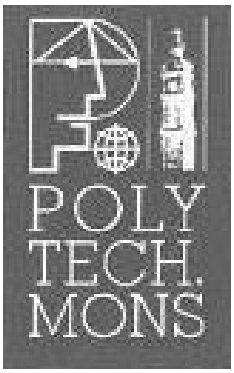
Faculté Polytechnique de Mons



Laboratoire TCTS, Mons

Etude et développement d'architectures multi-bandes et multi-modales pour la reconnaissance robuste de la parole

Dissertation originale présentée pour l'obtention du grade de Docteur en Sciences Appliquées par

Stéphane Dupont

Juin 2000

Membres du jury :

Professeur J. TRECAT, FPMs, Président

Professeur H. BOURLARD, EPFL, IDIAP
Professeur F. GRENEZ, ULB

Dr. B. GOSSELIN, FPMs
Professeur C. CONTI, FPMs, Doyen
Professeur H. LEICH, FPMs, Promoteur

Résumé

Le niveau de maturité actuel des technologies de reconnaissance vocale semble toujours insuffisant pour permettre leur application à une large gamme de produits et de conditions d'utilisation. Les principales limitations apparaissent pour de la parole spontanée et dans des conditions de bruit additif, où l'on observe des taux d'erreurs plus de dix fois supérieurs à ceux des humains. Ces faibles performances résultent de l'inadéquation des modèles utilisés aux conditions de terrain, les paramètres de ces modèles étant estimés sur base de corpus de parole enregistrés dans des conditions de laboratoire. Ce travail s'oriente vers l'étude, le développement et l'évaluation de techniques permettant d'améliorer la reconnaissance vocale sur de la parole perturbée par du bruit additif indépendant et dont les propriétés ne sont pas connues a priori.

Dans un premier temps, ce rapport présente les technologies de base intervenant dans la reconnaissance automatique de la parole. L'analyse acoustique par traitement du signal, l'utilisation de corpus d'apprentissage, la modélisation statistique par modèles de Markov cachés (HMMs), les réseaux de neurones artificiels, et les grammaires stochastiques sont abordés d'un point de vue théorique. Ensuite, un aperçu bibliographique des différentes techniques de reconnaissance vocale en milieu bruité est proposé. L'influence du bruit et l'intérêt pratique de certaines des techniques présentées sont également étudiés.

Les domaines étudiés plus particulièrement concernent d'abord des méthodes robustes au bruit appliquées à un signal acoustique unique. Une approche originale faisant intervenir une décomposition en bandes de fréquence (stratégie multi-bande) est développée. Elle consiste à mener une partie du processus de reconnaissance vocale indépendamment dans plusieurs bandes de fréquences. On combine ensuite les résultats des différentes bandes sur base de formalismes statistiques ou de stratégies heuristiques. Cette approche est essentiellement motivée par la redondance spectrale de l'information acoustico-phonétique. De nombreuses méthodes d'analyse acoustique et de combinaison des résultats sont envisagées et comparées.

L'étude porte aussi sur des techniques faisant intervenir des signaux provenant de sources d'information alternatives (stratégies multi-modales), plus particulièrement le mouvement des lèvres. Le processus appelé "lecture labiale" est bien connu pour augmenter l'intelligibilité de la parole, surtout en milieu bruité. Les aspects visuels et acoustiques sont ici intégrés dans un formalisme de reconnaissance automatique de la parole.

Un des points communs entre les stratégies de décomposition en bandes de fréquence et multi-modale réside dans l'intervention de multiples canaux d'information. Un formalisme général est alors introduit. Celui-ci, appelé "multi-stream", permet la coopération de plusieurs HMMs traitant les différents canaux. Cette approche originale permet en outre l'indépendance des régimes stationnaires des différents canaux. Ceux-ci peuvent s'aligner sur des portions temporelles différentes et donc se désynchroniser les uns par rapport aux autres, la transition d'un état (régime stationnaire) au suivant ne se produisant pas nécessairement au même instant pour les différents canaux. Nos travaux ont conduit au développement d'algorithmes de reconnaissance optimaux et de stratégies d'élagage permettant de réduire la complexité du processus de reconnaissance.

L'évaluation expérimentale de l'approche de décomposition en bandes de fréquence est

réalisée sur un corpus de nombres connectés ainsi que sur le corpus de gestion de ressources navales "Resource Management", tous deux en anglais et multi-locuteurs. Elle donne donc une indication des performances pour des tâches de complexité faible à moyenne. Notons cependant que même ces tâches simples n'ont pas encore trouvé de solution suffisamment efficace dans des situations où le bruit additif est important. Les protocoles expérimentaux font appel à divers types de bruits, artificiels et réels, stationnaires et non stationnaires. Les bruits sont ajoutés au signal de parole de façon à couvrir une plage raisonnable de rapports signal/bruit (jusqu'à 5 dB), étant donné les mesures que l'on peut trouver dans la littérature. Les techniques classiques de soustraction spectrale et de filtrage log-RASTA ou J-RASTA sont utilisées comme référence ou en complément aux méthodes étudiées. Les résultats indiquent une robustesse accrue dans le cas de bruits fortement colorés. Dans le cas de bruits large-bande cependant, l'apport de la décomposition en bandes de fréquence est négligeable. Une nouvelle stratégie est alors introduite. Elle est basée sur l'observation que, si l'on considère une bande de fréquence relativement étroite, les bruits ne différent essentiellement que par leur niveau. De ce fait, des modèles associés à chacune des bandes de fréquence du système peuvent être entraînés après contamination du corpus d'apprentissage par un bruit quelconque; ces modèles demeurent relativement insensibles à d'autres types de bruits. Par rapport aux systèmes de référence, on observe globalement une réduction du taux d'erreur proche de 30%.

L'étude concernant les stratégies multi-modales conduit à la présentation d'un système comprenant trois composantes: un module d'analyse visuelle, un module d'analyse acoustique, et un module d'intégration et de reconnaissance. Le module d'analyse visuelle (fournit par l'IDIAP) suit le mouvement des lèvres du locuteur et extrait des paramètres représentant leur contour et leur luminosité. Le module d'analyse acoustique extrait des paramètres représentatifs du signal audio. Finalement, le module d'intégration et de reconnaissance est responsable de la modélisation temporelle des deux sources d'information. Il utilise notamment l'approche "multi-stream" qui permet la modélisation de l'asynchronisme des deux sources. L'approche est évaluée sur une tâche de reconnaissance multi-locuteurs de chiffres connectés en français et pour deux types de bruits: un bruit blanc stationnaire et un bruit réel non stationnaire. Nos résultats expérimentaux ne justifient pas l'intérêt de l'approche "multi-stream". Ils confirment cependant l'intérêt de la stratégie multi-modale sur base d'un grand corpus multi-locuteurs de parole continue. Dans certaines conditions de bruit, cette stratégie conduit à une réduction du taux d'erreur supérieure à 50% par rapport à un système purement acoustique.

D'autre part, le problème de l'estimation du spectre de bruit est également introduit et une méthode originale est proposée et utilisée dans le cadre des stratégies multi-bande et multi-modale ainsi que dans les systèmes de référence.

Mots-clés: reconnaissance vocale en milieu bruité, bandes de fréquence, reconnaissance vocale audiovisuelle, débruitage de la parole, modèles de Markov cachés coopératifs, modèles multi-stream.

Remerciements

Je tiens tout d'abord à adresser mes remerciements à Monsieur le Professeur Henri Leich, pour m'avoir accordé sa confiance et permis de réaliser dans son service les recherches qui ont conduit à cette thèse. Je souhaite également remercier les membres du jury, Messieurs Jacques Trecat, Hervé Bourlard, Calogero Conti, Bernard Gosselin et Francis Grenez, pour s'être intéressés à ce travail et pour avoir accepté de consacrer une partie de leur temps à son évaluation.

Je remercie le Fonds National pour la Recherche dans l'Industrie et dans l'Agriculture, principale source de financement de ces travaux.

Les résultats présentés ici doivent beaucoup à l'aide de nombreuses personnes du laboratoire TCTS, plus particulièrement celles du groupe de reconnaissance vocale: Jean-Marc Boite, Leila Cheboub, Laurent Couvreur, Olivier Deroo, Geoffrey Durou, Vincent Fontaine, Christophe Ris et Laurent Zanoni. Christophe, Jean-Marc et Olivier, le noyau dur, ont particulièrement marqué ces cinq années par leur disponibilité, leurs conseils judicieux, leurs critiques, leur rigueur constructive,... et les kilomètres de code informatique qui sont à la base d'une majorité des expérimentations que j'ai réalisées. Leur relecture de ce rapport de thèse a également contribué à en améliorer la qualité. Je les en remercie. Je dois également beaucoup aux qualités scientifiques de partenaires étrangers. Je pense notamment aux personnes travaillant à l'International Computer Science Institute, en Californie.

Je suis particulièrement reconnaissant à Monsieur Hervé Bourlard pour m'avoir orienté dans mes travaux. Il m'a également donné la possibilité de séjourner quelques temps à l'IDIAP, en Suisse. Je souhaite souligner l'importance des collaborations qui y sont nées, particulièrement avec Monsieur Juergen Luettin.

Les méthodes développées ici ont conduit à un dépôt de brevet. Je voudrais témoigner ma gratitude à Messieurs Yves Dehon, Thierry Dutoit et Olivier Van Der Vrecken pour leur support à ce sujet.

Merci finalement à ma famille, et en particulier à ma mère, pour sa participation à la relecture de pans importants de ce rapport.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Les technologies vocales et de traitement du langage naturel ont pour but de faciliter l'interaction entre l'homme et la machine. Elles visent également à faciliter l'encodage, l'indexation et la recherche "intelligente" (basée sur le contenu) d'informations dans des bases de données à caractère vocal ou audiovisuel ou à partir de requêtes vocales. L'objectif de la technologie de reconnaissance vocale proprement dite consiste généralement à obtenir une transcription orthographique d'une séquence de parole préalablement enregistrée. Seul cet aspect sera traité dans cette thèse. Notons cependant que bien souvent, il ne s'agit là que d'un maillon, certes fondamental, dans une chaîne de traitement plus complexe. Ainsi, des modules de traitement du langage naturel permettront d'apporter une interprétation sémantique au texte obtenu en vue de formuler une réponse sensée ou bien d'indexer le document vocal sur base de classes conceptuelles adaptées à la tâche considéré. Dans le domaine médical par exemple, on peut souhaiter organiser des rapports médicaux dictés sur base de concepts médicaux appartenant à une nomenclature bien définie. Dans le domaine des journaux radiodiffusés, on souhaitera classer les enregistrements sur base des thèmes abordés.

Tout système de reconnaissance vocale repose sur la définition d'unités lexicales élémentaires, généralement des phonèmes. Un dictionnaire permet ensuite de représenter les mots à partir de ces unités. Finalement, une grammaire décrit la façon dont peuvent être organisés les mots pour former des phrases. Le système opère sur base d'un signal de parole échantillonné. Un module d'analyse faisant appel à des routines de traitement du signal fournit une séquence de vecteurs de petite dimension. Chaque vecteur représente une portion de courte durée (quelques dizaines de millisecondes typiquement) pendant laquelle le signal est supposé stationnaire. Ces vecteurs sont ensuite utilisés par un module de classification statistique dont les paramètres ont été estimés sur base de corpus de parole de quelques heures. Ce module permet d'associer des probabilités aux différentes unités lexicales. Ces probabilités sont ensuite utilisées par un algorithme de décodage générant la ou les phrases les plus probables étant donnés le dictionnaire et la grammaire.

Cette tâche n'est pas simple. Les problèmes sont de plusieurs ordres. Citons en quelques uns. Tout d'abord, deux prononciations différentes de la même phrase ne sont jamais identiques. Si les différences sont mineures lorsqu'il s'agit de la même per-

sonne, elles peuvent être très importantes si l'on considère deux personnes différentes. Ces variabilités intra-locuteur et inter-locuteur sont à la source de l'utilisation de modèles statistiques dans le cadre de la reconnaissance vocale. Les microphones, les interfaces d'acquisition ainsi que les conditions d'enregistrement sont d'autres sources de variabilités importantes. La nature même du processus de phonation introduit une autre source de variabilités appelée coarticulation. Ce phénomène correspond à la modification du contenu spectral d'un son par les sons voisins. Il est dû à l'inertie de l'appareil vocal. Finalement, la phase de décodage est très coûteuse en temps de calcul et impose souvent l'utilisation d'approches sous-optimales au sens statistique.

Bien que ce domaine de recherche soit actif depuis plus de 30 ans, le niveau de maturité actuel semble toujours insuffisant pour permettre l'application de cette technologie à une large gamme de produits et de situations. Dans des conditions d'enregistrement claires, les taux d'erreur des systèmes automatiques sont bien souvent jusqu'à dix fois supérieurs à ceux des humains. Dans des conditions de bruits et de distorsions, l'écart se creuse encore. Une direction de recherche importante concerne donc la *robustesse* des systèmes de reconnaissance aux variations des conditions de bruit et du canal de transmission. Les taux d'erreur importants obtenus pour la *parole spontanée* constituent une faiblesse supplémentaire. Ces points font bien entendu l'objet de recherches importantes comme en témoignent les articles publiés ainsi que les sessions spéciales qui leurs sont consacrées lors de conférences internationales. Le problème de la robustesse aux bruits environnementaux sera envisagé plus en détail dans la suite de cette thèse. Nous y étudierons notamment une approche de reconnaissance originale faisant intervenir une décomposition en bandes de fréquence. Nous avons également travaillé sur la reconnaissance audiovisuelle de la parole: il s'agit d'utiliser le mouvement des lèvres comme source d'information additionnelle au signal acoustique. Nous confirmerons des résultats récents indiquant que cette stratégie conduit à une robustesse nettement accrue. Un formalisme commun pour la reconnaissance sur base de canaux d'information distincts est aussi introduit, le "multi-stream". Nous nous sommes également intéressé à l'estimation automatique du niveau de bruit, problème important dans le cadre de plusieurs méthodes de reconnaissance vocale robuste ainsi que des méthodes originales étudiées dans cette thèse.

Ce document est organisé comme suit. Après une brève description des application de la reconnaissance vocale et de ses domaines connexes, le Chapitre 2 fournit une présentation des technologies de base. Les techniques d'analyse acoustique, les modèles de Markov cachés (HMM - Hidden Markov Models) ainsi que les méthodes de modélisation du langage sont ainsi abordées de manière synthétique. Finalement, ce chapitre rappelle les limitations de la technologie actuelle. Comme déjà signalé, la robustesse aux bruits ambiants reste ainsi un domaine de recherche important. Ce problème fait l'objet du Chapitre 3. Les difficultés posées par le bruit sont d'abord rappelées. Ensuite, les techniques classiques de débruitage, d'adaptation et de reconnaissance partielle sont traitées brièvement. Notons tout de suite que ces différentes techniques ne sont en général pas exclusives et peuvent être combinées, dans la mesure où elles sont complémentaires. Finalement, des résultats de reconnaissance illustrant l'influence d'un bruit additif seront présentés. Ces résultats visent également à

montrer l'intérêt de quelques unes des méthodes de reconnaissance robustes traitées dans ce chapitre. Seules les méthodes testées seront utilisées par la suite comme stratégies robustes de référence. Nous avons cependant trouvé utile de les placer dans un cadre général évoquant de façon relativement complète les différentes approches décrites dans la littérature.

Les approches classiques de reconnaissance vocale contiennent un classificateur statistique. Celui-ci opère soit par modélisation statistique de classes phonétiques (ou plus généralement d'unités lexicales élémentaires), soit par estimation des probabilités a posteriori de ces classes. Dans les deux cas, les paramètres des modèles utilisés sont estimés sur base d'un corpus d'apprentissage comprenant des séquences de parole représentatives de la tâche de reconnaissance qui nous intéresse. L'Annexe A présente les corpus qui ont été utilisés dans le cadre de cette thèse. On y trouvera aussi la description de corpus contenant des enregistrements de bruits divers. Ces bruits ont été ajoutés artificiellement aux signaux de parole pour évaluer la robustesse des méthodes développées au bruit additif. Finalement, nous avons également utilisé une base de donnée audiovisuelle. Ce corpus contient des séquences vidéo du visage du locuteur, synchronisées avec le signal audio.

Les méthodes de reconnaissance robuste et de débruitage présentées au Chapitre 3 nécessitent souvent une estimation du spectre de bruit. Une des méthodes choisies comme stratégies de référence y fait notamment appel. Nous aborderons ce point fondamental au Chapitre 4. Nous y présenterons notamment une méthode originale tirant parti du caractère périodique des voyelles. Cette méthode est finalement appliquée au problème de la reconnaissance robuste de la parole. Ces techniques d'estimation du spectre de bruit seront également appliquées ultérieurement: (1) dans le cadre d'une approche originale de décomposition en bandes de fréquence, en vue d'estimer le rapport signal/bruit dans chacune des bandes de fréquence et (2) dans le cadre de la reconnaissance audiovisuelle de la parole, pour l'estimation du rapport signal/bruit global du signal acoustique.

Au Chapitre 5, nous abordons une méthode de reconnaissance vocale originale appelée "multi-stream". Elle consiste à diviser l'information disponible en plusieurs canaux. Chaque canal est alors traité indépendamment des autres sur base de son propre modèle de Markov caché. Les contributions de ces différents HMM sont ensuite combinées de façon à imposer un certain niveau de synchronisme aux différents canaux. Alors que les HMMs peuvent essentiellement être utilisés pour modéliser un seul processus, ou plusieurs processus dépendants, les modèles "multi-stream" pourront être utilisés pour des processus qui évoluent indépendamment au sein d'unités linguistiques pré-définies. La suite de cette thèse visera notamment à évaluer l'intérêt de la méthode dans le cadre de notre approche originale de décomposition en bandes de fréquence et dans le cadre de la reconnaissance audiovisuelle de la parole.

Au Chapitre 6, nous nous intéresserons à l'approche de reconnaissance faisant intervenir une décomposition en bandes de fréquence, ou multi-bande. Son principe est le suivant:

- Analyse, estimation de probabilités phonétiques et éventuellement catégorisation phonétique dans différentes bandes de fréquence, sur base de "sous-reconnaisseurs" dont l'implémentation repose sur des méthodes classiques.
- C ombinaison des résultats d'analyse, d'estimation ou de catégorisation corres-

pondant aux différentes bandes de fréquence et intégration sur des intervalles de temps plus larges.

Ce chapitre présentera d'abord les motivations de cette approche: études psycho-acoustiques et robustesse au bruit notamment. Un état de l'art des travaux touchant au domaine concerné sera également présenté. Nous nous attacherons ensuite à rapporter nos derniers résultats de recherche. Ceux-ci portent sur les paramètres représentatifs des différentes bandes de fréquence et sur les méthodes de combinaison des résultats provenant de l'utilisation indépendante des différentes bandes de fréquence. Les résultats expérimentaux concernent la reconnaissance de nombres connectés en anglais. Nous introduirons également l'utilisation de l'approche "multi-stream" comme stratégie de reconnaissance dans le cadre de l'approche multi-bande. Ce chapitre contient notamment un compte rendu relativement exhaustif concernant les méthodes d'ensemble. Celles-ci visent à combiner plusieurs classificateurs ou plusieurs estimateurs de probabilité dans le but d'améliorer les capacités de classification ou de généralisation de l'ensemble. Elles sont donc d'un intérêt particulier dans le cadre de la décomposition en bandes de fréquence.

Au Chapitre 7, nous décrirons un système complet pour la reconnaissance vocale audiovisuelle. Il s'agit d'un système qui utilise conjointement l'information acoustique et l'information concernant le contour et la luminosité des lèvres du locuteur. Ce système fera notamment appel à l'approche "multi-stream" ainsi qu'aux techniques d'estimation du niveau de bruit étudiées au Chapitre 4. Par rapport à un système de reconnaissance acoustique, même si l'on utilise des paramètres acoustiques robustes, le système audiovisuel conduira à une réduction importante du taux d'erreur qui peut être divisé par deux en présence de bruit additif.

Finalement, le dernier chapitre est dédié à une nouvelle méthode de reconnaissance robuste. Celle-ci est basée sur une architecture multi-bande dont l'apprentissage est réalisé sur de la parole bruitée. De nombreux résultats expérimentaux seront rapportés dans ce chapitre. Ils concernent la reconnaissance de nombres connectés et la tâche de gestion de ressources navales "Resource Management", plus compliquée. Divers bruits artificiels et réels, stationnaires et non-sationnaires, ont été utilisés. Les résultats indiquent une réduction du taux d'erreur de reconnaissance de l'ordre de 30% par rapport à des stratégies de reconnaissance robustes faisant partie de l'état de l'art.

Chapitre 2

Reconnaissance automatique de la parole

2.1 Introduction

Le but de ce chapitre est d'illustrer l'intérêt de la technologie de la reconnaissance de la parole en présentant certaines de ses applications. Il est aussi de présenter l'état de la technologie et de rappeler l'intérêt de ce travail de recherche en mettant l'accent sur les limitations actuelles.

De façon générale, le problème de la reconnaissance de la parole consiste à extraire la transcription orthographique d'un signal de parole. Le signal est échantillonné à 8 kHz dans le cas de lignes téléphoniques ou plus dans le cas de saisie directe par microphone.

Une bonne technologie de reconnaissance automatique de la parole permettrait aux humains d'interagir de façon plus naturelle avec les machines. Chacun peut facilement imaginer de nouvelles applications de ce type de technologie. Il serait par exemple possible de contrôler des ordinateurs ou des équipements de communication par la voix, sous forme de mots isolés ou de séquences de mots. Dans une application interactive utilisant la reconnaissance vocale, le système de reconnaissance constitue seulement une partie de l'application. Généralement, le système complet fait appel à un module de dialogue (avec l'intervention d'un synthétiseur de parole) et de contrôle (grâce à des interfaces avec le monde extérieur) qui effectue l'action demandée par l'utilisateur. Pour pouvoir être utilisée efficacement, une telle application doit avoir une ergonomie bien étudiée, elle doit être fiable, facile à utiliser et il est impératif qu'elle réponde en temps réel.

La technologie de reconnaissance vocale permet également d'envisager l'encodage et l'indexation de grandes quantités de données vocales ou audiovisuelles. Il s'agit ici d'applications non interactives qui peuvent notamment intéresser les chaînes de télévision et de radiodiffusion. Cette forme d'encodage peut également faciliter la création de bases de données et de statistiques dans le domaine médical: les rapports des médecins ainsi que les interventions médicales pourraient faire l'objet de ce type

de traitement.

Ces taches ne sont cependant pas simples, notamment à cause de nombreux facteurs humains, du caractère hautement variable des sons constituant la parole ainsi que des perturbations possibles dues à l'environnement. A ce titre, l'oreille humaine est un outil très performant pouvant s'adapter presque instantanément à tous les types de locuteurs (même dans le cas d'accents très prononcés) et à tous les types d'environnements.

2.1.1 Applications

Concentrons nous sur les applications de la reconnaissance de la parole en télécommunications. La technologie vocale réduit les coûts et permet de fournir des services 24h/24h en remplaçant un opérateur humain par un système de reconnaissance associé à un système de synthèse de la parole. D'autre part, la technologie disponible permet d'envisager de nouveaux services. Citons quelques applications typiques:

- Automatisation des services de renseignement téléphoniques.
- Composition vocale du numéro d'appel. Cet aspect est intéressant lorsqu'il est préférable d'avoir les mains libres (pour conduire une voiture par exemple).
- Remplacement des touches pour des services basés sur des menus, ou accès par la voix pour les téléphones sans touches.
- Opérations bancaires à domicile ("Voice banking"),
- Service d'accès facile à des informations (cours des actions de bourse, horaires des transports en commun, prévisions météorologiques...) et bases de données,
- Services de réservation ou d'achat par téléphone.

Parmi les applications les plus représentatives des services non liés aux télécommunications, nous noterons:

- Transcription/Indexation automatique d'émissions de radio ou de télévision. L'indexation permet une recherche ultérieure sur base de mots clés, de termes ou de concepts apparaissant dans les émissions.
- Transcription/Indexation automatique d'interventions ou de rapports médicaux.
- Remplissage de formulaires par la voix.
- Remplacement de l'utilisation d'un clavier pour les tâches d'encodage de texte.
- Accès à des bases de données dans le monde professionnel.
- Contrôle vocal de processus de fabrication.
- Aide aux handicapés.
- Apprentissage d'une langue ou de la lecture.

Les possibilités de cette technologie sont sans limites. Nous pouvons imaginer, pour le siècle prochain, un équipement autonome qui combinerait les capacités de calcul et de communication de nos équipements actuels avec une interface en langage naturel. D'abord du domaine du rêve et souvent utilisé dans les films de science-fiction, cette vision n'est plus très loin de la réalité. Les techniques de traitement de la parole sont en pleine expansion et font l'objet de recherches dans de nombreux laboratoires. De nouvelles sociétés ayant pour but d'exploiter les techniques

de reconnaissance dans les différentes applications mentionnées ci-dessus se créent régulièrement. Actuellement, beaucoup d'applications de reconnaissance de la parole peuvent tourner en temps réel sur un seul circuit ou sur un ordinateur personnel.

Cependant, les systèmes de reconnaissance actuels souffrent généralement d'un manque de robustesse face aux bruits additifs, aux bruits dûs au canal de transmission (téléphone, microphone), à la réverbération... Ce manque de robustesse justifie la nécessité de poursuivre de nouvelles directions de recherche.

2.2 Domaines connexes

D'autres types d'applications font également appel aux technologies déployées dans le domaine de la reconnaissance vocale. Leur but n'est cependant pas d'obtenir la transcription sous forme de texte:

- **Identification de la langue**: Il s'agit ici de déterminer automatiquement la langue d'un utilisateur d'application vocale. Il est ainsi possible d'aiguiller l'utilisateur vers un opérateur parlant la même langue ou vers un module de dialogue adapté.
- **Identification et vérification du locuteur**: L'identification consiste à déterminer l'auteur d'un signal de parole, parmi un ensemble de personnes ayant préalablement participé à une phase d'entraînement. Le nombre de décisions envisageables est au moins égal à la taille de la population. Par conséquent, les performances d'une tâche d'identification se dégraderont avec la taille de cette population. La vérification, quant à elle, consiste à authentifier ou à rejeter un locuteur proclamant son identité. Dans ce cas, uniquement deux décisions peuvent être envisagées : acceptation ou rejet. Les performances d'une tâche de vérification seront donc a priori insensibles à la taille de la population. La vérification vocale du locuteur peut être couplée à d'autres approches de vérification dans le cadre d'applications de sécurité et de vérification d'identité: vérification d'empreintes digitales, vérification sur base d'une image du visage ou de l'iris...
- **Segmentation en locuteurs** ("Speaker Tracking"): Dans le cadre d'applications de transcription et d'indexation automatique, il peut être utile de déterminer automatiquement les tours de parole et le nombre d'intervenants. Ces données enrichissent l'index automatique. Elles permettent de plus d'envisager une adaptation du système de reconnaissance vocale aux différents locuteurs, augmentant ainsi l'efficacité du système.

2.3 Production de la parole

2.3.1 Quasi-stationnarité

On peut considérer que les changements dans la configuration de l'appareil articulatoire sont relativement lents comparés aux fréquences intervenant dans le signal. Les plosives sont des exceptions à cette règle. On considère néanmoins que pendant

de cours intervalles de temps, de l'ordre de quelques dizaines de millisecondes, l'appareil articulatoire ne change pas de position et conduit à la production d'une portion quasi-stationnaire de signal acoustique.

Nous développerons plus en détail ce point dans la suite de ce texte, mais signalons dors et déjà que la plupart des systèmes de reconnaissance automatique de la parole utilisent une technologie basée sur l'analyse de fenêtres temporelles de signal de l'ordre de 20 à 30 ms. De ces fenêtres sont extraits des paramètres représentatifs permettant de distinguer les différents sons du signal de parole. Cette approche est bien entendu basée sur l'hypothese de quasi-stationnarité du signal. Les fenêtres d'analyse sont ensuite concaténées dans le but de pouvoir s'accomoder des différences dans la vitesse d'élocution et donc dans la durée des différentes phases stationnaires du signal de parole; le nombre de répétitions est déterminé de façon optimale par le moteur de reconnaissance. Notons finalement que les fenêtres se recouvrent quelque peu car l'analyse est effectuée avec une période inférieure (autour de 10 ms) à la durée des fenêtres. Ceci permet une plus grande précision quant à la détermination de la frontière temporelle entre les différentes phases stationnaires.

