui ignore même si bruit il y a.

Les modèles les plus proches du Multi-Bandes, mis à part les HMM, sont sans doute ceux utilisant des Champs de Markov Cachés, mais aucun résultat n'est encore disponible concernant ceux-ci, qui sont en développement [Gravier98]. Les systèmes hybrides, dans lesquels les mélanges de gaussiennes dans les états des HMM sont remplacés par des perceptrons [Bourlard94], appartiennent également à la même catégorie que le Multi-Bandes. En ce qui les concerne, la comparaison entre ces systèmes hybrides et un système Multi-Bandes a déjà été réalisée et s'est montrée positive pour le Multi-Bandes [Bourlard96].

Une autre utilité de la comparaison est de valider un nouveau système en montrant qu'il permet d'obtenir des résultats comparables avec ceux obtenus par d'autres systèmes remplissant la même tâche de reconnaissance. Or, pour tester la validité du paradigme Multi-Bandes, il est nécessaire et suffisant de prouver qu'il obtient de meilleurs résultats que le système de base sur lequel il s'appuie : c'est ce qui est fait tout au long de ce mémoire lorsque nous comparons les résultats obtenus par le système Multi-Bandes avec ceux correspondant au système dit de référence.

Toutefois, dans un tel but de validation, et toujours dans un soucis de généralité, il faut tester le paradigme Multi-Bandes avec différents types de reconnaisseurs de base, et pas seulement avec les modèles de Markov du second ordre. Nous l'avons fait dans le cadre de l'identification de la langue, en remplaçant nos HMM2 par de simples centroïdes, et nos résultats montrent que le paradigme Multi-Bandes s'est avéré efficace également dans ce cas. Enfin, nous pouvons noter que dans la littérature, d'autres systèmes de base ont été utilisés pour le Multi-Bandes, comme le modèle hybride dont nous avons déjà parlé, ou encore un classifieurs basé sur une simple distance entre trames dans le cadre de la reconnaissance du locuteur [Besacier98a]. Nous avons déduit de tous ces résultats, joints à ceux fournis par nos différents systèmes, que le paradigme Multi-Bandes est un principe valide et qui mérite d'avoir sa place parmi les systèmes aujourd'hui considérés comme robustes.

6.10.4. Expériences en milieu non bruité et passage au mode continu

Nous avons montré, dans la deuxième partie de ce chapitre, que contrairement aux hypothèses initialement émises concernant le Multi-Bandes, celui-ci se comporte également bien en milieu non bruité. De plus, après avoir présenté dans les chapitres 2 et 3 plusieurs algorithmes permettant d'utiliser le Multi-Bandes en mode continu, nous avons testé certains de ces algorithmes dans le présent chapitre. Nous avons ainsi montré qu'il est possible d'utiliser un étage de pré-segmentation au système Multi-Bandes, mais que cette solution ne permet pas de bénéficier pleinement des avantages du Multi-Bandes, et qu'elle pose quelques problèmes en milieu bruité. Nous avons de même montré la possibilité d'utiliser un algorithme de programmation dynamique. Les premiers résultats que nous avons obtenus avec cet algorithme sont encourageants, même si certaines modifications sont encore nécessaires avant de pouvoir utiliser une recombinaison par perceptron.

✓ *Remarque concernant le corpus TIMIT*

Nous avons présenté dans ce chapitre essentiellement les résultats que nous avons obtenus sur TIMIT, car cette base d'apprentissage est une base de référence en reconnaissance de la parole. Toutefois, nous avons également réalisé quelques expériences similaires sur BREF-80, un corpus semblable à TIMIT mais en français [Gauvain91]. L'interprétation des résultats étant la même sur les deux corpus, nous nous sommes contentés de ne présenter que les résultats sur TIMIT dans ce mémoire.

Chapitre 7
Apprentissage global

7.1. Motivations

Comme nous l'avons déjà mentionné, les trois premières motivations pour le paradigme Multi-Bandes sont, tout d'abord l'inspiration du modèle de l'audition humaine, puis la robustesse au bruit limité fréquentiellement, et enfin le fait que l'information qui caractérise un phonème est souvent principalement contenue dans une zone limitée du spectre de fréquences. Le dual de cette idée est qu'il n'y a pas assez d'information dans une seule bande de fréquence pour identifier tous les phonèmes. Ce qui signifie que les phonèmes ne sont pas des classes[12] adaptées à la reconnaissance dans une sous-bande. C'est pourquoi l'idée d'utiliser de nouvelles classes dans les sous-bandes est récemment apparue dans un contexte Multi-Bandes. Nous pouvons présenter quatre avantages principaux à l'utilisation de telles classes :

- Si nous supposons, comme nous venons de le montrer, que les phonèmes ne sont pas des classes pertinentes dans une sous-bande, essayer tout de même de les modéliser entraîne obligatoirement des recouvrements entre les classes, et donc un plus grand nombre d'erreurs dans le processus de classification. Il vaut donc beaucoup mieux modéliser des classes qui représentent effectivement l'information contenue dans chaque sous-bande.

- Comme il y a moins d'information dans une sous-bande que dans le spectre complet, il doit également y avoir moins de classes qu'il n'y a de phonèmes. Ceci signifie qu'un plus grand nombre d'exemples du corpus d'apprentissage sont attribués à chaque classe, ce qui permet une meilleure modélisation de chacune de ces classes.

- Le fait qu'il y ait moins de confusion possible entre les classes entraîne obligatoirement une baisse des erreurs de reconnaissance dans chaque bande. Ceci signifie que les entrées du module de recombinaison contiennent moins d'erreurs, et donc que la tâche de celui-ci en est d'autant plus facilitée.

- Comme il y a moins d'entrées dans le module de recombinaison, l'espace dans lequel peuvent évoluer ses paramètres est de dimension moins grande et son apprentissage est donc plus facile. Ceci résulte du fait que les imprécisions dues au trop célèbre problème des « grandes dimensions », plus connu sous le nom de « *dimensionnality curse* », sont réduites.

12 Dans la suite, le terme *classe* désigne la zone de l'espace acoustique modélisée par un HMM d'une bande. Jusqu'à présent, ce terme était équivalent à celui de *phonème*. Mais dans ce chapitre, l'algorithme d'apprentissage global rend cette équivalence fausse, et c'est pourquoi nous distinguons maintenant ces deux termes. Le terme *modèle* représente le HMM lui-même.

Une première idée pour construire ces nouvelles classes dans les bandes est d'étudier la confusion entre les phonèmes dans chaque bande et de regrouper les classes phonétiques qui sont le plus souvent confondues. Un travail basé sur cette idée a été réalisé par Mirghafori [Mirghafori99], au cours duquel elle a étudié trois méthodes pour regrouper les phonèmes : la première utilise les matrices de confusion des reconnaisseurs dans chaque bande, la deuxième utilise un critère d'information mutuelle entre les phonèmes et la troisième tente de minimiser le taux d'erreur en choisissant tous les groupes de phonèmes possibles. Malheureusement, les résultats de ces expériences ne sont pas aussi bons que l'on aurait pu s'y attendre. Par exemple, si on regroupe les phonèmes qui sont le plus souvent confondus entre eux, des super-classes contenant tous les phonèmes apparaissant le moins fréquemment dans le corpus d'apprentissage sont construites. Nous voyons que des facteurs autres que phonétiques interviennent et perturbent les résultats. Nous pouvons en déduire qu'il est difficile de construire ces nouvelles classes *a priori*.

Une autre possibilité pour construire ces nouvelles classes est de laisser le système ajuster lui-même les classes phonétiques dont il a besoin dans les bandes. En fait, dans un système Multi-Bandes, il n'est pas nécessaire d'avoir une classification en phonèmes à la sortie des HMM ! Celle-ci ne doit intervenir qu'à la sortie du module de recombinaison. L'algorithme d'apprentissage global présenté dans ce chapitre construit les classes dans chaque bande en fonction du taux de reconnaissance final, et n'oblige en rien ces classes à être des phonèmes. Dans ce cadre, nous ne nous intéressons donc plus du tout au taux de reconnaissance dans chaque bande, mais uniquement au taux de reconnaissance final. Un apprentissage de ce type aura comme effet « secondaire » de construire de nouvelles classes qui représentent effectivement l'information contenue dans chaque bande, même s'il ne s'agit pas de phonèmes.

L'idée que nous développons dans ce chapitre consiste donc à entraîner en une seule étape le système Multi-Bandes *complet*, c'est-à-dire aussi bien l'étage composé d'un HMM par bande que le module de recombinaison. L'apprentissage classiquement utilisé dans les systèmes Multi-Bandes, et qui l'est par tous les systèmes Multi-Bandes que nous connaissons, utilise en fait deux étapes : la première consiste à entraîner les classifieurs dans les bandes, en maximisant les taux de reconnaissance dans celles-ci, tandis que la seconde consiste à apprendre les paramètres du module de recombinaison en fonction des sorties des reconnaisseurs. Ceux-ci ne changent plus au cours de la deuxième étape. Or, l'optimisation globale d'un système, lorsqu'elle est possible, est toujours meilleure que la somme des optimisations individuelles de ses composants, car elle fait intervenir non seulement les composants eux-mêmes mais aussi les interactions qui peuvent exister entre ces composants.

7.2. Choix du critère d'optimisation

Nous utilisons le critère de minimisation de l'erreur finale de classification, ou critère MCE, pour calculer les paramètres des HMM et du module de recombinaison. Nous avons choisi ce critère car il correspond au but recherché, à savoir obtenir le taux de reconnaissance final maximum. De plus, il aurait été difficile d'utiliser le critère de maximisation de l'estimation de la vraisemblance (MLE), car les sorties du module de recombinaison ne sont pas des vraisemblances [Besacier98a]. Nous aurions pu bien entendu adapter ce critère et maximiser les « scores finaux par modèle », mais nous aurions alors perdu le pouvoir discriminant du critère MCE. Or, il nous semble très important d'avoir un critère discriminant dans le Multi-Bandes, car nous nous attendons intuitivement à ce que les reconnaisseurs se « spécialisent » dans l'identification de traits phonétiques qui sont différents d'une classe à l'autre. Enfin, le critère MCE a l'avantage d'être simple à utiliser et à implémenter, comme nous le verrons, ce qui n'est pas le cas du critère de maximisation de la probabilité *a posteriori* (MAP), ni du critère de maximisation de l'estimation de l'information mutuelle (MMIE). Ces deux derniers critères sont déjà très difficiles à implémenter pour un seul HMM, ce qui les rend presque impossible à adapter lorsque plusieurs HMM sont considérés avec un module de recombinaison.

7.3. Principe

Le critère MCE a déjà été décrit dans la partie 5.2. Le principe de l'algorithme d'apprentissage global consiste, pour chaque exemple du corpus d'apprentissage, à appliquer le critère MCE tout d'abord au module de recombinaison, comme cela est expliqué dans la partie 5.2, puis à l'appliquer aux HMM qui composent le premier étage du système Multi-Bandes. Nous étudions dans la suite de quelle manière ce critère peut s'appliquer à un HMM.

7.3.1. Modification des HMM

Rappelons la définition de la fonction sigmoïde approchant l'erreur de classification qui a déjà été donnée au chapitre 5 :

$$l_{C(x)}(x)=\frac{1}{1+\exp(-\gamma\, d_{C(x)}(x))}$$

Considérons pour simplifier l'exposé que nous utilisons une recombinaison linéaire des densités de probabilités calculées par les HMM. Nous avons donc :

$$d_{C(x)}=-g_{C(x)}(x)+g_{\overline{C(x)}}(x) \qquad \text{(Eq-4)}$$

Et :

$$g_j(x)=\sum_{b=1}^{B}\alpha_{b,M}P(x|M,b)$$

La première remarque que nous pouvons faire concernant la modification des HMM est que, pour chaque exemple du corpus d'apprentissage, tous les HMM de toutes les bandes doivent être modifiés. Nous voyons tout de suite que ceci est extrêmement coûteux en temps de calcul, d'autant plus que dans un grand nombre de cas, la modification à réaliser, qui est proportionnelle au gradient de l'erreur de classification d'après le critère MCE, est presque nulle. Ceci peut survenir par exemple lorsque deux classes sont déjà bien séparées et qu'un exemple d'une de ces deux classes est présenté au système. Celui-ci modifie alors l'autre classe de manière infime, ce qui amène à penser qu'il n'est pas nécessaire en pratique de réaliser cette modification lorsque le gradient de l'erreur de classification est très proche de zéro. La précision de l'algorithme d'apprentissage est peut-être alors moins grande, mais son temps d'exécution, déjà suffisamment long, en est considérablement réduit. Nous avons déjà utilisé auparavant cette première heuristique, qui se traduit formellement par deux équations : l'équation Eq-4 exprimée ci-dessus, qui ne fait intervenir que deux modèles pour chaque exemple du corpus d'apprentissage, le bon phonème et le meilleur « mauvais » phonème, et l'équation suivante :

$$\frac{1}{1+\exp(-\gamma\, d_{C(x)}(x))} \backsimeq 0 \quad \text{lorsque} \quad d_{C(x)}(x) < -\delta < 0$$

Cette formule fait intervenir un seuil noté δ qui, intuitivement, permet au système de ne pas modifier les modèles lorsque l'exemple concerné est clairement affecté à la bonne classe par le système et que le « meilleur mauvais » modèle donne un score nettement inférieur à celui du bon modèle. Nous aurions pu donner à δ une valeur nulle, auquel cas l'algorithme n'aurait modifié un modèle que lorsque l'exemple est mal reconnu. Néanmoins, nous avons préféré lui donner une petite valeur négative afin que, même lorsque l'exemple est bien reconnu, si un autre modèle fournit un score très proche de celui du bon modèle, alors l'ajustement de ces deux modèles est néanmoins réalisé. Ceci permet d'assurer une certaine stabilité à l'algorithme d'apprentissage.

Revenons maintenant à la modification des paramètres du HMM. Si nous appelons λ un paramètre d'un HMM, le critère MCE précise qu'il faut modifier λ selon la formule :

$$\lambda(t+1) = \lambda(t) - \epsilon \frac{\partial l_{C(x)}(x)}{\partial \lambda}$$

Puisque nous avons supposé pour les besoins de l'exposé que la recombinaison est linéaire, la dérivée de la fonction sigmoïde dépend directement de la dérivée du score final associé à $C(x)$ (ou à $\overline{C(x)}$), c'est-à-dire de :

$$\frac{\partial g_{C(x)}(x)}{\partial \lambda} = \alpha_{b,C(x)} \frac{\partial P(x|b,C(x))}{\partial \lambda}$$

Dans le cas où seul le module de recombinaison est modifié, nous avons vu que cette dérivée aboutit à une formule exacte. Malheureusement, ce n'est plus le cas ici, car la fonction $f(\lambda)=P(x|b,M)$, où λ est un paramètre du modèle *M*, n'est pas dérivable. En effet, cette fonction dépend du meilleur chemin choisi par le HMM, ce qui signifie que $f(\lambda)$ n'est que dérivable par morceaux, chaque « morceau » correspondant à un même meilleur chemin dans le HMM. Ainsi, aux valeurs de λ pour lesquelles un autre meilleur chemin est choisi par le HMM, la courbe de $f(\lambda)$ peut changer brutalement de direction, comme le montre la figure 7.1.

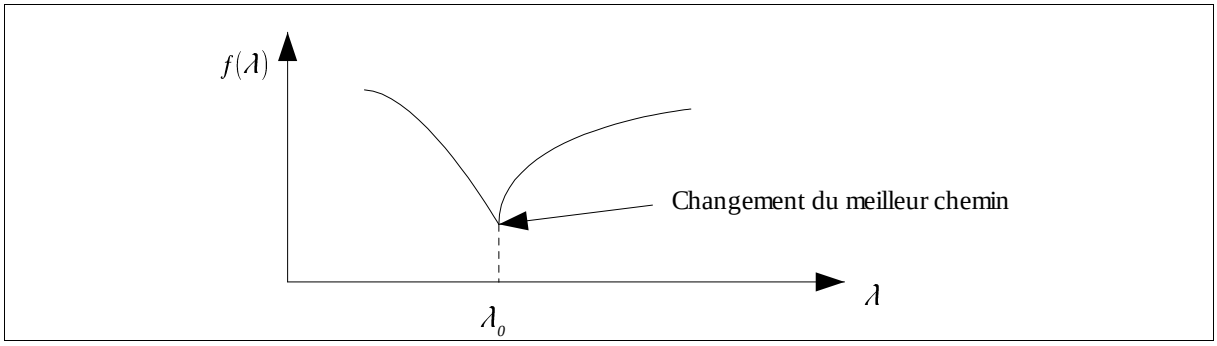

Figure 7.1 : Courbe extraite de la vraisemblance de sortie d'un HMM en fonction d'un de ses paramètres.