2.3.2 Les classes de sons

Les sons des langues européennes peuvent être classifiés en 6 groupes:

1. **les voyelles**: elles proviennent de l'excitation du conduit vocal par un train d'impulsions quasi-périodique produit par la vibration des cordes vocales. La forme du conduit vocal, déterminée essentiellement par la position de la langue, impose des résonances caractéristiques permettant la production de différents sons. On classe généralement les voyelles en trois catégories suivant la position de la constriction principale du conduit vocal: à l'avant, au milieu ou au fond de celui-ci.
2. **les diphtongues**: il s'agit de la combinaison temporelle de deux voyelles, la configuration du conduit vocal variant d'un point à un autre.
3. **les semi-voyelles**: celles-ci sont très proches des voyelles. Elles sont produites par la transition du conduit vocal d'une position correspondant au phonème précédent vers une position correspondant au phonème suivant, en passant par une position correspondant au son à produire. Il s'agit toujours de sons provenant de la vibration des cordes vocales.
4. **les nasales**: cette fois, le flux d'air passe par le conduit nasal. Le conduit oral, bien qu'obstrué, agit toujours comme une cavité permettant de modifier les résonances du signal.
5. **les fricatives**: les fricatives non-voisées sont produites grâce à un flux d'air qui devient turbulent suite à une constriction du conduit vocal. La position de cette constriction (lèvres, dents, milieu ou fond du conduit vocal) détermine la nature du son produit. Pour les fricatives voisées, la vibration des cordes vocales produit un son périodique qui se superpose au son fricatif.
6. **les plosives**: les plosives sont produites par le relâchement d'une constriction obstruant le conduit vocal. Elles sont fortement dynamiques. La position de la constriction (lèvres, dents, fond du palais) détermine la nature du son produit. Les plosives voisées s'accompagnent en plus d'une vibration des cordes vocales.

Lorsque le son qui suit est voisé, les plosives voisées et non voisées diffèrent par le délai entre l'instant de relâchement et de début de vibration des cordes vocales visant à produire ce son voisé.

2.3.3 Le phénomène de coarticulation

On considère généralement que les phrases sont constituées d'une séquence de sons élémentaires (phonèmes) faisant partie de l'une ou l'autre des classes introduites à la section précédente.

L'inertie de l'appareil articulatoire ne lui permet cependant pas de se placer instantanément à chacune des positions caractéristiques des sons à produire. Certains sons (semi-voyelles) correspondent même par nature au mouvement continu d'une position à une autre. La réalisation d'un son dépend donc des phonèmes qui l'entourent: c'est le phénomène de coarticulation.

Pour faire face au problème posé par la coarticulation, les systèmes de reconnaissance vocale les plus évolués font appel à une représentation des mots sous forme de séquence d'allophones, c'est-à-dire de phonèmes dans un contexte particulier. Un triphone par exemple correspond à un phonème donné dans une contexte phonétique gauche-droit particulier.

2.4 Etat de l'art

Le schéma bloc général d'un système de reconnaissance automatique de la

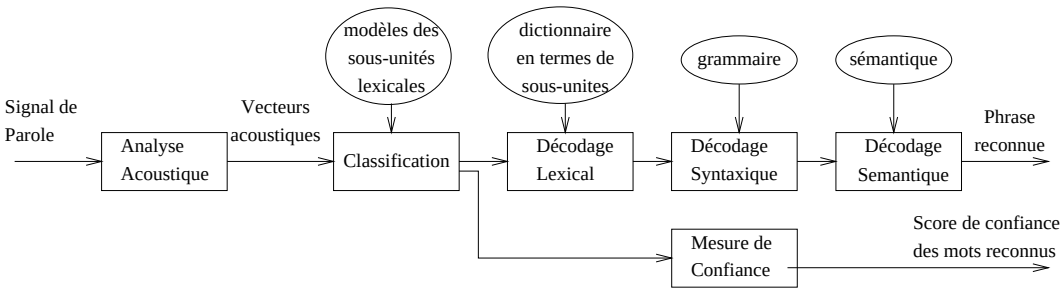


FIG. 2.1 – *Schéma bloc d'un système de reconnaissance automatique de la parole.*

parole est présenté à la figure 2.1. Le système reçoit en entrée un signal de parole et, idéalement, fournit en sortie la phrase qui a été prononcée ou le mot qui a été prononcé. Dans le premier cas, on parle de *système de reconnaissance de parole continue* ou de *système de reconnaissance de mots connectés*, et dans le deuxième cas, de *système de reconnaissance de mots isolés*. Ces systèmes sont entraînés sur base d'une grande quantité de parole pré-enregistrée et stockée sous forme de *base de données*. On parlera alors de base de données de parole continue ou de base de données de mots isolés.

Outre la transcription orthographique de la phrase reconnue, le système peut également fournir une mesure de confiance à chacun des mots de cette phrase. Idéalement, ces scores de confiance indiquent la probabilité que les mots soient correctement reconnus.

Ces scores de confiance peuvent être utilisés pour signaler certaines anomalies à l'utilisateur (ou à des modules de traitement ultérieurs). Ces anomalies peuvent être liées aux conditions d'enregistrement (qui peuvent être différentes des conditions utilisées pour développer le système), à la présence de bruit ou à la présence de mots n'appartenant pas au vocabulaire défini dans l'application de reconnaissance vocale.

Voyons quelles sont les étapes nécessaires à la reconnaissance automatique de la parole. Le signal de parole est fortement non stationnaire. Par conséquent, son analyse le décompose en une succession de tranches élémentaires supposées stationnaires. Ces tranches sont appelées *fenêtres d'analyse* ou *trames* ("frames"). Typiquement, une analyse est appliquée toutes les 10 ms sur des fenêtres d'analyse de 30 ms (par glissement et recouvrement des fenêtres d'analyse) pour générer un *vecteur acoustique*, c'est l'étape d'*analyse acoustique* (voir également la Section 2.7). Les paramètres représentatifs (vecteurs acoustiques) ainsi extraits du signal sont significatifs car ils sont calculés à partir d'un segment de parole relativement stationnaire. De nombreuses méthodes d'analyse ont déjà été étudiées. Parmi les paramètres utilisés avec beaucoup de succès ces dernières années, citons les cepstres calculés à partir d'un modèle auto-régressif du signal de parole, les paramètres PLP (Perceptual Linear Prediction) [83] calculés à partir d'un spectre représentant le contenu fréquentiel du signal suivant l'échelle des Bark (correspondant à l'échelle des bandes critiques du système auditif humain), ou finalement les paramètres MFCC (Mel Frequency Cepstral Coefficients) [163], cepstres calculés à partir d'une représentation fréquentielle suivant l'échelle des Mels, liée également aux propriétés de l'oreille humaine.

L'étape d'analyse acoustique convertit donc un mot ou une phrase en une séquence de vecteurs acoustiques $X = \{x_1,...,x_N\}$. Dès lors, la reconnaissance automatique de la parole peut être vue comme un problème de reconnaissance de formes qui peut lui-même être résolu par *classification statistique*. Il convient en effet d'associer à chaque vecteur acoustique, la sous-unité lexicale (phonème, allophone[1]...) à laquelle il appartient. Cette classification est réalisée sur base de modèles représentant les différentes sous-unités lexicales, chaque sous-unité lexicale correspondant à une classe. En réalité, le problème n'est pas aussi simple car notre but n'est pas de reconnaître des sous-unités lexicales mais plutôt de reconnaître des mots ou des phrases entières. Comme il est pratiquement irréalisable d'utiliser un modèle pour chaque phrase possible, les modèles de phrases seront constitués par concaténation de modèles de sous-unités lexicales, impliquant l'utilisation d'approches de décodage particulières.

Tous les systèmes de reconnaissance de la parole actuels sont basés sur la théorie des modèles de Markov cachés à temps discret ("Hidden Markov Models" – HMMs). Typiquement, chaque sous-unité lexicale (souvent le phonème) est modélisée par un ou plusieurs états stationnaires. Les mots sont ensuite construits en terme de séquences de phonèmes (à partir du *dictionnaire*) et les phrases en terme de séquences de mots (*syntaxe* et *sémantique*). Des algorithmes de décodage (reconnaissance) particuliers permettent la reconnaissance de phrases. Chaque état stationnaire est

1. Voir Section 2.3.3.

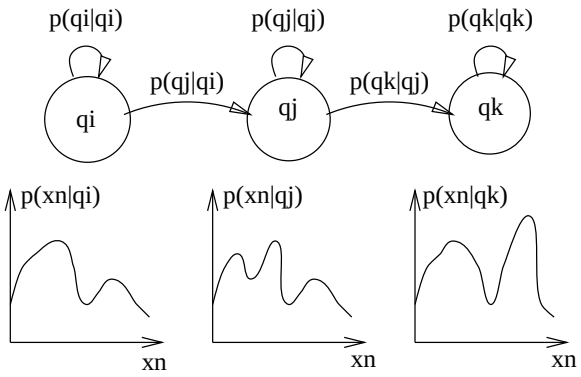


FIG. 2.2 – *Modèle de Markov caché à trois états (q_i,q_j et q_k). Chaque état est caractérisé par une distribution de probabilité pour les vecteurs d'observation (ex. $p(x_n|q_j)$). Les transitions d'un état à un autre sont caractérisées par une probabilité de transition (ex. $p(q_j|q_i)$).*

représenté par les paramètres de fonctions statistiques invariables, par exemple la moyenne et la variance d'une distribution gaussienne. Un modèle de Markov caché est donc un automate probabiliste construit sur base d'un ensemble de K états $Q = \{q_1,...,q_K\}$ stationnaires. Chaque état de cet automate est représenté par une distribution de probabilité décrivant la probabilité d'observation des différents vecteurs acoustiques[2] (voir figure 2.2). Les transitions entre les états sont instantanées. Elles sont caractérisées par une probabilité de transition. Remarquons que si chaque état du modèle permet de modéliser un segment de parole stationnaire, la séquence d'états permet quant à elle de modéliser la structure temporelle de la parole comme une succession d'états stationnaires, permettant le décodage cité précédemment. Les modèles utilisés en reconnaissance automatique de la parole sont généralement du type gauche-droite. Les transitions permises par ce type de modèles sont soit des boucles sur un même état, soit le passage d'un état à l'état qui le suit directement. L'aspect séquentiel du signal de parole est ainsi modélisé. Rappelons finalement que ces modèles sont dits *cachés* car la séquence d'état n'est pas directement observable. Seule la séquence X de vecteurs acoustiques est observée.

Lors du décodage (reconnaissance) basé sur les modèles de Markov cachés, il est possible de faire intervenir des grammaires (ou modèles de langage) permettant de contraindre la recherche de la meilleure phrase, sachant que toutes les séquences de mots ne sont pas possibles, ou que certaines sont moins probables que d'autres. Cette phase de décodage peut être vue comme indépendante de la phase de décodage lexical. En effet, la probabilité des hypothèses de reconnaissance (phrases possibles) est généralement estimée comme le produit d'une probabilité issue du modèle acoustique[3] et d'une probabilité issue du modèle de langage.

Un aspect "décodage sémantique" pourrait également intervenir dans le système de reconnaissance vocale, par exemple par l'intermédiaire de modèles de langage

2. Une approche alternative consiste à estimer les probabilités a posteriori des différents états (par exemple sur base de réseaux de neurones artificiels).

3. Les mots du lexique étant fixés par les topologies des modèles de Markov cachés.

faisant intervenir des aspects conceptuels. Ce point ne sera pas traité ici.

Finalement, certaines approches permettent de fournir, pour chacun des mots reconnus, une estimation de la confiance que le système a en sa décision. Cet aspect ne sera pas traité ici. Le lecteur se référera à [204] pour de plus amples informations.

Signalons pour terminer que les techniques de reconnaissance robuste s'intègrent généralement dans le schéma bloc proposé ici. Elles ne seront étudiées qu'au Chapitre 3.

2.5 Les modèles de Markov cachés

Soit M, le modèle de Markov caché représentant une phrase particulière. Chaque phrase possède son propre modèle. Rappelons que ces modèles sont constitués de la concaténation de modèles élémentaires représentant des sous-unités lexicales. Connaissant la séquence de vecteurs acoustiques X, le problème de la reconnaissance consiste alors à déterminer le meilleur modèle M, c'est-à-dire celui qui maximise la probabilité a posteriori $P(M|X)$. On peut montrer que ce critère est discriminant et qu'il minimise donc le taux d'erreur (cfr. Théorie de la classification [167]).

En utilisant la loi de Bayes, on obtient

$$P(M|X) = \frac{P(X|M)P(M)}{P(X)} \tag{2.1}$$

$P(X|M)$ est la vraisemblance de la séquence de vecteurs acoustiques X étant donné un modèle M (probabilité que le modèle M émette la séquence de vecteurs acoustiques X). $P(M)$ est la probabilité a priori du modèle M et $P(X)$ la probabilité a priori de la séquence de vecteurs acoustiques X.

L'entraînement permet difficilement le calcul de $P(M|X)$. Pour faciliter la tâche, on émet alors deux hypothèses:

- $P(M)$ est indépendant de la séquence d'observation X. En reconnaissance de parole continue, M représente une séquence de modèles de mots et $P(M)$ peut alors être estimé à partir du modèle de langage, celui-ci étant généralement exprimé sous forme de *grammaire stochastique* (voir Section 2.6).
- $P(X)$ est supposé constant car la séquence d'observation X est indépendante des modèles de Markov cachés[4].

L'estimation de 2.1 revient finalement à calculer $P(X|M)$. On peut montrer que le critère consistant à maximiser $P(X|M)$ n'est plus discriminant. L'augmentation de la probabilité associée au modèle correct ne se fait plus au détriment des modèles concurrents. Ce critère est donc sous-optimal.

Cette formulation étant cependant admise, les trois problèmes fondamentaux en reconnaissance automatique de la parole sont les suivants:

- *Estimation des probabilités*: étant donné une séquence d'observation X et un modèle M, comment calculer $P(X|M)$?

4. Remarquons que cela n'est vrai que si les paramètres des modèles sont fixés.

- *Entraînement*: étant donné un ensemble de séquences d'observations X_i et leurs modèles de Markov cachés respectifs M_i (entraînement supervisé), comment estimer les paramètres des modèles de façon à maximiser la probabilité $P(X_i|M_i)$ que chaque modèle génère la séquence d'observation qui lui est associée? (Un système de reconnaissance de parole continue à grand vocabulaire peut demander plusieurs millions de vecteurs d'observation soit quelques heures de parole). Il s'agit d'un problème d'estimation par maximum de vraisemblance [139, 163] ("Maximum Likelihood Estimation" - critère MLE), faisant appel à l'algorithme EM [38, 163]. Signalons finalement qu'il existe des approches permettant l'estimation des paramètres maximisant la probabilité a posteriori $P(M_i|X_i)$ ("Maximum A Posteriori" - critère MAP) [19].
- *Décodage* ou *Reconnaissance*: étant donné un ensemble de modèles élémentaires M_k et une séquence d'observation X, comment déterminer la meilleure séquence de modèles élémentaires M_k de façon à maximiser la probabilité que cette séquence de modèles ait émis la séquence d'observation X (c'est-à-dire comment déterminer la séquence de modèles qui 'explique' le mieux possible la séquence d'observation)?

Dans la suite, nous développerons uniquement le problème de l'estimation des probabilités. Ceci permettra de mettre en évidence les hypothèses généralement posées en reconnaissance automatique de la parole, pour finalement aboutir au modèle illustré à la figure 2.2. Le lecteur trouvera la résolution des problèmes d'entraînement dans les ouvrages [20] et [163].

Le problème du décodage fait appel à l'estimation des probabilités. Il s'agit en effet de déterminer la séquence de mots M_k maximisant la probabilité $P(M_k|X)$ (ou $P(X|M_k)$). On utilisera donc les récurrences présentées à la section suivante. Ces méthodes optimales sont cependant trop lourdes pour un fonctionnement en temps réel, particulièrement pour les tâches de reconnaissance à grand vocabulaire. On aura donc recours à des méthodes sous-optimales basées sur un élagage ("pruning") ou sur une recherche en faisceau ("beam search") au sein des hypothèses possibles. Le principe de ces méthodes est de ne conserver à chaque instant que les hypothèses les plus vraisemblables [157, 166].

2.5.1 Estimation des probabilités

Le problème est de calculer $P(X|M)$ étant donné une séquence d'observation $X = \{x_1,...x_N\}$ de durée N et un modèle M constitué d'une séquence de L états stationnaires. Nous appellerons chemin C toute séquence de N états permise par le modèle M. La notation q_k^n signifiera que l'état q_k ($\in Q$) est visité par le modèle à l'instant n.

Une façon de calculer $P(X|M)$ est d'énumérer tous les chemins C permis par le modèle. On obtient alors:

$$P(X|M) = \sum_C P(C,X|M) \tag{2.2}$$

On peut alors écrire:

$$P(X|M) = \sum_{l_1=1}^{L} ... \sum_{l_N=1}^{L} P(q_{l_1}^1,...,q_{l_N}^N,X|M) \tag{2.3}$$

comme les événements q_l^n sont mutuellement exclusifs, on obtient:

$$P(X|M) = \sum_{l=1}^{L} P(q_l^n,X|M),\forall n \in [1,N] \tag{2.4}$$

Chaque terme de cette somme exprime la probabilité que X soit émis par le modèle M en passant par l'état q_l à l'instant n. Ces termes peuvent être décomposés de la sorte:

$$P(q_l^n,X|M) = P(q_l^n,X_1^n|M)P(q_l^n,X_{n+1}^N|q_l^n,X_1^n,M) \tag{2.5}$$

où X_m^n représente une séquence partielle de vecteurs d'observation $\{x_m,...,x_n\}$. Le calcul de $P(X|M)$ se ramène alors à une somme de produits de deux probabilités (algorithme dit "forward-backward"):

- la probabilité-avant $P(q_l^n,X_1^n|M)$ ("forward probability").
- la probabilité-arrière $P(q_l^n,X_{n+1}^N|q_l^n,X_1^n,M)$ ("backward probability").

Moyennant certaines hypothèses, ces deux probabilités peuvent aisément être calculées par récurrence. Montrons quelles sont les hypothèses généralement émises. Considérons pour cela la probabilité-avant $P(q_l^n,X_1^n|M)$. La récurrence suivante est utilisée:

$$P(q_l^n,X_1^n|M) = \sum_{k=1}^{L} P(q_k^{n-1},X_1^{n-1}|M)P(q_l^n,x_n|q_k^{n-1},X_1^{n-1},M) \tag{2.6}$$

La probabilité conditionnelle correspondant au deuxième terme du produit est généralement appelée *contribution locale* ou *probabilité locale* car elle représente la contribution du n-ème vecteur acoustique dans le calcul de $P(X|M)$. Lorsqu'on prend le *logarithme* de ces grandeurs, on ne parle plus de probabilités mais plutôt de distances. On obtient alors des *distances locales* et des *distances accumulées* ($log(P(X|M))$). Le deuxième terme du produit peut finalement être décomposé de la sorte:

$$P(q_l^n,x_n|q_k^{n-1},X_1^{n-1},M) = P(q_l^n|q_k^{n-1},X_1^{n-1},M)P(x_n|q_l^n,q_k^{n-1},X_1^{n-1},M) \tag{2.7}$$

Pour faciliter le calcul de cette probabilité, on pose généralement les hypothèses suivantes:

- on suppose que les modèles de Markov cachés sont des modèles d'ordre 1,

$$P(q_l^n|q_k^{n-1},X_1^{n-1},M) = P(q_l^n|q_k^{n-1},M) \tag{2.8}$$

- on suppose que les vecteurs d'observation sont non corrélés,

$$P(x_n|q_l^n,q_k^{n-1},X_1^{n-1},M) = P(x_n|q_l^n,q_k^{n-1},M) \tag{2.9}$$

– on suppose que les vecteurs d'observation ne dépendent que de l'état du modèle à l'instant considéré,

$$P(x_n|q_l^n,q_k^{n-1},M) = P(x_n|q_l^n,M) \tag{2.10}$$

– on suppose que les fonctions de densité de probabilité définies par $P(x_n|q_l^n,M)$ sont indépendantes du modèle M,

$$P(x_n|q_l^n,M) = P(x_n|q_l^n) \tag{2.11}$$

Si ce n'était pas le cas, chaque modèle M aurait ses propres fonctions de densité de probabilité associées aux K états stationnaires. Cela serait équivalent à multiplier le nombre d'états par le nombre de modèles dans lesquels ils apparaissent. Le nombre d'états serait alors beaucoup trop important rendant la méthode impraticable.

On suppose aussi que les modèles sont stationnaires. Les indices en exposant peuvent donc être supprimés. Bien que les fonctions de densité de probabilité ne dépendent pas de l'instant considéré, il faut cependant se rappeler que q_k représente l'état du modèle à l'instant $n-1$. On obtient donc:

$$P(q_l,x_n|q_k,X_1^{n-1},M) = P(q_l|q_k,M)P(x_n|q_l) \tag{2.12}$$

Le premier terme du produit représente la probabilité de transition de l'état q_k à l'état q_l. Le deuxième terme représente la probabilité d'émission calculée à partir d'une distribution de probabilité dans l'espace des x. Ces termes sont calculés à partir des paramètres des modèles, modèles qui ont finalement la forme de celui présenté à la figure 2.2.

Les paragraphes précédents décrivent le critère du maximum de vraisemblance. Le critère de Viterbi est une approximation de ce critère. Il revient à ne prendre en considération que le chemin C le plus probable. Les sommes de l'équation (2.3) sont alors remplacées par un opérateur "maximum" et on peut montrer que la récurrence 2.6 peut alors être formulée en termes de *programmation dynamique*. On utilise généralement cette approximation car elle permet de réduire très sensiblement la charge de calcul imposée par le critère du maximum de vraisemblance. La présentation sous forme de chaîne de Markov (figure 2.2) prend tout son sens dans le cadre de cette approximation. En effet, à chaque trame, le système se trouve dans un des états de la chaîne, et le passage d'un état à l'autre est guidé par la topologie du modèle.

Le calcul de $P(x_n|q_l)$ est la tâche du système de classification de la figure 2.1. Nous nous contenterons ici de présenter un bref résumé des techniques les plus utilisées dans le cadre de la reconnaissance automatique de la parole.

La vraisemblance $P(x_n|q_l)$ peut être calculée sur base d'une fonction de densité de probabilité en x_n. Comme x_n est un vecteur multidimensionnel (de dimension d) de composantes réelles, des hypothèses sur la forme des distributions sont nécessaires.

Dans de nombreux cas, on suppose que $P(x_n|q_l)$ a la forme d'une distribution multi-Gaussienne [102]. Les systèmes basés sur ce type d'approximation sont appelés *systèmes à densités d'observation continues* ou *systèmes multi-Gaussiens*. Une

seule distribution normale ne suffit généralement pas pour modéliser correctement $P(x_n|q_l)$. Les performances de reconnaissance obtenues sont alors médiocres, ce qui justifie le choix de distributions multi-Gaussiennes. Pour réduire le nombre de paramètres du modèle, on peut émettre l'hypothèse que les différentes composantes du vecteur x_n sont non corrélées. Les matrices de covariance des distributions sont alors diagonales. Dans ce cas, si le nombre de distributions Gaussiennes par état est G, le nombre de paramètres de l'ensemble des modèles (c'est-à-dire le nombre de paramètres permettant de définir complètement les distributions de probabilités des K états stationnaires qui sont à la base des modèles de Markov cachés) est de $2dGK$.

Une autre solution permet aisément le calcul de $P(x_n|q_l)$. Elle consiste à quantifier la séquence X de vecteurs acoustiques. On utilise généralement une quantification vectorielle et chaque vecteur x_n est remplacé par le vecteur prototype y_i le plus proche (il convient ici de définir une distance, par exemple la distance Euclidienne) dans un ensemble prédéterminé de I vecteurs prototypes $Y = \{y_1,...,y_I\}$. Cet ensemble, appelé "codebook" est déterminé par "clustering" des vecteurs de la base d'entraînement [163]. Cette technique consiste à grouper l'ensemble des vecteurs traités en I classes. Chaque classe contient des vecteurs relativement proches les uns des autres au sens de la distance choisie. Le centroïde de la classe i est défini comme la moyenne de l'ensemble des vecteurs de cette classe. Ce centroïde est utilisé comme vecteur prototype y_i. La fonction de densité de probabilité que nous souhaitons calculer devient donc une distribution de probabilité à une seule variable discrète:

$$P(x_n|q_l) = P(y_i|q_l) \tag{2.13}$$

Il n'est donc plus nécessaire de formuler des hypothèses quant à la forme des distributions de probabilité. Le nombre de paramètres de l'ensemble des modèles est IK. Les systèmes basés sur une quantification des vecteurs acoustiques sont appelés *systèmes discrets*.

2.5.2 Modèles hybrides HMM/ANN

Dans le but d'éviter certaines des hypothèses sous-jacentes aux modèles HMMs, une alternative a été proposée dans [20] et s'est récemment révélée particulièrement efficace. Elle consiste à utiliser des HMMs conjointement avec des réseaux de neurones artificiels (artificial neural networks – ANNs) qui ne sont utilisés que pour estimer les statistiques locales requises par les HMMs qui continuent à modéliser le caractère séquentiel du signal. Ce type de système a été dénommé "système hybride HMM/ANN". On aboutit à un système discriminant au niveau de la fenêtre d'analyse (discrimination locale plutôt qu'au niveau global du mot ou de la phrase). Les avantages de ces modèles sont multiples. Ils résultent notamment de l'approche de classification utilisée, qui est une application de la méthode du diagnostic: méthode de classification basée sur l'estimation de probabilités a posteriori et non sur des distributions de probabilité des paramètres représentatifs. Les avantages principaux de ces modèles sont les suivants:

- modèle ne nécessitant pas d'hypothèses sur la forme des distributions (gaussienne ou multigaussienne) statistiques associées à chaque état des HMMs. En

effet, il a été démontré en théorie et en pratique que l'entraînement du réseau de neurones permettait d'estimer des distributions statistiques de n'importe quelle forme.

- du fait de l'entraînement discriminant des réseaux de neurones (ce qui est une de leurs propriétés majeures), on aboutit à des HMMs avec discrimination locale (au niveau de la fenêtre d'analyse).

D'autre part, l'utilisation de l'information temporelle est plus aisée avec ce type de système: il est facile de fournir plusieurs vecteurs acoustiques à l'entrée du réseau de neurones. Une information contextuelle est donc prise en compte dans les probabilités estimées et la corrélation entre des fenêtres successives n'est pas négligée. Pour diverses raisons, cela n'est pas possible avec les HMMs classiques[5]. Des résultats, confirmés par plusieurs équipes de recherche, ont montré que l'ordre de grandeur du contexte optimal est proche de 100 millisecondes, soit la longueur moyenne du phonème.

Plusieurs résultats récents (obtenus sur différentes bases de données allant des petits vocabulaires aux très grands vocabulaires) ont montré que ces systèmes conduisent généralement à des performances de reconnaissances significativement meilleures que celles des systèmes classiques utilisés dans les mêmes conditions [189].

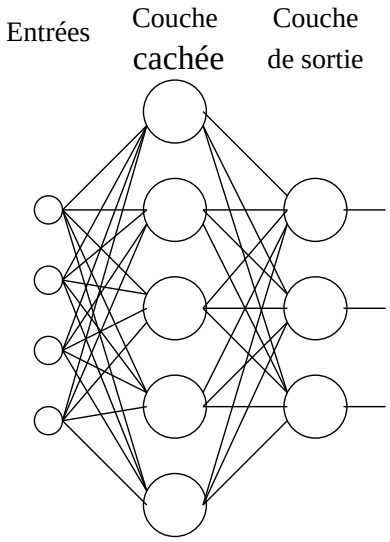


FIG. 2.3 – *Architecture d'un perceptron multicouche.*

Les réseaux de neurones artificiels généralement utilisés en reconnaissance automatique de la parole sont des perceptrons multicouches (multilayer perceptrons – MLPs) ayant une seule couche cachée. L'architecture générale de ce type de réseau est schématisée à la figure 2.3. Chaque noeud (ou neurone) du réseau présenté est appelé perceptron. L'entrée du perceptron est un vecteur de dimension d. Le perceptron (figure 2.4) réalise une opération de somme pondérée des différentes composantes du vecteur d'entrée et ajoute à cette somme un seuil w_o. Une fonction non-linéaire

5. le contexte peut cependant être introduit grâce aux dérivées temporelles premières et secondes des paramètres acoustiques

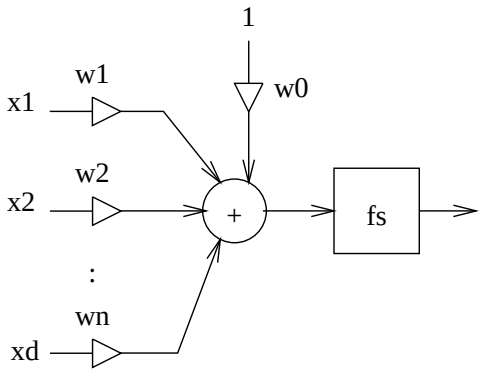


FIG. 2.4 – *Architecture du perceptron.*

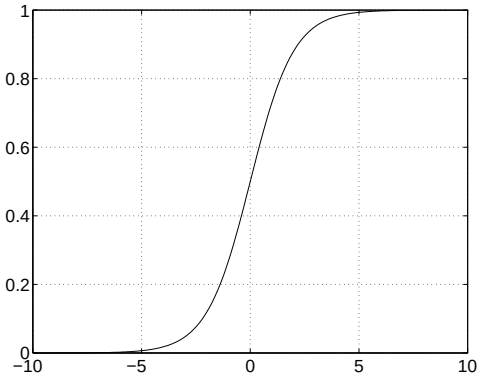


FIG. 2.5 – *Sigmoïde,* $\beta = 1.0$

f_s est finalement appliquée au résultat. Cette fonction non-linéaire est généralement une sigmoïde (figure 2.5) dont le gain β vaut 1:

$$f_s(y) = \frac{1}{1 + e^{-\beta y}} \tag{2.14}$$

Les différents termes de la somme pondérée des noeuds du réseau constituent les paramètres (ou poids) de celui-ci. Ils seront estimés lors de l'entraînement sur base de données.