Ces points pour lesquels $f(\lambda)$ n'est pas dérivable nous empêchent d'utiliser la procédure de descente de gradient telle qu'elle est définie. En effet, supposons que le paramètre λ que nous voulons modifier ait pour valeur $\lambda_0-\epsilon$, avec ϵ une constante positive très proche de zéro. Si l'algorithme veut faire décroître $f(\lambda)$, alors il déplacera λ vers la droite ; le risque que la nouvelle valeur de λ soit $\lambda+\epsilon$, et donc que le nouveau $f(\lambda)$ soit en réalité supérieur au précédent, est alors grand, ce qui est contraire aux objectifs de l'algorithme d'apprentissage. Nous voyons donc avec cet exemple qu'il est impossible d'utiliser la procédure classique de descente du gradient afin d'optimiser les paramètres des HMM.

Cependant, nous n'avons pas vraiment besoin d'une fonction dérivable pour augmenter ou diminuer $f(\lambda)$. En effet, les modèles de Markov sont initialement entraînés grâce à l'algorithme de Baum-Welch, qui utilise les comptes du nombre de fois qu'une transition est empruntée. Ce qui signifie qu'il suffit d'augmenter ou de diminuer ces comptes afin que $f(\lambda)$ reproduise le même comportement. Cette idée provient originellement de l'algorithme d' « apprentissage correctif » proposé par Bahl et al. en 1988 [Bahl88] dans le but d'améliorer l'apprentissage des HMM. Cependant, il ne suffit pas simplement d'augmenter ou de diminuer $f(\lambda)$, mais il faut aussi que cette augmentation ou cette diminution soit proportionnelle au gradient de l'erreur si nous voulons respecter le critère MCE. Ceci est réalisé en calculant le gradient de l'erreur de classification finale jusqu'aux sorties des HMM, puis en multipliant ce gradient par une constante et en additionnant cette valeur aux comptes des HMM correspondants. La constante est déterminée empiriquement. Il est intéressant de remarquer que le gradient de l'erreur est calculé jusqu'aux sorties des HMM, c'est-à-dire une étape « plus loin » que lors de l'apprentissage du module de recombinaison seul.

Ce raisonnement est identiquement applicable lorsque le module de recombinaison est un perceptron, car le gradient de l'erreur de classification finale est également calculé pour un tel module de recombinaison.

Notre algorithme de modification des paramètres des HMM diffère donc de celui de Bahl dans le sens où la modification des comptes est dépendante du gradient de l'erreur de classification. Nous avons tout de même utilisé dans notre algorithme une heuristique proposée par Bahl destinée à accélérer la convergence de notre système. Il s'agit en fait d'ajouter une constante au gradient de l'erreur de classification pour le bon modèle, de façon à toujours augmenter la probabilité de ce modèle, même lorsque l'erreur de classification est nulle.

7.3.2. Initialisation des paramètres

L'initialisation des paramètres peut être aussi simple que ce qui suit :

1. Utiliser comme HMM initiaux le même modèle vide, c'est-à-dire contenant une gaussienne nulle par état et dont les transitions sont équiprobables.
2. Dans le cas de la recombinaison linéaire, fixer tous les coefficients de la recombinaison à 1 ;
3. Dans le cas de la recombinaison neuronale, donner des valeurs aléatoires aux coefficients du perceptron.

Cette initialisation est intéressante dans la mesure où aucune connaissance *a priori* n'est utilisée par exemple pour indiquer aux modèles dans les sous-bandes qu'ils doivent être semblables à des modèles de phonèmes, ou de la même manière pour indiquer au module de recombinaison que la cinquième bande fournit globalement les meilleurs résultats, car ceci est préjudiciable au bon apprentissage du système, comme nous l'avons montré dans la partie 6.3.3. Cependant, l'apprentissage global d'un système Multi-Bandes est très long, et afin d'obtenir rapidement les premiers résultats, nous avons choisi pour les paramètres des valeurs initiales identiques à celles fournies par un apprentissage séparé. Ainsi, nous avons choisi d'initialiser notre système de la manière suivante :

1. Un apprentissage initial des HMM est réalisé ;
2. Dans le cas de la recombinaison linéaire, les coefficients sont choisis empiriquement. En fait, ils sont fixés à 0,1 pour les sous-bandes et à 0,6 pour le spectre complet ;
3. Dans le cas de la recombinaison neuronale, un apprentissage du perceptron seul est initialement réalisé.

7.4. Résumé de l'algorithme

Finalement, l'algorithme d'apprentissage global que nous avons utilisé est le suivant :

1. Un apprentissage initial des HMM est réalisé. Les comptes des transitions sont sauvegardés dans $\Gamma(b,i)$ pour le modèle i de la bande b.
2. Le module de recombinaison est initialisé comme cela est indiqué dans la partie précédente.
3. Pour chaque exemple u du corpus d'apprentissage :
   4. Les scores finaux associés à chaque modèle pour l'exemple u sont calculés. Notons $S(u,i)$ le score associé au modèle i pour l'exemple u. Notons de même $C(u)$ le véritable modèle de u.
   5. Une approximation $E(u)$ de l'erreur de classification finale est calculée à partir des $S(u,i)$.
   6. Les paramètres du module de recombinaison sont ajustés en fonction du gradient de $E(u)$, grâce au critère MCE qui peut se dériver de deux manières possibles selon les cas :
      7. Dans le cas d'une recombinaison linéaire, l'ajustement des coefficients de la recombinaison est réalisée comme précédemment grâce à l'algorithme MCE, qui utilise une procédure de descente du gradient de l'erreur de classification.
      8. Dans le cas d'une recombinaison neuronale, l'ajustement des poids est réalisé grâce à l'algorithme de rétro-propagation, qui utilise également une procédure de descente du gradient de l'erreur de classification.
   9. Dans les deux cas, la valeur du gradient de l'erreur de classification est calculée jusqu'aux sorties des HMM.
   10. Pour chaque modèle w tel que $S(u,w) > S(u,g) - \delta$,
      11. Pour chaque bande b,
         12. Une itération de l'algorithme « *forward-backward* » (utilisé dans l'algorithme de Baum-Welch) est réalisée sur le HMM correspondant au modèle w de la bande b, avec l'exemple u. Les comptes des transitions $\Gamma(b,u,w)$ sont conservés.
         13. Les comptes globaux $\Gamma(b,w)$ sont modifiés par : $\Gamma(b,w) = \Gamma(b,w) - c\ \Gamma(b,u,w)$, où c est le gradient de $E(u)$ multiplié par une constante.
   14. Pour chaque bande b,
      15. Une itération de l'algorithme *forward-backward* est réalisée sur le HMM de la bande b correspondant au modèle $C(u)$ avec l'exemple u. Soit $\Gamma(b,u,C(u))$ les comptes correspondants.
      16. Les comptes globaux $\Gamma(b,C(u))$ sont modifiés par :
         $\Gamma(b,C(u)) = \Gamma(b,C(u)) + c'\ \Gamma(b,u,C(u))$.
   17. Tous les paramètres des HMM des bandes sont recalculés avec les nouveaux comptes $\Gamma(b,i)$.
   18. L'algorithme itère à partir de l'étape 3 jusqu'à ce que tout le corpus d'apprentissage soit utilisé.

7.5. Résultats expérimentaux

7.5.1. Expériences dans un environnement non bruité

En réalité, l'apprentissage global du système est très long, et nous n'avons pas pu conduire cet apprentissage jusqu'à convergence de l'algorithme, comme il aurait été préférable de le faire. C'est pourquoi nous ne montrons pas seulement le taux de reconnaissance final du système, mais aussi l'évolution de ce taux de reconnaissance en fonction du nombre d'itérations réalisées. Cette courbe nous permet d'avoir une idée plus précise du comportement de l'algorithme, et ainsi de mieux régler les paramètres d'apprentissage. Nous avons simultanément entraîné le système de référence et affiché la courbe correspondante afin de pouvoir étudier des résultats comparables. Il est important de noter que cet apprentissage du système de référence a été réalisé avec le même algorithme d'apprentissage que celui du Multi-Bandes. Les résultats de ce système de référence, du Multi-Bandes avec recombinaison linéaire (MCE) et de celui avec recombinaison neuronale (PMC) sont affichés sur la figure 7.2.

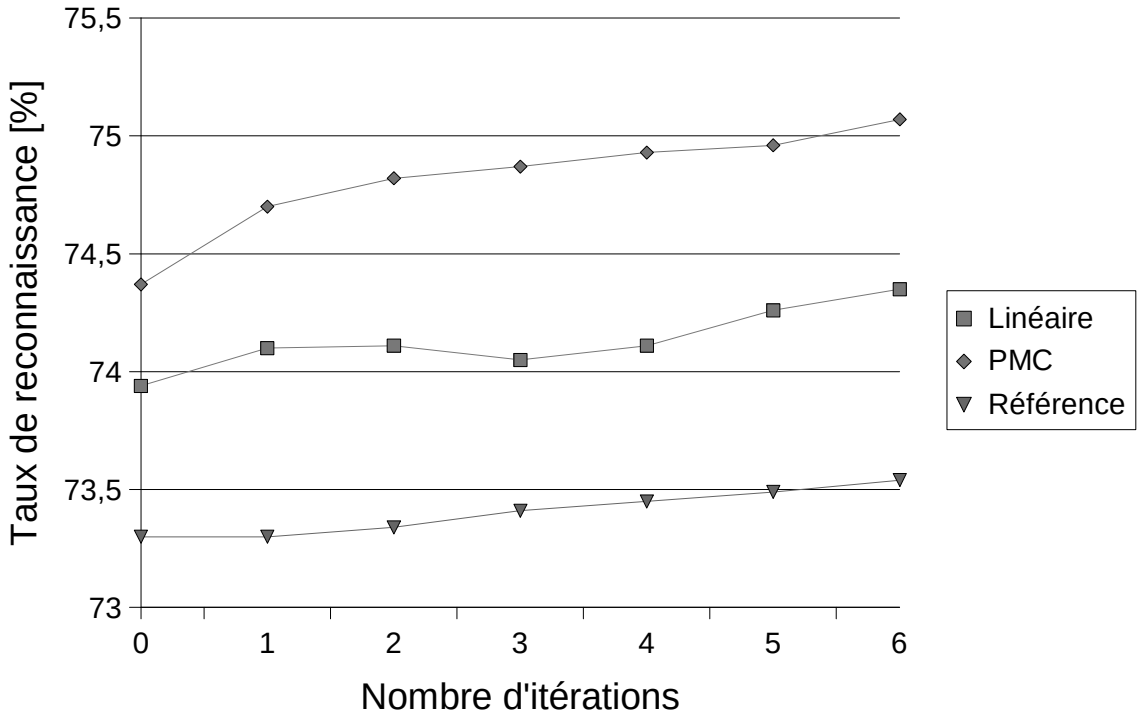

Figure 7.2 : Évolution du taux de reconnaissance des systèmes Multi-Bandes en fonction du nombre d'itérations de l'apprentissage global

Comme nous le voyons sur cette courbe, l'apprentissage global a un effet nettement positif, notamment sur le système Multi-Bandes utilisant un perceptron. L'évolution du système linéaire est certes assez bonne, mais elle semble un peu plus chaotique. Nous pensons que ce phénomène est tout simplement dû au réglage très approximatif des paramètres de l'algorithme. Ceux-ci mériteraient en effet une étude plus longue afin que leurs valeurs soient mieux ajustées.

7.5.2. Expériences dans un environnement bruité

Nous allons maintenant étudier le comportement du système Multi-Bandes lorsque du bruit est ajouté au corpus de test. Dans les expériences qui suivent, seul le système Multi-Bandes avec une recombinaison par un perceptron issu de la quatrième itération de l'apprentissage global est considéré.

Les résultats du système dans différents types de bruits sont donnés dans le tableau 7.1. Les taux de reconnaissance appartenant à la colonne « Référence » correspondent aux taux obtenus avec le système de référence ayant subi un apprentissage supplémentaire identique à celui du système Multi-Bandes.

environnement	***App. global***	***App. MB classique***	***Référence***
Bruit filtré grave (RSB = 20 dB)	31,9 %	27,1 %	23,6 %
Bruit blanc (RSB = 20 dB)	43,8 %	43,8 %	40,7 %
Bruit de voiture (RSB = -10 dB)	27,5 %	29,1 %	24,5 %
Bruit de cantine (RSB = 20 dB)	38,4 %	38,5 %	36,2 %

Tableau 7.1 : Taux de reconnaissance du système Multi-Bandes avec recombinaison par PMC dans plusieurs environnements bruités après quatre itérations d'apprentissage global.

Comme nous pouvons le constater, le système Multi-Bandes après un apprentissage global est toujours plus robuste aux bruits que ne l'est le système de référence, quel que soit l'environnement considéré. Nous pouvons donc en conclure que le système Multi-Bandes avec apprentissage global est performant aussi bien dans le bruit que dans un environnement non bruité : il s'agit donc d'un algorithme particulièrement intéressant pour le Multi-Bandes.

7.5.3. Apprentissage dans un environnement bruité

Nous avons voulu dans cette partie compléter l'étude du système en observant son comportement après un apprentissage global dans un milieu bruité, comme nous l'avons fait pour un système Multi-Bandes « classique », i.e. sans apprentissage global. Le bruit ajouté au corpus d'apprentissage est comme précédemment du bruit blanc à 25 dB. Seules deux itérations en apprentissage global bruité ont été réalisées, à partir des modèles issus de la cinquième itération de l'algorithme d'apprentissage global en milieu non bruité.

Les résultats de ce système dans différents environnements sont donnés dans le tableau 7.3. Les résultats correspondants du système Multi-Bandes avec apprentissage dans le bruit, mais sans apprentissage global, sont également donnés dans le tableau 7.2, afin de pouvoir comparer les augmentations de la robustesse dans les deux cas.

environnement	***MB avant app. indépendant bruité***	***MB après app. indépendant bruité***
Environnement non bruité	74,4 %	72,6 %
Bruit filtré grave (RSB = 20 dB)	27,1 %	28,8 %
Bruit blanc (RSB = 20 dB)	43,8 %	56,8 %
Bruit de voiture (RSB = -10 dB)	29,1 %	31,3 %
Bruit de cantine (RSB = 20 dB)	38,5 %	37,4 %

Tableau 7.2 : Taux de reconnaissance du système Multi-Bandes avec recombinaison par PMC avant et après un apprentissage « classique » (i.e. indépendant) dans un milieu bruité.

environnement	***MB avant app. global bruité***	***MB après app. global bruité***
Environnement non bruité	75,0 %	73,6 %
Bruit filtré grave (RSB = 20 dB)	30,3 %	31,9 %
Bruit blanc (RSB = 20 dB)	44,0 %	59,8 %
Bruit de voiture (RSB = -10 dB)	28,1 %	26,4 %
Bruit de cantine (RSB = 20 dB)	38,6 %	38,4 %

Tableau 7.3 : Taux de reconnaissance du système Multi-Bandes avec recombinaison par PMC avant et après un apprentissage global dans un milieu bruité.

Nous voyons que, à la différence du cas où seul le module de recombinaison est entraîné dans un milieu bruité, l'apprentissage global donne des résultats nettement meilleurs lorsque le bruit apparaissant pendant le test correspond à celui utilisé pendant l'apprentissage, mais que les résultats sont moins bons lorsque les deux environnements diffèrent. Ceci provient du fait que l'apprentissage, lorsqu'il est global, occasionne une dépendance plus grande vis-à-vis de l'environnement que lorsqu'il est indépendant. En fait, ceci est parfaitement compréhensible, car l'apprentissage global modifie non seulement le module de recombinaison, mais également les modèles dans les bandes : il crée donc des modèles de phonèmes bruités. Cette expérience montre qu'il n'est donc pas possible d'améliorer la robustesse « générale » du système Multi-Bandes par un apprentissage global, comme cela est le cas avec un apprentissage du module de recombinaison seul.

7.6. Étude des nouvelles classes phonétiques

7.6.1. Introduction

Nous avons jusqu'à présent essentiellement étudié les performances du système Multi-Bandes après son apprentissage global, mais nous avons déjà signalé en introduction que cet algorithme est également très intéressant car son « effet secondaire » principal est de modifier les classes modélisées par les HMM dans l'espace phonétique des sous-bandes. Nous abordons cette question ici et nous présentons ce qui peut être considéré comme un embryon d'analyse concernant ces nouvelles classes. Nous n'osons parler d'analyse complète, car, autant il est relativement facile d'exhiber les déplacements réalisés par les classes phonétiques au cours de l'apprentissage global, autant ceux-ci n'ont pas grand intérêt si nous ne décrivons pas intuitivement ce qu'ils représentent. Il ne faut pas oublier que le but de cette partie est avant tout de comprendre ce qui se passe dans l'espace phonétique des bandes. De plus, le déplacement des classes observé avec notre système est assez lent, ce qui est dû au faible nombre d'itérations que nous avons réalisées avec notre système. Nous n'avons donc pas pu observer de bouleversements majeurs dans l'espace phonétique, mais seulement des petites variations qui n'en demeurent pas moins intéressantes.

Idéalement, une telle analyse des nouvelles classes phonétiques devrait être menée dans les conditions suivantes :

1. Le système Multi-Bandes est initialement dépourvu de toute information *a priori*, c'est-à-dire que les HMM doivent être initialement vides et que les paramètres du module de recombinaison doivent être initialisés aléatoirement. De plus, *seul* l'apprentissage global doit être utilisé pour entraîner le système.
2. Le nombre d'itérations réalisées au cours de l'apprentissage est suffisamment grand pour s'assurer que le système Multi-Bandes a atteint un état stable.

7.6.2. Principe

La méthode la plus simple pour étudier ce que deviennent les classes consiste à considérer que les classes phonétiques définies par la segmentation constituent un « pavage » de l'espace acoustique. Ces classes phonétiques incluent des erreurs de la segmentation manuelle du signal, et sont donc certainement légèrement différentes des phonèmes tels que l'on peut les imaginer. Nous avons utilisé ce pavage comme base de notre étude, car, d'une part, les phonèmes idéaux n'existent pas, et, d'autre part, parce qu'il est facile de mesurer l'intersection des classes effectivement modélisées par une bande et de ces classes définies manuellement. Cette mesure est réalisée à partir de la matrice de confusion d'une bande qui est calculée après apprentissage global du système Multi-Bandes. L'étude de cette matrice de confusion est en effet riche en indications quant aux éventuels déplacements des classes dans une même sous-bande. Les matrices de confusion ont été calculées sur le corpus d'apprentissage qui est beaucoup plus volumineux que le corpus de test, car les nombres apparaissant dans la matrice de confusion calculée sur le corpus de test sont trop petits pour que leurs différences puissent être interprétées significativement.

La méthode pour analyser ce que deviennent les classes phonétiques peut donc être résumée dans les étapes suivantes :

1. Avant l'apprentissage global, les matrices de confusion des différentes bandes sont calculées à partir des modèles ayant subi un apprentissage « classique », i.e. indépendant des autres bandes. Soit $M_0(b,m)$ cette matrice pour la bande *b* et le modèle *m*.
2. Le système Multi-Bandes est ensuite construit, et quatre itérations de l'algorithme d'apprentissage global sont réalisées sur ce système.
3. Les bandes sont ensuite à nouveau séparées, et la matrice de confusion de chacune d'elle est calculée : soit $M_1(b,m)$ cette matrice pour la bande *b* et le modèle *m*.

Une fois ces matrices obtenues, nous calculons à partir d'elles un premier indice qui indique l'accroissement ou la réduction de la taille de la classe considérée. En effet, il est possible de connaître la taille de toute classe en sommant la colonne correspondante de la matrice de confusion, en sachant que chaque ligne de la matrice de confusion représente la classe prononcée et chaque colonne la classe reconnue. Nous calculons donc la taille de la classe avant (soit T_{avant} cette valeur) et après (soit $T_{après}$ cette valeur) l'apprentissage global. Nous pouvons alors déduire le taux d'accroissement de la taille de la classe en divisant $T_{après}$ par T_{avant}. Nous avons de plus soustrait l'unité à ce taux afin de le rendre négatif lorsque la classe se rétrécit et positif lorsqu'elle s'agrandit.

L'indice associé à la taille de la classe est donc :

$$\frac{T_{après}}{T_{avant}} - 1$$

Ce taux indique effectivement l'accroissement ou la diminution de la taille de la classe, car nous avons initialement considéré que l'espace phonétique est « pavé » par les phonèmes issus de la segmentation manuelle. Chaque exemple du corpus d'apprentissage définit donc un point de l'espace phonétique discrétisé. La taille d'une classe augmente si le nombre de points qui lui appartiennent augmente.

Le deuxième indice que nous avons calculé représente l'augmentation ou la diminution du taux de reconnaissance de la nouvelle classe par rapport à l'étiquetage manuel du corpus d'apprentissage. Cet indice nous permet notamment de vérifier que les nouvelles classes phonétiques sont effectivement différentes des phonèmes, auquel cas il devrait diminuer. Si au contraire il augmente, alors ceci signifie que le phonème concerné est bien adapté à la sous-bande considérée. Cet indice se calcule de la même manière que le précédent :

$$\frac{T_{après}}{T_{avant}} - 1$$

où $T_{après}$ et T_{avant} représentent respectivement le taux de reconnaissance après et avant l'apprentissage global.

Grâce à ces deux indices, nous pouvons d'ores et déjà déduire un certain nombre de résultats concernant les nouvelles classes phonétiques. Notre interprétation, très intuitive, peut se résumer dans les quatre cas suivants :

1. *Le taux de reconnaissance d'une classe augmente, ainsi que sa taille*. Ceci se produit lorsque l'information phonétique permettant d'identifier le phonème et qui se trouve dans la bande est suffisante. En effet, nous pouvons représenter la modification de la classe modélisée comme l'indique la figure 7.3. Sur ce schéma, le phonème « idéal », c'est-à-dire défini par la segmentation manuelle, est représenté en traits pointillés tandis que la classe modélisée apparaît en trait continu. Respectivement à gauche et à droite se trouve une disposition possible des classes avant et après l'apprentissage global.

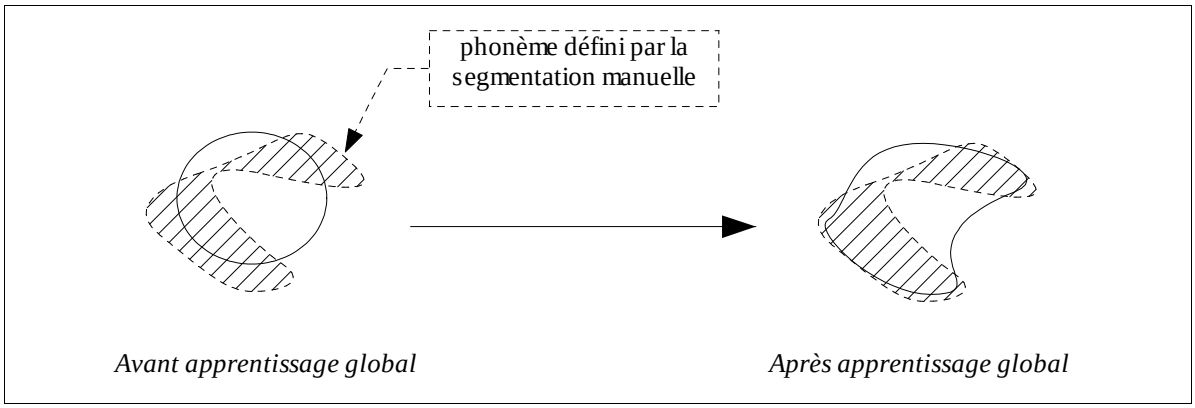

Figure 7.3 : Exemple de comportement d'une classe dans le cas 1.

Nous pouvons associer ce cas à un apprentissage classique qui présente le même comportement. Ce qui nous intéresse spécialement dans ce cas, c'est qu'il met en évidence le fait que de l'information phonétique permettant de distinguer le phonème considéré *est présente* dans la bande.

2. *Le taux de reconnaissance d'une classe diminue, ainsi que sa taille*. Ceci se produit lorsque l'information phonétique permettant de reconnaître le phonème n'est pas présente dans la bande, ou du moins qu'elle n'est pas suffisante pour que le système décide d'utiliser cette bande dans l'identification dudit phonème. Ce cas correspond donc vraisemblablement à une disparition progressive de la classe considérée. En fait, cette affirmation est un peu radicale, dans le sens où le système pourrait très bien « déplacer » la classe afin de lui faire modéliser un autre indice acoustique. Cependant, cet indice est certainement mineur, c'est-à-dire qu'il concerne peu d'exemples du corpus, à cause notamment du fait que la taille de la classe diminue. Or, notre étude concerne essentiellement les mouvements généraux des grandes classes et n'est pas suffisamment développée pour prendre en compte ces indices mineurs. Ce cas est donc identique, pour nous, à la disparition de la classe.

3. *Le taux de reconnaissance d'une classe diminue, mais pas sa taille*. Ceci se produit lorsque le phonème considéré n'est pas adapté à la bande, mais que le système « translate » la classe afin de lui faire modéliser un autre indice acoustique. La figure 7.4 présente un exemple d'un tel comportement.

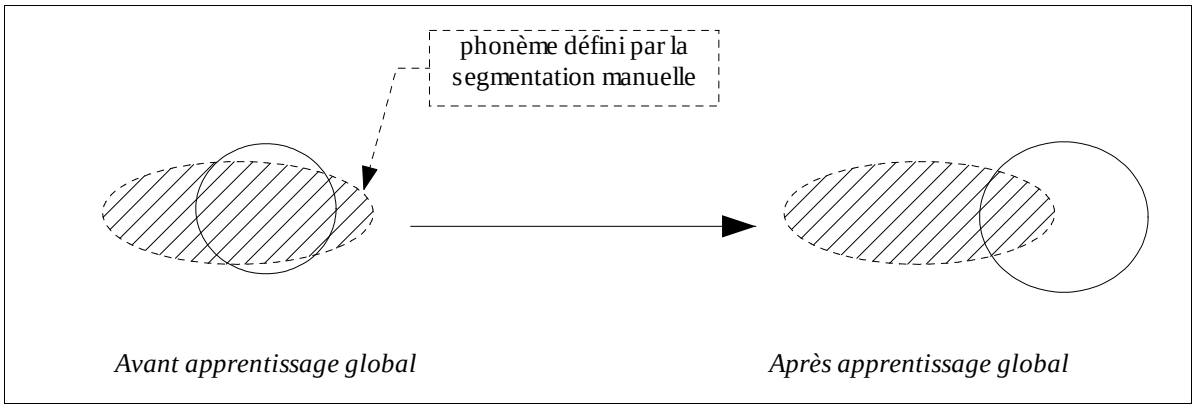

Figure 7.4 : Exemple de comportement d'une classe dans le cas 3.

4. *Le taux de reconnaissance d'une classe augmente tandis que sa taille diminue*. Nous nous trouvons alors en présence d'une classe plus petite, mais également moins confuse. Nous pouvons associer ce phénomène soit à du bruit qui a été introduit lors de la segmentation manuelle qui a associé par exemple de manière erronée des exemples du corpus à ce phonème, soit à des prononciations extravagantes d'un phonème, soit tout simplement à une grande confusion possible entre deux phonèmes dans la bande considérée. Dans tous les cas, il apparaît que le système « purifie » en quelque sorte la classe, en éliminant de celle-ci les exemples qui posent problème. De plus, le fait que le taux de reconnaissance augmente montre clairement que des indices acoustiques permettant d'identifier le phonème concerné se trouvent dans la bande.

Nous pouvons imaginer encore d'autres cas de figures, mais il s'agit en fait de comportements marginaux qui n'apparaissent que très rarement et que nous pouvons négliger dans un premier temps. Enfin, lorsque nous avions des doutes concernant certains phonèmes, ou lorsque nous voulions tout simplement analyser plus précisément le comportement d'un phonème donné, nous avons utilisé un troisième indice, qui n'est d'ailleurs pas vraiment un indice numérique. En fait, nous avons observé pour chaque classe, c'est-à-dire pour chaque colonne de la matrice $M_1(b,m) - M_0(b,m)$, les valeurs les plus grandes, les plus petites, mais aussi le nombre de valeurs positives comparé au nombre de valeurs négatives. Bref, nous avons chercher à visualiser pour chaque classe de quels phonèmes elle se rapproche ou au contraire s'éloigne. Ceci permet d'exhiber le « déplacement » de la classe dans l'espace phonétique.

Voilà les grandes lignes de la méthodologie que nous avons utilisée pour étudier ces nouvelles classes phonétiques. Passons maintenant aux résultats eux-mêmes.

7.6.3. Résultats expérimentaux

✓ ***Étude de la première bande***

Nous avons présenté ci-dessus des méthodes générales permettant d'étudier les modifications apportées aux classes phonétiques. Ces méthodes mesurent essentiellement l'importance accordée à ces nouvelles classes par le système, c'est-à-dire intuitivement le rôle que joue la classe dans la décision finale. Pour résumer, nous avons vu que si la taille de la classe et son taux de reconnaissance diminuent en même temps, ceci signifie que la classe a tendance à disparaître, et donc que son importance diminue également. Inversement, lorsque le taux de reconnaissance augmente, la classe peut être très correctement reconnue dans la bande considérée. Afin de chercher à faire apparaître ces deux phénomènes, nous avons affiché sur la figure 7.5 les deux indices définis ci-dessus : la courbe en trait continu représente le taux de modification de la taille des classes et la courbe en traits pointillés représente le taux de modification du taux de reconnaissance des classes. L'axe des abscisses contient les 48 classes modélisées dont la liste peut être consultée en annexe 2. Ces courbes correspondent à l'étude sur la première sous-bande uniquement.

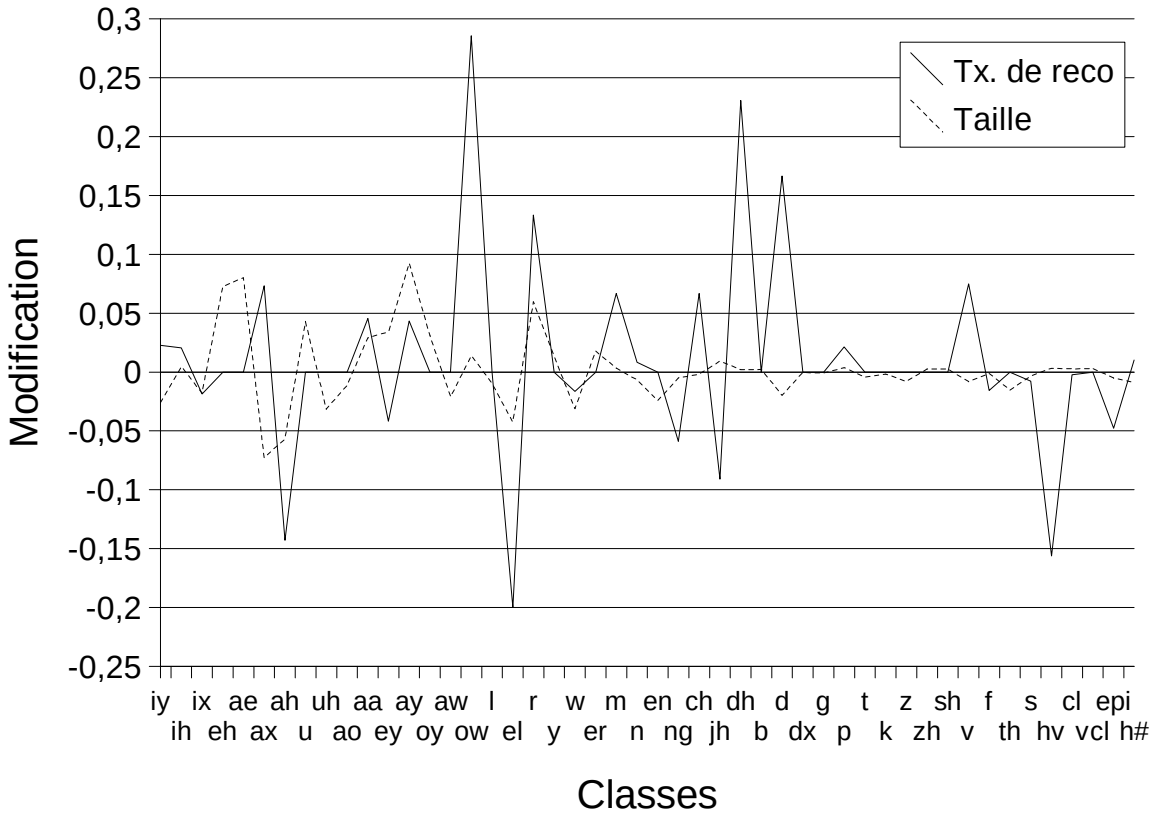

Figure 7.5 : Étude comparée de la modification de la taille des classes et de leur taux de reconnaissance pour la première sous-bande

La première remarque que nous pouvons faire sur cette figure est que la modification de la taille des classes est beaucoup plus grande pour les voyelles que pour les consonnes. Ceci suggère que les classes représentant les voyelles disposent d'un grand nombre d'indices acoustiques dans la première sous-bande et que le système réarrange dans une large mesure celles-ci.