Ces MLP sont utilisés comme classificateurs statistiques. Ils permettent de classifier des vecteurs acoustiques en différentes classes, chaque classe étant associée à un état stationnaire de l'ensemble Q (voir formalisme des modèles de Markov cachés). Le vecteur d'observation x_n est introduit aux entrées du MLP. Si l'ensemble Q contient K états stationnaires, le réseau présentera K noeuds de sortie. Etant donné x_n à l'entrée du MLP, on peut montrer que la sortie k de ce réseau est une estimation de la probabilité locale $P(q_k|x_n)$. En utilisant la loi de Bayes:

$$P(q_k|x_n) = \frac{P(x_n|q_k)P(q_k)}{P(x_n)} \tag{2.15}$$

Il suffit de diviser cette probabilité locale par la probabilité a priori $P(q_k)$ pour obtenir un rapport de vraisemblance ("scaled likelihood") $P(x_n|q_k)/P(x_n)$. Comme

pendant la reconnaissance, $P(x_n)$ est constant et ne modifie en rien la classification, on se ramène au formalisme des modèles de Markov cachés présentés précédemment. Ce formalisme permet alors de modéliser le caractère séquentiel du signal de parole. Remarquons que l'architecture du réseau permet aisément d'introduire plusieurs vecteurs acoustiques consécutifs. Il suffit pour cela d'augmenter le nombre d'entrées du MLP. Cela a simplement pour conséquence d'augmenter la dimension des vecteurs d'entrées des perceptrons de la couche cachée.

Si NE, NC et NS correspondent respectivement au nombre d'entrées du MLP, au nombre de noeuds cachées et au nombre de noeuds de sortie, le nombre de paramètres permettant le calcul de $P(x_n|q_k)$ est alors de $NC(NE+NS)+NC+NS$.

Durant l'entraînement, les vecteurs d'observation x_n de l'ensemble d'entraînement sont consécutivement présentés aux entrées du MLP. L'entraînement est dit supervisé car on présente également au MLP les sorties désirées de celui-ci. La sortie associée à l'état stationnaire correspondant au vecteur d'entrée est forcée à 1 alors que les autres sorties sont à 0. L'algorithme opère alors par rétro-propagation de l'erreur d'estimation du vecteur de sortie du MLP et utilise la méthode itérative du gradient pour estimer les poids du réseau [96].

Les modèles hybrides HMM/ANN sont à la base de la plupart des systèmes présentés dans ce travail.

2.6 Les modèles de langage

Comme indiqué à la section précédente, le problème de la reconnaissance vocale consiste à déterminer la séquence de mots dont la probabilité a posteriori est maximale étant donné la séquence d'observations acoustiques. L'utilisation de la loi de Bayes conduit alors à l'expression suivante pour cette probabilité a posteriori:

$$\underset{M}{\text{argmax}}\ P(M|X) = \underset{M}{\text{argmax}}\ \frac{P(X|M)P(M)}{P(X)} \tag{2.16}$$

Le terme $P(X|M)$ correspond au modèle acoustique. Le terme $P(M)$ représente la probabilité a priori de la séquence de mots M. Il correspond au modèle de langage. Ces modèles de langage sont importants comme information a priori permettant de contraindre l'espace de recherche. Ils conduisent généralement à de très nettes améliorations.

Pour une séquence $M = m_1,...,m_N$ de N mots, la probabilité liée au modèle de langage peut s'écrire:

$$P(M) = \prod_{i=1}^{N} P(m_i|m_0,...,m_{i-1}) \tag{2.17}$$

La probabilité de la séquence M est donc le produit des probabilités associées à chacun des mots, ces probabilités étant conditionnées sur l'historique jusqu'au mot considéré.

Pour simplifier ce modèle, on impose généralement un historique limité. Les modèles *n-grammes*[6] par exemple, consistent à n'utiliser que les $n-1$ mots qui

6. un n-gramme est un modèle de Markov non-caché d'ordre $n-1$

précèdent dans l'historique de chacun des mots:

$$P(M) = \prod_{i=1}^{N} P(m_i|m_{i-n+1},..,m_{i-1}) \tag{2.18}$$

Ces probabilités ($P(m_i|m_{i-n+1},..,m_{i-1})$) sont estimées par comptage sur base de très grands corpus de texte. La taille de ces corpus va de 1 million à 500 million de mots [7]. En dépit de la taille importante de ces bases de données, certains $n-grammes$ (groupes de n mots) n'y apparaissent pas. Pour résoudre ce problème, qui conduirait à des probabilités nulles pour certains groupes de mots, on a généralement recours à un estimation lissée de la probabilité des $n-grammes$. Pour un tri-grammes par exemple, on utilise:

$$P(m_i|m_{i-2},m_{i-1}) = \lambda_3 f(m_i|m_{i-2},m_{i-1}) + \lambda_2 f(m_i|m_{i-1}) + \lambda_1 f(m_i) + \lambda_0 \frac{1}{V} \tag{2.19}$$

où V est la taille du vocabulaire et f sont les estimations obtenues sur base du corpus de texte. Les coefficients de pondération λ_i peuvent être optimisés par maximum de vraisemblance sur base de données qui n'ont pas été utilisées pour l'estimation des $n-grammes$. L'algorithme EM peut être utilisé à cette fin [99]. Il paraît également préférable de favoriser le terme λ_3 lorsque la fréquence d'apparition de l'historique du tri-gramme considéré est élevée et de défavoriser ce terme dans le cas contraire. On peut par exemple utiliser plusieurs vecteurs λ, chacun étant optimisé sur base d'historiques dont les fréquences d'apparitions appartiennent à une plage donnée.

D'autres types de modèles de langage peuvent également être utilisés. Les grammaires en paire de mots ("wordpair") contiennent, pour chacun des mots du vocabulaire, l'ensemble des mots qui peuvent suivre. Il s'agit en quelque sorte d'une simplification des grammaires bi-grammes.

Pour des tâches caractérisées par un langage fort contraint, on peut également utiliser une grammaire représentée sous forme d'automate d'états fini ("Finite State Automaton" ou "Finite State Grammar" - FSG) ou d'automate d'états fini pondéré. Celui-ci impose généralement des contraintes sur base d'un historique allant au delà des quelques mots caractéristiques du $n-gramme$.

2.6.1 Perplexité

Différents modèles de langage sont généralement comparés sur base de leur perplexité.

La théorie de l'information [36] nous apprend que la quantité d'information émise par une source peut être mesurée par l'entropie. Pour une source choisissant des mots appartenant à un vocabulaire de L mots, indépendamment les uns des autres, et si la probabilité d'émission du mot m est $P(m)$, l'entropie vaut:

$$H = -\left[\sum_{m=1}^{L} P(m) log_2(P(m))\right] \tag{2.20}$$

7. Ils correspondent par exemple aux transcriptions phonétiques d'émission de radio/télévision, aux textes d'articles de presse...

Une source d'entropie H [8] peut donc être vue comme émettant autant d'information qu'une source choisissant des mots de façon équiprobable dans un vocabulaire de 2^H mots.

Par extension, l'entropie d'une source générale est:

$$H = - \lim_{n \to \inf} \left(\frac{1}{n} \right) \sum_{m_1,...,m_n} P(m_1,...,m_n) log_2(P(m_1,...,m_n)) \tag{2.21}$$

la somme étant étendue à toutes les séquences de mots. Si la source émet des mots indépendamment les uns des autres, on retrouve l'expression précédente. Si d'autre part, la source est ergodique, on a:

$$H = - \lim_{n \to \inf} \left(\frac{1}{n} \right) log_2(P(m_1,...,m_n)) \tag{2.22}$$

où $m_1,...,m_n$ représente une séquence de mots réellement observée.

Cette théorie s'applique directement aux modèles de langage. L'équation (2.22) peut être utilisée pour estimer l'entropie sur base d'un corpus de texte suffisamment long. La mesure d'entropie obtenue permet de juger de la difficulté de la tâche de reconnaissance vocale. En effet, pour chaque mot à transcrire, ce sont en moyenne H bits qui doivent être extraits du signal. En pratique, la probabilité intervenant dans l'expression (2.22) n'est pas connue. On l'estimera donc sur base du modèle de langage utilisé par le système de reconnaissance, par exemple par produit de trigrammes. Les textes utilisés pour estimer cette entropie doivent être différents de ceux utilisés pour estimer les paramètres du modèle de langage [9].

On définit alors la perplexité de la tâche de reconnaissance comme:

$$P = 2^H \tag{2.23}$$

Une tâche d'entropie H peut donc être vue comme aussi compliquée (c'est-à-dire nécessitant l'extraction d'autant d'information) que la reconnaissance d'un texte généré sur base d'un vocabulaire de P mots, choisis indépendamment les uns des autres suivant une distribution de probabilité uniforme. Le nombre moyen de branchements d'un mot vers les mots suivants est donc de P. Notons ici que la perplexité (ou l'entropie) estimée de cette façon sera en général supérieure à la perplexité (l'entropie) réelle de la source. Ainsi, même une source d'entropie faible peut conduire à une tâche de reconnaissance compliquée si le modèle de langage est mal choisi. Un bon modèle de langage devra donc avoir une perplexité faible, tout en permettant de représenter n'importe quelle séquence de mot prononçable dans le cadre de la tâche choisie.

La perplexité d'un modèle de langage dépend fortement du domaine du discours. Ainsi, si le domaine est restreint, on peut s'attendre à une perplexité relativement faible. C'est le cas par exemple pour la tâche de gestion de ressources navales illustrée par la base de données Resource Management (perplexité de 60, voir Annexe A), pour une tâche de dictée de rapport en médecine d'urgence ($P = 60$) ou en radiologie

8. Le logarithme en base deux fournit une mesure d'entropie qui s'exprime en "bits" d'information: il faut en moyenne H bits pour représenter un mot émis par une source d'entropie H

9. Sinon, il est toujours possible d'obtenir un modèle qui annule l'entropie.

($P = 20$). Par contre, pour l'anglais général, la plus faible perplexité publiée jusqu'à présent est de 247 [33].

Un des objectifs des recherches dans le domaine de la modélisation du langage consiste à diminuer la perplexité de ces modèles. On peut alors s'attendre à une réduction du taux d'erreur des systèmes de reconnaissance vocale.

2.6.2 Améliorations

Différentes améliorations ont été proposées récemment sur base des modèles $n-$ *grammes* décrits précédemment [33]:

- **Modèles basés sur des classes**. Plutôt que d'utiliser les mots précédents pour décrire l'historique des $n-$ *grammes*, il s'agit ici d'utiliser les classes de mots. Ces classes peuvent être constituées sur base de la nature grammaticale des mots, de leur morphologie ou même de leur nature sémantique (un mot pouvant généralement appartenir à plusieurs de ces classes). L'avantage potentiel de cette approche est de permettre le traitement efficace de mots très rares ou même de nouveaux mots, par simple spécification des classes auquelles ils appartiennent. On pourra par exemple considérer tous les noms propres de la même façon en définissant une classe adéquate.
- **Modèles dépendant du sujet**. Une détermination automatique du sujet correspondant au texte prononcé pourrait être utilisée pour sélectionner un modèle de langage particulier, développé sur base d'un corpus d'entraînement propre au sujet identifié.
- **Modèles dynamiques**. Une alternative à l'approche précédente consiste à utiliser l'historique du document (vocal) en cours de traitement pour adapter les probabilités des $n-$ *grammes*. Dans cette section par exemple, les mots "modèle" et "langage" apparaissent plus souvent que dans les autres sections. Ceci justifierait l'utilisation d'un modèle de langage adaptatif.
- **Modèle structurés**. Il s'agirait ici d'utiliser des parseurs grammaticaux, ou des modèles de langage plus évolués, comme les grammaires probabilistes libres de contexte ("Probabilistic Context Free Grammars"). L'inconvénient de ces approches est de nécessiter des corpus de texte annotés.

Dans l'état actuel des connaissances, la supériorité du modèle tri-gramme classique, développé sur base de très grands corpus de texte, est très rarement mise en défaut.

2.6.3 Exemple

Pour illustrer l'importance des modèles de langage, prenons comme exemple la tâche de reconnaissance "Resource Management" (voir Annexe A). Lorsqu'on utilise le modèle de langage en paire de mots fourni avec la base de donnée, notre système de référence conduit à un taux d'erreur au niveau du mot proche de 5%. Sans modèle de langage par contre, le taux d'erreur obtenu est proche de 20%.

2.7 Analyse du signal de parole

Le problème de la reconnaissance de la parole est notamment axé sur une classification des divers sons intervenant dans la construction des mots et des phrases. Depuis de nombreuses années, les recherches ont montré l'importance de l'enveloppe spectrale pour la classification de ces sons. Cette enveloppe spectrale fait apparaître certaines "bosses" appelées *formants* résultant des résonances imposées par la configuration du conduit vocal à l'instant considéré. Ces constatations ont guidé l'utilisation de représentations paramétriques du signal dans les systèmes de reconnaissance automatique de la parole. A partir des échantillons d'une portion de signal considérée comme stationnaire, un module de traitement de signal extrait un nombre réduit de paramètres représentatifs, qui peuvent généralement être assimilés à une représentation compacte de l'enveloppe spectrale de la portion considérée. Parmi les méthodes les plus courantes, il convient de citer ici celles basées sur l'utilisation d'un banc de filtres, ainsi que celles utilisant une modélisation autorégressive du signal de parole. Ces deux types de méthodes sont parfois combinées [83]. Différents auteurs proposent également d'utiliser certains aspects du fonctionnement de l'oreille, par exemple pour définir les spécifications du banc de filtres. Il est également possible d'aller plus loin encore dans l'utilisation des propriétés physiologiques et psychoacoustiques en effectuant un traitement non-linéaire à la sortie des différents filtres de façon à obtenir des paramètres représentant les impulsions transmises au cerveau par les nerfs auditifs. Les sections suivantes dressent un aperçu sommaire des méthodes utilisées dans ce travail.

2.7.1 Pré-accentuation

L'étape de pré-accentuation (ou pré-emphase) consiste à accentuer les hautes fréquences. On fait généralement appel à un filtre de la forme:

$$H(z) = 1 - az^{-1} \text{ avec } 0.9 < a < 1.0 \tag{2.24}$$

où a est généralement égal à 0.95. L'intérêt de cette pré-emphase est d'applatir le spectre du signal de parole et de filtrer la composante continue de façon à se placer dans des conditions "optimales" vis-à-vis des traitement ultérieurs, notamment le calcul d'un modèle autorégressif [159].

2.7.2 Décomposition en trames et fenêtrage

Le signal de parole est ensuite décomposé en trames dont la durée est proche de 30 ms. Chaque trame correspond à une portion sur laquelle le signal de parole peut être considéré comme stationnaire. Ces trames sont généralement extraites à une fréquence de 100 Hz, c'est-à-dire toutes les 10 ms.

L'intérêt de ce dernier choix nous parait incertain vu les observations récentes [78] indiquant que le spectre de modulation des paramètres représentatifs (spectre du signal correspondant à l'évolution de chacun des paramètres représentatifs) est fortement atténué au delà de 20 Hz. Des résultats récents vont dans ce sens [87]. Il est cependant de pratique courante d'utiliser un échantillonnage à 100 Hz.

Ensuite, on applique une fenêtre qui a pour fonction d'atténuer le signal au début et à la fin de chaque trame. Le choix se porte généralement sur les fenêtres de *Hanning* ou de *Hamming*:

$$\begin{aligned} &\text{Hanning}(n) = 0.5 + 0.4.cos\left(2\pi\tfrac{n}{N-1}\right) \\ &\text{Hanning-généralisée}(n) = \alpha + (1-\alpha).cos\left(2\pi\tfrac{n}{N-1}\right) \\ &\text{Hamming}(n) = 0.54 + 0.46.cos\left(2\pi\tfrac{n}{N-1}\right) \end{aligned} \tag{2.25}$$

N étant la largeur de la fenêtre et α un paramètre. Dans le domaine spectral, ce fenêtrage permet d'atténuer les lobes secondaires associés aux différentes composantes fréquentielles du signal.

Dans le cadre de la reconnaissance automatique de la parole, ce fenêtrage nous parait cependant peu important.

2.7.3 Modèle autorégressif - Analyse LPC

Le principe du modèle autorégressif du signal de parole est de modéliser le processus phonatoire par un système de synthèse élémentaire comprenant un module d'excitation à gain variable G, suivi par un filtre tout-pôles d'ordre p (approche LPC: "Linear Predictive Coding"). Les coefficients du filtre sont considérés constants (hypothèse de quasi-stationnarité) pendant des intervalles de temps réduits de l'ordre de 30 ms. L'excitation u est soit périodique (train d'impulsions, ou plus généralement signal périodique dont le spectre d'amplitude est un train d'impulsions, ce qui permet de modéliser les déphasages entre les différentes harmoniques), soit stochastique (bruit blanc), et éventuellement mixte, de façon à pouvoir modéliser les sons voisés ainsi que les sons non-voisés. Remarquons que pour le cas des sons purement voisés, l'excitation du système représentera l'action opérée par la vibration des cordes vocales, alors que le filtre représentera l'action du conduit vocal. Pour le cas de sons partiellement non-voisés par contre, le signal acoustique est le résultat d'un processus plus complexe faisant intervenir la frication, c'est à dire les perturbations créées par le passage de l'air au travers des constrictions du conduit vocal ou des lèvres. L'interprétation du modèle n'est donc plus aussi simple. Ce modèle reste cependant très utilisé en pratique car, quel que soit la nature périodique ou apériodique du signal, la fonction de transfert du filtre sera un bon modèle de l'enveloppe spectrale du signal, caractéristique essentielle pour la distinction des sons linguistiques.

Un échantillon $s(n)$ est calculé de la sorte:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{2.26}$$

En effectuant la transformation en z, on obtient:

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z) \tag{2.27}$$

La fonction de transfert du filtre est bien évidemment exprimée par:

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2.28}$$

et devra idéalement avoir un ordre suffisamment élevé pour modéliser avec précision la structure en formants du spectre du signal. L'ordre ne sera cependant pas trop élevé, et ce pour éviter la modélisation de détails spectraux au contenu linguistique négligeable. On estime en général avoir besoin d'une paire de pôles par kHz de bande passante, plus 3 ou 4 pôles pour l'excitation glottique et la radiation des lèvres. Pour une fréquence d'échantillonnage de 8 kHz, on choisira donc un ordre de 11 ou 12. Les expériences de reconnaissance vocale montrent que ces valeurs sont raisonnables.

Les paramètres de ce modèle, à savoir le gain, l'excitation et les coefficients a_i peuvent être estimées par des méthodes d'analyse. Une interprétation de ces méthodes d'analyse est de séparer la source et la structure, et donc d'obtenir des paramètres de structure a_i relativement "propres" car débarrassés de données moins importantes comme la fréquence fondamentale du son, les déphasages entre les harmoniques et les petites variations dans l'enveloppe spectrale. Ces données sont généralement considérées comme du bruit pour la reconnaissance automatique de la parole.

A partir du modèle qui vient d'être décrit, une estimation de l'échantillon $s(n)$ peut-être calculée de la sorte:

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{2.29}$$

L'erreur de prédiction $\hat{s}(n) - s(n)$ vaut donc:

$$s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{2.30}$$

Une estimation des paramètres a_i peut être obtenue par minimisation de la somme des carrés des erreurs de prédiction sur une trame de parole provenant des étapes de traitement précédentes, ce qui conduit à un système linéaire de p équations à p inconnues faisant intervenir la fonction de covariance du signal s. En limitant l'ordre de la somme des erreurs de prédiction par définition d'une fenêtre de signal de durée limitée, on peut montrer que les éléments intervenant dans le systèmes d'équation sont les $p+1$ premiers éléments de la fonction d'autocorrélation du signal. De plus, la matrice du système est une matrice de Toeplitz (les éléments de toutes les diagonales sont égaux) symétrique. Cette particularité permet l'utilisation d'une méthode de résolution particulièrement efficace appelée récursion de Durbin. Une description de cette méthode peut être trouvée dans [14, 163]

Divers jeux de paramètres

A partir des $p+1$ coefficients d'autocorrélation, une récursion permet donc d'obtenir les p coefficients du modèle autorégressif. D'autres récursions permettent alors d'obtenir divers types de paramètres. Il est ainsi possible de travailler avec les coefficients de Parcor k_i, aussi appelés coefficients de réflexion. Ceux-ci sont liés aux rapports de section de tubes cylindriques modélisant le conduit vocal, des cordes vocales jusqu'aux lèvres. Une récursion permet également d'obtenir les coefficients cepstraux c_i (pt. (2), figure 2.7). L'utilisation courante de ces derniers est justifiée

par leur caractère décorrélé ainsi que par les bonnes performances et la robustesse généralement observée en pratique dans le cadre de la reconnaissance automatique de la parole. Ces méthodes sont décrites dans [13, 163]. Sur base des paramètres cepstraux c_i, il est également possible d'obtenir une représentation de type banc de filtres, sachant que les cepstres correspondent à la transformée de Fourier discrète inverse du logarithme de la transformée de Fourier discrète du signal. Cette représentation spectrale (pt. (1), figure 2.7) correspondrait à une version lissée (par conséquence de la modélisation autorégressive) d'une représentation spectrale classique.

2.7.4 Analyse par banc de filtres

Sur base des trames d'analyse, il s'agit ici de calculer les énergies dans un ensemble de bandes de fréquence couvrant l'ensemble du spectre utile. Ce calcul peut être effectué dans le domaine temporel sur base de filtres définissant les différentes bandes de fréquence choisies. Il peut également être effectué dans le domaine fréquentiel, par exemple à partir de la transformée de Fourier discrète de la trame de signal.

Le nombre de filtres sera suffisamment important pour représenter avec précision l'enveloppe spectrale du signal, mais suffisamment réduit pour éviter de représenter des détails spectraux n'ayant que peu d'intérêt pour l'identification des sons linguistiques. En pratique, le nombre de filtres est généralement inférieur à 32 [163]. Le banc de filtres à la base de l'analyse PLP ("Perceptual Linear Prediction") par exemple [83] comprend 15 filtres pour couvrir la plage de 0 à 4000 Hz. Un banc de filtres typiquement utilisé pour le calcul de paramètres MFCC ("Mel Frequency Cepstral Coefficients) [163] comprend 20 filtres pour couvrir la plage de 0 à 4000 Hz.

Les filtres sont généralement contigus et leurs fréquences centrales peuvent suivre une échelle linéaire. Cependant, des études psychoacoustiques ont montré que la perception du contenu fréquentiel d'un signal semble suivre une loi non-linéaire. Ces études ont notamment conduit à la définition de deux échelles fréquentielles très semblables: l'échelle des Mel et l'échelle des Bark. Il semble d'autre part qu'en prenant 24 bandes de fréquence contiguës dont les fréquences centrales suivent l'échelle des Bark, chaque bande de fréquence correspond à approximativement 1.3 mm le long de la cochlée de l'oreille humaine. Ces résultats ont mené à l'élaboration de méthodes d'analyse visant à obtenir une représentation spectrale qu'on pourrait qualifier de "spectre auditif" ou de "spectre subjectif". Il s'agit simplement d'utiliser un banc de filtres dont les fréquences centrales suivent l'une ou l'autre des échelles perceptuelles.

Divers jeux de paramètres

Sur base de la représentation issue du banc de filtres, il est possible d'effectuer une analyse par prédiction linéaire et d'en déduire divers jeux de paramètres. Il suffit en effet d'effectuer une transformée de Fourier inverse pour obtenir une représentation temporelle, et ensuite utiliser les méthodes citées à la Section 2.7.3. Le calcul de cepstres par cette méthode (pt. (4), figure 2.7) est à la base de l'analyse PLP [83] qui permet de combiner l'intérêt d'un banc de filtres suivant une échelle non-linéaire avec le lissage opéré par le modèle autorégressif.

Il est également possible de calculer directement les cepstres par transformée de Fourier inverse du logarithme de la représentation en banc de filtres. Cette méthode (pt. (6), figure 2.7) est à la base de l'approche MFCC [163].

La représentation issue du banc de filtres (pt. (5), figure 2.7) peut également être utilisée directement. Dans ce travail, elle a été utilisée dans le cadre de l'approche de reconnaissance multi-bande pour calculer les paramètres représentatifs des différentes bandes de fréquence.

Il est également possible de combiner les avantages du banc de filtres non-linéaire avec l'analyse LPC pour obtenir une représentation de type banc de filtres (pt. (3), figure 2.7).

Analyse PLP ("Perceptual Linear Prediction")

Insistons ici sur l'analyse en bandes critiques et sur la méthode PLP, qui ont été utilisées dans ce travail.

Les fréquences centrales du banc de filtres suivent une échelle perceptuelle dont l'unité est le Bark. La fréquence en Bark B peut être obtenue par l'expression:

$$B = 6 \ln \left(\frac{f}{600} + \sqrt{\left(\frac{f}{600} \right)^2 + 1} \right) \tag{2.31}$$

où f est la fréquence en Hertz. La figure 2.6 montre que cette loi est quasi-logarithmique pour les fréquences supérieures à 1000 Hz et également que 15 bandes de fréquence de 1 Bark permettent de couvrir la plage de fréquence de 0 à 4000 Hz. Un filtrage équi-énergie est ensuite appliqué aux sorties des filtres. Il consiste grossièrement à amplifier les sorties des filtres à haute fréquence centrale. Il s'agit d'une implémentation fréquentielle de la pré-accentuation, typiquement réalisée dans le domaine temporel (voir figure 2.7). Finalement, les valeurs obtenues sont compressées par une fonction racine cubique. Ce traitement est basé sur les conclusions d'études psychoacoustiques relatives à la perception auditive et aux caractéristiques fonctionnelles de l'oreille moyenne. Cette analyse conduit donc à une représentation en banc de filtres. Celle-ci peut être utilisée comme paramètres représentatifs ou peut servir de point de départ à une analyse plus adaptée au problème de la reconnaissance de la parole. Une transformée de Fourier discrète inverse [10] peut être appliquée aux bandes critiques, de façon à obtenir des coefficients d'autocorrélation qui seront alors utilisés de façon classique pour effectuer une analyse LPC et finalement extraire des cepstres. L'algorithme de transformée de Fourier rapide n'est pas utilisé car le nombre de points n'est pas forcément une puissance de 2 et qu'il est de toute façon très faible. Rappelons ici l'expression de la transformée de Fourier discrète (DFT) d'une fenêtre de signal comprenant N échantillons:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{(-j2\pi \frac{kn}{N})}, k = 0,..,N \tag{2.32}$$

10. En toute rigueur, il s'agit d'une transformée en cosinus discrète (DCT) car l'information de phase n'est plus utilisée à ce stade.