La deuxième remarque est que le taux de reconnaissance diminue pour plusieurs phonèmes au cours de l'apprentissage global, ce qui signifie que celui-ci est bien différent d'un apprentissage classique, pour lequel les taux de reconnaissance des phonèmes ne font essentiellement que croître. Nous sommes donc bien en présence d'un début de transformation des classes phonétiques dans les bandes.

La troisième remarque concerne le fait que, comme prévu, il est rare de constater simultanément une diminution conséquente de la taille et une augmentation du taux de reconnaissance d'une classe : ceci confirme le fait que les classes ont tendance à modéliser de moins en moins d'indices acoustiques, mais de mieux en mieux, du moins nous pouvons le supposer au vu du taux de reconnaissance final. Intuitivement, cela signifie que, à la limite, nous pourrions par exemple obtenir les classes suivantes : voyelle, dentale, labiale, plosive, etc.

La quatrième remarque générale que nous pouvons faire sur la forme de ces courbes est que la modification du taux de reconnaissance est beaucoup plus importante que celle de la taille des classes. Ceci tendrait à montrer que les classes se « déplacent » dans l'espace phonétique plus qu'elles ne grossissent ou ne se réduisent. Nous pouvons donc considérer qu'au cours des quelques itérations d'apprentissage global que le système a subi, il a essentiellement recentré ses classes sur de nouvelles positions dans l'espace acoustique. Mais nous ne pouvons savoir si ce mouvement continuera ou si la taille des classes sera ensuite modifiée à son tour, une fois qu'un plus grand nombre d'itérations de l'algorithme se seront écoulés.

En ce qui concerne l'analyse plus détaille de cette première courbe, les phonèmes suivants semblent adaptés à la sous-bande considérée :

- /ow/, qui apparaît dans « boat »
- /dh/, qui apparaît dans « then »
- /d/, qui apparaît dans « day »
- /r/, qui apparaît dans « ray »

En effet, ceux-ci voient leur taux de reconnaissance augmenter considérablement, ce qui signifie que le système réalise un apprentissage presque « classique » sur eux, et donc ne les laisse pas *a fortiori* disparaître. Nous pouvons remarquer que ce ne sont que des phonèmes voisés.

Inversement, les phonèmes suivants ont un taux de reconnaissance qui chute beaucoup :

- /el/, qui apparaît dans « bottle »
- /hv/, qui apparaît dans « ahead »
- /ah/, qui apparaît dans « but »
- /jh/, qui apparaît dans « joke »

Ceci peut être dû, soit à un repositionnement radical de la classe, soit à sa disparition progressive. Cependant, l'analyse de la modification de la taille permet de classer ces quatre phonèmes en deux catégories : /el/ et /ah/, pour qui la taille diminue, et /hv/ et /jh/, pour qui la taille reste constante. Les deux premiers auraient donc tendance à disparaître, et les deux derniers à translater.

Nous ne poursuivons pas plus en détails l'étude de cette première sous-bande, non pas parce que celle-ci manque d'intérêt, mais parce que cette analyse, très délicate, ne constitue pas le sujet principal de la thèse. De plus, il est difficile de réaliser des interprétations plus fines des résultats sans avoir recours à plusieurs avis d'experts en phonétique.

L'étude des autres bandes est poursuivie en annexe 4.

7.7. Conclusion

Nous avons proposé dans ce chapitre un algorithme original réalisant l'apprentissage global du système Multi-Bandes. Les expériences ayant été réalisées avec cet algorithme ont montré que son utilisation permet d'améliorer les performances du système Multi-Bandes. De plus, il a pour conséquence de modifier la configuration des classes phonétiques modélisées par les HMM des bandes. Ceux-ci ne représentent alors plus les phonèmes, mais de nouvelles classes qui dépendent notamment de l'information acoustique présente dans chaque bande.

Nous avons réalisé une première analyse de ces nouvelles classes. Les conclusions de cette analyse sont les suivantes :

- L'apprentissage global se différencie d'un apprentissage classique par le fait qu'il modifie les classes phonétiques modélisées dans les bandes. Cette conclusion est motivée par différentes raisons, dont la plus importante est la décroissance des taux de reconnaissance de certaines classes dans toutes les bandes.

- Le réajustement des classes dans les bandes n'est pas seulement dû à des critères phonétiques, mais également à d'autres, plus stratégiques, comme la séparation dans différentes bandes des classes qui peuvent facilement être confondues. Ceci aboutit par exemple à la « perte » de certaines voyelles dans la première bande, voyelles qui se retrouvent ainsi dans les autres bandes. Cette conclusion montre que si les indices phonétique sont importants lors de la modélisation, ils ne sont cependant pas les seuls à entrer en compte, et des considérations stratégiques, comme la distance entre les classes ont également une grande influence sur les résultats.

Nous n'avons cependant fait qu'effleurer l'analyse des classes phonétiques, analyse qui mériterait des tests plus nombreux et un plus grand développement que ce qui a été présenté ici. Notre but était essentiellement de démontrer la différence entre l'apprentissage « classique » et l'apprentissage global, et de montrer qu'il est ensuite possible d'analyser ces données phonétiques pour en extraire des informations sur le fonctionnement interne d'un système.

Chapitre 8
Conclusions et Perspectives

8.1. Objectifs initiaux et conclusions de notre étude

8.1.1. Résumé des points principaux du mémoire

Le but de notre travail était d'évaluer en quoi le principe Multi-Bandes pouvait être utile à la reconnaissance automatique de la parole. Nous avons montré qu'il permet d'améliorer trois caractéristiques des systèmes de RAP :

1. Il offre une plus grande robustesse aux bruits additifs ;
2. Il augmente également les taux de reconnaissance dans les environnements non bruités ;
3. Il permet de créer de nouvelles unités phonétiques qui représentent mieux que les phonèmes l'information acoustique présente dans chaque bande.

Ces trois points résument les apports essentiels du principe Multi-Bandes et nous permettent de conclure positivement quant à l'efficacité d'un tel principe dans des systèmes réels de reconnaissance de la parole.

Néanmoins, un certain nombre de problèmes nouveaux sont issus de ce paradigme. Le plus contraignant d'entre eux est certainement l'impossibilité d'utiliser l'algorithme de Viterbi, dans le cas où l'asynchronisme entre les bandes est conservé à l'intérieur d'un phonème et dans des applications de reconnaissance de la parole en mode continu. Nous avons toutefois vu qu'un certain nombre d'autres algorithmes avaient été proposés dans la littérature pour remplacer l'algorithme de Viterbi, mais que ceux-ci sont souvent trop complexes ou trop coûteux pour être réellement efficaces. Aussi, la plupart des systèmes Multi-Bandes qui sont utilisés dans les autres laboratoires suppriment l'asynchronisme entre les bandes afin de pouvoir continuer à utiliser l'algorithme de Viterbi.

Cependant, nous pensons que l'hypothèse d'indépendance entre les bandes, qui est à la base du principe Multi-Bandes, est intimement liée à cet asynchronisme entre les bandes, et que le supprimer contraint le système à adopter une structure proche de celle d'un HMM classique. Nous pensons que le principe Multi-Bandes peut aller beaucoup plus loin que ça, et c'est ce que nous avons voulu démontrer dans le chapitre 7. C'est pourquoi nous avons proposé notre propre algorithme d'alignement, basé sur une récurrence de programmation dynamique associée à une recherche en faisceau, et qui permet de réaliser une reconnaissance en mode continu tout en conservant l'indépendance temporelle entre les bandes à l'intérieur de chaque phonème. Celui-ci a également l'avantage d'avoir une complexité asymptotiquement égale à celle de l'algorithme de Viterbi, même si la constante multiplicative de cette complexité est nettement plus grande.

De même, et toujours dans le but de préparer la conception d'un système complet de reconnaissance de la parole fondé sur le principe Multi-Bandes, nous avons considéré un par un tous les autres problèmes qui peuvent survenir lors de la réalisation dudit système. La plupart de ceux-ci relèvent d'ailleurs des choix d'implémentation plutôt que des difficultés de conception. Nous avons ainsi étudié en détails les différents modules de recombinaison possibles et avons mis en évidence leurs avantages et leurs défauts respectifs, en essayant toujours de comprendre les raisons de leurs comportements. Finalement, nous avons conçu un algorithme d'apprentissage global qui permet d'unifier les différents modules qui composent un système Multi-Bandes, car de tels systèmes souffraient jusqu'à présent de cette absence d'optimalité globale.

8.1.2. Originalité de notre travail

Les points qui justifient l'originalité de notre travail sont les suivants :

- Un algorithme d'apprentissage global pour le Multi-Bandes est présenté au chapitre 7, ce qui n'a encore jamais été réalisé à notre connaissance. Le premier apport de cet algorithme est de permettre d'approcher l'optimum global du système au cours de l'apprentissage, et donc d'accroître les performances de celui-ci. Le deuxième avantage est la création de nouvelles classes phonétiques dans chaque bande, et par là même de permettre une réflexion plus approfondie sur l'information phonétique effectivement contenue dans chaque zone du spectre.

- Nous avons été parmi les premiers à associer les sous-bandes, i.e. les bandes n'utilisant qu'une partie du spectre, avec le spectre complet à l'intérieur d'un même système. Cette association permet d'améliorer les taux de reconnaissance du système de référence, i.e. du spectre complet. La même idée est ensuite apparue dans d'autres laboratoires et gagne maintenant en ampleur.

- De même, nous avons été les premiers à appliquer l'algorithme de minimisation de l'erreur de classification au module de recombinaison linéaire.

- Une méthode permettant d'accroître la robustesse du Multi-Bandes aux bruits stationnaires, grâce à un apprentissage sélectif du module de recombinaison, est proposée dans la partie 6.5.

- Le Multi-Bandes a été appliqué à une nouvelle tâche en traitement automatique de la parole, l'identification du langage, dans la partie 6.7. Notre but était de montrer que le principe Multi-Bandes est un principe général de modélisation de la parole pouvant avoir des applications en reconnaissance de la parole mais aussi dans d'autres domaines.

- Un nouvel algorithme autorisant l'asynchronisme entre les bandes à l'intérieur de chaque modèle, et en même temps réalisant une reconnaissance de la parole en mode continu est présenté dans la partie 3.4. La plupart des autres travaux dans le domaine du Multi-Bandes ont contourné ce problème en obligeant les bandes à être synchrones à l'intérieur des modèles, mais nous pensons que cela n'est pas conforme au principe de base du Multi-Bandes et que cela ne permet pas d'exploiter au mieux celui-ci.

- De la même manière, un autre algorithme ne recombinant les bandes qu'en fin de phrase est présenté dans la partie 3.3. Cet algorithme impose des contraintes au module de recombinaison qui rendent celui-ci beaucoup moins performant. Néanmoins, nous avons montré qu'il devrait être possible d'obtenir de bons résultats malgré ces contraintes, bien que nous n'ayons pu construire expérimentalement une telle recombinaison. Nous espérons que les recherches dans le domaine de la fusion d'information permettront un jour de créer le module de recombinaison recherché.

- Nous avons systématiquement cherché à comprendre les raisons de tel ou tel comportement du système. Cette réflexion nous a, par exemple, amené à proposer une nouvelle explication à la robustesse du Multi-Bandes, qui est basée sur une résistance plus grande de chaque sous-bande que du spectre complet au bruit. De même, nos expériences nous ont amené à classer les différents types de bruits en deux catégories selon leur effet sur le Multi-Bandes, et donc à mieux analyser, prévoir et choisir une architecture en fonction de l'application voulue.

- D'autre part, la plupart des recherches sur le Multi-Bandes ont développé toute une série de tests permettant de mesurer la robustesse de ce principe dans différents environnements. Nous avons apporté notre contribution à ces expériences dans de nouveaux environnements et avec d'autres démarches expérimentales.

- Une étude approfondie du module de recombinaison est réalisée dans le chapitre 5, étude qui a également été menée ailleurs, mais avec des méthodologies et des conclusions différentes. Nous avons par exemple montré que la recombinaison linéaire permet d'obtenir des résultats très bons lorsque le milieu est très bruité.

- Finalement, notre système Multi-Bandes diffère des autres par son architecture (Utilisation de modèles de Markov du second ordre, ajout du spectre complet, ...), ce qui contribue à la diversité des systèmes Multi-Bandes existants.

8.2. Perspectives

8.2.1. Perspectives à court terme

Les problèmes liés au Multi-Bandes ont été suffisamment traités dans ce mémoire et dans les travaux antérieurs pour pouvoir envisager maintenant la création d'un système complet de reconnaissance de la parole Multi-Bandes. Cependant, le travail restant à faire sur ce système pour qu'il soit le meilleur possible tout en fonctionnant en temps réel est encore considérable.

Tout d'abord, il faut optimiser les différents algorithmes que nous avons proposé tout au long de ce mémoire. Ainsi, il est nécessaire de mieux ajuster l'algorithme d'apprentissage global afin de construire des modèles fiables et performants. Cette étape risque d'ailleurs d'être assez longue, car ces modèles ne représentant plus les phonèmes, comme cela est le cas depuis longtemps dans la littérature, leur configuration optimale reste encore à découvrir.

Ensuite, les paramètres de l'algorithme utilisant la programmation dynamique et qui permet la reconnaissance en mode continu doivent encore être correctement ajustés et des tests exhaustifs doivent être réalisés en ce qui les concerne. Cet algorithme doit également être adapté afin d'autoriser la reconnaissance des mots en mode continu. Le principe fondamental reste le même, mais de nouveaux problèmes se posent : il faut par exemple implanter un lexique, ce qui n'est pas toujours très facile en pratique, mais surtout résoudre certains problèmes liés à la complexité des graphes d'alignement, et inclure ainsi de nouvelles recherches en faisceau.

Enfin, il est indispensable d'ajouter un étage linguistique au système afin de pouvoir construire un système complet de reconnaissance de la parole grand vocabulaire. Il faudra alors prendre en compte le contexte dans notre système, aussi bien au niveau des modèles des bandes qu'au niveau sémantique. Il faudra peut-être ainsi reconsidérer le choix du phonème comme unité de base, et le remplacer par des di-phones, voire des syllabes. Cependant, nous ignorons l'importance du contexte pour les modèles issus d'un apprentissage global : cette étude devra préalablement être menée, mais il est difficile de savoir actuellement si des unités incluant le contexte devront être utilisées, ou si l'apprentissage global pourra être adapté afin d'être employé en mode continu, et s'il sera capable alors de gérer lui-même l'influence du contexte, ce qui semble probable.

8.2.2. Perspectives à plus long terme

Nous développons dans la suite des propositions d'évolution qui concernent le principe Multi-Bandes lui-même, et plus seulement des extensions pour notre système.

Nous pouvons ainsi remarquer que les architectures Multi-Bandes développées dans ce mémoire sont presque toutes basées sur une recombinaison utilisant les scores retournés par les HMM. Or, nous avons vu qu'il devrait être possible de réaliser des recombinaisons également performantes en utilisant beaucoup moins d'information, par exemple les rangs des modèles dans chaque bande. Des recombinaisons de ce type recommencent d'ailleurs à voir le jour, notamment dans le champ de la théorie des données manquantes qui suggère l'utilisation de recombinaisons dites « par sélection dynamique ». Nous en avons cité quelques exemples dans ce mémoire. L'avantage obtenu lorsque l'on n'utilise pas les scores des HMM est de pouvoir, par exemple, recombiner les bandes à la fin de la phrase, comme cela est montré dans la partie 3.3. Ceci permet donc de tester de nouvelles architectures très intéressantes pour les systèmes Multi-Bandes. De même, il serait profitable pour notre système d'utiliser dans le module de recombinaison de nouveaux réseaux de neurones plus adaptés à la reconnaissance de la parole, comme les réseaux récurrents.

En ce qui concerne la rapidité d'exécution, nous avons montré que nos algorithmes suivent la même loi de croissance que l'algorithme de Viterbi, mais qu'ils sont cependant plus lents à cause des constantes apparaissant dans le calcul du coût. Ce problème n'est pas critique, dans la mesure où la puissance sans cesse croissante des machines devrait permettre de combler très rapidement le retard occasionné par ces « constantes », mais il serait néanmoins très intéressant de développer également des heuristiques supplémentaires pour les différents algorithmes élaborés dans ce mémoire. Il faut en effet laisser le plus de temps de calcul possible disponible pour les modules linguistiques des systèmes de reconnaissance et de compréhension de la parole.

Bibliographie

[Allen94] Allen J. B. : How do Humans Process and Recognize Speech ? *IEEE Trans. on Speech and Audio Processing*, Vol. 2, pp. 567-576, Octobre 1994.

[Bahl88] Bahl L. R., Brown P. F., de Souza P. V. et Mercer R. L. : A New Algorithm for the Estimation of Hidden Markov Model Parameters. *ICASSP'88*, New York, USA, Avril 1988.

[Bengio92] Bengio Y., De Mori R., Flammia G. et Kompe Ralf : Global Optimization of a Neural Network-Hidden Markov Model Hybrid. *IEEE Trans. on Neural Networks*, Vol. 3, N. 2, Mars 1992.

[Berger90] Berger M. : *Géométrie*. Vol. 1 et 2, NATHAN, 1990.

[Berthommier98] Berthommier F., Glotin H., Tessier E. et Bourlard H. : Interfacing of CASA and partial recognition based on a multistream technique. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Berthommier99] Berthommier F. et Glotin H. : A measure of speech and pitch reliability from voicing. Ed. F. Klassner, *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI), Computational Auditory Scene Analysis (CASA) workshop*, pages 61-70, Stockholm, July 1999.

[Besacier98a] Besacier L. : Un Modèle Parallèle pour le Reconnaissance Automatique du Locuteur. *Thèse de Doctorat de l'Université d'Avignon et des Pays de Vaucluse*, Avignon, France, Avril 1998.

[Besacier98b] Besacier L., Bonastre J.-F. et Fredouille C. : Architecture en sous-bandes pour la reconnaissance automatique du locuteur en milieu bruité. *RFIA'98*, Clermont-Ferrand, France, Janvier 1998.

[Boll79] Boll S. F. : Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. ASSP-27, N°2, pp.113-120, Avril 1979.

[Bourlard94] Bourlard H. et Morgan N. : *Connectionnist Speech Recognition: A Hybrid Approach*. Kluwer Press, 1994.

[Bourlard96] Bourlard H. et Dupont S. : A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands. *ICSLP'96*, Philadelphia, USA, Octobre 1996.

[Bourlard97] Bourlard H. et Dupont S. : Subband-based speeh recognition. *ICASSP'97*, Münich, Allemagne, pp. 1251-1254, 1997.

[Breiman96] Breiman L. : Stacked Regressions. *Machine Learning*, Vol. 24, pp. 49-64, 1996.

[Calliope89] Calliope. *La parole et son traitement automatique*. Masson, 1989.

[Cerisara96] Cerisara C., Gong Y. et Haton J.-P. : Reconnaissance de la parole continue par le modèle STM polynomial. *Journées d'Étude de la Parole JEP'96*, Avignon, France, 1996.

[Cerisara97] Cerisara C., Haton J.-P., Mari J.-F. et Fohr D. : Multi-band continuous speech recognition. *EUROSPEECH'97*, Rhodes, Grèce, Septembre 1997.

[Cerisara98a] Cerisara C., Haton J.-P., Mari J.-F. et Fohr D. : A recombination model for multi-band speech recognition. *ICASSP'98*, Seattle, USA, Mai 1998.

[Cerisara98b] Cerisara C., Afify M. et Haton J.-P. : Étude de la recombinaison de plusieurs classifieurs appliqués à deux tâches de reconnaissance de la parole. *Journées d'Étude sur la Parole JEP'98*, Martigny, Suisse, Juin 1998.

[Cerisara99a] Cerisara C.. Dealing with Loss of Synchronism in Multi-Band Continuous Speech Recognition. *Computational Models of Speech Pattern Processing*, Ed. Keith M. Ponting, NATO ASI Series F, Springer Verlag, 1999.

[Cerisara99b] Cerisara C., Haton J.-P. et Fohr D. : Towards a Global Optimization Scheme for Multi-Band Speech Recognition. *EUROSPEECH'99*, Budapest, Hongrie, Septembre 1999.

[Cerisara99c] Cerisara C., Fohr D. et Haton J.-P. : Robust Behavior of Multi-Band Paradigm. *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finlande, Mai 1999.

[Chu98] Chu S. M. et Zhao Y. : Robust Speech Recognition using Discriminative Stream Weighting and Parameter Inerpolation. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Cooke97] Cooke M., Morris A. et Green P. : Missing Data Techniques for Robust Speech Recognition. *ICASSP'97*, Münich, Allemagne, Avril 1997.

[Dasarathy94] Dasarathy B.V. : *Decision Fusion*. IEEE Computer Society Press, 1994.

[DeMori98] De Mori R. : *Spoken Dialogues with Computers*. Ed. Renato De Mori, Academic Press, 1998.

[Duchnowsky93] Duchnowsky P. : A New Structure for Automatic Speech Recognition. *PhD. thesis, Massachusetts Institute of Technology*, Cambridge, USA, Septembre 1993.

[Duda73] Duda R.O. et Hart P.E. : *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[Dupont97] Dupont S. et Bourlard H. : Using Multiple Time Scales in a Multi-Stream Speech Recognition System. *EUROSPEECH'97*, Rhodes, Grèce, Septembre 1997.

[Dupont98] Dupont S. : Missing Data Reconstruction for Robust Automatic Speech Recognition in the Framework of Hybrid HMM/ANN Systems. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Ephraïm95] Ephraïm Y. et Van Trees H. L. : A Signal Subspace Approach for Speech Enhancement. *IEEE Trans. on Speech and Audio Processing*, Vol. 3, N°4, Juillet 1995.

[Fohr97] Fohr D., Haton J.-P., Mari J.-F., Smaïli K. et Zitouni I. : MAUD: Un prototype de machine à dicter vocale. *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, pp. 25-30, 1997.

[Forney73] Forney G. D. : The Viterbi Algorithm. *IEEE Transactions*, Vol. 61, pp. 268-278, Mars 1973.

[French47] French F. R. et Steinberg J. C. : Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, Vol. 19, N°1, pp. 90-119, Janvier 1947.

[Gales98] Gales M. J. F. : Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, Vol. 25, pp. 49-74, 1998.

[Garofolo93] Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J. G., Pallet D. S. et Dahlgren N. L. : The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, 1993.

[Gauvain91] Gauvain J.-L., Lamel L.-F. et Eskénazi M. : BREF, a large vocabulary spoken corpus for french. *EUROSPEECH'91*, Genova, Italie, pp. 505-508, 1991.

[Ghahramani97] Ghahramani Z., Jordan M. : Factorial Hidden Markov Models. *Machine Learning*, Vol. 29, pp. 245-273, 1997.

[Ghitza94] Ghitza O. : Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Trans. on Speech and Audio processing*, Vol. 2, N°1, Janvier 1994.

[Glotin98a] Glotin H., Berthommier F., Tessier E. et Bourlard H. : Interfacing of CASA and multistream recognition. Éditeur : Petr Sojka, *TSD'98-Text, Speech and Dialog International Workshop*, Brno, République Tchèque, Septembre 1998.

[Glotin98b] Glotin H., Tessier E., Bourlard H. et Berthommier F. : Reconnaissance robuste de la parole par segmentation signal/bruit en sous-bandes. *IX ème Journées Neurosciences et Sciences de l'Ingénieur*, Munster, France, Mai 1998.

[Glotin99] Glotin H., Berthommier F. et Tessier E. : A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition. *EUROSPEECH'99*, Budapest, Hongrie, Septembre 1999.

[Gong97] Gong Y. : Stochastic Trajectory Modeling and Sentence Searching for Continuous Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 5, N°1, pp. 33-44, Janvier 1997.

[Gravier98] Gravier G., Sigelle M. et Chollet G. : Toward Markov Random Field Modeling of Speech. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Greenberg96] Greenberg S. : Understanding Speech Understanding : Towards a Unified Theory of Speech Perception. *ESCA Workshop on the « Auditory Basis of Speech Perception »*, Keele University, pp. 1-8, 1996.

[Greenberg98a] Greenberg S. et Arai T. : Speech Intelligibility is Highly Tolerant of Cross-Channel Spectral Asynchrony. *Joint Proceedings of the Acoustical Society of America and the International Congress on Acoustics,* Seattle, USA, 1998.

[Greenberg98b] Greenberg S., Arai T. et Silipo R. : Speech Intelligibility Derived from Exceedingly Sparse Spectral Information. *ICSLP'98,* Sydney, Australie, Décembre 1998.

[Hagen99] Hagen A., Morris A. et Bourlard H. : Different Weighting Schemes in the Full Combination Subbands Approach in Noise Robust ASR. *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finlande, Mai 1999.

[Haton74] Haton J.-P. : Contribution à l'Analyse, la Paramétrisation et la Reconnaissance Automatique de la Parole. *Thèse de Doctorat de l'Université de Nancy I*, Nancy, 1974.

[Haton91] Haton J.-P., Bouzid N., Charpillet F., Haton M.-C., Lâasri B., Lâasri H., Marquis P., Mondot T. et Napoli A. : *Le raisonnement en intelligence artificielle : Modèles, techniques et architectures pour les systèmes à bases de connaissance*. InterEditions, Paris, 1991.

[Haton95] Haton J.-P. : Les modèles neuronaux et hybrides en reconnaissance automatique de la parole : états des recherches. Méloni H., éditeur, *Actes École Thématique sur les Fondements et Perspectives en Traitement Automatique de la Parole*, pp. 139-153. Université d'Avignon et des pays de Vaucluse, Marseille, Juillet 1995.

[Hermansky94] Hermansky H. et Morgan N. : RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, Vol. 2, N°4, pp. 578-589, Octobre 1994.

[Hermansky96] Hermansky H., Tibrewala S. et Pavel M. : Towards ASR On Partially Corrupted Speech. *ICSLP'96*, Philadelphia, USA, Octobre 1996.

[Hermansky98] Hermansky H. et Sharma S. : TRAPS - Classifiers of Temporal Patterns. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Ho94] Ho T. K., Hull J. J. et Srihari S. N. : Decision Combination in Multiple Classifier Systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 16, N. 1, Janvier 1994.

[Husson98] Husson J.-L. : Une approche hiérarchique de la segmentation du signal de parole. *Thèse de Doctorat, Université Henri Poincaré I*, Nancy, France, Novembre 1998.

[Illina97] Illina I. : Extension du modèle stochastique des mélanges de trajectoires pour la reconnaissance automatique de la parole continue. *Thèse de Doctorat, Université Henri Poincaré I*, Nancy, France, Octobre 1997.

[Juang92] Juang B.-H. et Katagiri S. : Discriminative Learning for Minimum Error Classification. *IEEE Trans. on Signal Processing*, Vol. 40, N°12, pp. 3043-3054, Décembre 1992.

[Juang97] Juang B.-H., Chou W. et Lee C.-H. : Minimum Classification Error Rate Methods for Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 5, N° 3, pp. 257-265, Mai 1997.

[Junqua96] Junqua J.-C. et Haton J.-P. : *Robustness in Automatic Speech Recognition : Fundamentals and Applications*. The Kluwer International Series in Engineering and Computer Science, 1996.

[Lee88] Lee K. F. : Large Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX system. *PhD. thesis, CMU*, Pittsburgh, USA, 1988.

[Lee89] Lee K. F. et Hon H. W. : Speaker independent phon recognition using hidden Markov models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 37, N° 11, Novembre 1989.

[Lee97] Lee T-W., Bell A.J. et Orglmeister R. : Blind Source Separation of Real World Signals, *IEEE International Conference on Neural Networks*, Houston, pp 2129-2135, Juin 1997.

[Levinson86] Levinson S. : Continuous Variable Duration Hidden Markov Models for Automatic Speech Recognition. *Computer Speech and Language*, Vol. 1, pp. 29-45, 1980.

[Lippmann97a] Lippmann R. P. et Carlson B. A. : Robust Speech Recognition with Time-Varying Filtering, Interruptions, and Noise. Ed. Furui S., Juang B.-H. et Chou W., *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 365-372, Santa Barbara, USA, Décembre 1997.

[Lippmann97b] Lippmann R. : Using Missing Feature Theory to Actively Select Features for Robust Speech Recognition with Interruptions, Filtering and Noise. *EUROSPEECH'97*, Rhodes, Grèce, Septembre 1997.

[Mari96] Mari J.-F. : Perception de signaux complexes et intéraction homme-machine. *Habilitation à diriger des recherches*, Université Henri Poincaré I, Nancy, France, 1996.

[Mari97] Mari J.-F., Haton J.-P. et Kriouile A. : Automatic word recognition based on second-order hidden Markov Models. *IEEE Trans. on Speech and Audio Processing*, Vol. 5, pp. 22-25, Janvier 1997.

[McMahon98] McMahon P., McCourt P. et Vaseghi S. : Discriminative Weighting of Multi-Resolution Sub-band Cepstral Features for Speech Recognition. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Mella94] Mella O. : Extraction of formants of oral vowels and critical analysis for speaker characterization. *ESCA Workshop on Automatic Speaker Recognition, Identification & Verification*, pp. 193-196, Suisse, 1994.

[Miller55] Miller G. A. et Nicely P. E. : An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, Vol. 27, N°2, pp. 338-352, Mars 1955.

[Mirghafori97] Mirghafori N. M. : Multi-Band Speech Recognition: A Summary of Recent Work at ICSI. *Rapport Technique TR-97-051, ICSI*, Berkeley, Septembre 1997.

(ftp://ftp.icsi.berkely.edu/pub/techreports/1997/tr-97-051.ps.gz)

|Mirghafori98] Mirghafori N. M. et Morgan N. : Combining Connectionist Multi-Band and Full-Band Probability Streams for Speech Recognition of Natural Numbers. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Mirghafori99] Mirghafori N. N. : A Multi-Band Approach to Automatic Speech Recognition. *PhD. thesis, ICSI*, Berkeley, USA, Janvier 1999.

[Mitchell97] Mitchell T.M. : *Machine Learning*, Mc Graw-Hill, 1997.

[Mokhtari94] Mokhtari P. et Clermont F. : Contributions of Selected Spectral Regions to Vowel Classification Accuracy. *ICSLP'94*, Yokohama, Japon, Septembre 1994.