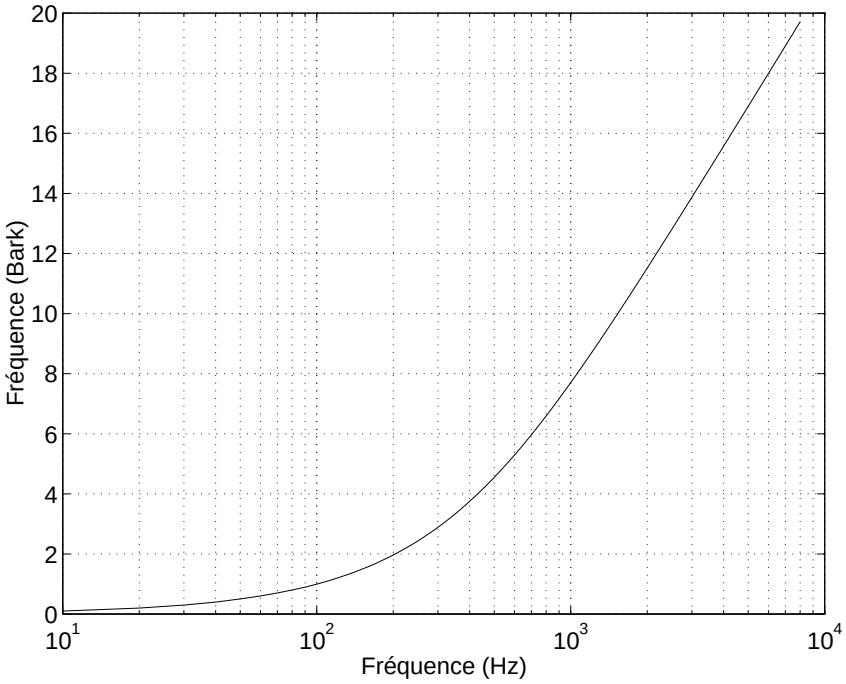


FIG. 2.6 – *Loi Bark en fonction de la fréquence en Hz.*

et de la transformée de Fourier discrète inverse:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{(j2\pi \frac{kn}{N})}, t = 0,..,N \tag{2.33}$$

On peut également appliquer la transformée de Fourier discrète inverse (DCT en pratique) aux logarithmes des racines carrées des énergies des bandes critiques [11]. On obtient alors des paramètres qu'on peut qualifier de cepstres perceptuels. La différence avec les paramètres des paragraphes précédents est qu'aucun lissage par modélisation autorégressive n'est appliqué. On peut cependant supposer que ces paramètres sont également de bonne qualité vu qu'un lissage fréquentiel est déjà obtenu grâce au filtrage en bandes critiques.

2.7.5 Paramètres dynamiques - Contexte

Le vecteur de paramètres issus des méthodes précédentes peut être complété par un vecteur correspondant aux dérivées temporelles premières et secondes de ces paramètres. Ces dérivées sont estimées sur base de plusieurs trames adjacentes [62]. L'approche permet d'introduire une information concernant le contexte temporel de la trame courante.

Une approche plus directe consiste à utiliser plusieurs trames successives en entrée du système de reconnaissance. Cette approche est courante lorsque le système

11. Par définition du cepstre.

de classification est un réseau de neurones artificiels [20]. Des expériences ont montré un optimum autour de 9 à 15 trames (décalées de 10 ms) pour plusieurs tâches différentes.

2.7.6 Schéma complet d'analyse du signal de parole

La figure 2.7 donne un schéma représentant les méthodes d'analyses classiques. Il fait appel aux modules décrits aux sections précédentes, auquelles on se référera pour plus de détails et de liens vers d'autres publications. Toutes ces méthodes sont fondamentalement similaires. Elles visent à extraire des paramètres de structure représentant l'enveloppe spectrale de courtes trames de signal.

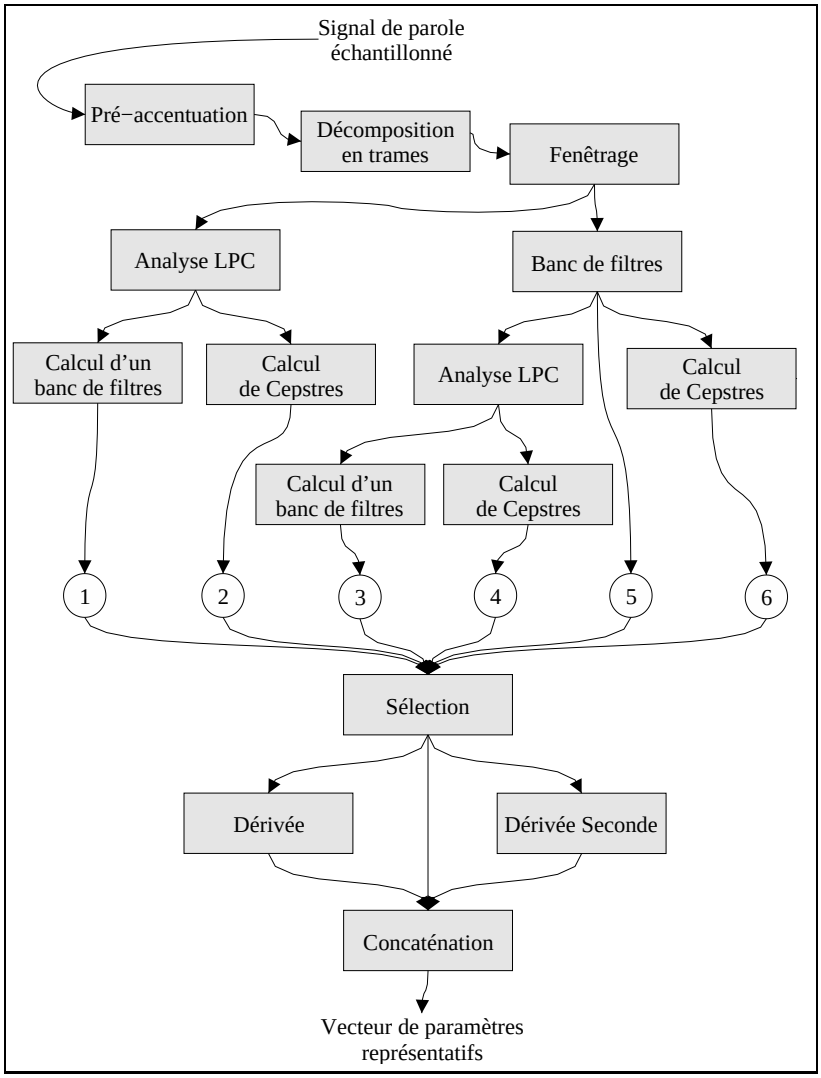


FIG. 2.7 – *Schéma général d'analyse du signal de parole.*

2.7.7 Modèle auditif

Une approche alternative aux méthodes précédentes consiste à s'inspirer du fonctionnement de l'appareil auditif humain. Les modèles qui en résultent comprennent notamment un modèle de cellules ciliées caractérisé par un contrôle automatique de gain, une saturation de l'activité neuronale et un phénomène de masquage temporel. Ces phénomènes n'étaient pas modélisés par les algorithmes utilisés précédemment. Ces propriétés fonctionnelles pourraient être une des raisons de l'efficacité et de la robustesse du système auditif humain. Des deux premières propriétés découle une relative indépendance de la perception par rapport au niveau sonore et au canal de transmission. La troisième propriété, quant à elle, entraîne une certaine robustesse par rapport aux bruits additifs.

Ces modèles, basés sur les travaux de Martens [130, 131], de Gao [65] et de Ghitza [69] ne seront cependant pas étudiés ici.

2.8 Bilan

Un compte rendu récent [118] visait à comparer les performances des systèmes automatiques et des humains. Celui-ci constitue un bilan éclairant des limitations pratiques de la technologie de reconnaissance vocale. Il nous parait important d'en rappeler les conclusions principales. Dans des conditions d'enregistrement claires, les taux d'erreur des systèmes automatiques sont bien souvent jusqu'à dix fois supérieurs à ceux des humains. Dans des conditions de bruits et de distorsions, lorsque les caractéristiques du canal de transmission changent, ou encore pour de la parole spontanée, l'écart se creuse.

Notons que ces résultats peuvent être observés non seulement pour des tâches de reconnaissance complexes mais également pour des tâches ne pouvant faire appel à aucune information linguistique ou sémantique de haut niveau (par exemple reconnaissance de syllabes sans signification "nonsense syllables" ou de chiffres connectés). Ces résultats suggèrent donc que l'*acoustico-phonétique* reste un domaine de recherche important, au même titre que la modélisation du langage et de la sémantique.

Une deuxième direction de recherche identifiée par cet article concerne la ***robustesse*** et la *capacité d'adaptation rapide* des systèmes de reconnaissance aux variations des conditions de bruit et du canal de transmission.

Des expériences de reconnaissance humaine démontrent également que seulement trois syllabes sont nécessaires pour s'adapter à un nouveau locuteur et obtenir des performances similaires à celles obtenues sur base d'un seul locuteur. L'*adaptation au locuteur* est également un sujet de recherche important. En effet, les systèmes d'adaptation actuels ne présentent un intérêt que lorsque plusieurs minutes de données d'adaptation sont disponibles, cette adaptation étant généralement réalisée en mode supervisé, c'est-à-dire en connaissant le texte qui a été prononcé.

Les taux d'erreur élevés obtenus pour la *parole spontanée* constituent une faiblesse supplémentaire. Des expériences concernant la *modélisation du langage* ont également montré que les performances humaines en termes de perplexité sont 3 fois supérieures à celles de modèles tri-grammes.

Finalement, en plus des problèmes précédents, l'auteur considère que les systèmes

de reconnaissance actuels ont deux limitations supplémentaires. Ils sont incapables d'*identifier les mots inconnus* dans le dictionnaire. Ensuite, ils sont incapables de *distinguer les sons environnementaux non stationnaires* des sons de parole.

Tous ces points font bien entendu l'objet de recherches importantes comme en témoignent les articles publiés ainsi que les sessions spéciales qui leurs sont consacrées lors de conférences internationales. Le problème de la robustesse aux bruits environnementaux sera envisagé plus en détail dans la suite de cette thèse.

2.9 Test d'hypothèse

Clôturons ce chapitre par un bref rappel de la théorie des tests d'hypothèse. Ceux-ci seront utilisés pour comparer des approches de reconnaissance vocale. Ils permettent d'estimer à partir de quel niveau une différence de résultats est significative. Les tests d'hypothèse consistent à émettre des hypothèses statistiques concernant certaines populations et à vérifier leur validité. On peut par exemple vouloir vérifier que deux populations sont identiques. On peut également vouloir vérifier qu'elles sont différentes (par exemple en rejetant l'hypothèse qu'elles sont identiques), ou que l'une à une moyenne supérieure à l'autre. On peut associer un niveau de signification α à ces tests (par exemple $\alpha = 0.05$). Ce niveau de signification, qui sera choisi librement, est la probabilité que nous acceptons de faire une erreur de première espèce, c'est-à-dire le fait de rejeter une hypothèse alors qu'elle devrait être acceptée. Les tests d'hypothèse peuvent être bilatéraux si l'on s'intéresse aux valeurs extrêmes des distributions. Ils peuvent également être unilatéraux si l'on ne s'intéresse qu'à une seule branche des distributions, de façon à tester par exemple l'hypothèse qu'un processus est plus performant qu'un autre.

Dans notre cas, il s'agit de comparer des systèmes de reconnaissance vocale sur base des résultats découlant de tests sur un nombre limité de données. La distribution de la variable indiquant si un mot est correctement reconnu peut être supposée binomiale $B(n,p)$ où n est le nombre d'échantillons (nombre de mots) et p la probabilité qu'un mot soit correctement reconnu par le système (moyenne de la binomiale). On sait également que l'écart-type de cette distribution vaut $\sqrt{p(1-p)}$. Deux systèmes différents seront caractérisés par deux binomiales: $B(n,p_1)$ et $B(n,p_2)$.

Les taux de reconnaissance sont estimés par les moyennes d'échantillon $\hat{p}$ sur un ensemble de test de taille n. On peut montrer, suivant la loi forte des grands nombres que la distribution de ces moyennes d'échantillon (pour n suffisament grand) tend vers une normale de moyenne p et d'écart-type $\sqrt{\sigma^2/n}$, où σ est l'écart type de la distribution binomiale, soit $\sqrt{p(1-p)}$. Généralement, on utilisera l'écart-type observé pour estimer cet écart-type de distribution. Sur cette base, on peut émettre l'hypothèse que la moyenne de la distibution de $\hat{p_2}$ est identique à celle de la distribution $\hat{p_1}$, et tenter de la mettre en défaut (en utilisant un test unilatéral) de façon à montrer que les deux taux d'erreur sont différents et donc qu'un des systèmes est meilleur que l'autre. Ce test conduit à la variable réduite suivante:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{(\sigma_1)^2}{n} + \frac{(\sigma_2)^2}{n}}} \tag{2.34}$$

dont la distribution est une normale réduite $N(0,1)$. Il suffit alors de vérifier que la valeur de cette variable réduite est située dans un domaine correspondant à une des extrémités de la normale réduite et caractérisé par une aire égale au niveau de signification choisi. Dans ce cas, on peut affirmer que les deux systèmes qui font l'objet de la comparaison sont significativement différents, avec un niveau de signification α [12].

2.10 Conclusions

Après une description des applications principales de la reconnaissance vocale et de ses domaines connexes, ce chapitre s'est attaché à présenter un bref état de l'art des différentes technologies intervenant en reconnaissance automatique de la parole. L'objectif était de proposer une vue synthétique des éléments à la base des systèmes actuels.

La reconnaissance de la parole est basée sur la représentation du signal de parole sous forme d'une succession de vecteurs de petite taille. Cette représentation est obtenue par **analyse** du signal de parole suivant des méthodes bien connues en traitement du signal ou en traitement automatique de la parole.

Les **modèles de Markov cachés**, avec leurs différentes alternatives en ce qui concerne l'estimation des probabilités des classes phonétiques, restent un outil puissant pour la reconnaissance vocale. La topologie de ces modèles, ainsi que les algorithmes de décodage évolués que l'on trouvera dans la littérature, permettent d'envisager aussi bien le décodage de parole continue que la reconnaissance de mots-clés ou de mots isolés.

Les **modèles de langage**, généralement basés sur des grammaires stochastiques, permettent d'améliorer fortement les performances des systèmes de reconnaissance de parole continue. Ces modèles, dépendant du domaine de langage, imposent des contraintes stochastiques aux hypothèses de reconnaissance.

Finalement, il nous a paru utile de rappeler les principaux thèmes de recherches actuels. Nous avons ainsi pu constater que les recherches concernant la reconnaissance purement acoustique sont tout autant d'actualité que les recherches concernant la modélisation grammaticale et le traitement du langage naturel. La robustesse à des conditions de bruit variables passera obligatoirement par des avancées dans ces deux domaines. Le chapitre 3 s'intéressera au problème de la reconnaissance robuste de la parole.

12. Pour $\alpha = 0.01$, Z doit être supérieur à 2.33 et pour $\alpha = 0.05$, Z doit être supérieur à 1.65

Chapitre 3

Reconnaissance vocale robuste

3.1 Introduction

Si les systèmes de reconnaissance automatique de la parole atteignent de bons niveaux de performances [1] dans des conditions bien contrôlées, il n'en est pas de même lorsqu'ils sont utilisés dans certaines applications, et particulièrement en présence de bruit additif. On constate dans ce cas une chute des performances, limitant l'usage de ces technologies.

De façon générale, l'effet néfaste du bruit additif est double:

- d'une part, il s'ajoute au signal utile et modifie ainsi le contenu spectro-temporel du signal résultant, entraînant des distorsions dans les vecteurs représentatifs extraits d'une analyse de ce signal.
- ensuite, il perturbe le locuteur en s'ajoutant a son retour ("feedback") vocal, ce qui a pour effet de modifier sa façon de prononcer les sons. Schématiquement, le locuteur aura tendance à parler plus fort, voire à crier, en vue d'accroître l'efficacité de la communication. Ce phénomène, qui ne correspond pas simplement à une augmentation du volume sonore (on peux facilement se convaincre qu'élever la voix entraîne également une modification du contenu spectral du signal), est connu sous le nom d'*effet Lombard*. Bien que négligeable dans le cas de bruits présents dans un bureau [37], cet effet pourrait cependant entraîner de fortes dégradations dans des conditions de bruits plus sévères. Il ne fait malheureusement pas l'objet d'importantes recherches. Ce phénomène ne sera pas considéré dans ce travail. Cependant, si on dispose d'une méthode efficace permettant d'obtenir une estimation du signal de parole clair, il serait possible de faire face au problème de l'effet Lombard par entraînement sur des données claires plus ou moins affectées par ce phénomène. De telles

1. Bien que toujours nettement inférieures à celles des humains, les performances des machines ont considérablement progressé ces 10 dernières années. Elles permettent actuellement d'envisager des applications utiles soit dans des conditions d'utilisations très spécifiques, soit basées sur un vocabulaire limité ou finalement destinées à un public très ciblé et conscient des limitations de la technologie.

données peuvent être obtenues par utilisation d'un retour casque bruité (suivant une plage de niveaux de bruits susceptibles d'apparaitre à l'utilisation) lors de l'enregistrement de la base de données. Le base de données *BDBRUIT* disponible par l'intermédiaire d'ELRA [51] propose ce type d'enregistrements vocaux. On peut également envisager une génération automatique de parole stressée sur base de parole naturelle [16]. Dans [156], les auteurs présentent différentes techniques d'amélioration de la robustesse au stress, incluant l'effet Lombard. L'effet Lombard correspond essentiellement à une augmentation du niveau de parole et à un aplatissement du spectre, avec augmentation du contenu haute-fréquence par rapport au contenu basse-fréquence. Dans [94], une mesure du niveau de bruit au début de la phrase est utilisée pour prédire le niveau de parole et y appliquer ainsi une correction. Finalement, une approche d'adaptation non paramétrique pourrait également être envisagée (voir Section 3.6).

D'autre part, il est important de remarquer que bien souvent, un niveau de bruit élevé par rapport au niveau du signal de parole va de pair avec un niveau de *réverbération* élevé. En effet, lorsque le microphone est très proche de la source, l'énergie du signal de parole, capté par le microphone, est très élevée et celui-ci n'est que peu perturbé par les bruits ambiants ou par la réverbération. C'est par contre lorsque le microphone est à plus grande distance (plusieurs centimètres ou dizaines de centimètres) de la source que la réverbération, tout comme le bruit ambiant, peuvent devenir gênants. Comme pour le bruit additif, le problème de l'effet de la réverbération sur les systèmes de reconnaissance automatique de la parole bénéficie d'un engouement important. Signalons quelques travaux récents concernant ce problème [105, 213]. Notons également que les techniques faisant intervenir des réseaux de microphones, par leur capacité à se focaliser dans une direction déterminée (celle de la source utile), ainsi que par l'intermédiaire de techniques de déconvolution aveugles, peuvent fournir une solution élégante au problème posé par la réverbération. Dans la suite de ce travail, on ne considérera plus ce problème de réverbération. On supposera donc qu'on se trouve dans une pièce peu réverbérante, ou à l'air libre, ou bien encore que le microphone est proche du locuteur ou que l'on dispose de techniques permettant de minimiser l'importance du phénomène.

En bref, l'effet d'un bruit additif imprévisible est de conduire à un environnement d'utilisation différent de l'environnement ayant servi à l'entraînement du système. La validité des modèles statistiques intervenant dans le système est donc mise en défaut et conduit à une chute des performances. Insistons sur le fait que ce n'est pas le bruit en lui-même qui est néfaste mais bien le fait que le système soit utilisé dans des conditions de bruit différentes que celles ayant servi à l'entraînement du système. Cela conduit à deux constations importantes:

- si les performances d'un système entraîné sous un certain rapport signal/bruit diminuent lorsque le rapport signal/bruit diminue, on constatera qu'elles diminuent également lorsque le rapport signal/bruit augmente.
- si le bruit est prévisible et pas trop variable, on aura peut-être intérêt à développer le système sur base de données acoustiques enregistrées dans les conditions de bruits réelles ou même volontairement bruitées avec un niveau de bruit équivalent au niveau de bruit d'utilisation. Si le type de bruit est connu

mais que son niveau n'est pas prévisible, il sera également possible d'effectuer l'entraînement du système sur base de **données volontairement bruitées** couvrant une plage raisonnable de rapports signal/bruit. On s'approche ainsi des méthodes d'entraînement multi-locuteurs (voire multi-styles) couramment utilisées. Ces méthodes font l'objet de la Section 3.3.

Malheureusement, cette dernière condition (type de bruit connu) ne sera que très rarement remplie. Comme il est difficilement envisageable de considérer tous les types de bruits susceptibles d'affecter le système de reconnaissance durant son utilisation, il convient de s'intéresser à des approches plus souples.

La littérature foisonne de travaux concernant la parole en milieu bruité. Historiquement, les recherches se sont orientées vers le **débruitage** ("speech enhancement"), dont le but est d'obtenir un signal de parole sensiblement plus intelligible et plus agréable à écouter, notamment en vue de la transmission ou du codage de la parole. Ces méthodes ont par la suite été utilisées dans le cadre de la reconnaissance automatique de la parole. Elle permettent d'espérer une diminution de l'écart entre les conditions d'entraînement et les conditions d'utilisation. Ces méthodes seront discutées à la Section 3.4.

Viennent ensuite les approches concernant uniquement la reconnaissance automatique de la parole. Nous avons indiqué dans l'introduction de cette section que l'effet du bruit est de modifier le contenu spectro-temporel du signal, entraînant des changements dans la distribution des paramètres représentatifs du signal. Par exemple, l'ajout d'un bruit blanc stationnaire entraîne une diminution de la variance de paramètres log-spectraux. Une façon de réduire l'écart entre les conditions d'entraînement et d'utilisation est donc d'obtenir des **paramètres représentatifs intrinsèquement robustes**, c'est-à-dire relativement insensibles à l'environnement. Ceci fait l'objet de la Section 3.5.

Si les méthodes de débruitage s'attachent à transformer les paramètres représentatifs du signal de façon à ce qu'ils s'approchent des paramètres idéaux (transformation de l'espace d'utilisation vers l'espace d'entraînement), une autre classe de méthodes s'attache plutôt à transformer les paramètres de modèles statistiques de façon à les faire coller aux conditions d'utilisation (transformation de l'espace d'entraînement vers l'espace d'utilisation). Ces méthodes d'**adaptation des modèles** font l'objet de la Section 3.6. Notons immédiatement que ce type d'approches concernera essentiellement les modèles génératifs (systèmes à distributions de probabilité discrètes ou continues).

Les méthodes de débruitage et d'adaptation des modèles introduites précédemment sont qualifiées de méthodes paramétriques. Elles font généralement appel à des connaissances quant à l'effet créé par la perturbation sur les paramètres représentatifs ou sur les paramètres de modèles statistiques. On considère par exemple que le bruit est non-corrélé avec la parole et qu'il agit donc par simple addition dans le domaine spectral. Dans le cas d'une méthode de débruitage, il suffira alors d'effectuer une opération de soustraction spectrale. Dans le cas d'une méthode d'adaptation, on compensera les moyennes et les variances des modèles statistiques sur base des statistiques du bruit. Ces connaissances sont utiles et permettent d'envisager une adaptation rapide. Cependant, ces *connaissances sont souvent incomplètes* et peuvent parfois être limitatives. Par exemple, l'hypothèse de non-corrélation ne tient

pas du tout compte des influences possibles de la réverbération et de l'effet Lombard, influences qui peuvent être difficiles à modéliser correctement. De plus, ces méthodes paramétriques font généralement appel à une estimation extérieure de paramètres intervenant dans le modèle de connaissances. On aura par exemple besoin d'une estimation du niveau de bruit. Ici encore, une estimation incorrecte aura un effet néfaste sur les performances du système. Une approche alternative serait de "modéliser" l'effet du perturbateur sous forme d'une transformation "non paramétrique" (linéaire ou non-linéaire) qui pourrait être optimisée sur base d'une petite quantité de données perturbées. Ces méthodes d'**adaptation non-paramétique** font l'objet de la Section 3.7. Elles peuvent être appliquées soit sur l'espace des paramètres représentatifs, où elle s'apparentent à un *débruitage*, soit sur l'espace des paramètres des modèles statistiques, où elles s'apparentent alors à une *adaptation des modèles.*

Dans certains cas, il est peut être préférable d'ignorer certains éléments des vecteurs de caractéristiques pour ne baser les décisions de reconnaissance que sur les éléments qui semblent les moins perturbés (on peut aussi envisager d'exploiter la redondance du signal pour obtenir une version reconstruite des éléments perturbés, sur base des éléments non perturbés). C'est le cas lorsque, suite à un filtrage du signal de parole, sa bande passante est différente de celle qui a été utilisée lors du développement du système. C'est également le cas lors de courtes interruptions du signal ou lorsque le niveau de bruit local (dans le plan spectro-temporel) est bien supérieur au niveau du signal utile. D'autre part, il apparaît que l'audition humaine est relativement insensible à des dégradations de ce type, ce qui suggère qu'elle opère par distinction des portions spectro-temporelles fortement perturbées ou absentes, des autres portions. Ces approches de **reconnaissance partielle** sont étudiées à la Section 3.8.

L'utilisation de sources d'informations alternatives et/ou complémentaires peut également être envisagée. La **reconnaissance audiovisuelle** (cf. Section 3.9) de la parole ([1, 125, 126, 124, 133, 160, 178, 190, 194]...) par exemple, offre certaines potentialités. Plusieurs études ont montré que l'utilisation d'une image du visage du locuteur, ou plus spécifiquement du mouvement des lèvres , en plus de l'acoustique, permet d'améliorer significativement les performances de reconnaissance dans le cas de parole bruitée[2]. De plus, on considère généralement que le mouvement des lèvres comporte une information complémentaire à l'acoustique, qui pourrait éventuellement conduire à une amélioration des performances en parole claire. Par exemple, la discrimination entre les phonèmes /t/ et /p/ peut être plus facile sur base de l'information visuelle que sur base de l'information acoustique. Des études plus approfondies de ces problèmes peuvent être trouvées dans des publications adressant spécifiquement la reconnaissance audiovisuelle de la parole [190, 124]. En anticipant quelque peu, on pourrait également envisager l'utilisation de sources d'information articulatoires, obtenues sur base de systèmes de transduction permettant

2. Ces conclusions sont valides pour les systèmes de reconnaissance automatique mais également pour la reconnaissance humaine. Par exemple, une expérience de perception [169] sur un vocabulaire de sept voyelles montre qu'à un rapport signal/bruit de 0 dB, l'audiovisuel permet d'obtenir un taux de reconnaissance supérieur à 90% (100% en parole claire) alors que le signal acoustique seul conduit à un taux de reconnaissance inférieur à 80% (100% en parole claire). Ceci confirme donc l'intérêt de recherches concernant les aspects visuels et audiovisuels de la reconnaissance de la parole.

des mesures directes des états du système de production humain. Ceci va donc dans le sens d'une connexion directe entre l'homme et la machine. Notons que certains paramètres articulatoires sont déjà utilisés dans le but d'obtenir un aperçu plus précis du fonctionnement du système vocal humain [4]. Ces recherches pourraient éventuellement déboucher sur des **modèles de production plus précis et robustes**, conduisant à un gain en performances des systèmes de reconnaissance utilisant des paramètres représentatifs purement acoustiques.

Une dernière classe de méthodes consiste à mettre **tout en oeuvre en vue d'obtenir un signal le plus clair possible** (voir par exemple [5]). Utiliser un microphone de proximité et/ou directionnel, et isoler physiquement le locuteur et le microphone d'un environnement bruyant sont des solutions à envisager lors de la conception d'un produit utilisant la reconnaissance vocale. Un bref aperçu permettra de se faire une idée plus précise du gain que l'on peut espérer par l'utilisation de microphones directionnels. Les microphones peuvent être caractérisés par un diagramme polaire représentant leur sensibilité dans différentes directions. Les microphones les plus simples sont sensibles à la pression sur un diaphragme et sont donc *omnidirectionnels*. Lorsque le son peut atteindre le diaphragme des deux côtés, on obtient des capteurs sensibles à la différence de pression et une caractéristique en *dipôle*. D'autres solutions de conception physique permettent d'obtenir les caractéristiques classiques de type *cardioïde*, *hypercardioïde* ou même *quadripôle* [5, 140], ces caractéristiques étant très peu sensibles à la fréquence. Pour mieux se rendre compte de la directivité effective d'un microphone particulier, on utilise la notion de facteur de directivité. Celui-ci représente l'atténuation d'un bruit diffus (provenant de toutes les directions) par rapport à un son (utile) arrivant dans la direction de plus grande sensibilité. Pour les microphones de type omnidirectionnel, cardioïde, hypercardioïde et quadripôle, les facteurs de directivité valent respectivement 0 dB, 4.8 dB, 6 dB et 10 dB. Clairement, ceci signifie que si le bruit est diffus, le gain en rapport signal sur bruit sera de 6 dB lorsqu'on utilise un microphone hypercardioïde plutôt qu'un microphone omnidirectionnel, ce qui n'est pas négligeable. Dans cette optique consistant à augmenter la directivité du système de prise de son, on peut faire appel à l'utilisation de **réseaux de microphones** qui permettent de focaliser la prise de son dans une direction précise et de façon adaptative. Quelques références en vrac sur le sujet: [56, 93, 103, 128, 179, 213].

Une autre approche à l'utilisation de plusieurs microphones, qui n'a cependant aucun lien avec les solutions précédentes consiste à utiliser un microphone de référence, non orienté vers la source utile, et permettant d'obtenir soit une estimation du spectre de bruit intervenant dans un processus de soustraction spectrale [93], soit un signal de référence intervenant dans un processus d'égalisation adaptative [82, 104]. Les processus d'égalisation adaptative et de soustraction spectrale peuvent également être combinés [73].