[Mokhtari96] Mokhtari P. et Clermont F. : A Methodology for Investigating Vowel-Speaker Interactions in the Acoustic-Phonetic Domain. *VIth Australian International Conference on Speech Science and Technology*, pp. 127-132, Australie, Décembre 1996.

[Muthusamy92a] Muthusamy Y.K. et Cole R. A. : Automatic segmentation and identification of ten languages using telephone speech. *ICASSP'92*, pp. 1007-1010, Octobre 1992.

[Muthusamy92b] Muthusamy Y. K., Cole R. A. et Oshika B. T. : The OGI multi-language telephone speech corpus. *ICSLP'92*, Alberta, Canada, Octobre 1992.

[Myers81] Myers C. S. et Rabiner L. R. : Connected Digit Recognition using a Level-Building DTW Algorithm. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 29, N°3, pp. 351, 1981.

[NOISE-ROM90] CDROM NOISE-ROM-0 : NATO : AC243/(Panel 3)/RSG-10. ESPRIT: Project N° 2589-SAM. *Institute for Perception –TNO*, The Netherlands. Speech Research Unit, RSRE, United Kingdom, Mars 1990.

[Okawa98] Okawa S., Bocchieri E. et Potamianos A. : Multi-Band Speech recognition in noisy environments. *ICASSP'98*, Seattle, USA, Mai 1998.

[Ostendorf92] Ostendorf M., Kannan A., Kimball O. et Rohlicek J. R. : Continuous Word Recognition Based on the Stochastic Segment Model. *Proc. DARPA Workshop Continuous Speech Recognition*, 1992.

[Ostendorf96] Ostendorf M., Digalakis V. et Kimball O. : From HMM's to Segments Models: a Unified View of Stochastic Modelling for Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, Vol. 4, N°5, pp. 360-378, 1996.

[Pican95] Pican N. : Approches Statique et Dynamique de la Modulation des Efficacités Synaptiques dans les Réseaux de Neurones. *Thèse de Doctorat, Université Henri Poincaré I*, Nancy, France, Janvier 1995.

[Rabiner78] Rabiner L. R. et Schafer R. W. : *Digital Processing of Speech Signals*. Ed. A. V. Oppenheim, Prentice-Hall signal processing series, 1978.

[Rabiner93] Rabiner L. R. et Juang B.-H. : *Fundamentals of Speech Recognition*. Prentice Hall International Editions, 1993.

[Raj98] Raj B., Singh R. et Stern R. M. : Inference of Missing Spectrographic Features for Robust Speech Recognition. *ICSLP'98*, Sydney, Australie, Décembre 1998.

[Ravishankar96] Ravishankar M. K. : Efficient Algorithms for Speech Recognition. *PhD. thesis, CMU*, Pittsburgh, USA, Mai 1996.

[Reyes94] Reyes A. A., Seino T. et Nakagawa S. : Three Language Identification Methods based on HMM. *ICSLP'94*, pp. 1895-1898, Yokohama, Japon, Septembre 1994.

[Saporta90] Saporta G. : *Probabilités, Analyse des données et Statistique*. Ed. Technip, 1990.

[Simonnard62] Simonnard M. : *Programmation linéaire*. DUNOD, 1962.

[Siohan95] Siohan O. : Reconnaissance automatique de la parole continue en environnement bruité : Application à des modèles stochastiques de trajectoire. *Thèse de Doctorat, Université Henri Poincaré I*, Nancy, France, Septembre 1995.

[Siohan98] Siohan O. : Sub-Band Censoring for Noisy Speech Recognition. *Communication personnelle*. 1998.

[Su94] Su K.-Y. et Lee C.-H. : Speech Recognition Using Weighted HMM and Subspace Projection Approaches. *IEEE Trans. on Speech and Audio Processing*, Vol. 2, N. 1, Part 1, Janvier 1994.

[Tibrewala97] Tibrewala S. et Hermansky H. : Sub-band based recognition of noisy speech. *ICASSP'97*, pp. 1255-1258, Münich, Allemagne, 1997.

[Tumer99] Tumer K. et Gosh J. : Linear and Order Statistics Combiners for Pattern Classification. Ed. A. Sharkey, *Combining Artificial Neural Nets*, pp. 127-162, Springer-Verlag, 1999.

[Verlinde99] Verlinde P., Chollet G. et Acheroy M. : Mulit-Modal Identity Verification Using Decision Fusion. *International Journal on Information Fusion*, 5 février 1999.

[Wolpert92] Wolpert D. H. : Stacked Generalization. *Neural Networks*, Vol. 5, pp. 241-259, 1992.

[Xu96] Xu D., Fancourt C. et Wang C. : Multi-channel HMM. *ICASSP'96*, pp. 841-844, Atlanta, USA, Mai 1996.

[Zissman96] Zissman M. A. : Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. *IEEE Trans. on Speech and Audio Processing*, Vol. 4, N°1, pp. 31-44, Janvier 1996.

Annexe 1

Résultats chiffrés de toutes les expériences présentées dans le mémoire

Tout au long du mémoire, nous avons étayé nos commentaires d'un nombre appréciable de courbes, car celles-ci ont l'avantage d'être facilement et rapidement interprétables. Toutefois, afin que le lecteur qui le désire puisse consulter de manière plus précise les chiffres qui sont présentés dans ces courbes, nous avons décidé de présenter en Annexe 1 tous les tableaux chiffrés qui correspondent à ces courbes. Au dessus de chaque tableau, nous pouvons trouver la référence de la courbe correspondante dans le mémoire.

✓ ***Taux de reconnaissance des systèmes linéaires sans apprentissage du module de recombinaison dans le bruit filtré aigu.***

Figure correspondante : Figure 6.4.

RSB	***Moyenne 4 bandes***	***Moyenne 5 bandes***	***0,1 x 4 + 0,6***	***0,05 x 4 + 0,8***	***Référence***
20 dB	47,5 %	51,1 %	47,4 %	45,7 %	44,8 %
30 dB	54,0 %	59,4 %	56,6 %	55,7 %	54,8 %
38 dB	59,8 %	66,0 %	65,3 %	64,8 %	64,1 %
48 dB	63,4 %	70,6 %	70,5 %	70,2 %	69,8 %
53 dB	64,8 %	72,6 %	73,0 %	72,7 %	72,5 %

✓ ***Taux de reconnaissance des systèmes linéaires sans apprentissage du module de recombinaison dans le bruit filtré grave.***

Figure correspondante : Figure 6.5.

RSB	***Moyenne 4 bandes***	***Moyenne 5 bandes***	***Référence***
20 dB	30,0 %	27,7 %	19,9 %
30 dB	46,7 %	48,8 %	39,3 %
38 dB	59,2 %	65,9 %	56,9 %
48 dB	64,8 %	72,1 %	65,3 %
53 dB	65,6 %	73,0 %	67,2 %

✓ ***Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans le bruit filtré aigu***

Figure correspondante : Figure 6.6.

RSB	***PMC***	***Référence***	***Moyenne 5 bandes***	***MCE***
20 dB	49,2 %	44,8 %	51,1 %	46,3 %
30 dB	58,6 %	54,8 %	59,4 %	56,6 %
38 dB	67,0 %	64,1 %	66,0 %	65,3 %
48 dB	71,5 %	69,8 %	70,6 %	71,1 %
53 dB	73,8 %	72,5 %	72,6 %	73,5 %

✓ ***Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans le bruit filtré grave***

Figure correspondante : Figure 6.7.

RSB	***PMC***	***Référence***	***Moyenne 5 bandes***
20 dB	27,1 %	19,9 %	27,7 %
30 dB	48,0 %	39,3 %	48,8 %
38 dB	65,4 %	56,9 %	65,9 %
48 dB	72,9 %	65,3 %	72,1 %
53 dB	74,2 %	67,2 %	73,0 %

✓ ***Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans du bruit blanc.***

Figure correspondante : Figure 6.8.

RSB	***0,05 x 4 + 0,8***	***0,1 x 4 + 0,6***	***Référence***	***Moyenne 4 bandes***	***Moyenne 5 bandes***
20 dB	41,2 %	42,2 %	40,6 %	42,7 %	44,7 %
30 dB	60,0 %	61,0 %	59,4 %	56,6 %	61,9 %
38 dB	69,9 %	70,5 %	69,5 %	63,6 %	70,8 %
48 dB	73,2 %	73,6 %	72,9 %	65,8 %	72,8 %
53 dB	73,5 %	73,9 %	73,2 %	65,7 %	73,4 %

✓ ***Taux de reconnaissance des systèmes Multi-Bandes « complexes » dans du bruit blanc.***

Figure correspondante : Figure 6.9.

RSB	***PMC***	***Référence***	***Moyenne 5 bandes***	***MCE***
20 dB	43,8 %	40,6 %	44,7 %	42,3 %
30 dB	63,1 %	59,4 %	61,9 %	61,0 %
38 dB	71,9 %	69,5 %	70,8 %	70,5 %
48 dB	74,2 %	72,9 %	72,8 %	73,6 %
53 dB	74,2 %	73,2 %	73,4 %	73,9 %

✓ ***Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans le bruit de voiture.***

Figure correspondante : Figure 6.11.

RSB	***0,05 x 4 + 0,8***	***0,1 x 4 + 0,6***	***Référence***	***Moyenne 4 bandes***	***Moyenne 5 bandes***
-10 dB	25,8 %	28,0 %	24,4 %	35,3 %	35,7 %
-5 dB	33,0 %	35,5 %	31,3 %	39,5 %	41,8 %
5 dB	48,2 %	50,1 %	46,5 %	48,1 %	53,2 %
10 dB	55,5 %	57,2 %	54,1 %	52,5 %	58,1 %

✓ ***Taux de reconnaissance, dans le bruit de voiture, des systèmes Multi-Bandes avec apprentissage du module de recombinaison***

Figure correspondante : Figure 6.12.

RSB	***PMC***	***Référence***	***MCE***	***Moyenne 5 bandes***
-10 dB	29,1 %	24,4 %	28,0 %	35,7 %
-5 dB	35,4 %	31,3 %	35,4 %	41,8 %
5 dB	50,0 %	46,5 %	50,1 %	53,2 %
10 dB	56,8 %	54,1 %	57,2 %	58,1 %

✓ ***Taux de reconnaissance des systèmes Multi-Bandes sans apprentissage du module de recombinaison dans le bruit de cantine***

Figure correspondante : Figure 6.14.

RSB	***0,05 x 4 + 0,8***	***0,1 x 4 + 0,6***	***Référence***	***Moyenne 4 bandes***	***Moyenne 5 bandes***
20 dB	36,9 %	37,9 %	36,1 %	36,4 %	39,6 %
30 dB	57,8 %	58,0 %	57,1 %	51,0 %	58,5 %

RSB	***0,05 x 4 + 0,8***	***0,1 x 4 + 0,6***	***Référence***	***Moyenne 4 bandes***	***Moyenne 5 bandes***
38 dB	68,9 %	69,4 %	68,6 %	61,0 %	69,3 %
48 dB	72,9 %	73,2 %	72,5 %	64,6 %	72,6 %
53 dB	73,3 %	73,7 %	73,0 %	65,3 %	72,8 %

✓ ***Taux de reconnaissance, dans le bruit de cantine, des systèmes Multi-Bandes avec apprentissage du module de recombinaison***

Figure correspondante : Figure 6.15.

RSB	***PMC***	***Référence***	***Moyenne 5 bandes***	***MCE***
20 dB	38,5 %	36,1 %	39,6 %	38,4 %
30 dB	58,8 %	57,1 %	58,5 %	58,9 %
38 dB	69,9 %	68,6 %	69,3 %	69,6 %
48 dB	73,9 %	72,5 %	72,6 %	73,7 %
53 dB	74,1 %	73,0 %	72,8 %	74,2 %

✓ ***Évolution du taux de reconnaissance des systèmes Multi-Bandes en fonction du nombre d'itérations de l'apprentissage global***

Figure correspondante : Figure 7.2.

Nb. d'itérations	***Linéaire***	***PMC***	***Référence***
0	73,9 %	74,4 %	73,3 %
1	74,1 %	74,7 %	73,3 %
2	74,1 %	74,8 %	73,3 %
3	74,1 %	74,9 %	73,4 %
4	74,1 %	74,9 %	73,5 %
5	74,3 %	75,0 %	73,5 %
6	74,4 %	75,1 %	73,5 %

Annexe 2

Liste des phonèmes utilisés

La liste des phonèmes que nous avons utilisés tout au long de ce mémoire est celle fournie avec le corpus TIMIT, et elle correspond donc aux phonèmes de l'américain standard. Nous rappelons cette liste ci-dessous, en indiquant le numéro de chaque phonème, avant et après regroupement, ainsi qu'un exemple de prononciation.

Phonème	***Exemple***	***Indice***	***Indice après regroupement***
iy	beet	0	0
ih	bit	1	1
ix	debit	2	1
eh	bet	3	2
ae	bat	4	3
ax	about	5	4
ah	but	6	4
u	boot	7	5
uh	book	8	6
ao	bought	9	7
aa	bott	10	7
ey	bait	11	8
ay	bite	12	9
oy	boy	13	10
aw	bout	14	11
ow	boat	15	12
l	lay	16	13
el	bottle	17	13
r	ray	18	14
y	yacht	19	15
w	way	20	16
er	bird	21	17
m	mom	22	18

Phonème	***Exemple***	***Indice***	***Indice après regroupement***
n	noon	23	19
en	button	24	19
ng	sing	25	20
ch	choke	26	21
jh	joke	27	22
dh	then	28	23
b	beet	29	24
d	day	30	25
dx	muddy	31	26
g	gay	32	27
p	pea	33	28
t	tea	34	29
k	key	35	30
z	zone	36	31
zh	azure	37	32
sh	she	38	32
v	van	39	33
f	fin	40	34
th	thin	41	35
s	sea	42	36
hv	ahead	43	37
cl		44	38
vcl		45	38
epi		46	38
h#		47	38

Annexe 3

Étude individuelle du comportement phonétique des sous-bandes

Cette annexe poursuit l'étude du comportement phonétique des sous-bandes, étude qui a débuté dans la partie 4.4.2. Nous rappelons que celle-ci concerne les sous-bandes considérées isolément, et ne fait pas intervenir de module de recombinaison. De même, les HMM utilisés pour modéliser les phonèmes dans ces sous-bandes ont subi un apprentissage classique indépendant des autres bandes.

✓ ***Étude de la deuxième sous-bande : [461 ... 1000 Hz]***

La différence des taux de reconnaissance entre le spectre complet (noté SC) et la deuxième sous-bande est présentée sur la figure B.1.

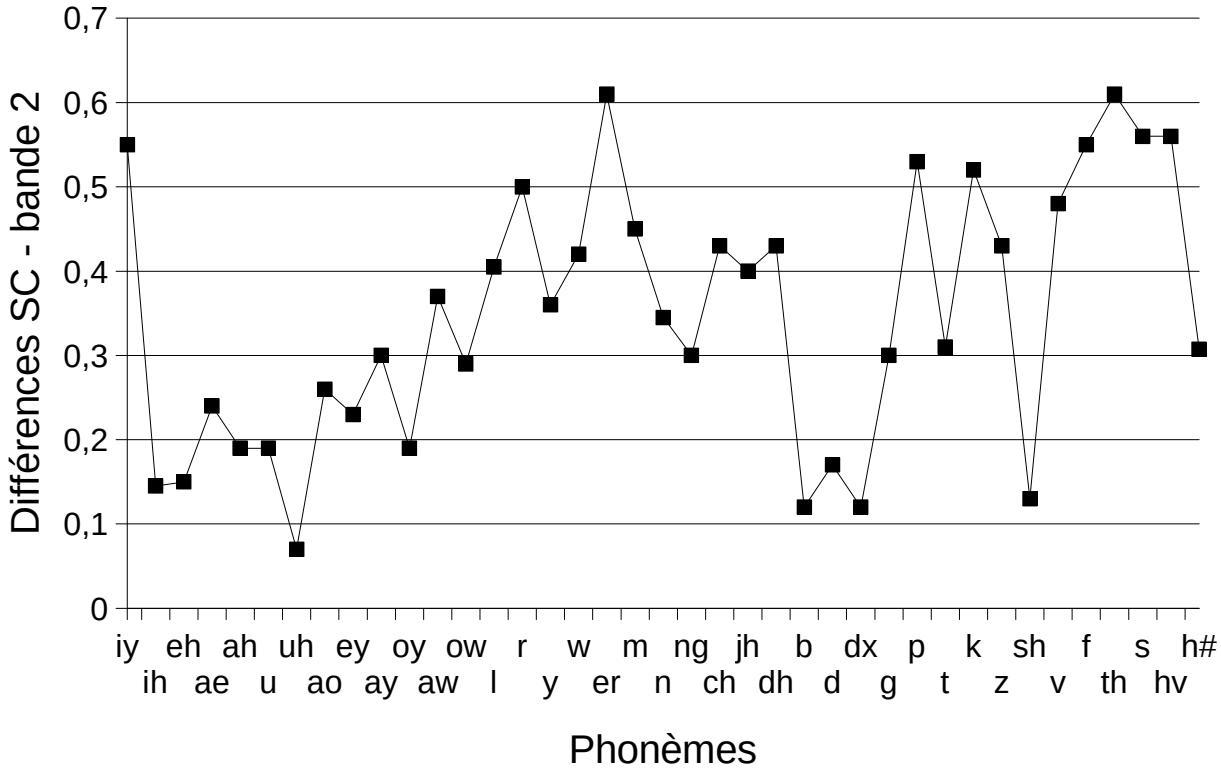

Figure B.1 : Différence des taux de reconnaissance entre le spectre complet et la deuxième sous-bande.