Il est important de remarquer que les différentes classes de techniques envisagées dans cette introduction ne sont pas exclusives. Citons quelques exemples: certains travaux [154, 174] proposent l'utilisation d'une approche de combinaison parallèle de modèles (méthode d' adaptation des modèles) en vue de compenser les distorsions introduites suites à une soustraction spectrale (méthode de débruitage). Comme autre exemple, certains auteurs [49] proposent d'utiliser conjointement la soustrac-

tion spectrale généralisée (débruitage) et une technique de reconnaissance partielle basée sur l'identification de données manquantes. Comme ces auteurs, nous pensons qu'une approche efficace réside peut-être dans l'*utilisation combinée de méthodes différentes*, ayant chacune des domaines d'action, des avantages et des limitations différentes. Il est en effet clair qu'une solution basée sur l'implémentation d'une des approches rappelées ici (et qui font partie de l'état de l'art) ne conduira pas à des performances satisfaisantes dans toutes les situations. Citons par exemple des mesures récentes, effectuées par la société *Babel Technologies* [7], montrant qu'une application de type borne d'information interactive dans un centre commercial peut conduire à un signal de parole de qualité médiocre (rapport signal/bruit proche de 0 dB) lorsque le client est à plus de 50 cm d'un microphone de type cardioïde (voir Section 3.2). Dans ces conditions, et pour une application nécessitant un vocabulaire relativement important, l'utilisation d'une technique de débruitage n'est vraisemblablement pas suffisante. L'utilisation d'un microphone de proximité, d'un réseau de microphones, ou une étude approfondie de la configuration physique de la borne interactive dans son environnement sont peut-être à envisager. Comme dernier exemple, considérons le cas des techniques de débruitage et d'adaptation des modèles. Ces approches utilisent une estimation du niveau de bruit. Or, le problème de l'estimation du niveau de bruit n'est pas simple et les méthodes actuelles permettent au mieux d'obtenir une estimation robuste pour les bruits très stationnaires par rapport à la parole. Par contre, d'autres méthodes comme l'utilisation de paramètres robustes, ou l'approche multi-bande, ne demandent pas nécessairement d'estimation du niveau de bruit. Elles sont cependant inférieures aux méthodes précédentes lorsque le niveau de bruit est connu. La combinaison des deux types de techniques pourrait donc conduire à une robustesse accrue.

L'étendue considérable des travaux dans le domaine de la reconnaissance robuste de la parole (ou même de l'analyse du signal de parole en général) ne nous permet pas d'étudier les détails de chacune des méthodes existantes, ni même d'envisager tous les concepts à la base de ces méthodes. Nous nous contenterons ici de présenter les techniques de base les plus utilisées, celles qui nous semblent les plus prometteuses ainsi que celles qui ont été implémentées et/ou utilisées dans le cadre de ce travail. Le lecteur pourra cependant obtenir une apperçu plus complet de l'état de la recherche en consultant les références citées ici ainsi que les études [74, 159] et les actes d'un congrès récent dédié à la reconnaissance robuste de la parole [153]. Ils font le bilan des recherches dans le domaine, avec de nombreuses références à l'appui.

Dans la suite, le débruitage par soustraction spectrale ainsi que le filtrage de type RASTA seront utilisés comme stratégies de référence.

3.2 Fixons les idées

Avant de passer à la description des différentes méthodes de reconnaissance robuste, nous nous proposons d'envisager deux exemples qui permettrons de fixer les idées en ce qui concerne les rapport signal/bruit que l'on est susceptible d'obtenir dans des conditions très défavorables. Le premier exemple concerne une application de borne interactive dans un centre commercial. Le niveau de bruit mesuré est de

l'ordre de 66 dBA [7]. Comme deuxième exemple, considérons le cas du bruit dans l'habitacle d'une voiture. Dans [5], on trouvera des mesures indiquant des niveau de bruit variant de 60 dBA à 70 dBA pour des vitesses allant de 80 à 130 km/h et avec un microphone directionnel bien placé[3]. Ces deux conditions sont donc assez similaires quant au niveau de bruit perturbateur.

En ce qui concerne le niveau de parole moyen, on peut trouver des valeurs de l'ordre de 70 dBA à 20 cm du microphone [7] ou de 74 dBA à 50 cm du microphone [5]. A ces distances raisonnables, on peut donc s'attendre à obtenir des rapports signal/bruit allant de 5 dB à 15 dB (et pouvant même descendre à 0 dB si le niveau de voix est relativement faible). Un rapport signal/bruit plus élevé peut bien entendu être obtenu en rapprochant le microphone de la bouche de l'utilisateur, aux prix d'une souplesse d'utilisation moindre cependant.

Ces valeurs seront utilisées comme ordre de grandeur pour nos expériences (voir Section 3.10 et autres chapitres).

3.3 Contamination

Cette classe de techniques consiste a contaminer (par bruitage à plusieurs niveaux de bruits différents) la totalité ou une partie du corpus d'apprentissage et à estimer les paramètres des modèles intervenant dans le système ASR sur base de ce corpus [148] contaminé. L'intérêt de cette méthode est de présenter des performances quasi-optimales[4] lorsque le bruit caractérisant les conditions d'utilisation est similaire au bruit utilisé pour contaminer le corpus d'apprentissage. Dans le cas contraire, la méthode a peu d'intérêt. Son domaine d'application est donc limité car il est difficilement envisageable d'effectuer la contamination sur base de bruits diversifiés couvrant toutes les situations qui pourraient être rencontrés lors de l'utilisation.

Cette technique peut bien entendu être combinée librement avec d'autres techniques de reconnaissance robuste de la parole, comme la soustraction spectrale par exemple.

Dans [188], il est proposé d'entraîner un réseau de neurones artificiels sur base de paramètres représentatifs bruités, dans le but d'estimer une version débruitée de ces paramètres représentatifs. L'idée est d'apprendre la relation non-linéaire existant entre les paramètres issus du signal bruités et les paramètres issus du signal clair. Finalement, l'approche constitue une méthode de débruitage.

Ces techniques de contamination sont rappelées ici car elles sont à la base d'une approche originale qui sera proposée au Chapitre 8.

3.4 Méthodes de débruitage

Dans leur article résumant l'état de l'art en débruitage de la parole, Lim et Oppenheim [115] proposent une classification des techniques de débruitage existantes

3. Un microphone omnidirectionnel donnera des niveaux de bruit allant jusqu'à 74 dB.

4. Elle pourra donc servir à établir une borne supérieure des performances qui peuvent être attendues d'un système de reconnaissance robuste

en trois catégories:

- les méthodes faisant appel au caractère périodique de certaines portions du signal de parole,
- les systèmes qui se basent sur une modélisation du signal de parole,
- la soustraction spectrale, les filtres de Wiener et les techniques dérivées.

Dans cet article, les auteurs suggèrent également de comparer les méthodes de débruitage sur base de l'usage qui est fait de connaissances concernant le signal de parole et le bruit perturbateur. Ils signalent ainsi que, plus on tente de modéliser finement le signal de parole, plus il sera possible de le séparer du bruit ambiant. Par contre, cette modélisation fine et les hypothèses qu'elle peut impliquer risquent de rendre le système sensible aux déviations par rapport au modèle. De même, il est possible d'incorporer une information détaillée concernant le bruit additif, au risque de rendre le système inefficace face à des bruits ne respectant pas le modèle utilisé. L'incorporation ou non d'information de ce type est un compromis qui peut être observé dans les systèmes de débruitage. Par exemple, les systèmes utilisant le caractère périodique des portions voisées utilisent une hypothèse qui n'est généralement pas faite dans le cadre des systèmes basés sur la soustraction spectrale.

Dans les paragraphes qui suivent, la classification proposée dans [115] est utilisée pour donner un aperçu des principales méthodes de débruitage.

3.4.1 Utilisation du caractère périodique des portions voisées

Un débruitage des portions voisées peut être effectué grâce à un filtre en peigne. Un filtre en peigne est un filtre dont la réponse impulsionnelle est un train d'impulsions fini. Dans le cas qui nous intéresse, la période de celui-ci est choisie égale à la période du signal vocal. Dans le domaine fréquentiel, ce filtre à la forme d'un peigne atténuant les composantes fréquentielles entre les harmoniques du signal. Dans le domaine temporel, il correspond à calculer une moyenne du signal sur plusieurs périodes [114].

Plusieurs difficultés apparaissent en pratique:

- Un signal de parole n'est jamais parfaitement périodique. De ce fait, le filtre en peigne atténuera également le signal de parole. En pratique, on constate que les harmoniques d'ordre élevé sont élargies et atténuées.
- Si le rapport signal/bruit est augmenté par l'opération de filtrage en peigne, il n'en est rien pour l'intelligibilité du signal vocal, qui est en général affaiblie [114].
- Une détermination précise de la période du signal vocal est requise.
- Cette méthode n'est pas applicable aux transitions entre phonèmes et aux portions non-voisées (fricatives...) où elle risque d'entraîner des distorsions du signal: une localisation des portions purement voisées est donc requise.

Sur base d'un modèle de parole purement périodique, il est montré dans [108] que l'estimateur MAP[5] pour la parole claire est obtenu par application d'un filtre

5. estimateur qui maximise la probabilité a posteriori des paramètres du modèle de parole, étant donnés les observations bruitées et le niveau de bruit connu.

de Wiener (voir Section 3.4.3) suivi d'un filtre en peigne. Lorsque la parole n'est pas purement voisée, l'estimateur optimal devient simplement le filtre de Wiener.

3.4.2 Utilisation d'un modèle du signal de parole

Le modèle autorégressif (voir Section 2.7.3) est l'un des fondements des techniques de traitement de la parole. L'utilisation de ce modèle dans le cadre du débruitage conduit à la procédure de filtrage décrite dans [113]. Elle est itérative et consiste à débruiter le signal en utilisant un filtre de Wiener (voir paragraphes suivants). Ensuite, le modèle autorégressif du signal débruité est calculé. Ce modèle autorégressif sert à construire un spectre hypothétique qui est alors utilisé pour estimer le filtre de Wiener qui interviendra à l'itération suivante.

3.4.3 Soustraction spectrale et approches dérivées

Lorsqu'un signal de parole $x(t)$ est dégradé par un bruit additif $n(t)$, le signal bruité s'obtient suivant l'expression[6]:

$$y(t) = x(t) + n(t). \tag{3.1}$$

Si le bruit n'est pas corrélé avec le signal de parole, la spectre d'énergie d'une trame de signal bruité s'exprime de la sorte:

$$|Y(\omega)|^2 = |X(\omega)|^2 + |N(\omega)|^2 + X(\omega)N^*(\omega) + X^*(\omega)N(\omega) \tag{3.2}$$

où $X(\omega)$ et $N(\omega)$ sont respectivement les transformées de Fourier du signal de parole et du bruit additif. $X^*(\omega)$ et $N^*(\omega)$ sont les complexes conjugués des transformées de Fourier. Malheureusement, $|N(\omega)|^2$ n'est pas connu. On a alors recours à une estimation $|\widehat{N(\omega)}|^2$ de ce spectre, exprimée comme le spectre d'énergie moyen calculé pendant les portions de signal exemptes de parole, ou sur base des méthodes étudiées au Chapitre 4. De même, comme conséquence de l'indépendance stochastique des signaux de parole et de bruit, on suppose que les deux derniers termes de l'expression précédente sont nuls[7].

La soustraction spectrale dans le domaine des spectres d'énergie (aussi appelée soustraction spectrale en puissance), qui s'exprime alors suivant l'expression:

$$|\widehat{X(\omega)}|^2 = |Y(\omega)|^2 - |\widehat{N(\omega)}|^2 \tag{3.3}$$

permet d'obtenir une estimation $|\widehat{X(\omega)}|^2$ du spectre d'énergie du signal de parole claire.

Si le spectre d'énergie est très important pour la perception des sons, il est généralement considéré que la phase n'a que peu d'importance. De ce fait, la reconstruction du signal débruité s'effectue en utilisant la phase du signal bruité, conduisant à l'algorithme schématisé à la figure 3.1.

Le signal est traité par trames successives se recouvrant sur une longueur correspondant à une demi-trame. Une fenêtre de Hanning est appliquée à chaque trame

6. Le signal de parole étant non stationnaire, on considère en fait de courtes fenêtres de signal.
7. Théoriquement, seule l'espérance mathématique de ces termes est nulle.

avant traitement. Le signal débruité est obtenu par addition des trames traitées, opération rendue possible grâce aux propriétés de la fenêtre de Hanning. Remarquons que lorsque le but est d'obtenir des paramètres acoustiques en vue de la reconnaissance vocale, il n'est pas forcément nécessaire de passer par cette phase de resynthèse d'un signal temporel, l'extraction de certains types de paramètres pouvant s'opérer à partir de la représentation spectrale débruitée. De même, l'opération de soustraction spectrale (hormis la phase de reconstruction d'un signal temporel) peut s'effectuer sur base d'une représentation fréquentielle autre que le spectre en énergie du signal. On peut par exemple l'effectuer sur les énergies de bandes de fréquence obtenues par l'intermédiaire d'un banc de filtres quelconque [154, 180]. C'est cette approche qui sera appliquée dans le cadre de ce travail.

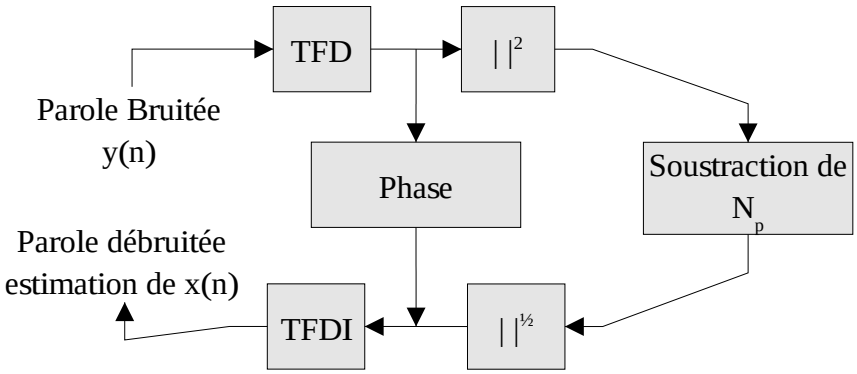


FIG. 3.1 – *Schéma bloc de l'approche de soustraction spectrale. TFD est la transformée de Fourier discrète et TFDI est la transformée de Fourier discrète inverse.*

Une approche différente au problème de débruitage consiste à appliquer au signal bruité un filtre optimal destiné à obtenir une estimation du signal non bruité. Il s'agit en fait d'une approche duale aux techniques de soustraction spectrale, qui peuvent également s'exprimer sous forme de filtrage du signal, la différence résidant dans la façon dont on obtient les paramètres du filtre et dans le filtre ainsi obtenu. On peut exprimer l'estimateur (par soustraction spectrale) du signal non bruité par un filtre à phase nulle dont la réponse en fréquence est la suivante:

$$H(\omega) = \begin{cases} \frac{|\widehat{X(\omega)}|}{|Y(\omega)|} = \sqrt{1 - \frac{1}{REB(\omega)}} & \text{si } 1 - \frac{1}{REB(\omega)} > 0 \\ 0 & \text{sinon} \end{cases} \tag{3.4}$$

$$\text{avec } REB(\omega) = \frac{|Y(\omega)|^2}{|\widehat{N(\omega)}|^2} \tag{3.5}$$

$$\tag{3.6}$$

où REB est le rapport signal d'entrée sur bruit. Cette formulation permet de tracer des courbes de gain et ainsi de nous éclairer sur l'effet du débruitage (voir figure 3.2). Nous y reviendrons par la suite.

Le filtrage de Wiener consiste à obtenir une estimation du signal de parole claire suivant le critère du minimum de l'erreur quadratique moyenne. Cela conduit au

filtre suivant (filtre de Wiener non-causal):

$$H(\omega) = \frac{|X(\omega)|^2}{|Y(\omega)|^2} = \frac{|X(\omega)|^2}{|X(\omega)|^2 + |N(\omega)|^2} \tag{3.7}$$

Malheureusement, ni $|X(\omega)|^2$ (spectre d'énergie du signal clair), ni $|N(\omega)|^2$ (spectre d'énergie du bruit) ne sont connus et le filtre ne peut donc être appliqué tel quel. Généralement, $|N(\omega)|^2$ est remplacé par sa moyenne estimée pendant les portions de silence, tout comme pour la soustraction spectrale. $|X(\omega)|^2$ quant à lui, est estimé par soustraction de l'estimation de $|N(\omega)|^2$ au spectre d'énergie courant. Ceci conduit donc au filtre suivant:

$$H(\omega) = \begin{cases} 1 - \frac{1}{REB(\omega)} & \text{si } 1 - \frac{1}{REB(\omega)} > 0 \\ 0 & \text{sinon} \end{cases} \tag{3.8}$$

$$\tag{3.9}$$

La courbe de gain du filtre de Wiener est présentée à la figure 3.2, en superposition avec la courbe de gain résultant de la soustraction spectrale dans le domaine des spectres d'énergie.

Les filtres de Wiener paramétriques [115]:

$$H(\omega) = \left(\frac{|X(\omega)|^2}{|X(\omega)|^2 + A|N(\omega)|^2} \right)^B \tag{3.10}$$

où A et B sont des paramètres ajustables, permettent d'obtenir des solutions de filtrage de Wiener précisément équivalentes à la soustraction spectrale dans le domaine des spectres d'énergie.

Un des inconvénients de cette méthode de soustraction spectrale/filtre de Wiener est qu'il est nécessaire d'avoir une bonne estimation du niveau de bruit. Nous reviendrons sur ce point au Chapitre 4.

Problème du bruit musical

La soustraction spectrale introduit généralement des "pic" et des "creux" dans le spectre d'énergie du signal. Les "pics", distribués de façon aléatoire sur le spectre du signal, sont perçu comme des sons purs dont les fréquences et les amplitudes varient au cours du temps. Cet effet est donc appelé *bruit musical*. La soustraction spectrale conduit également à des distorsions du signal vocal. Ces effets pourraient avoir un impact néfaste sur les performances d'un système de reconnaissance. Naturellement, on préférera développer les modèles acoustiques sur base de données d'entraînement ayant subi l'opération de soustraction spectrale. Ceci permet de modéliser correctement les distorsions résultant de cette opération. Ces distorsions sont néanmoins gênantes car elles dépendent du niveau du bruit additif. Les sections suivantes présentent différentes approches utilisées pour diminuer l'influence du bruit musical et des distorsions entraînées par la soustraction spectrale. D'autres méthodes de débruitage sensiblement différentes, mais inspirée de la soustraction spectrale, font également l'objet de ces sections.

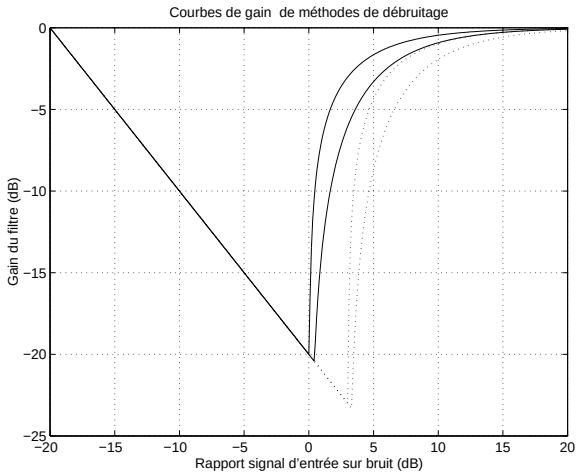


FIG. 3.2 – *Courbes de gain pour la méthode de soustraction spectrale dans le domaine des spectres d'énergie, en traits pleins ($\beta = 0.01$ et α valant 1 ou 2) et pour la méthode de soustraction spectrale de Wiener en pointillés ($\beta = 0.01$ et α valant 1 ou 2). Voir le paragraphe concernant la soustraction spectrale généralisée pour une explication des paramètres α et β.*

Compensation des modèles

Il s'agit ici de considérer les distorsions qui résultent de la soustraction spectrale comme des perturbations pouvant être compensées (voir Section 3.6) au niveau de modèles génératifs (par exemple multi-gaussiens). Dans [154, 174], les auteurs utilisent une approche de combinaison parallèle de modèles (PMC: Parallel Model Combination).

Filtrage des trajectoires spectrales

L'erreur spectrale (cause du bruit musical) est égale à la différence entre le spectre d'énergie de bruit et l'estimation du spectre d'énergie de bruit, utilisée comme terme de soustraction. Cette estimation est généralement une estimation de la moyenne du spectre de bruit. Pour un bruit stationnaire, l'erreur spectrale provient donc de la variance du bruit et peut finalement être réduite en filtrant passe-bas les trajectoires des énergies des bandes de fréquence du spectre du bruit ou du signal. Dans [15] par exemple, le spectre d'énergie du signal bruité est remplacé par sa moyenne calculée sur plusieurs (typiquement 3) trames adjacentes avant d'appliquer la soustraction spectrale.

Filtrage médian

Il s'agit ici de tirer parti du caractère impulsionnel aléatoire du bruit musical. Un filtre médian est appliqué pour chaque fréquence à la sortie du module de soustraction. Ce filtrage consiste simplement à générer une liste ordonnée des valeurs de la fenêtre de filtrage. La sortie du filtre est alors la valeur de l'élément central de cette liste. Ce type de filtrage permet l'élimination d'impulsions de longueur inférieure ou

égale à l ($2l+1$ étant la longueur de la fenêtre de filtrage), et ce, quelle que soit l'amplitude de ces impulsions [116].

Cette approche est une amélioration d'une méthode proposée dans [15] consistant à remplacer l'énergie de chaque fréquence par la valeur minimum choisie sur plusieurs trames adjacentes.

Débruitage Ephraim-Malah

Nous avons constaté que le bruit musical provient essentiellement du fait que les méthodes de soustraction spectrale opèrent par soustraction d'une énergie de bruit moyenne. De ce fait, l'atténuation (en dB) résultant de la soustraction présente des variations importantes d'une trame à l'autre, même pendant les portions exemptes de signal vocal, ces variations étant d'autant plus élevées que la variance du bruit est elle-même élevée. Plus précisément, l'atténuation sera plus faible pour les trames dont l'énergie (ou le rapport signal/bruit) est élevée et plus élevée si l'énergie de la trame est plus faible, avec pour conséquence un augmentation de la variance (en dB) temporelle du spectre d'énergie du signal, conduisant à la perception de bruit musical.

L'approche proposée dans [53] consiste à obtenir une estimation plus consistante du rapport signal/bruit RSB (et par conséquent du rapport signal d'entrée sur bruit REB) par lissage non-linéaire d'estimations successives. Le but étant finalement de réduire les sauts d'atténuation dus aux sauts d'énergie, (essentiellement pendant les portions exemptes de parole). L'approche est caractérisée par l'utilisation d'un rapport signal/bruit a posteriori $RSB_{post}(n,\omega)$ et d'un rapport signal bruit a priori $RSB_{prio}(tn,\omega)$, n étant l'indice temporel de la trame traitée et ω étant l'indice fréquentiel. Le premier est calculé de façon classique sur base du spectre de bruit estimé et sur base du spectre de la trame considérée. $RSB_{prio}(n,\omega)$ par contre est calculé sur base de la trame courante et du résultat du traitement de la trame précédente suivant l'expression:

$$RSB_{prio}(n,\omega) = (1-\alpha) f(RSB_{post}(n,\omega)) + \alpha \frac{(H(n-1,\omega))^2 \, |Y(n-1,\omega)|^2}{|N(n,\omega)|^2} \quad (3.11)$$

avec:

$$RSB_{post}(n,\omega) = \frac{|Y(n,\omega)|^2}{|N(n,\omega)|^2} - 1 \quad (3.12)$$

et où $H(n-1,\omega)$ est le gain de soustraction à la trame précédente, $|Y(n-1,\omega)|^2$ est le spectre d'énergie de la trame précédente et $|N(n,\omega)|^2$ le spectre d'énergie du bruit et α (choisi égal à 0.98 [53]) est un paramètre permettant de pondérer l'influence du rapport signal/bruit de la trame courante et celle de la trame précédente (deuxième terme de l'expression (3.11)). Finalement, la fonction f est définie comme suit: $f(x) = x$ si $x > 0$ et $f(x) = 0$ dans les autres cas. Une analyse de cette expression permet de constater que lorsque le niveau de signal est élevé, on introduit un simple délai d'une trame par rapport a $RSB_{post}(n,\omega)$. Par contre, lorsque le niveau de signal est faible, $RSB_{prio}(n,\omega)$ correspond à une version fortement lissée de $RSB_{post}(n,\omega)$. Les auteurs proposent finalement une expression du gain spectral faisant intervenir ces deux mesures de rapport signal/bruit. Tracé en fonction du

rapport signal/bruit a priori, la courbe de gain présente une allure classique, similaire à celle de la figure 3.2. Elle est cependant modulée par le rapport signal/bruit a posteriori. Lorsque celui-ci augmente, l'atténuation augmente également, réduisant ainsi la variance temporelle du spectre débruité. Les auteurs constatent aussi qu'il est possible d'utiliser avec succès une courbe de gain de type *filtre de Wiener* basé uniquement sur le rapport signal/bruit a priori, présentant moins de fluctuations que la rapport signal/bruit a posteriori. Le *filtre de Wiener* étant préféré par rapport à la *soustraction spectrale dans le domaine des spectres d'énergie* pour son atténuation plus importante (voir figure 3.2).

Atténuation du signal en l'absence de parole

En l'absence de parole, il est possible d'appliquer un gain très faible mais constant, n'introduisant pas de bruit musical [15]. L'inconvénient de cette méthode est de nécessiter une décision logique parole/silence.

Cette approche est à la base des méthodes de débruitage basées sur les modèles de Markov cachés, qui visent à appliquer une opération de débruitage dépendant des états des HMMs. Nous y reviendrons par la suite.

D'autre part, on constatera par la suite que certaines méthodes plus évoluées d'estimation spectrale conduisent également à des courbes de gain (de soustraction) caractérisées par un gain relativement constant pendant les portions de signal à faible énergie (voir figure 3.3).

Soustraction spectrale généralisée

Cette méthode de soustraction spectrale initialement proposée dans [11] consiste à soustraire une surestimation du spectre de bruit (facteur de surestimation $\alpha > 1$). Ensuite, on empêche les composantes spectrales traitées de descendre sous un seuil (seuil spectral). Ce seuil est exprimé comme une fraction (β) du spectre de bruit. La surestimation permet de réduire l'amplitude des pics caractérisant le bruit musical. Le seuil spectral quant à lui permet de remplir les vallées présentes entre ces pics. En pratique, pour un rapport signal/bruit de 0 dB, les auteurs proposent pour α des valeurs entre 3 et 6 et pour β des valeurs entre 0.005 et 0.1. Comme le signalent les auteurs, une valeur de l'ordre de 4 pour α n'est pas spécialement inquiétante. Elle correspond à supposer que la bruit à soustraire est supérieur à l'estimation du bruit d'environ 6 dB. Ce facteur de surestimation représente donc le fait qu'à chaque trame, la variance des composantes spectrales du bruit est pratiquement égale à l'énergie du bruit.

Par conséquent, la soustraction spectrale est implémentée comme suit:

$$|\widehat{X(\omega)}|^2 = \begin{cases} |O(\omega)|^2 & si\ |O(\omega)||^2 > \beta|N(\omega)|^2 \\ \beta|N(\omega)|^2 & sinon \end{cases} \tag{3.13}$$

$$avec\ |O(\omega)|^2 = |Y(\omega)|^2 - \alpha|N(\omega)|^2 \tag{3.14}$$

$$et\ \alpha \geq 1,\ et\ 0 < \beta \ll 1 \tag{3.15}$$

où $|Y(\omega)|^2$ est le spectre d'énergie du signal bruité, $|N(\omega)|^2$ est le spectre d'énergie

du bruit, estimé quand la parole est absente du signal, et $|\widehat{X(\omega)}|^2$ est le spectre d'énergie débruité.

Certains auteurs [200] proposent une implémentation légèrement différente:

$$|\widehat{X(\omega)}|^2 = \begin{cases} |O(\omega)|^2 & si\ |O(\omega)|^2 > \beta|Y(\omega)|^2 \\ \beta|Y(\omega)|^2 & sinon \end{cases} \tag{3.16}$$

$$avec\ |O(\omega)|^2 = |Y(\omega)|^2 - \alpha|N(\omega)|^2 \tag{3.17}$$

$$et\ \alpha \geq 1,\ et\ 0 < \beta \ll 1 \tag{3.18}$$

L'utilisation de valeurs élevées pour α conduit cependant à des distorsions du signal pendant les portions de parole. De manière à réduire ces distorsions, la valeur du facteur α est adaptée d'une trame à l'autre en fonction du rapport signal/bruit local et de façon à être proche de 1 dans le cas où il y a très peu de bruit. La loi suivante est proposée:

$$\alpha = \alpha_0 - RSB/s \quad pour - 5dB \leq RSB \leq 20dB \tag{3.19}$$

$$\alpha = \alpha_0 + 5/s \qquad pour RSB \leq -5dB \tag{3.20}$$

$$\alpha = 1 \qquad pour RSB \geq 20dB \tag{3.21}$$

où α_0 est la valeur désirée de α pour $RSB = 0dB$, RSB est l'estimation du rapport signal/bruit local de la trame considérée et s est choisi de façon à obtenir $\alpha = 1$ lorsque $RSB = 20dB$. La même valeur de α est donc utilisée pour les différentes fréquences.