D'une manière générale, nous pouvons remarquer que cette bande reconnaît mieux les voyelles que la première sous-bande. Ceci est normal, car le premier formant des voyelles, excepté pour le /u/ et le /iy/, se situe directement dans cette deuxième bande. Toutefois, en ce qui concerne le /u/, son deuxième formant est contenu dans la bande, ce qui explique les bons taux de reconnaissance obtenus. Une seule voyelle fait exception à cette règle, le /iy/ de « b**ee**t ». Or, comme nous l'avons signalé pour ce phonème, le premier formant est en-dessous de 460 Hz tandis que le deuxième formant est au-dessus de 1000 Hz : il est donc logique qu'il soit mal reconnu.

Les autres phonèmes bien reconnus sont les suivants :

- La consonne /b/ ;
- La consonne /d/ ;
- La consonne /dx/ de "mu**dd**y" ;
- La consonne /sh/ de "**sh**e" ;

Comme précédemment, les plosives voisées possèdent des indices importants permettant de les distinguer des autres plosives dans les basses fréquences, ce qui pourrait bien expliquer la bonne reconnaissance de celles-ci dans la deuxième bande. Enfin, la sifflante /sh/ affecte généralement une grande région située dans les hautes fréquences, mais qui s'étend relativement bas, parfois à partir de 1000 Hz. Ceci peut donc expliquer que certains indices pour ce phonèmes se situent à la limite de la deuxième bande.

Les phonèmes les moins bien reconnus sont :

- La voyelle /iy/ de "b**ee**t" ;
- La consonne /r/ ;
- La consonne /er/ ;
- La consonne /p/ de "**p**ea" ;
- La consonne /k/ de "**k**ey" ;
- La consonne /v/ de "**v**an" ;
- La consonne /f/ de "**f**in" ;
- La consonne /th/ de "**th**in" ;
- La consonne /s/ ;
- La consonne /hv/ de "a**h**ead" ;

Par rapport à la première sous-bande, nous voyons que cette bande reconnaît mieux les diphtongues et les liquides, ce qui peut se comprendre car elle est située plus près du centre du spectre et elle est donc plus à même de déceler les mouvements fréquentiels des indices spectraux. Les fricatives aiguës sont toujours mal reconnues tout comme elles l'étaient dans la première sous-bande. Toutefois, nous voyons cette fois que les plosives non voisées comme /p/ ou /k/ sont moins bien reconnues dans cette bande que dans la première, les barres spectrales caractéristiques des bursts de ces plosives apparaissant moins clairement dans cet intervalle de fréquences que dans le premier.

✓ *Étude de la troisième sous-bande : [923 ... 2823 Hz]*

La différence des taux de reconnaissance entre le spectre complet et la troisième sous-bande est présentée sur la figure B.2.

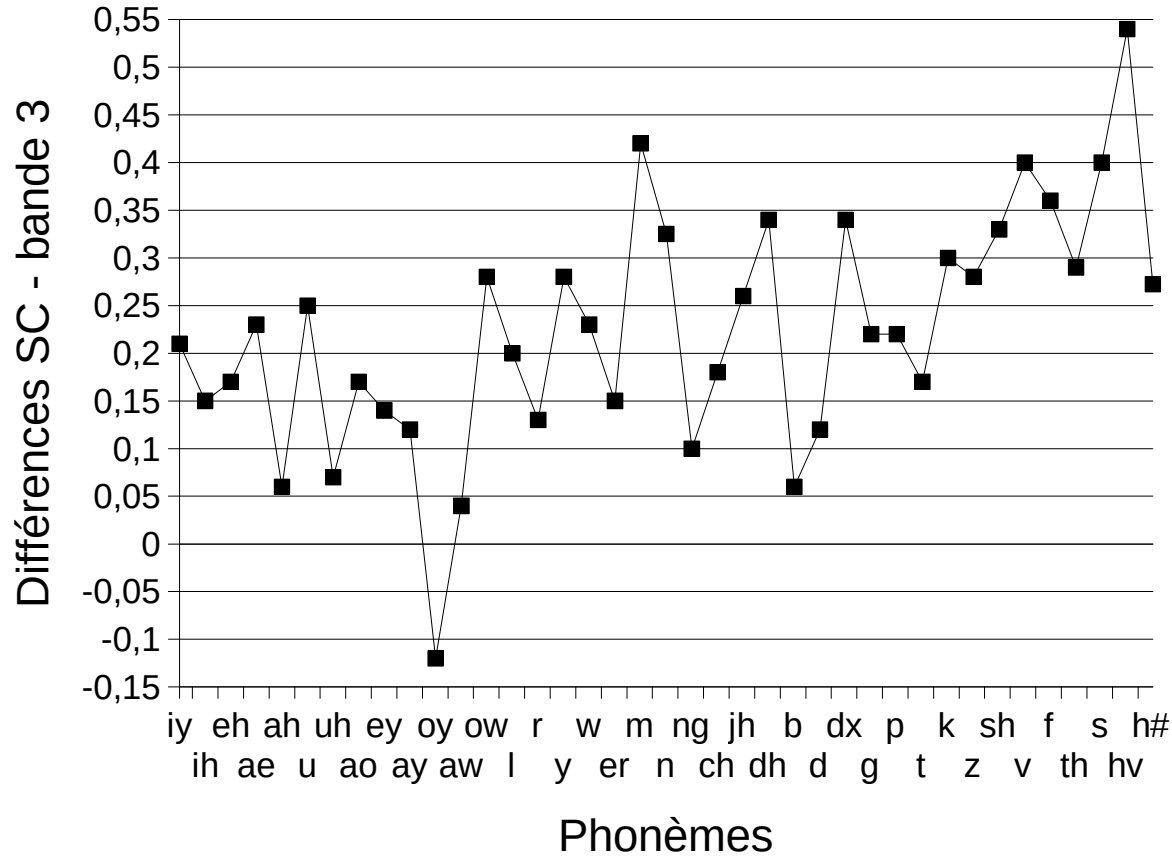

Figure B.2 : Différence des taux de reconnaissance entre le spectre complet et la troisième sous-bande.

Les phonèmes les mieux reconnus dans cette sous-bande sont les suivants :

- La voyelle /ah/ ;
- La voyelle /uh/ ;
- Mais surtout la diphtongue /oy/ de "b**oy**" ;
- La diphtongue /aw/ de "b**ou**t" ;
- La consonne /b/ ;

Nous voyons cette fois que les diphtongues sont bien adaptées à cette bande, toujours pour les mêmes raisons que celles indiquées pour la deuxième bande, à savoir la possibilité, en étant placé au milieu du spectre, de mieux détecter les changements de fréquences. La bonne reconnaissance du /uh/ provient certainement de l'identification du troisième formant de cette voyelle.

Les phonèmes les moins bien reconnus sont les suivants :

- La consonne /m/ de "**m**om" ;
- La consonne /dh/ de "**th**en" ;
- La consonne /dx/ ;
- La consonne /v/ ;
- La consonne /s/ ;

- La consonne /hv/ ;

La fricative /s/ n'est toujours pas correctement reconnue dans cette bande, ce qui signifie que probablement seules les hautes fréquences permettent de l'identifier. En revanche, cette fois-ci, les indices acoustiques semblent suffisants pour distinguer la plupart des plosives, à l'exception peut-être du /dx/. En ce qui concerne ces plosives, il faut également noter que plus nous faisons progresser les limites des sous-bandes vers les hautes fréquences, plus la bande devient « large ». Or, les plosives se caractérisant par une barre affectant toutes les fréquences, une telle barre se verra sans doute plus facilement lorsque la bande fréquentielle est large, ce qui pourrait également expliquer qu'elles soient mieux reconnues dans les hautes fréquences, au moins en ce qui concerne les plosives non voisées.

✓ ***Étude de la quatrième sous-bande : [2374 ... 7983 Hz]***

La différence des taux de reconnaissance entre le spectre complet et la quatrième sous-bande est présentée sur la figure B.3.

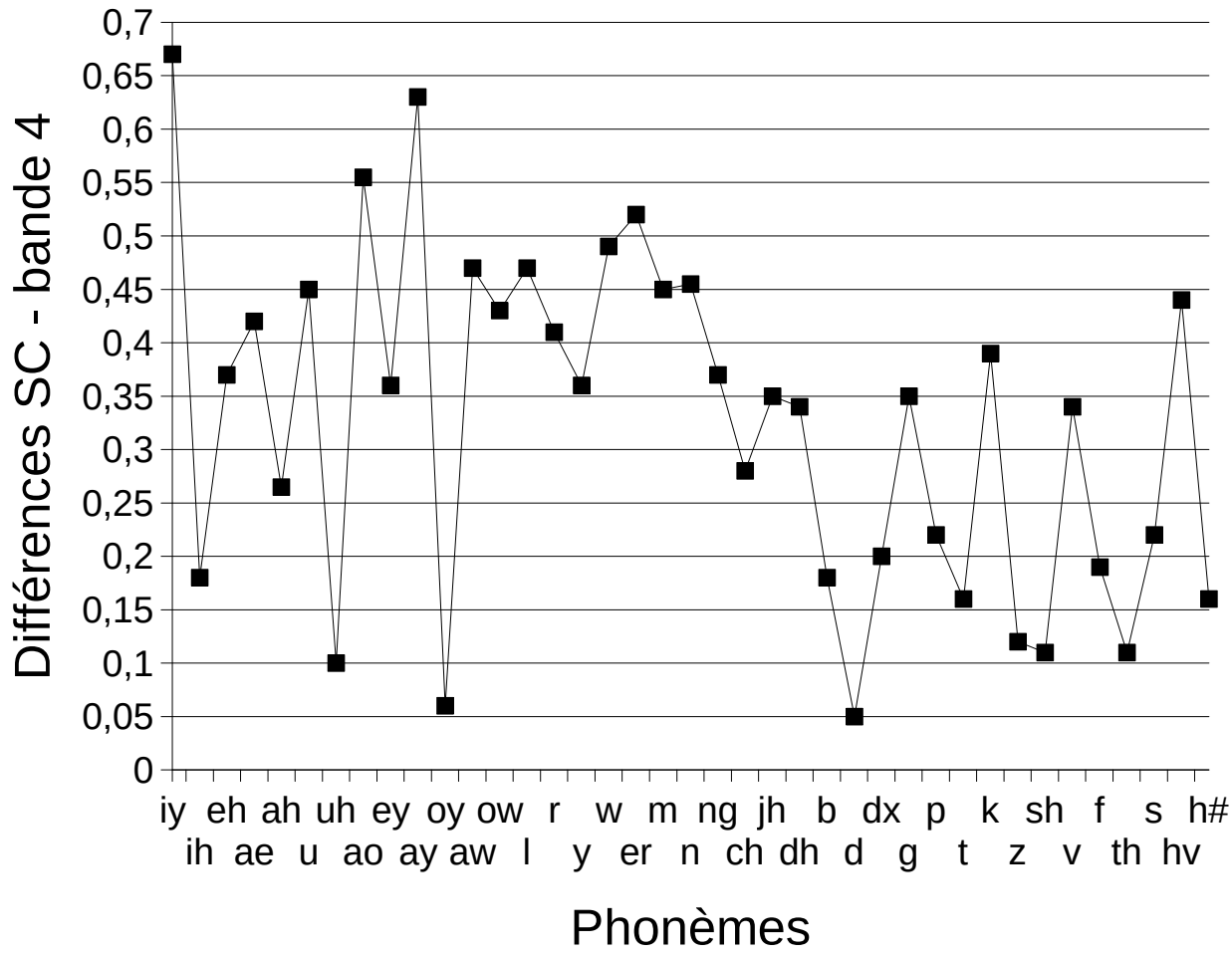

Figure B.3 : Différence des taux de reconnaissance entre le spectre complet et la quatrième sous-bande.

Les phonèmes les mieux reconnus dans cette bande sont les suivants :

- La voyelle /uh/ ;
- La diphtongue /oy/ ;
- La consonne /d/ ;
- La consonne /z/ ;
- La consonne /zh/ ;
- La consonne /th/ ;

Nous voyons cette fois apparaître essentiellement les sifflantes /z/ et /zh/, ce qui est parfaitement normal car leurs indices fréquentiels se situent essentiellement dans les aigus.

Les phonèmes les moins bien reconnus sont les suivants :

- La voyelle /iy/ ;
- La voyelle /aa/ ;
- La diphtongue /ay/ ;
- La consonne /er/ ;

Les sons voisés sont apparemment mal reconnus dans cette bande, ce qui semble plutôt compréhensible car les fréquences de résonance du conduit vocal n'apparaissent que très peu dans cette zone fréquentielle.

Annexe 4

Étude individuelle de la modification des classes dans chaque bande pendant l'apprentissage global

Cette étude poursuit celle menée dans la partie 7.6. Un apprentissage global, qui a été réalisé sur le système Multi-Bandes complet, a modifié les classes phonétiques modélisées par les HMM des sous-bandes. Nous tentons ici de donner quelques voies de recherches permettant de comprendre quelle est cette modification.

✓ ***Étude de la deuxième bande [461 ... 1000 Hz]***

Les modifications de la taille des phonèmes et de leur taux de reconnaissance pour la deuxième bande sont représentées sur la figure C.1.