Dans [120], outre sa dépendance par rapport au rapport signal/bruit, le facteur de surestimation α devient dépendant d'une estimation de la variance du bruit.

Finalement, dans [174], il est montré que l'optimisation de la soustraction spectrale demande également une adaptation du seuil spectral β avec le rapport signal/bruit. Les auteurs proposent une variation linéaire de 0.15 à 1.0 dans la plage de 0 à $20dB$.

Soustraction spectrale et masquage du bruit résiduel

Plutôt que d'utiliser un seuil de soustraction spectrale calculé comme une fraction du niveau de bruit, on utilise ici [187] un seuil constant, masquant le bruit résiduel et permettant de réduire le décalage entre deux conditions d'utilisation différentes. Le choix du seuil est un problème important. S'il est trop faible, le masquage devient inefficace. S'il est trop élevé, il introduit des distorsions au niveau du signal de parole. En pratique, ce seuil de masquage gagnerait à être adapté en fonction du niveau de bruit. L'approche proposée consiste a développer un système basé sur des modèles multiples, chaque modèle étant spécialisé sur un niveau de masquage particulier. Lors de l'utilisation, le modèle le plus approprié pour le niveau de bruit courant est alors sélectionné. L'idée étant que la quantité de bruit résiduel résultant de la soustraction spectrale dépend de la variance du bruit, qui, tout comme le niveau de bruit moyen, peut être estimée pendant les portions de silence.

Estimation dans le domaine log-spectral

L'idée est ici d'utiliser une estimation du spectre de parole claire minimisant l'erreur quadratique moyenne dans le domaine log-spectral [209, 54]. En comparaison aux méthodes classiques de filtrage de Wiener et de soustraction spectrale, qui visent à minimiser l'erreur quadratique moyenne dans le domaine temporel ou dans le domaine du spectre d'énergie, cette nouvelle approche est plus consistante avec le fonctionnement du système auditif humain.

Le but est d'estimer le spectre (dans le domaine logarithmique) de parole claire sur base du spectre de parole bruitée et de statistiques concernant les spectres pour la parole claire et pour le bruit. On effectue généralement cette estimation indépendamment pour chaque fréquence du spectre. L'estimateur minimisant l'erreur quadratique moyenne est la moyenne du spectre clair conditionnée sur l'observation bruitée, ce qui conduit à l'expression suivante (où l'indice fréquentiel est sous-entendu):

$$\widehat{X_l} = E\{X_l|Y_l\} = \int X_l p(X_l|Y_l) dX_l \tag{3.22}$$

où X_l et Y_l sont les logarithmes des spectres d'énergie du signal clair et du signal bruité et où $p(X_l|Y_l)$ est la fonction de densité de probabilité conditionnelle de X_l étant donné Y_l.

Malheureusement, même en faisant des hypothèses sur la forme des distributions de probabilité pour la parole et du bruit, cette expression n'a pas de solution analytique. Dans [209], les auteurs proposent d'établir le lien entre l'estimation $\widehat{X_l}$ et l'observation bruitée sur base d'une simulation de Monte-Carlo. La distribution de probabilité pour le bruit est choisie log-normale et celle pour la parole est log-uniforme. Bien évidement, cette simulation ne vaut que pour certaines valeurs de paramètres des modèles statistiques de la parole et du bruit et doit donc être répétée de façon à couvrir une plage raisonnable de valeurs. Finalement, les tables obtenues sont approximées par une fonction continue faisant intervenir les différents paramètres des modèles et fournissant une estimation de la courbe de gain du système.

Comme on pourra s'en rendre compte à la figure 3.3, cette approche conduit à des courbes de gain dont l'allure est sensiblement différente de celle des courbes de gain caractérisant la soustraction spectrale classique. On observe en effet un gain relativement constant dans la région dominée par le bruit (rapport signal d'entrée sur bruit proche de 0) et que la plage concernée s'élargi lorsque l'écart type du bruit augmente. Ces régions seront simplement atténuées sans introduction de bruit musical. Pour rappel, la soustraction spectrale, de part la pente élevée de sa courbe de gain, conduit quant à elle à une amplification des excursions spectrales (dans le domaine logarithmique), ce qui se traduit par l'apparition de bruit musical, dès que la variance du bruit n'est pas nulle, et d'autant plus élevé que cette variance est grande.

Utilisation d'estimateurs adaptatifs

Très tôt, certain auteurs ont constaté l'intérêt d'utiliser des estimateurs de parole claire (utilisant par exemple la soustraction spectrale) adaptatifs. Dans [15] par exemple, il est proposé d'utiliser un estimateur conditionné sur une détection

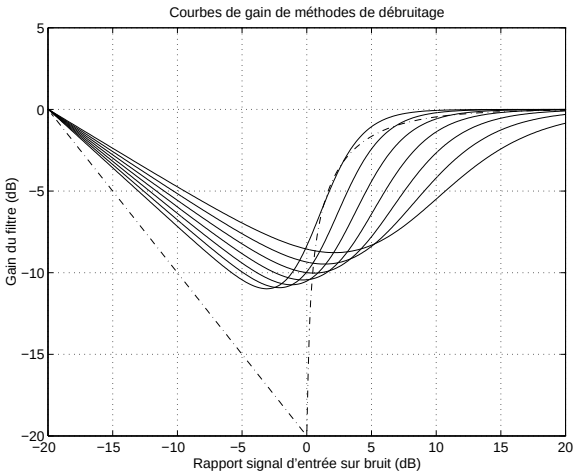


FIG. 3.3 – *Courbes de gain pour la méthode de soustraction spectrale dans le domaine des spectres d'énergie ($\alpha = 1$ et $\beta = 0.01$) et pour la méthode de soustraction spectrale non-linéaire. Pour cette méthode, la moyenne de la distribution pour la parole = 10 dB, l'écart type de la distribution pour la parole = 17.3 dB (comme proposé dans [209]) et l'écart type de la distribution du bruit varie de* 0 *à* 6 *dB, par pas de 1 dB. Le niveau de bruit moyen est choisi comme facteur de normalisation (0 dB). Dans le cas présenté ici, le rapport signal sur bruit moyen est de 10 dB. Pour une valeur différente du rapport signal sur bruit moyen, les courbes seront différentes.*

parole/silence, permettant d'atténuer plus fortement le signal bruité pendant les portions de silence. Dans [136], les auteurs se basent sur la même idée mais vont plus loin en proposant d'estimer la parole claire comme une moyenne pondérée de deux estimations, l'une pour la parole et l'autre pour le silence. Comme coefficient de pondération, on utilise une estimation de la probabilité a posteriori que la trame considérée contienne de la parole.

La caractéristique fondamentale des ces méthodes est de conditionner l'estimation de parole claire sur certains paramètres, comme la détection parole/silence ou la probabilité que la trame contienne de la parole. Ainsi, l'approche de soustraction spectrale généralisée présentée précédemment, avec son paramètre de surestimation dépendant du rapport signal/bruit [11] est également à classer parmi ces méthodes.

Les recherches dans ce sens ont finalement conduit à une méthode très évoluée où l'estimation est conditionnée sur les classes de sons ou sur les différents modes d'une distribution multi-Gaussienne [52]. En schématisant, la méthode consiste à développer un modèle statistique décrivant les différents régimes stationnaires (par exemple un modèle multi-Gaussien ou un modèle de Markov caché). Ce modèle permet alors d'associer un filtre de Wiener (donc optimal au sens des moindres carrés) à chaque régime stationnaire[8]. A l'utilisation, les probabilités a posteriori de chacune des classes, obtenues à partir du modèle statistique, sont utilisées pour

8. bien entendu, ce filtre dépendra également de l'instant considéré, voir pour cela le début de cette Section 3.4.3 décrivant les méthodes de soustraction spectrale.

obtenir une moyenne pondérée des estimations fournies par les différents filtres de Wiener. L'estimation du signal x à partir du signal bruité y est obtenue par:

$$\begin{aligned} \hat{x} &= E\{x|y\} \\ &= \int x p(x|y)dx \\ &= \int x \sum_i p(x|y,q_i)p(q_i|y) \\ &= \sum_i p(q_i|y)E\{x|y,q_i\} \end{aligned} \tag{3.23}$$

où les q_i indiquent les différents régimes stationnaires totalement exclusifs et où $p(x|y,q_i)$ est la fonction de densité de probabilité conditionnelle de x. Le débruitage est effectué de façon itérative, une version débruitée du signal permettant de réestimer les probabilités a posteriori et de lancer l'itération suivante. L'arrêt de la procédure est obtenu grâce à un critère de convergence. La preuve de convergence est fournie dans [52]. Dans [176], les performances de cette méthode ont été évaluées. Les test effectués indiquent un gain de l'ordre de 3 à 6 dB, soit, dans leur cas, 2,5 dB de mieux qu'une méthode basée sur un filtrage de Wiener classique. Au chapitre 8, nous comparerons des résultats obtenus à partir d'une approche originale de décomposition en bandes de fréquences à ceux obtenus sur base de cette stratégie (résultats de la littérature).

Dans [9], les auteurs proposent une approche similaire mais orientée vers la reconnaissance automatique de parole plutôt que vers un simple débruitage. L'idée est d'estimer la probabilité associée à chaque classe de sons (phonèmes) sur base d'une version filtrée du vecteur d'observation, le filtre (de Wiener) est dépendant du phonème considéré. Le décodage s'effectue alors normalement sur base des probabilités ainsi estimées. Notons cependant que cette technique est difficilement applicable aux systèmes utilisant des réseaux de neurones artificiels.

3.5 Paramètres robustes

Ces approches, qui visent à obtenir des paramètres représentatifs moins sensibles au bruit, ont les avantages suivants:

- Elles ne demandent que peu ou pas d'hypothèses concernant la nature du bruit, contrairement à l'approche de soustraction spectrale, par exemple, qui demande une estimation du spectre de bruit.
- Elles ne font pas appel à une phase d'adaptation sur base de données, comme par exemple pour les techniques introduites à la Section 3.7.

Si l'on connait la nature et/ou le type de bruit par contre, les approches de débruitage ou d'adaptation des modèles conduiront généralement à des performances supérieures à celles obtenues grâce aux paramètres robustes.

3.5.1 Normalisation des paramètres représentatifs

Cette approche [201] consiste à normaliser les paramètres spectraux dans le temps par soustraction de la moyenne et division par l'écart type. Ces statistiques sont calculées soit sur base d'un segment de quelques centaines de millisecondes

de parole, soit de façon récursive. Pour une tâche de reconnaissance de mot isoles dépendante du locuteur, les auteurs rapportent des performances supérieures à l'approche de combinaison parallèle de modèles [63] (CPM) utilisée conjointement à un détecteur de parole [59]. Un des avantages de cette méthode par rapport à la méthode CPM est de ne pas nécessiter de segmentation explicite parole/silence ni même de modèle de la perturbation.

Cette approche est également proposée dans le cadre de système utilisant un réseau de neurones artificiels comme estimateur des probabilités des classes phonétiques [193].

3.5.2 Filtrage des variations, Log-RASTA, J-RASTA, dérivées et spectre de modulation

Dans l'espace des paramètres représentatifs, les bruits additifs sont souvent caractérisés par des variations dont les fréquences sont plus faibles que celles du signal de parole. Ce sera le cas pour un bruit stationnaire mais également pour tout bruit quasi-stationnaire variant plus lentement que le signal de parole, par exemple un bruit de moteur. Cette constatation a conduit à l'utilisation de méthodes visant à atténuer l'effet de ces modulations basses fréquences. Par extension, certains auteurs se sont également penché sur l'analyse du spectre caractérisant ces modulations. Pour le signal de parole, ce *spectre de modulation* présente une caractéristique passe bande, conduisant à l'utilisation de méthodes visant à atténuer également les variations haute-fréquences, normalement absentes du signal de parole.

La technique RASTA ("Relative Spectra") [84, 85] consiste à appliquer un filtre passe-bande à chaque composante d'une représentation spectrale en bandes critiques. La fonction de transfert de ce filtre est la suivante:

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{3.24}$$

Le filtre consiste donc à réintégrer (dénominateur) une estimation de la dérivée temporelle (numérateur) de la composante spectrale considérée. Ce filtre atténue les composantes basses-fréquences ainsi que les composantes hautes-fréquences, au delà du spectre de modulation du signal de parole. D'autres types de filtres passe-bande pourraient être utilisés.

Ce filtre, lorsqu'il est appliqué dans le domaine log-spectral (approche *Log-RASTA*), conduit à l'atténuation du bruit de convolution: variations dues au microphone, au canal de transmission et à l'orientation de la tête du locuteur [71]. Pour atténuer un bruit additif, ce filtrage doit être effectué dans le domaine linéaire. Pour traiter le bruit additif et le bruit de convolution simultanément, les auteurs [85] proposent d'appliquer le filtre sur une fonction de la représentation spectrale (approche *J-RASTA*). Cette fonction est quasi-linéaire pour les faibles énergies (c'est-à-dire pour les portions significativement affectées par le bruit additif) et logarithmique pour les énergies plus élevées (significativement affectées par le bruit de convolution). Cette fonction:

$$f(x) = ln(1 + Jx) \tag{3.25}$$

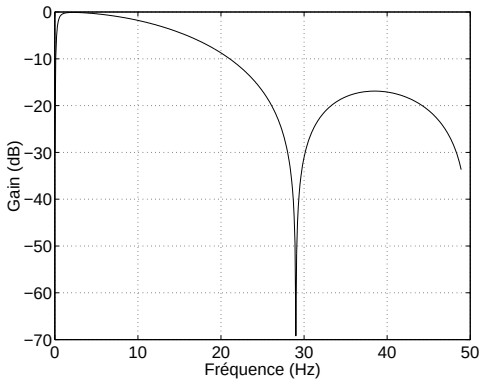


FIG. 3.4 – *Réponse en fréquence du filtre RASTA pour une fréquence d'analyse du signal de 100 Hz (une trame toutes les 10 ms).*

est appliquée à chaque composante spectrale et est contrôlée par le paramètre J qui peut dépendre du niveau de bruit. Lorsque le bruit augmente, la valeur de ce paramètre pourrait diminuer de façon à étendre la portion quasi-linéaire de cette fonction.

Pour traiter conjointement l'effet du bruit convolutif et du bruit additif, d'autres auteurs [184] proposent plutôt d'utiliser une méthode de soustraction spectrale, suivie d'un filtrage de type Log-RASTA.

Une approche similaire à la technique Log-RASTA consiste à supprimer la moyenne au niveau de la représentation cepstrale (CMS - cepstral mean subtraction) [61, 163], ce qui correspond effectivement à supprimer la moyenne dans le domaine log-spectral.

Une autre approche consiste à utiliser des paramètres dynamiques comme les dérivées premières et secondes de chacun des paramètres représentatifs [62, 163]. Les dérivées sont effectivement insensibles aux variations très lentes et entraînent donc un gain en robustesse. Elles augmentent également les performances en parole claire, indiquant que l'information apporté par l'évolution locale des paramètres représentatifs n'est pas négligeable.

3.5.3 Modèle auditif

Plusieurs auteurs [65, 69, 83, 130, 131] ont tenté d'implémenter des mécanismes justifiés par la neurophysiologie du système auditif ou par la psychoacoustique. Ces mécanismes comprennent le filtrage en bandes critiques, une sensibilité dépendant de la fréquence, une compression d'énergie non-linéaire et enfin le phénomène d'adaptation à court-terme des cellules ciliées de l'oreille interne. Ces propriétés fonctionnelles sont vraisemblablement une des raisons de l'efficacité et de la robustesse du système auditif humain. Les travaux dans ce sens visent à l'implémentation plus ou moins fine d'un ou de plusieurs de ces mécanismes.

La représentation en bandes critiques vise à modéliser le filtrage acoustique effectué par la cochlée. Celui-ci entraîne une précision fréquentielle plus élevée en basses fréquences qu'en hautes fréquences. Différentes approches peuvent être envi-

sagées pour obtenir la représentation en bandes critiques. Ainsi, dans [83], l'auteur utilise une échelle de *Bark*. De nombreux auteurs envisagent l'utilisation d'une échelle fréquentielle de *mel* [95, 159]. Enfin, une transformation en ondelettes peut également être utilisée [202]. Actuellement, la plupart des méthodes d'analyse rencontrées dans la littérature utilisent l'une ou l'autre de ces variantes.

La sensibilité du système auditif humain est simulée par une pré-emphase des hautes fréquences [83]. Cette opération de pré-emphase est présente dans la plupart des modules d'analyse. Ensuite, une loi de compression [83, 202] est parfois utilisée dans le but de simuler la relation non-linéaire existant entre l'intensité d'un son et le niveau perçu de ce son.

Un des aspects le moins utilisé des propriétés connues du système auditif est vraisemblablement l'adaptation à court-terme (ou masquage temporel) caractérisant les cellules ciliées. Ces cellules sont très sensibles après une longue période durant laquelle elles n'ont pas été stimulées et inversement, elles deviennent temporairement moins sensibles après une longue période de forte stimulation. De cette propriété découle une implémentation sous forme de contrôle automatique de gain permettant notamment le masquage des faibles signaux qui suivent une stimulation importante. Cette propriété est utilisée dans [65, 69, 130, 131]. Remarquons également que dans ces travaux, contrairement aux autres contributions utilisant certaines des propriétés auditives, les propriétés décrites précédemment sont modélisées de façon plus fine, c'est-à-dire pour coller aux mieux aux données psychoacoustiques et physiologiques. A notre connaissance, l'intérêt d'une telle approche n'a malheureusement pas fait l'objet d'études suffisamment approfondies, pour chacun des éléments intervenant dans le système d'analyse (voir également la Section 2.7.7).

3.6 Méthodes d'adaptation des modèles

3.6.1 Décomposition de HMMs - Combinaison parallèle de modèles

Cette technique, proposée dans [64, 199] notamment, consiste à définir des modèles de Markov cachés pour le signal de parole, mais également pour le bruit. Elle implique donc généralement l'estimation des paramètres des modèles de bruit. Ensuite, des modèles composites sont constitués. Chacun des états de ces modèles correspond à une paire d'états des modèles initiaux. Ainsi, si N est le nombre d'états d'un modèle de parole et si le modèle de bruit est constitué de M états (permettant éventuellement de modéliser des bruits non-stationnaires), le signal de parole bruité peut être modélisé par un modèle de Markov caché composite comprenant $M * N$ états. La reconnaissance simultanée de parole et de bruit peut alors être effectuée grâce à l'algorithme de Viterbi classique sur base de ce HMM composite.

La probabilité d'observation d'un signal résultant de la combinaison de parole et de bruit est évaluée comme suit:

$$P(x_n|M_1 \otimes M_2) \tag{3.26}$$

où x_n est le vecteur d'observation à l'instant n, M_1 et M_2 sont les paramètres des modèles indépendants des deux composantes du signal et $\otimes$ est un opérateur de combinaison indiquant la façon dont les deux signaux se combinent (bruit additif, bruit convolutif...). Par exemple, si le signal observé est composé de parole et de bruit additif non corrélés, et si l'étage de traitement de signal est un banc de filtres générant le niveau d'énergie dans chaque canal, alors, l'opérateur $\otimes$ est une addition pour les moyennes des fonctions de densité de probabilité définissant les modèles. En général, l'observation x_n est donc la combinaison de $x1_n$ et de $x2_n$ qui sont les vecteurs (non observables) correspondant aux deux composantes du signal. Par conséquent:

$$x_n = x1_n \otimes x1_n \tag{3.27}$$

Le décodeur Viterbi classique peut alors être étendu à un processus tri-dimensionnel (qui, rappelons le, peut être vu comme un décodage Viterbi classique utilisant des HMMs composites):

$$P_n(i,j) = \max_{u,v} P_{n-1}(u,v).a1_{u,i}.a2_{v,j}.b1_i \otimes b2_j(x_n) \tag{3.28}$$

où $P_n(i,j)$ est la probabilité, à l'instant n, que la première composante soit dans l'état i et que la seconde soit dans l'état j; $a1_{u,i}$ est la probabilité de transition de l'état u vers l'état i, pour la première composante; $a2_{v,j}$ est la probabilité de transition de l'état v vers l'état j pour la seconde composante[9] et $b1_i \otimes b2_j(x_n)$ est la probabilité d'émission. Comme $x1_n$ et $x2_n$ ne sont pas observables, la probabilité d'émission aura la forme générale suivante:

$$b1_i \otimes b2_j(x_n) = \int P(x1_n,x2_n|i,j)d(x1_n,x2_n), \tag{3.29}$$

l'intégration étant étendue à tous les couples $(x1_n,x2_n)$ qui vérifient $x_n = x1_n \otimes x1_n$. Si les deux composantes ne sont pas corrélées:

$$b1_i \otimes b2_j(x_n) = \int P(x1_n,i).P(x2_n,j)d(x1_n,x2_n) \tag{3.30}$$

Il est bien évidemment impossible d'énumérer tous les couples possibles de l'équation 3.29. Une solution à ce problème est d'estimer la fonction de densité de probabilité d'émission de chaque couple d'état (i,j) du modèle combiné, étant donnés les probabilités d'émission des deux composantes et connaissant le rapport signal à bruit. Bien entendu, la combinaison des deux composantes dépendra de l'espace des paramètres utilisés. Nous obtenons finalement:

$$b1_i \otimes b2_j(x_n) = \int P(x_n|i,j)dx_n \tag{3.31}$$

Nous pouvons également faire l'hypothèse que le signal résultant est soit de la parole, soit du bruit. L'expression (3.29) est alors fortement simplifiée.

Cette technique peut également être utilisée en combinaison avec d'autres approches de reconnaissance robuste. Dans [154], elle est utilisée pour compenser les distorsions introduites par un module de soustraction spectrale.

9. Ces probabilités de transition sont entraînées indépendamment pour les deux composantes.

3.7 Méthodes d'adaptation "non-paramétriques"

Signalons d'abord que ces techniques concernent l'adaptation des modèles ainsi que le débruitage. Ce dernier peut en effet être vu comme une adaptation des paramètres représentatifs.

Comme signalé dans l'introduction, les méthodes d'adaptation faisant l'objet de la section précédente (de même que les méthodes de débruitage traitées précédemment) font appel à des hypothèses concernant l'effet du perturbateur sur les paramètres représentatifs du signal et/ou sur les paramètres des modèles statistiques. Ces connaissances pouvant être imprécises et/ou incomplètes, il est peut-être préférable d'envisager des méthodes consistant à adapter (par exemple sur base d'une transformation linéaire ou non-linéaire) ces paramètres à partir d'une 'petite' quantité de données perturbées, évitant ainsi le recours à des hypothèses erronées[10]. Ces techniques visent donc à obtenir une transformation entre l'espace des paramètres bruités et l'espace des paramètres non bruités. L'apprentissage de cette transformation se fait généralement de façon supervisée sur base d'observations correspondant aux conditions originales et aux conditions bruitées, ou sur base d'étiquettes indiquant l'appartenance des observation bruitées aux différentes classes phonétiques. Ces méthodes peuvent être similaires aux techniques utilisées en adaptation au locuteur.

3.7.1 Régression linéaire

Une première classe de méthodes consiste à effectuer une adaptation sur base d'une transformation, généralement exprimée sous forme de transformation affine. On parle alors de régression linéaire. Cette matrice d'adaptation agit soit sur les paramètres représentatifs, soit sur les paramètres des modèles statistiques. Dans ce dernier cas, il est possible de définir plusieurs matrices de transformation, chacune agissant sur une classe de modèles phonétiques.

Dans le cas d'une adaptation des modèles, les paramètres des matrices d'adaptation peuvent être estimés sur base du critère du maximum de vraisemblance [112]. Soit A l'ensemble des données d'adaptation, soit M la matrice correspondant aux paramètres (fixés) de modèles statistiques et soit μ la matrice d'adaptation. La méthode consiste alors à estimer les éléments de μ suivant:

$$\mu' = \underset{\mu}{\text{argmax}}\ P(A|M,\mu) \tag{3.32}$$

L'algorithme EM est utilisé à cette fin. Les modèles faisant partie de la classe de modèles considérée sont alors adaptés suivant:

$$M' = \mu' M \tag{3.33}$$

Le processus d'estimation peut également être basé sur le critère du maximum de la probabilité a posteriori. On a alors:

$$\mu' = \underset{\mu}{\text{argmax}}\ P(\mu|A,M) \tag{3.34}$$

10. Comme aucune connaissance a priori n'est introduite, ces systèmes impliquent généralement une plus grande quantité de données d'adaptation

et par conséquent:

$$\mu' = \underset{\mu}{\text{argmax}}\ P(A|M,\mu)P(\mu) \tag{3.35}$$

où $P(\mu)$ est la distribution a priori des paramètres de la matrice d'adaptation. Plusieurs alternatives concernant cette distribution a priori sont proposées et discutées dans [181]. Ici aussi, l'algorithme EM est utilisé.

Dans le cadre des systèmes utilisant des réseaux de neurones artificiels, la régression linéaire à également été proposée dans [151] en vue d'adapter les paramètres représentatifs à l'entrée d'un perceptron multicouche. La matrice d'adaptation correspond simplement à une couche d'entrée (linéaire) supplémentaire dont les paramètres sont ajustés par descente de gradient grâce à l'algorithme de rétro-propagation de l'erreur à la sortie du réseau de neurones. Les paramètres du réseau initial ne sont pas modifiés.

3.7.2 Adaptation directe de paramètres

Une autre classe de méthodes correspond à compenser directement les paramètres des modèles par réapprentissage en utilisant le critère du maximum de la probabilité a posteriori (MAP). L'optimisation fait donc appel à des distributions a priori des paramètres des modèles. On peut alors montrer que l'estimateur MAP de la moyenne d'une distribution correspond à une moyenne pondérée de la moyenne non adaptée de la distribution (issue du modèle non adapté) et de l'estimation par maximum de vraisemblance de la moyenne étant donné les vecteurs d'adaptation. Nous n'entrerons pas dans plus de détails ici. Le lecteur intéressé pourra se référer à [66] notamment.

3.7.3 Méthodes d'adaptation rapide

Les méthodes d'adaptation qui font l'objet des sections précédentes sont en fait des méthodes d'apprentissage. Elle utilisent des algorithmes d'entraînement/estimation classiques pour adapter directement ou indirectement (via une transformation) les paramètres des modèles statistiques ou les paramètres représentatifs du signal. L'objectif de ces techniques est de réduire les différences entre les conditions d'entraînement et les conditions d'utilisation. Elles peuvent donc être envisagées pour l'adaptation au locuteur ou même pour l'adaptation au canal de transmission et/ou aux bruits perturbateurs. Les recherches actuelles portent une attention particulière aux méthodes permettant une adaptation très rapide c'est-à-dire sur base d'une très petite quantité de données d'adaptation. On parle ainsi d'adaptation sur base d'une seule phrase, voire d'un seul mot, alors que les méthodes classiques demandent plusieurs minutes de parole. L'idée à la base de ces méthodes est souvent d'utiliser une information a priori concernant les propriétés statistiques des paramètres d'adaptation [111, 181, 46], cette information a priori pouvant provenir d'une analyse des données d'entraînement. Le lecteur trouvera de plus amples informations dans [153].

3.8 Méthodes de reconnaissance partielle

Les contributions au domaine de la reconnaissance de la parole dans des conditions de bruit sont principalement orientées vers deux grandes idées. La soustraction spectrale (et les méthodes dérivées comme le filtrage RASTA) permet de réduire les distorsions entre les conditions d'utilisation et les conditions d'entraînement par soustraction d'une estimation du spectre de bruit. Étant donné un modèle de bruit caractérisant les conditions d'utilisation, les techniques de compensations fournissent des méthodes d'adaptation dynamique des paramètres des modèles statistiques associés aux états des modèles de Markov cachés.

Dans certains cas cependant, il pourrait être préférable de réduire l'importance, voir d'ignorer les composantes des vecteurs de paramètres caractéristiques représentant des régions spectrales fortement perturbées par du bruit additif. De même, les composantes représentant des régions filtrées devraient également être négligées par le processus de classification.

3.8.1 Données manquantes

Dans ce cadre, les composantes filtrées ou bruitées sont qualifiées de **manquantes** (par opposition aux composantes **présentes**). Des études récentes [35, 119] ont tenté de développer une architecture de reconnaissance basée sur ces idées. Les résultats obtenus montrent que, dans certains cas, on peut ignorer jusqu'à 90% de la représentation spectro-temporelle sans impact significatif sur les performances de reconnaissance. Ces travaux sont basés sur le paradigme d'échantillonnage: les fonctions de densité de probabilité des états des HMM sont connues. Des distributions multi-gaussiennes ("Gaussian Mixture Models" - GMMs) ont été utilisées pour représenter ces distributions de probabilité. Clairement, cela permet de calculer les vraisemblances associées aux états sur base des distributions marginales associées aux composantes présentes, fournissant donc une façon d'ignorer les composantes manquantes.

Cependant, les réseaux de neurones artificiels, combinés avec les capacités de modélisation des HMMs, présentent une alternative intéressante aux méthodes basées sur des distributions multi-gaussiennes (voir la Section 2.5.2). Les réseaux de neurones artificiels, contrairement aux modèles multi-gaussiens, ne permettent cependant pas de traiter les composantes manquantes.

Dans ce cas, on peut envisager une reconstruction des composantes manquantes, comme dans [43]. Les valeurs reconstruites sont calculées comme la moyenne des composantes manquantes, étant donné les composantes présentes. Des fonctions de densité de probabilité simples (comprenant un nombre limité de paramètres) sont utilisées pour modéliser les données, permettant donc un calcul rapide des moyennes conditionnelles. Les vecteurs reconstruits sont utilisés comme entrées des réseaux de neurones artificiels.

Formalisme

Les vecteurs d'observation x sont supposés indépendants et le processus générateur est supposé stationnaire. Ils sont distribuées suivant une fonction de densité de pro-

babilité constituée de K gaussiennes multidimensionnelles caractérisées par les paramètres suivants: w^i, le poids associés à la gaussienne i, μ^i, sa moyenne et C^i, sa matrice de covariance. Certains éléments du vecteur x sont identifiés comme manquants et x peut alors être réorganisé comme suit:

$$x = (x_p x_m), \tag{3.36}$$

x_p pour les composantes présentes et x_m pour les composantes manquantes. De la même façon, on peut réorganiser les éléments des vecteurs de moyenne et des matrices de covariance caractérisant la fonction de densité de probabilité de x:

$$\mu^i = (\mu^i_p \mu^i_m), C^i = \begin{bmatrix} C^i_{pp} C^i_{pm} \\ C^i_{mp} C^i_{mm} \end{bmatrix} \tag{3.37}$$

Nous souhaitons reconstruire un vecteur d'observation complet sur base des composantes présentes. La reconstruction sera basée sur la distribution conditionnelle des composantes manquantes étant données les composantes présentes. Cette distribution est de forme gaussienne. Les éléments reconstruits seront les moyennes de cette distribution, c'est à dire, pour la gaussienne i:

$$x^i_{m|p} = \mu^i_m + (C^i_{pm})^t (C^i_{pp})^{-1} (x_p - \mu^i_p) \tag{3.38}$$

En considérant la distribution multi-gaussienne, la valeur reconstruite est calculée comme suit:

$$x_{m|p} = \frac{\sum_{i=1}^{K} w^i \phi(x_p, \mu^i_p, C^i_{pp}) x^i_{m|p}}{\sum_{i=1}^{K} w^i \phi(x_p, \mu^i_p, C^i_{pp})} \tag{3.39}$$

où w^i est le poids associé à la distribution $\phi(x_p, \mu^i_p, C^i_{pp})$. Ce terme permet de pondérer les contributions des différentes gaussiennes étant donné la position des données présentes dans l'espace des paramètres.

Comme signalé ci-dessus, une approche alternative est d'ignorer les composantes manquantes et de n'utiliser que les composantes présentes pour calculer les vraisemblances des états des HMMs sur base de leurs distributions marginales. Pour un classificateur paramétrique génératif utilisant des distributions multi-gaussiennes, c'est possible, bien que cela implique une charge de calcul importante. En effet, des inversions de matrice de covariance sont nécessaires chaque fois qu'un changement est observé dans la répartition "données manquantes/données présentes". Cela serait également possible avec un classificateur utilisant des réseaux de neurones artificiels, bien que fort lourd également: il convient en effet de disposer d'autant de réseaux de neurones qu'il y a de configurations "données présentes/données manquantes".

Les expériences décrites dans [35] concernent une classification sur base de distributions multi-gaussiennes. Leurs résultats vont en faveur de l'approche par distributions marginales. Elle conduit à des performances sensiblement meilleures que celles de l'approche de reconstruction.

Cependant, cette dernière approche possède certains avantages qui n'ont pas encore été exploités. D'une part, elle permet d'utiliser un nombre très limité de gaussiennes dans la phase de reconstruction. Le système peut donc être fort compact (impliquant un nombre très limité d'inversions de matrice), sans dommage important

pour les performances globales, du moins pendant les portions de parole claire. D'autre part, l'approche permet d'obtenir des vecteurs reconstruits qui peuvent être utilisés comme entrées à n'importe quel système de reconnaissance automatique de la parole. Dans [43], nous nous sommes intéressé à un système hybride HMM/ANN.

La détection des composantes manquantes est un problème important dans le cadre de cette approche. Si le bruit est stationnaire, on peut simplement estimer le niveau de bruit sur base des trames qui précèdent les phrases prononcées. Un seuil sur le rapport signal/bruit sert alors à identifier les composantes manquantes. Dans [43], le module chargé de la détection des portions manquantes est basé sur une estimation automatique du niveau de bruit. Les approches d'estimation du niveau de bruit (voir [48] et le Chapitre 4) ainsi que l'analyse de scènes auditives [50] sont d'un intérêt particulier dans ce cadre.

3.8.2 Reconnaissance multi-bande

Il s'agit d'une approche faisant intervenir une décomposition en bandes de fréquence. Son principe est le suivant:

- Analyse, estimation de probabilités phonétiques et éventuellement catégorisation phonétique dans différentes bandes de fréquence, sur base de "sous-reconnaisseurs" dont l'implémentation repose sur des méthodes classiques.
- Recombinaison des résultats d'analyse, d'estimation ou de catégorisation correspondant aux différentes bandes de fréquence et intégration sur des intervalles de temps plus larges.

Cette approche fait l'objet du Chapitre 6.

3.9 Reconnaissance audiovisuelle et multimodale de la parole

L'utilisation de sources d'informations alternatives et/ou complémentaires peut également être envisagée. La reconnaissance audiovisuelle par exemple, offre certaines potentialités. Plusieurs études ont montré que l'utilisation d'une image du visage du locuteur, ou plus spécifiquement du mouvement des lèvres, en plus de l'acoustique, permet d'améliorer significativement les performances de reconnaissance dans le cas de parole bruitée.

Cette approche fait l'objet du Chapitre 7.

3.10 Comparatif - Influence du bruit

Des tests préliminaires ont été effectués en vue: (1) de souligner l'effet néfaste du bruit sur la reconnaissance automatique de la parole et (2) de comparer divers types de paramètres représentatifs. Pour toutes ces expériences, nous avons utilisé des données d'entraînement non bruitées et des données de test bruitées artificiellement suivant différents rapports signal/bruit.

Des systèmes de reconnaissance ont été développés sur base du corpus NUMBERS'93, correspondant à une tâche de reconnaissance de séquences de nombres en

anglais, sur ligne téléphonique. Les modèles de mots sont construits a partir de 33 modèles de Markov cachés représentant les phonèmes intervenant dans le vocabulaire de nombres. Les transcriptions de ces mots sont issues du dictionnaire *CMU 0.4*. La durée minimum des HMMs représentant chacun des phonèmes est égale à la moitié de la durée moyenne de ces phonèmes. Le système de classification phonétique est un perceptron multicouche (approche hybride HMM/ANN) comportant 400 noeuds cachés [11].

Différentes techniques d'analyse ont été comparées. Elles sont toutes basées sur le banc de filtres de l'approche PLP (voir la Section 2.7.4). Les paramètres basés sur ce type d'analyse (ainsi que les paramètres MFCC) semblent [135] surpasser les paramètres de type LPC pour la reconnaissance de parole bruitée: la tâche envisagée par l'auteur concernait la demande d'information pour les voyages en train. Sur base de la sortie du banc de filtres PLP, deux jeux de paramètres représentatifs ont été considérés:

- **CBE ("Critical Band Energies")**: les sorties du banc de filtres sont utilisées telles quelles. Ces sorties correspondent aux énergies dans 15 bandes critiques. Nous normalisons cependant ces paramètres par rapport à l'énergie de la trame (énergie mesurée après pré-traitement comme la somme des énergies des bandes critiques -voir plus bas-) et appliquons finalement une compression par racine cubique.
- **PLP ("Perceptual Linear Prediction")**: les énergies des bandes critiques (après compression par racine cubique) sont utilisées pour calculer des coefficients cepstraux sur base d'un modèle autorégressif d'ordre 10.

Les dérivées premières et secondes de ces paramètres ainsi que les dérivées premières et secondes de l'énergie de la trame sont adjointes au vecteur de paramètres représentatifs. De plus, neuf trames consécutives sont utilisées à l'entrée du perceptron multicouche. Ces choix font partie de l'état de l'art. Plusieurs approches de débruitage ont ensuite été envisagées, le but étant d'obtenir des paramètres de type **CBE** ou **PLP** plus robustes:

- **SPS ("Spectral Subtraction")**: une soustraction spectrale en puissance est appliquée aux 15 bandes critiques. L'approche de soustraction spectrale généralisée a été appliquée avec avec un facteur de surestimation α égal à 2 [12] et un facteur de seuil β égal à 0.001 [13]. Le spectre de bruit intervenant dans la méthode est estimé par moyenne sur les 10 premières trames de chaque phrase.
- **Log-RASTA ("Log Relative Spectra")**: l'approche log-RASTA est appliquée aux 15 bandes critiques. Cette approche est destinée a atténuer les composantes basse fréquence dans le domaine logarithmique et donc à réduire l'influence du canal de transmission. Dans le cas de bruit additif, elle conduira a une normalisation du signal mais introduira également des distorsions. Nous verrons cependant que globalement, son effet est favorable et entraîne une robustesse accrue.

11. Augmenter la taille de la couche cachée au delà de 400 n'apporte aucune amélioration.

12. Dans la littérature, on trouvera des valeurs allant de 1 à 5. Pour la tâche considérée, la valeur de 2 est celle qui fournit les meilleurs résultats parmi les valeurs 1, 2, 3, 4 et 5, et ce quel que soit le niveau de bruit.

13. Valeur habituellement rencontrée dans la littérature.

- **J-RASTA ("J Relative Spectra")**: L'approche J-RASTA est appliquée aux 15 bandes critiques. Elle est spécialement destinée à réduire l'influence du bruit de convolution ainsi que celle du bruit additif variant plus lentement que le signal de parole. Dans nos expériences, la valeur de J optimale est 10^{-6}, quel que soit le niveau de bruit.

Les versions débruitées des paramètres CBE sont ensuite utilisées pour obtenir des paramètres PLP. Dans les résultats qui vont suivre, nous noterons par exemple $PLP - SPS$ les paramètres obtenus par application de la soustraction spectrale et calcul de cepstres sur base des énergies des bandes critiques.

3.10.1 Résultats

Les données de test ont été bruitées artificiellement par ajout d'un bruit blanc gaussien à 5 rapports signal/bruit (SNR) différents. Le bruit est ajouté au signal de parole de façon à ce que l'énergie moyenne de chaque phrase soit SNR dB au dessus de l'énergie moyenne du bruit. Si certaines phrases de la base de données ont une énergie moyenne plus faible que les autres, le niveau de bruit ajouté à ces phrases sera plus faible également, de façon à conserver un rapport signal/bruit constant entre les différentes phrases. Cette approche, utilisée tout au long de cette thèse, a tendance à sous-estimer le niveau d'énergie du signal de parole et donc à sous-estimer le rapport signal/bruit. Comme les portions de silence représentent environ 30% des bases de données, et si l'on néglige l'énergie de ces portions, cette sous-estimation est de l'ordre de 1.5 dB seulement. Les résultats sont repris à la table 3.1 et à la figure 3.5. Le taux d'erreur au niveau du mot est défini comme suit:

$$\text{taux d'erreur} = \frac{S + I + D}{N} \tag{3.40}$$

où N est le nombre de mots de l'ensemble de test, S est le nombre d'erreurs de substitution, I est le nombre d'insertions et D est le nombre de suppressions. Les intervalles de confiance pour des taux d'erreur de 10%, 30% et 50% sont $\pm 1.6\%$, $\pm 2.5\%$ and $\pm 2.7\%$ respectivement ($\alpha = 0.05$).

On peut remarquer les dégradations importantes causées par l'ajout de bruit. On constatera également l'effet favorable des techniques de "débruitage", avec des performance similaires pour la soustraction spectrale et le J-RASTA. L'utilisation de cepstres de prédiction linéaire (PLP) ne conduit pas toujours à des améliorations significatives par rapport à l'utilisation des énergies des bandes critiques (CBE). On constatera également que le calcul de cepstres par prédiction linéaire sur base des bandes critiques débruitées par soustraction spectrale (PLP-SPS) conduit à une importante dégradation. Elle provient vraisemblablement des distorsions (bruit musical) causées par la soustraction spectrale. Ces distorsions, prises en compte par le modèle autorégressif, conduisent à une modélisation incorrecte des formants.

Des résultats concernant divers bruits réels seront présentés à la Section 8.

3.11 Conclusions

Dans ce chapitre, nous nous sommes intéressé aux techniques de reconnaissance robuste de la parole. Après un rappel des problèmes que pose le bruit dans le cadre

Rapport S/B (dB)	clair	20	15	10	5	0
CBE	13.2	16.2	24.2	44.7	73.8	83.6
PLP	11.0	14.4	22.0	42.6	73.3	83.0
CBE-Log-RASTA	12.4	15.8	18.6	25.4	41.1	65.5
PLP-Log-RASTA	**10.9**	14.5	17.0	24.8	42.5	68.5
CBE-SPS	12.3	**12.7**	**14.0**	**16.3**	**23.6**	**35.8**
PLP-SPS	12.1	14.3	18.4	23.9	31.2	43.3
CBE-J-RASTA	12.8	13.9	15.7	18.5	26.7	42.6
PLP-J-RASTA	11.9	13.6	15.0	18.9	27.4	45.3

TAB. 3.1 – *Taux d'erreur au niveau du mot (%) pour une tâche de reconnaissance de nombres connectés. Comparaison entre différents paramètres représentatifs. Influence d'un bruit additif: bruit blanc gaussien à différents niveaux.*

de la reconnaissance automatique de la parole, l'introduction du chapitre présente un aperçu critique très général des différentes méthodes traitées dans la littérature. Les autres sections fournissent l'information nécessaire à la compréhension des principes à la base des méthodes qui nous paraissent les plus importantes. On pourra utiliser les références données ici pour une étude plus approfondie des subtilités de chacune de ces méthodes.

Rappelons que ces méthodes ne sont en général pas exclusives. Par exemple, l'utilisation d'un réseau de microphones ne pourra éliminer l'intégralité des bruits ambiants. Une technique de débruitage pourrait alors se révéler fort utile, comme c'est le cas lors d'une acquisition utilisant un seul microphone.

A coté des techniques classiques de débruitage, d'adaptation et de reconnaissance partielle, l'utilisation de sources d'informations alternatives est également à envisager. Par exemple, les recherches en traitement d'image ont conduits à des algorithme d'analyse du mouvement des lèvres, fournissant des paramètres représentatifs qui peuvent être utilisés avec succès pour la reconnaissance de parole en milieu bruité[14].

Nous avons également constaté que pour certaines applications, il ne faut pas négliger l'aspect purement matériel de la conception: le choix du microphone et de la configuration physique du système peuvent peser fortement sur la qualité du signal destiné au système de reconnaissance.

Finalement, des résultats de reconnaissance illustrant l'influence d'un bruit additif ont été présentés à la Section 3.10. Ces résultats, qui seront utilisés ultérieurement, visent également à montrer l'intérêt de quelques unes des méthodes de reconnaissance robustes traitées dans ce chapitre.

14. Il semble que l'aspect visuel a également une importance non négligeable pour l'homme.

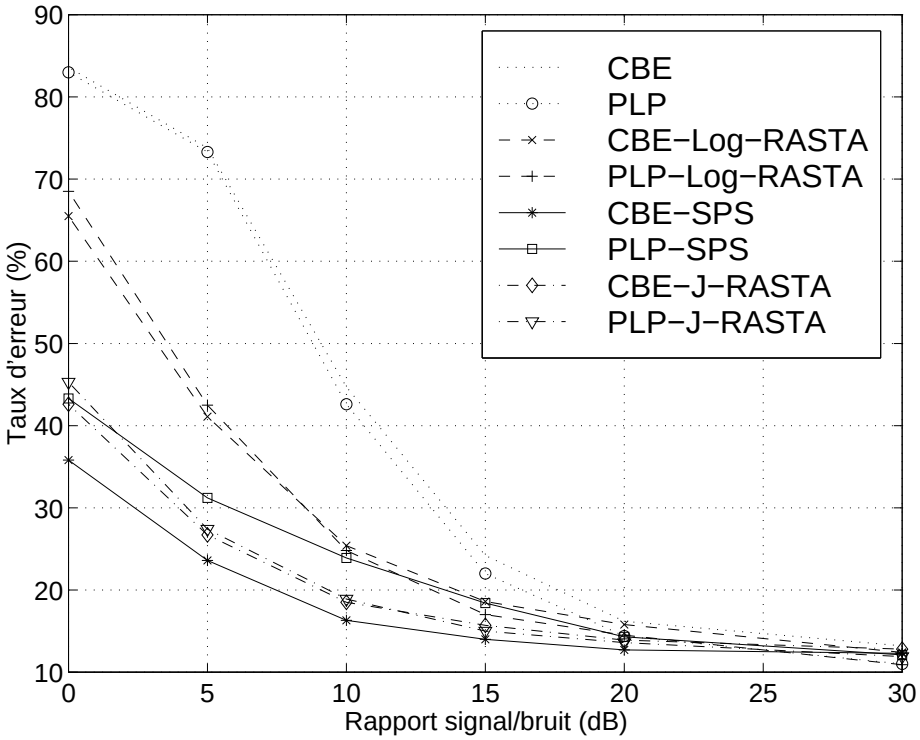


FIG. 3.5 – *Idem table 3.1*

Chapitre 4

Estimation du niveau de bruit

4.1 Introduction

Nous décrivons et testons ici différentes méthodes d'estimation du niveau de bruit. Une méthode originale a notamment été développée. Son principe est d'utiliser le caractère harmonique de certaines portions de parole pour estimer le spectre des bruits large-bandes. Les expériences rapportées ici confirment l'intérêt de la méthode lorsque les bruits sont non-stationnaires, notamment dans le cadre d'un débruitage par soustraction spectrale.

Ce chapitre synthétise les méthodes utilisées et les résultats obtenus. Pour une description plus détaillée, le lecteur se réfèrera à [168].

4.2 Intérêt de l'estimation du niveau de bruit

Plusieurs méthodes visant à augmenter la robustesse des systèmes de reconnaissance automatique de la parole requièrent une estimation locale du spectre du bruit perturbateur. C'est le cas des méthodes basées sur la soustraction spectrale par exemple. Pour les méthodes d'adaptation des HMMs, il convient même d'obtenir un modèle statistique (simple) du signal de bruit. Enfin, les méthodes de reconnaissance partielle exigent une localisation des composantes fréquentielles fortement perturbées. Cette localisation peut également passer par une estimation du spectre de bruit.

Toutes les méthodes pré-citées ont une importance considérable dans la littérature, soit comme approches de référence, soit comme sources de recherches plus approfondies. Malgré cela, on trouvera très peu de contributions relatives à l'estimation du spectre de bruit ou des paramètres d'un modèle de bruit. Ce point nous parait cependant fondamental. La difficulté du problème réside dans le caractère variable et imprévisible des bruits pouvant perturber la communication vocale. Bon nombre de travaux fournissent des résultats pour du bruit stationnaire dont le spectre est estimé sur base des trames initiales de chacune des phrases à reconnaître. Cette approche ne convient bien évidemment que pour les bruits stationnaires. D'autres

travaux sont plutôt orientés vers une détection parole/silence. Le spectre de bruit est alors mis a jour pendant les portions de silence [110, 137, 173].

Finalement, d'autres auteurs proposent des méthodes ne nécessitant pas de détection explicite des zones de silence [18, 90, 132]. Ces approches permettent de traiter certains bruits non-stationnaires. On suppose cependant que les bruits sont plus stationnaires que la parole, de sorte que le bruit peut être considéré comme stationnaire sur des segments de parole correspondant à plusieurs trames d'analyse consécutives. Plusieurs méthodes de ce type sont étudiées, développées et comparées ici.

4.3 Méthodes d'estimation du niveau de bruit

Les méthodes décrites ici sont basées sur l'utilisation d'une discrétisation en bandes de fréquence relativement étroites, permettant d'obtenir une bonne estimation du spectre de bruit. Chaque bande de fréquence est traitée indépendament des autres par analyse d'un segment de parole de plusieurs centaines de millisecondes.

4.3.1 Méthode d'estimation de Hirsch

Décrivons ici la méthode développée par Hirsch [89]. Cette méthode utilise des histogrammes d'énergie construits pour différentes bandes de fréquence du signal sur des segments de signal suffisament longs (plusieurs centaines de millisecondes). Elle est basée sur une constatation simple: la valeur la plus fréquente de l'énergie (maximum de l'histogramme) correspond souvent au niveau de bruit dans la bande considérée. Ceci est du au fait (1) qu'un segment de signal suffisament long comprend généralement de longues portions exemptes de parole et (2) que le bruit caractérisant ces portions est généralement plus stationnaire que le signal de parole lui-même.

Cependant, les histogrammes n'ont pas toujours ce comportement idéal. Il existe de nombreux cas où la valeur la plus fréquente de l'énergie est bien supérieure au niveau de bruit. Ce phénomène apparait souvent pour les basses fréquences où l'énergie de la parole prend souvent des valeur élevées. Pour remédier à ce problème, on utilise un niveau de bruit maximum admissible pour la bande considérée. Ce maximum est calculé comme la moyenne des minima d'énergie (du segment de signal) multipliée par un facteur constant suffisament faible mais supérieur à 1. La valeur la plus fréquente de l'energie jusqu'à ce maximum est alors prise comme estimation du niveau de bruit. Additionnellement, le niveau de bruit est limité à l'énergie moyenne dans la bande de fréquence considérée.

Le lecteur trouvera plus de détails sur cette méthode dans le rapport suivant: [89].

4.3.2 "Clustering" des énergies

Tout comme la méthode de Hirsch, cette méthode est basée sur l'observation des histogrammes d'énergie pour différentes bandes de fréquence. Ces histogrammes peuvent être utilisés pour estimer le niveau de bruit ainsi que le niveau de parole du signal acoustique. En effet, ils semblent présenter des propriétés systématiques

lorsqu'ils sont calculés sur des segments de signal suffisamment longs (plusieurs centaines de millisecondes), contenant de la parole et des instants de silence:

1. Tout histogramme, qu'il soit relatif à une bande contaminée par du bruit ou non, contient généralement deux modes:
   (a) Un mode à basse énergie (de moyenne E_1) représentant la contribution des trames de silence, avec éventuellement du bruit additif.
   (b) Un second mode correspondant aux plus hautes énergies (de moyenne E_2) représentant la contribution de la parole, éventuellement contaminée par du bruit.
2. Généralement, le mode à basse énergie est plus élevé et a une variance plus faible que le mode à haute énergie [1]. La raison en est que, dans les cas envisagés, le silence (ou le bruit) est plus stationnaire que le signal de parole.
3. Ces deux modes sont clairement séparés l'un de l'autre dans le cas de parole claire [2]: voir par exemple la figure 4.1.
4. Lorsque l'on ajoute du bruit dans la bande observée, ces deux modes se rapprochent, comme illustré au bas de la figure 4.1, pour finalement se joindre en un seul mode.

Il est alors possible de modéliser la distribution des énergies d'une bande de fréquence par une modèle comprenant deux modes, par exemple deux gaussiennes. On utilisera l'algorithme EM pour ajuster les paramètres de ces gaussiennes suivant le critère du maximum de vraisemblance.

Une alternative consiste à utiliser un algorithme de "clustering" à deux centroïdes permettant d'estimer les moyennes des deux modes de la distribution. La moyenne inférieure correspond au niveau de bruit et la moyenne supérieure correspond au niveau de parole bruitée.

4.3.3 Suivi d'enveloppe

La méthode décrite dans cette section est basée sur un suivi automatique de l'enveloppe inférieure de l'énergie du signal. Comme les méthodes précédentes, elle repose sur l'estimation de l'énergie dans différentes bandes de fréquence et sur des segments temporels de l'ordre de quelques centaines de millisecondes. Les minima d'énergie correspondent aux portions de signal exemptes de parole et donc à des fenêtres d'analyses ne contenant que du bruit. La moyenne de ces minima est utilisée comme estimation du niveau de bruit dans la bande de frequence considérée. L'utilisation d'une discrétisation en bandes de fréquence relativement étroites, quant à elle, permet d'obtenir une bonne estimation du spectre de bruit.

L'hypothèse à la base de cette méthode est que le bruit est stationnaire sur le segment temporel considéré et que celui-ci contient des portions sans parole. De ce fait, le mode correspondant au bruit présente une variance très faible et sa moyenne peut être estimée par une moyenne des minima d'énergie.

Si l'analyse est basée sur des séquences de N fenêtres, une diminution de la valeur de N permet de diminuer le délai du système ainsi que de suivre des bruits fortement

1. Rappellons que cette observation est à la base de la méthode d'estimation de Hirsch
2. Cette propriété est parfois utilisée pour faire de la distinction parole/silence.

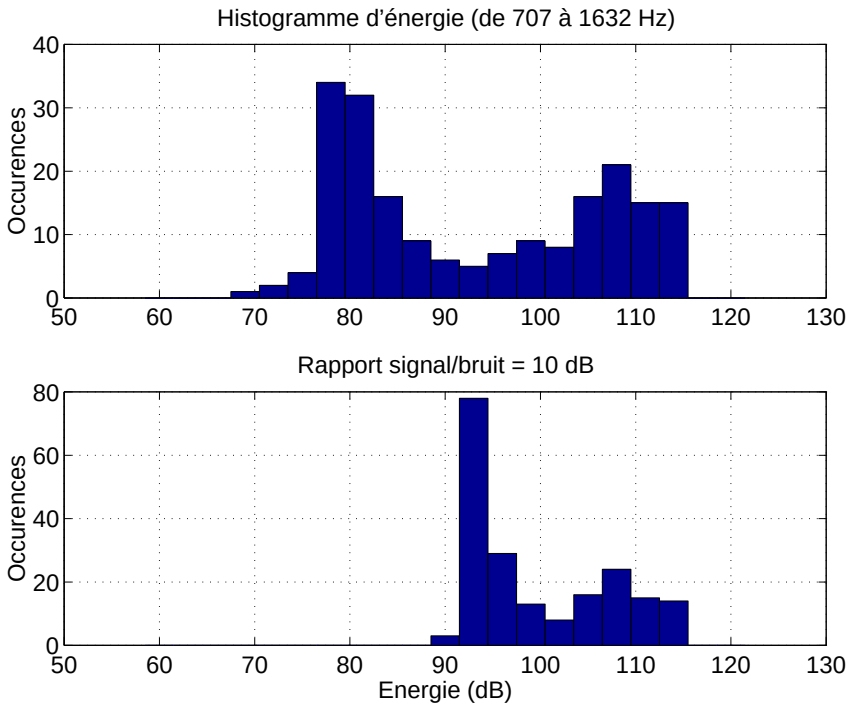


FIG. 4.1 – *Histogrammes d'énergie pour 2 secondes de parole dans la bande de fréquence 707-1632Hz. Figure du haut: parole claire. Figure du bas: parole bruitée, bruit blanc gaussien, rapport signal/bruit = 10 dB*

non stationnaires, au risque d'obtenir de nombreuses erreurs grossières pendant les portions de parole trop longues (de durée supérieure à N). La valeur de N résulte donc d'un compromis entre la robustesse et la possibilité de suivre des bruits non stationnaires. Ce compromis existe également pour les deux méthodes précédentes.

Correction des mesures

On observe généralement que cette dernière méthode fournit une sous-estimation du niveau de bruit. Ceci est dû au fait qu'on ne considère que les minima d'énergie dans le segment de parole considéré, et que la variance du bruit n'est pas négligeable. Pour compenser cet effet non désirable, nous avons utilisé des données perturbées par un bruit blanc non stationnaire connu. Cela nous a permis de définir une fonction de correction des valeurs fournies par l'estimateur, et cela pour une large plage de niveaux de bruit possibles [168].