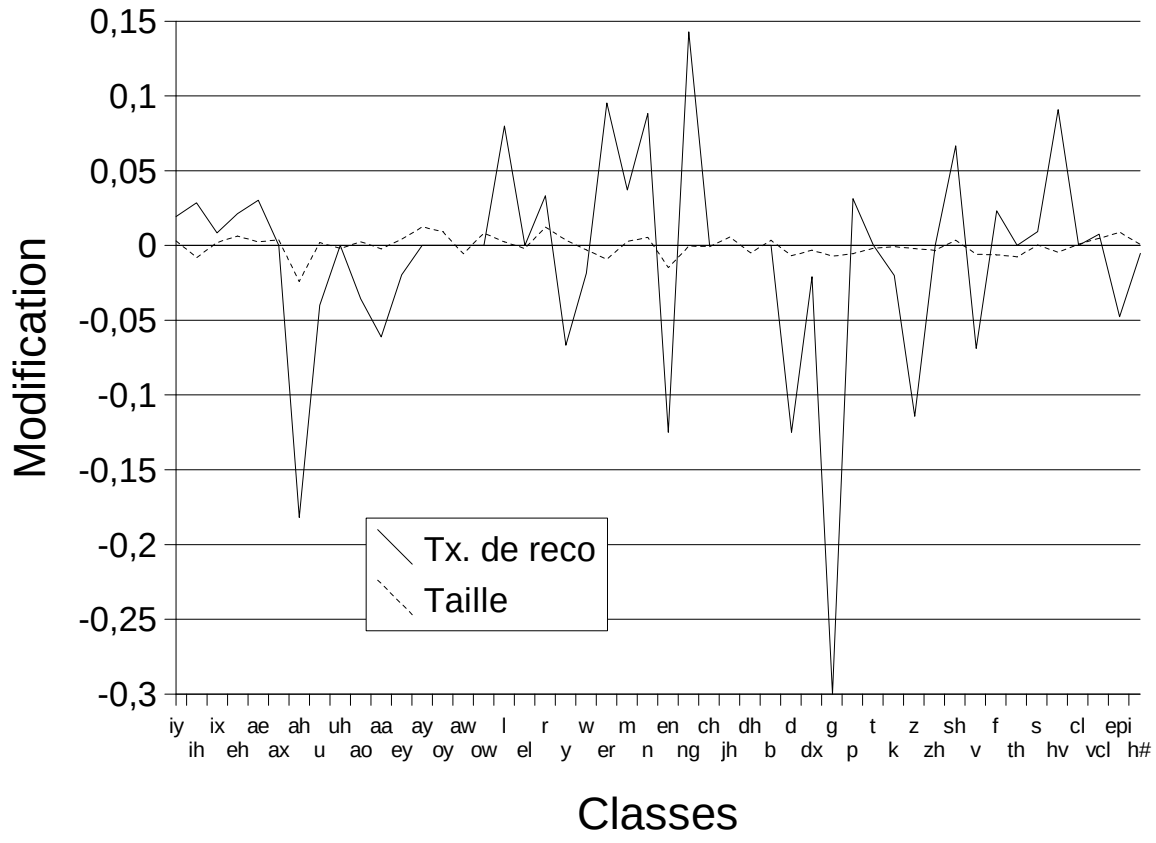

Figure C.1 : Modifications de la taille des phonèmes de la deuxième bande et de leur taux de reconnaissance après l'apprentissage global.

Nous pouvons tout d'abord remarquer que les modifications des tailles des classes sur cette bande (et les deux suivantes) sont beaucoup moins importantes que sur la première bande. Ceci peut signifier qu'aucune classe ne disparaît. Cependant, les fortes variations des taux de reconnaissance montrent qu'elles sont tout de même en pleine mutation, au moins en ce qui concerne leur position dans l'espace acoustique.

Nous allons donc nous intéresser essentiellement à la modification de leur taux de reconnaissance. Ainsi, les classes suivantes semblent relativement importantes pour le système, qui continue à les utiliser pour modéliser les phonèmes qui leur étaient initialement dédiés :

- /ng/, qui apparaît dans « sing »
- /er/, qui apparaît dans « bird »
- /hv/, qui apparaît dans « ahead »
- /n/, qui apparaît dans « noon »
- /l/, qui apparaît dans « lay »
- /sh/, qui apparaît dans « she »

Nous pouvons comprendre ce comportement en ce qui concerne les phonèmes voisés apparaissant dans cette liste, mais l'explication semble plus difficile en ce qui concerne /hv/ et /sh/. Une explication possible vient du fait que le système doit choisir pour chaque phonème au moins une bande dont le rôle est de le modéliser. Il se peut ainsi que certains indices existent dans cette bande qui permettent de mieux distinguer par exemple le /hv/ de ses plus proches concurrents.

Inversement, les phonèmes suivants voient leur taux de reconnaissance diminuer fortement :

- /g/, qui apparaît dans « gay »
- /ah/, qui apparaît dans « but »
- /en/, qui apparaît dans « button »
- /d/, qui apparaît dans « day »
- /z/, qui apparaît dans « zone »

✓ *Étude de la troisième bande [923 ... 2823 Hz]*

Les modifications de la taille des phonèmes et de leur taux de reconnaissance pour la troisième bande sont représentées sur la figure C.2.

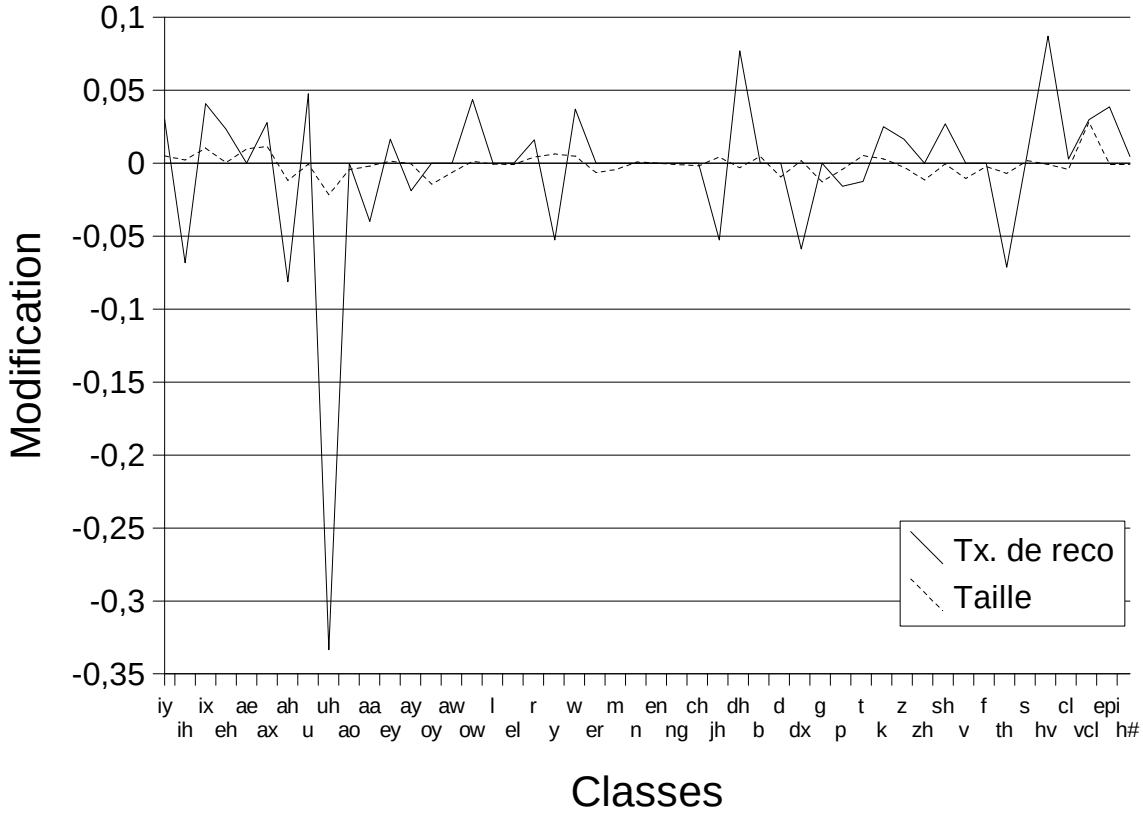

Figure C.2 : Modifications de la taille des phonèmes de la troisième bande et de leur taux de reconnaissance après l'apprentissage global.

Encore une fois, la taille des classes varie très peu. En ce qui concerne l'augmentation du taux de reconnaissance, les phonèmes suivants sont les plus performants :

- /hv/, qui apparaît dans « ahead »
- /dh/, qui apparaît dans « then »
- /u/, qui apparaît dans « boot »

Le son /u/ peut paraître surprenant ici, mais nous pouvons remarquer sur la courbe qu'il est entouré de deux phonèmes dont les taux de reconnaissance ont considérablement chuté au cours de l'apprentissage global. Nous voyons donc apparaître ici un nouveau phénomène, que nous expliquons par le fait que le système veut non seulement reconnaître au mieux certains phonèmes grâce à leurs indices acoustiques présents dans la bande, mais augmente également le pouvoir discriminant des bandes pour les phonèmes qui peuvent être facilement confondus. Ainsi, les phonèmes /ah/, /u/ et /uh/, qui sont relativement proches d'un point de vue acoustique, sont reconnus par des bandes différentes, de façon à ce que ces bandes ne possèdent pas de modèles qui puissent être facilement confondus. Ceci se traduit par une courbe en « dents de scie » que l'on retrouve dans toutes les bandes.

✓ ***Étude de la quatrième bande [2374 ... 7983 Hz]***

Les modifications de la taille des phonèmes et de leur taux de reconnaissance pour la quatrième bande sont représentées sur la figure C.3.

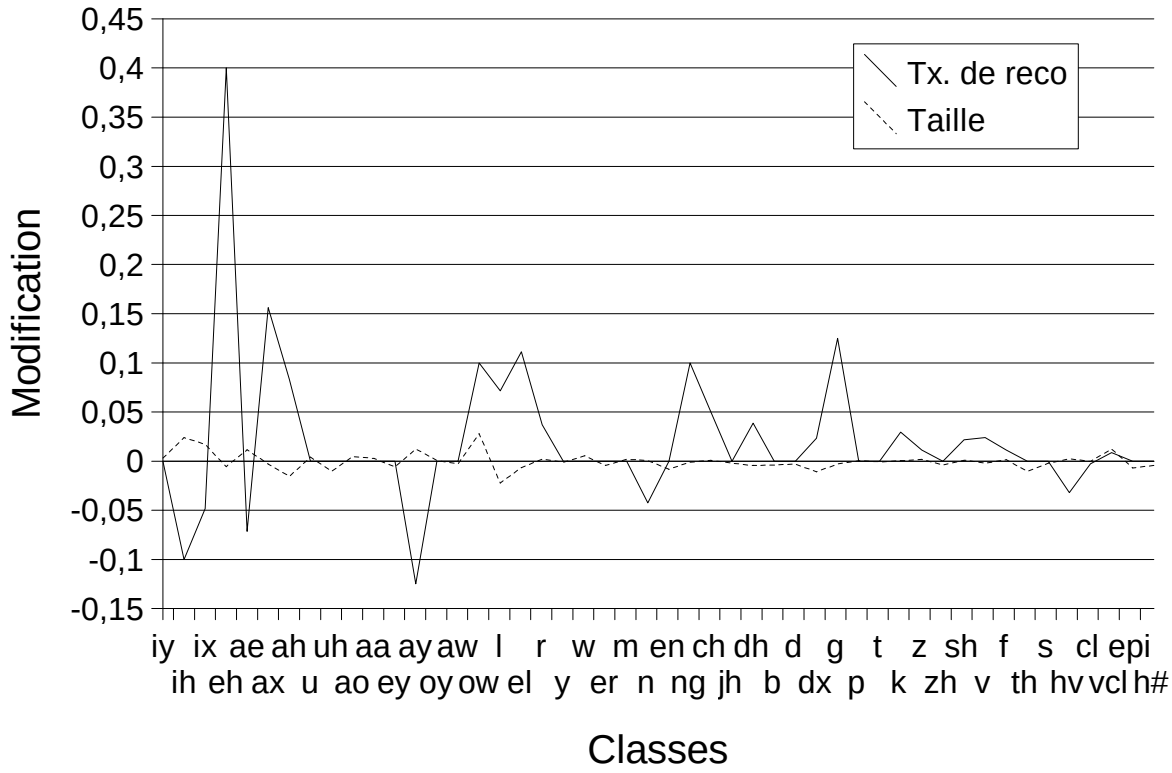

Figure C.3 : Modifications de la taille des phonèmes de la quatrième bande et de leur taux de reconnaissance après l'apprentissage global.

Les phonèmes qui augmentent leur taux de reconnaissance au cours de cet apprentissage global sont :

- /eh/, qui apparaît dans « bet »
- /ax/, qui apparaît dans « about »
- /g/, qui apparaît dans « gay »
- /el/, qui apparaît dans « bottle »

Inversement, les phonèmes dont le taux de reconnaissance diminue sont :

- /hv/, qui apparaît dans « ahead »
- /n/, qui apparaît dans « noon »
- /ix/, qui apparaît dans « debit »
- /ae/, qui apparaît dans « nat »
- /ih/, qui apparaît dans « bit »
- /ay/, qui apparaît dans « bite »

Encore une fois, les changements qui semblent prédominer dans cette bande sont bien plus liés à des notions stratégiques de répartition des classes proches entre plusieurs bandes qu'à des considérations phonétiques. La forme de la courbe en dents de scie illustre ce phénomène.

✓ *Étude de la cinquième bande (spectre complet)*

Les modifications de la taille des phonèmes et de leur taux de reconnaissance pour la cinquième bande sont représentées sur la figure C.4.

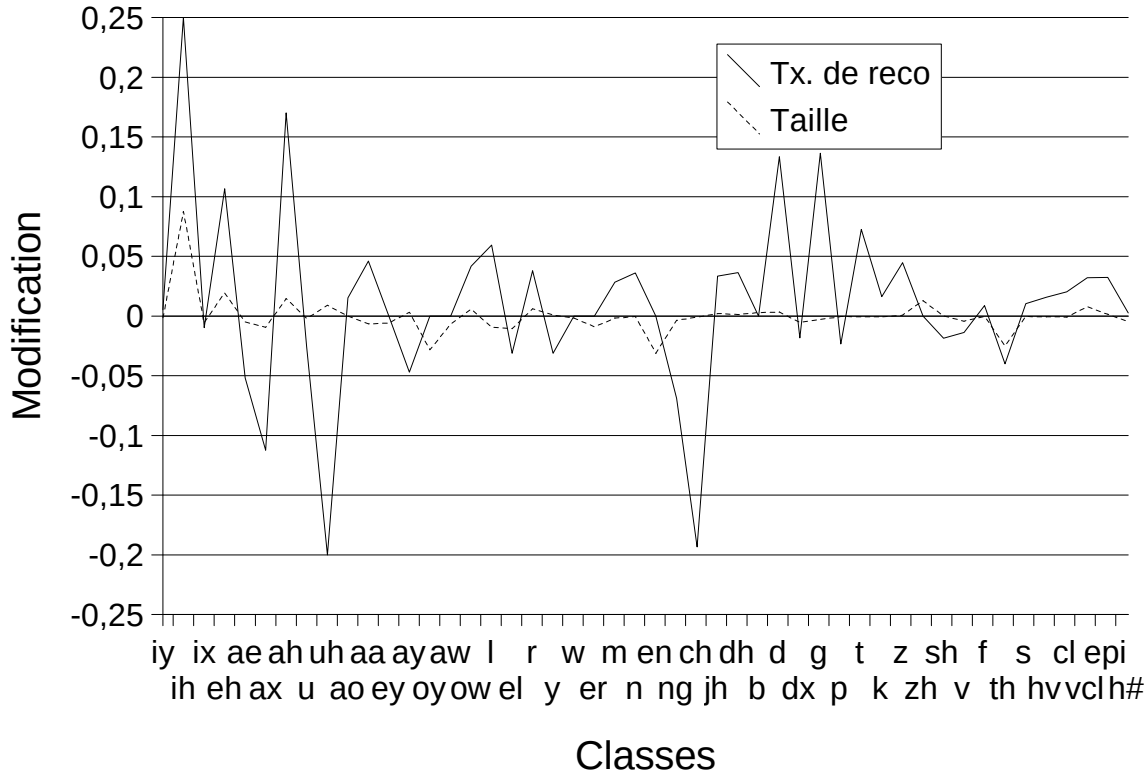

Figure C.4 : Modifications de la taille des phonèmes de la cinquième bande et de leur taux de reconnaissance après l'apprentissage global.

Le comportement des classes dans cette bande est assez similaire à celui des bandes précédentes, mis à part le fait que la taille de la classe modélisant /ih/ (bit) augmente beaucoup, par rapport à son évolution pour les autres phonèmes. Cette augmentation étant accompagnée d'une croissance du taux de reconnaissance, nous pouvons en conclure que ce phonème est spécialement important pour cette bande. Mis à part cela, nous retrouvons le même phénomène que précédemment, à savoir une répartition des classes proches entre plusieurs bandes. Sinon, les phonèmes dont le taux de reconnaissance progresse le plus sont :

- /ih/
- /ah/, qui apparaît dans « but »
- /g/, qui apparaît dans « gay »
- /d/, qui apparaît dans « day »
- /eh/, qui apparaît dans « bet »

Inversement, les phonèmes dont le taux de reconnaissance régresse le plus sont :

- /ax/, qui apparaît dans « about »
- /ch/, qui apparaît dans « choke »
- /uh/, qui apparaît dans « book »