4.3.4 Filtrage des harmoniques

Les méthodes décrites précédemment estiment le niveau de bruit sur base d'un segment de parole. Le segment sera suffisament long pour contenir des trames "silencieuses", faisant apparaître un mode à faible énergie dans les histogrammes d'énergie

des bandes de fréquence. Notons que même en parole continue, on trouvera ce genre de trames:

- juste avant les plosives,
- en basse fréquence pendant les fricatives,
- en haute fréquence pendant les sons voisés.

Nous savons cependant que les voyelles sont des sons périodiques conduisant à un spectre à bandes étroites contenant des raies harmoniques qui se superposent au spectre de bruit. Aux fréquences intermédiaires à ces harmoniques, la seule contribution provient du bruit (les conséquences du fenêtrage sont rappelées ci-dessous) . Aux fréquences harmoniques, le bruit se superpose à la parole. Les valeurs d'énergie dans une bande de fréquence peuvent être utilisées pour constituer un histogramme d'énergie. Si le bruit est quasi-blanc dans la bande de fréquence considérée, cet histogramme présentera deux modes:

- un mode à basse énergie correspondant aux valeurs d'énergie entre les harmoniques, valeurs principalement liées au bruit,
- un mode à haute énergie correspondant aux valeurs d'énergie des fréquences proches des harmoniques, correspondant à la superposition du bruit et de la parole.

L'hypothèse sur la forme du bruit est valide pour les bruits large-bandes et si la bande de fréquence est suffisament étroite.

Sur les segments de parole ne correspondant pas à des voyelles, ces histogrammes présentent malheureusement une allure monomode ne permettant pas l'estimation du niveau de bruit. Tout comme précédement, on aura donc recours à l'utilisation de segments de parole suffisament longs. Ces segments devront contenir, soit des portions de silence, soit des portions correspondant à des voyelles. De ce fait, ils pourront être plus court que les segments des méthodes précédentes. Cette approche, que nous appellerons "filtrage des harmoniques" permet donc d'éviter une surestimation du niveau de bruit pendant les longues portions de parole et/ou de diminuer la longueur des segments d'analyse (voir également la Section 4.3.7) et donc d'augmenter la qualité de l'estimation dans le cas de bruits non stationnaires[3].

Nous avons utilisé des bandes de fréquence d'une largeur de 1 Bark. Ces bandes sont suffisament étroites pour coller aux caractéristiques spectrales des bruits colorés et suffisament larges pour inclure des minima du spectre, sachant que le fenêtrage initial conduit à une convolution du spectre de raie idéal par une fonction "pieuvre" dont le lobe principal est relativement large. Nous avons utilisé la fenêtre de Hanning (la fenêtre de Hamming pourrait également convenir) car elle présente un bon compromis entre la largeur du lobe central et l'atténuation des lobes secondaires. Cette fenêtre conduit à un lobe principal de largeur égale à $4/d$, d étant la longueur (en secondes) de la fenêtre. Pour une fenêtre de 64 ms (soit 512 points pour une fréquence d'échantillonnage de 8 kHz), la largeur du lobe principal est de 62.5 Hz. Comme la fréquence fondamentale d'un signal vocal voisé est généralement supérieure à 100 Hz,

3. Rappellons qu'allonger les segments de parole permettrait d'augmenter la probabilité d'inclure suffisament de fenêtres d'analyses correspondant au portions de signal silencieuses, et donc la robustesse de l'estimation, au prix de l'impossibilité de suivre des bruits fortement non stationnaires et d'une augmentation du délai global du système.

les différentes harmoniques sont clairement séparées par une analyse sur base d'un fenêtre de 64 ms. On pourrait descendre jusqu'a 40 ms mais cela conduirait de toute façon au calcul d'une transformée de Fourier discrète à 512 points par FFT. Notons que pour certaines voix masculines particulièrement graves[4], une fenêtre d'analyse de longueur classique (30 ms) ne suffit pas pour obtenir une spectre dont les raies sont clairement séparées par des "vallées" spectrales.

L'atténuation du premier lobe secondaire dépend aussi de la longueur de la fenêtre. Pour une fenêtre de 64 ms, elle est de 29 dB. Lorsque le niveau de bruit est très faible par rapport au niveau de parole, la méthode risque donc de conduire à une surestimation due à l'énergie des lobes secondaires. A ce niveau cependant (-29 dB), le bruit n'est certainement pas gênant dans le cadre de la reconnaissance automatique de la parole.

Avec les méthodes classiques décrites précédemment, des vecteurs temporels de plusieurs centaines de millisecondes contenant les énergies moyennes dans des bandes de fréqences limitées sont utilisés pour l'estimation du spectre de bruit. Avec l'approche de filtrage des harmoniques, comme l'analyse est basée sur des spectres à bandes étroites, les énergies des régions inter-harmoniques peuvent également participer à l'estimation du spectre de bruit. Le filtrage des harmoniques est donc implémenté comme un pré-traitement des méthodes précédentes, auquelles on fournit des vecteurs temporels. Chaque élément d'un de ces vecteurs est le minimum du spectre d'énergie (bandes étroites) à l'instant considéré et dans la bande de fréquence limitée considérée.

Dans [168], l'implémentation est différente. Le filtrage des harmoniques est vu comme une extension aux méthodes classiques, sur base de matrices d'énergies provenant du spectrogramme à bandes étroites. Ces matrices couvrent plusieurs centaines de millisecondes et ont une plage de 200 Hz. Elles sont utilisées comme entrée d'une des trois méthodes décrites précédemment.

4.3.5 Approche hybride

La méthode décrite à la section précédente (que nous appellerons méthode harmonique) ne permet pas de mesurer le niveau de composantes sinusoïdales ou de bruits périodiques stationnaires, leurs harmoniques étant filtrées comme celles des sons voisés.

Les versions classiques des méthodes décrites aux sections 4.3.1, 4.3.2 et 4.3.3, quant à elles, permettent la mesure du niveau des bruits périodiques mais demandent des segments de parole plus longs, les bruits devant être stationnaires sur de plus longues portions de signal.

Les bruits périodiques peuvent cependant se superposer à des bruits large bande. L'annexe B présente les spectrogrammes de quelques types de bruits. A la figure B.1 par exemple, on peut constater la présence de plusieurs sinusoïdes superposées à un bruit large bande.

Pour pouvoir bénéficier des avantages des deux types d'approches, nous proposons de les combiner. Durant les portions identifiées comme silencieuses par la

4. ainsi que pour les passages de "vocal fry" correspondant à un doublement soudain de la période du signal.

méthode classique[5], la présence d'un bruit périodique conduit à une estimation du niveau de bruit plus élevée pour la méthode classique que pour la méthode harmonique, cette dernière filtrant systématiquement les composantes sinusoïdales. La différence entre les deux mesures est une estimation du niveau de bruit périodique dans la bande considérée. En supposant ce bruit périodique stationnaire, il est alors possible de corriger la mesure fournie par la méthode harmonique en y ajoutant cette estimation (voir figure 4.2). Rappellons qu'une soudaine augmentation du niveau de bruit large bande, comme illustré à la figure 4.2 n'est pas détectée par la méthode classique, les segments utilisés étant trop longs.

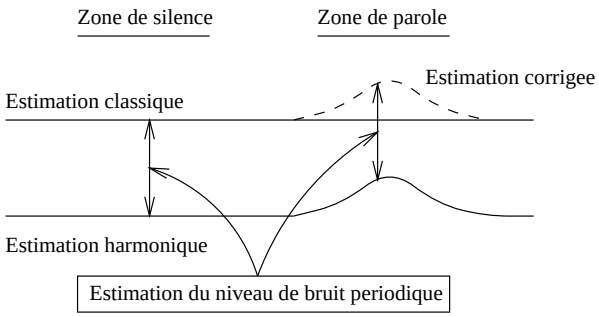


FIG. 4.2 – *Méthode d'estimation hybride*

Cette approche hybride sera utilisée dans les expériences qui vont suivre.

4.3.6 Positionnement de la mesure

Les méthodes décrites ici sont caractérisées par un délai important. Ce délai est essentiellement dû à la nécessité d'accumuler des segments de signal suffisament longs pour que la méthode soit robuste. De ce fait, le niveau de bruit mesuré à partir d'un segment de parole n'est pas assigné à la dernière trame de ce segment, mais bien à une des trames internes au segment. Dans un premier temps, nous avons choisi de l'assigner à la trame centrale. Or, ce niveau de bruit est calculé comme la moyenne de plusieurs minima d'énergie apparaissant dans le segment considéré (ou comme la moyenne du mode inférieur des histogrammes d'énergie), et ne correspond donc pas nécessairement au niveau de bruit au centre du segment. Par exemple, si tous les minima se produisent au début du segment, nous obtenons une mesure du niveau de bruit pour le début de celui-ci, plus précisément pour la trame moyenne des minima. Une interpolation linéaire permet finalement d'assigner une valeur aux autres trames.

Cette approche est utilisée dans toutes les expériences qui vont suivre.

5. Ces portions sont détectées comme un sous-produit de la procédure d'estimation du spectre de bruit, sur base d'un seuil sur l'énergie du signal.

4.3.7 Statistiques

En observant le spectrogramme d'un signal de parole, on constatera que:

1. l'énergie des fricatives est quasi nulle en basses fréquences,
2. l'énergie des voyelles est très faible en hautes fréquences.

Ces portions silencieuses permettent d'envisager l'estimation du niveau de bruit pour de la parole continue.

D'autre part, elles suggèrent que la durée optimale des segments intervenant dans les approches décrites ici pourrait dépendre de la bande de fréquence considérée. En vue d'étayer cette hypothèse, nous avons collecté des statistiques sur base des données d'entraînement du corpus RESOURCE MANAGEMENT. Une analyse en 28 bandes de fréquence de 250 Hz est effectuée à partir de trames de 30 ms décalées de 10 ms. Pour diverses longueurs de segments temporels (plusieurs trames adjacentes), nous estimons par comptage la probabilité que ces segments contiennent au moins une certaine proportion de trames silencieuses. Pour les méthodes décrites aux Sections 4.3.1, 4.3.2 et 4.3.3 (méthode classique), sont identifiées comme silencieuses:

1. les portions identifiées comme du silence par alignement Viterbi forcé d'un modèle HMM,
2. les portions dont l'énergie est inférieure à un seuil placé à 24 dB sous l'énergie moyenne du signal de parole dans la bande considérée. Nous considérons ainsi que les trames dont l'énergie est inférieure à ce seuil interviendront dans le mode à basse énergie des histogrammes. Ces portions correspondent essentiellement aux fricatives en basse fréquence et aux voyelles en haute fréquence.

Pour les méthodes basées sur un filtrage des harmoniques, sont en plus considérées comme silencieuses les trames correspondant aux voyelles, identifiées grâce à l'alignement Viterbi forcé d'un modèle HMM. Ces trames silencieuses interviennent vraisemblablement dans le mode inférieur des distributions discutées aux sections 4.3.1, 4.3.2 et 4.3.3 et permettent donc l'estimation du niveau de bruit.

Les résultats correspondant à la méthode classique sont présentés aux figures 4.3 et 4.4. La première figure donne la probabilité qu'un segment de x trames (en abscisse) contienne au moins 20% de silence; les 28 courbes correspondent aux 28 bandes de fréquence. La deuxième figure donne cette même probabilité pour les 28 bandes de fréquence (en abscisse); nous avons cette fois 20 courbes correspondant à des segments temporels allant de 10 à 200 trames (soit de 100 à 2000 ms), par pas de 10 trames. Pour une probabilité donnée, on peut constater une variance importante dans la longueur du segment à utiliser: de 200 à 800 ms pour une probabilité de 0.9 par exemple. Cela indique que le segment de parole peut être plus court pour certaines bandes de fréquence que pour d'autres.

En basses fréquences, les fricatives sont silencieuses alors qu'en hautes fréquences, se sont certaines voyelles. Aux fréquences moyennes cependant, ni les fricatives, ni les voyelles n'ont une énergie faible et seules les trames de silence sont réellement "silencieuses". Ceci explique la forme en "U" des courbes de la figure 4.4.

Ces expériences ont été répétées pour la méthode harmonique. Les résultats sont présentés aux figures 4.5 et 4.6. On constate une nette diminution de la durée nécessaire des segments de parole. Ainsi, pour que 90% des segments contiennent au moins 20% de silence, il suffira qu'ils aient une durée de 300 ms. Ici aussi, la variance

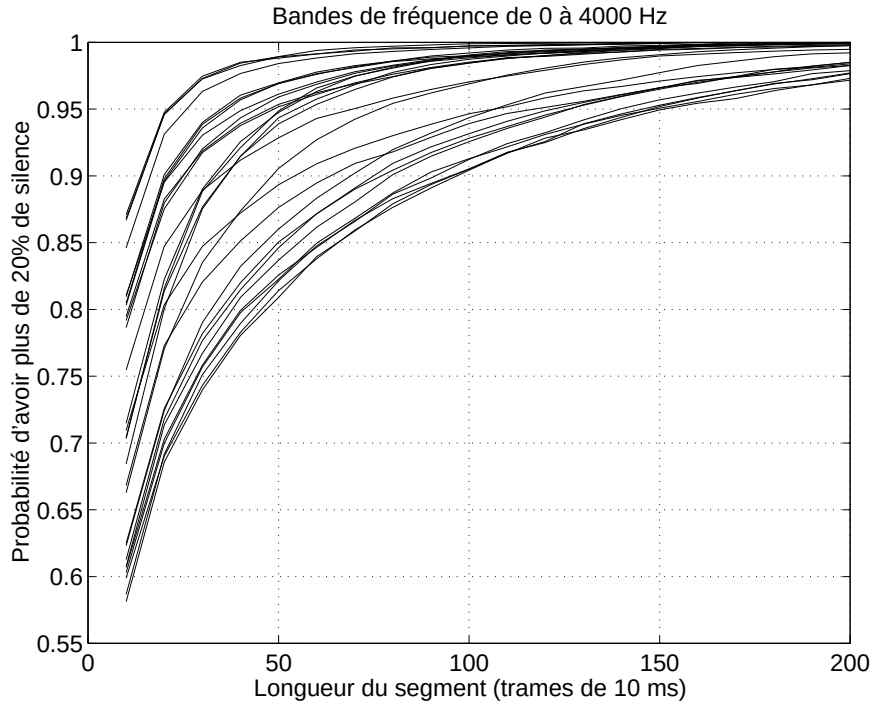


FIG. 4.3 – *Probabilité qu'un segment de x trames (en abscisse) contienne au moins 20% de silence. Les 28 courbes correspondent aux 28 bandes de fréquence.*

des résultats permet d'envisager d'ajuster la longueur des segments à la bande de fréquence considérée.

En basse fréquence, les voyelles et les fricatives sont "silencieuses" alors qu'en hautes fréquences, seules les voyelles le sont. Ceci explique l'allure des courbes de la figure 4.6. La probabilté d'avoir au moins 20% de trames silencieuses est élevée en basses fréquences et diminue progressivement pour aboutir à un minimum autour de 3000-3500 Hz. Elle augmente finalement pour être très élevée en hautes fréquences. De fait, les sons de parole ont une énergie très faible et l'énergie de ces bandes de fréquence est plus souvent proche du niveau de silence que celle des autres bandes.

Ces statistiques permettent de justifier (a posteriori) les ordres de grandeur habituellement rencontrés pour la longueur des segments temporels intervenant dans le méthodes étudiées ici. Elles pourraient cependant être utilisées dans le but d'optimiser la longueur des segments temporels en fonction de la bande de fréquence considérée. Cela n'a pas été envisagé dans les expériences qui vont suivre. Ce point pourrait cependant faire l'objet de recherches plus approfondies.

4.3.8 Détection parole/silence

D'autres techniques d'estimation du niveau de bruit se basent sur une détection parole/silence explicite [173]. Pour obtenir une borne supérieure des performances de cette classe de méthodes, nous avons utilisé un alignement parole/silence forcé sur

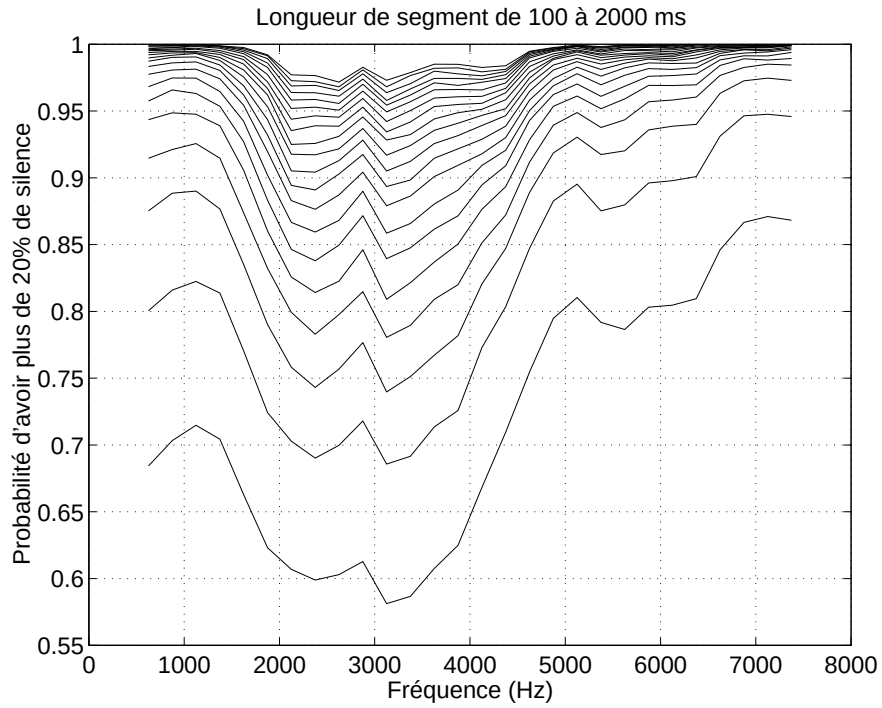


FIG. 4.4 – *Probabilité qu'une bande de fréquence (en abscisse) contienne au moins 20% de silence. Les 20 courbes correspondent à des segments temporels allant de 10 à 200 trames (soit de 100 à 2000 ms), par pas de 10 trames.*

base d'un modèle HMM. Le niveau de bruit est mis a jour pendant les portions de silence et est interpolé linéairement pendant les portions de parole. Les expériences de la section suivante visent notamment à comparer les performances des approches décrites précédement avec cette approche de détection parole/silence.

4.4 Comparaison des différentes méthodes

Trois méthodes d'estimation du niveau de bruit ont été comparées dans le cadre de nos recherches:

- la méthode des histogrammes de Hirsch,
- la méthode de clustering des énergies,
- la méthode de suivi d'enveloppe.

Pour la première méthode, nous avons utilisé le code fournit par l'ICSI (International Computer Science Institute) dans le cadre du programme d'analyse PLP et RASTA-PLP[6]. Nous avons utilisé les valeurs par défaut des paramètres de l'algorithme. Pour les autres méthodes, nous avons également utilisé la décomposition en bandes critiques de 1 Bark du programme d'analyse PLP.

6. Ce code est disponible à partir du site de l'ICSI [97].

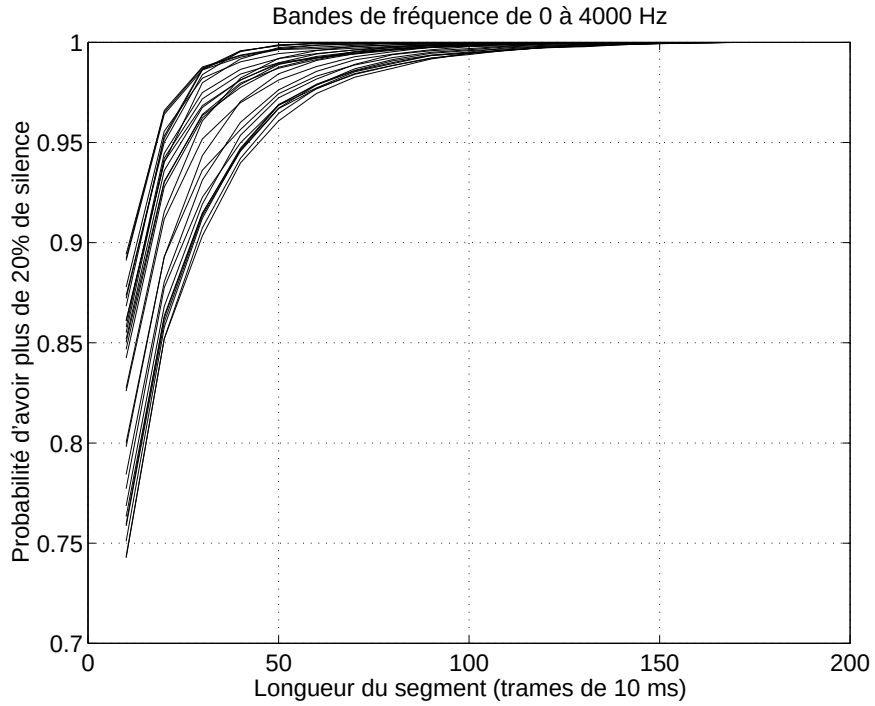


FIG. 4.5 – *Probabilité qu'un segment de x trames (en abscisse) contienne au moins 20% de silence. Les 28 courbes correspondent aux 28 bandes de fréquence.*

Nous avons évalué la qualité des estimations sur base de quatre types de bruits: un bruit blanc gaussien modulé en amplitude à 0.5 Hz (*N1*), un bruit blanc gaussien modulé à 1 Hz (*N2*), un bruit de voiture en mouvement du corpus MADRAS (*N3*) et un bruit d'hélicoptère du corpus NOISEX (*N4*). Les signaux de bruit ont été ajoutés au signal de parole claire (base de données NUMBERS'93) pour obtenir un rapport signal/bruit moyen de 15 dB. La longueur du segment temporel utilisé est fixée par le paramètre N indiquant le nombre de trames consécutives intervenant dans ce segment (trames décalées de 12.5 ms). Les résultats sont résumés à la table 4.1 pour une bande de fréquence allant de 707 à 1632 Hz. Rappelons que nous estimons d'abord le spectre de bruit en bandes critiques de 1 Bark et que la bande de fréquence considérée couvre en réalité 4 bandes critiques: son énergie est calculée comme la somme des énergies dans 4 bandes critques.

Comme on peut le constater, l'approche par filtrage des harmoniques conduit à de meilleurs résultats dans le cas de bruits non stationnaires ($N1$, $N2$ et $N3$). On peut également l'observer sur les figures 4.7 et 4.8. Pour le cas particulier du bruit $N2$, la méthode automatique surpasse même l'algorithme de détection parole/silence "idéal". Pour du bruit stationnaire ($N4$), toutes les méthodes se comportent bien.

Le lecteur trouvera plus de détails et d'autres résultats dans [168]. Notons cependant que dans cet article, l'implémentation du filtrage harmonique est différente (voir la section sur ce sujet) et également que l'estimation du spectre de bruit est

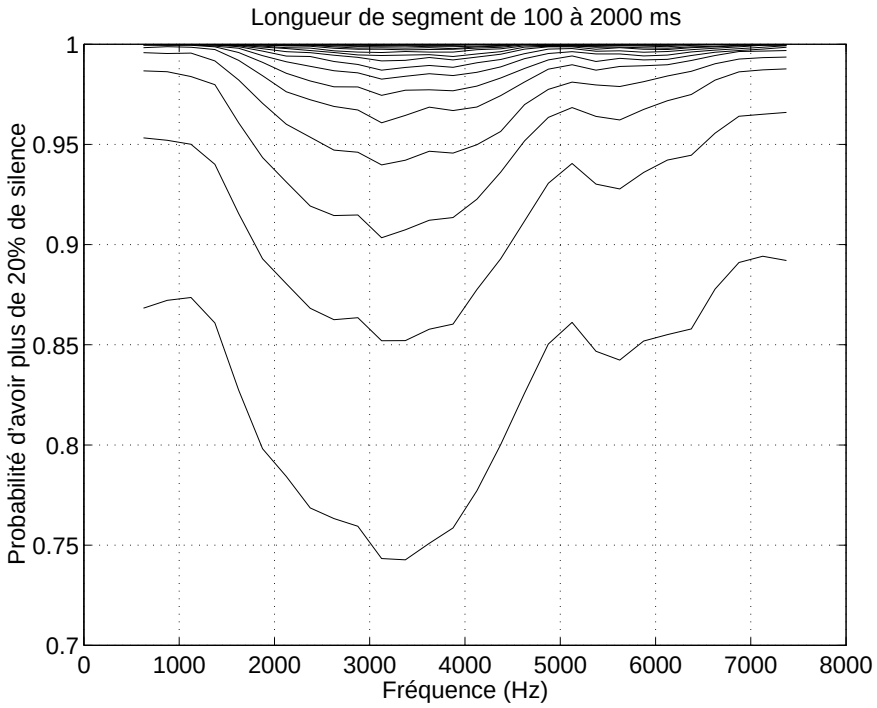


FIG. 4.6 – *Probabilité qu'une bande de fréquence (en abscisse) contienne au moins 20% de silence. Les 20 courbes correspondent à des segments temporels allant de 10 à 200 trames (soit de 100 à 2000 ms), par pas de 10 trames.*

basée sur une analyse en bandes de fréquences de 200 Hz, et nom plus en bandes critiques de 1 Bark.

4.5 Application en reconnaissance automatique de la parole

Nous avons effectué quelques expériences de reconnaissance vocale en vue d'illustrer l'intérêt d'une bonne méthode d'estimation adaptative du niveau de bruit. Les estimations fournies sont utilisées par un module de soustraction spectrale généralisée. Les paramètres utilisés sont les mêmes que pour les expériences de la Section 3.10. Les résultats obtenus sont comparés à une soustraction spectrale non-adaptative (le niveau de bruit est estimé pendant les 10 premières trames de chaque phrase) et aux méthodes de filtrage log-RASTA et J-RASTA.

Nous avons utilisé le corpus NUMBERS'93. Les données de test sont perturbées par le bruit de voiture de la base de données MADRAS.

Comme on peut le constater à la table 4.2, la soustraction spectrale, combinée à une bonne estimation du spectre de bruit local, conduit à des performances similaires à celles obtenues avec la technique J-RASTA.

Erreur quadratique moyenne	$N1$	$N2$	$N3$	$N4$
Hirsch N=25	68.3	148.5	**8.9**	5.8
Hirsch N=50	115.8	207.4	15.4	5.4
Clustering N=50	106.8	168.8	28.4	6.1
Enveloppe N=25	120.3	128.8	47.0	7.0
Enveloppe N=50	**31.8**	**94.5**	9.2	**2.6**
Envel. + harmo. N=25	49.2	**62.3**	15.1	**4.6**
Envel. + harmo. N=50	**22.5**	**81.5**	**7.6**	7.9
Détection parole/silence.	**17.3**	99.8	**5.1**	**2.9**

TAB. 4.1 – *Erreur quadratique moyenne (dB^2) de différentes méthodes d'estimation du niveau de bruit (bande de fréquence entre 707 et 1632 Hz) sur une portion de la base de donnée* NUMBERS'93. *N indique la longueur du segment en nombre de trames décalées de 12.5 ms.*

Taux d'erreur (%)	0 dB	10 dB
PLP-Log-RASTA	46.1%	18.4%
PLP-J-RASTA	33.6%	14.9%
CBE-SPS Stationaire	44.7%	18.5%
CBE-SPS Adaptatif N=25	36.8%	17.4%
CBE-SPS Adaptatif N=50	31.5%	16.3%

TAB. 4.2 – *Taux d'erreur au niveau du mot pour la reconnaissance de séquences de nombres avec du bruit de voiture ($N3$) à différents niveaux. $CEB-SPS$ Stationaire utilise une soustraction spectrale basée sur une estimation du niveau de bruit pendant les 10 premières trames de chaque phrase. $CBE-SPS$ Adaptatif utilise une estimation automatique du niveau de bruit: la méthode du suivi d'enveloppe avec filtrage des harmoniques a été appliquée sur des segments de 25 ou 50 trames de 12.5 ms.*

4.6 Conclusions

Les méthodes de reconnaissance robuste et de débruitage font souvent appel à une estimation des statistiques du bruit. Dans ce chapitre, différentes méthodes d'estimation du spectre de bruit ont été comparées. Elles découlent principalement de l'observation que les longs segments de signal contiennent des portions exemptes de parole qui peuvent être utilisées pour mettre à jour les estimations. Nous avons montré que ces méthodes sont utilisables pour l'estimation du spectre de bruits stationnaires.

Cependant, dans le cas de bruits non stationnaires, l'alternance entre les portions de parole et de silence risque d'être insuffisante pour obtenir une estimation fiable du niveau de bruit. Nous avons alors mis à profit le caractère périodique de certains sons de parole pour mettre à jour l'estimation. Tous les algorithmes envisagés dans ce chapitre peuvent tirer parti de cette méthode originale, conduisant à une estimation plus précise du niveau de bruit.

Nous avons également collecté des statistiques qui pourraient être utilisées dans

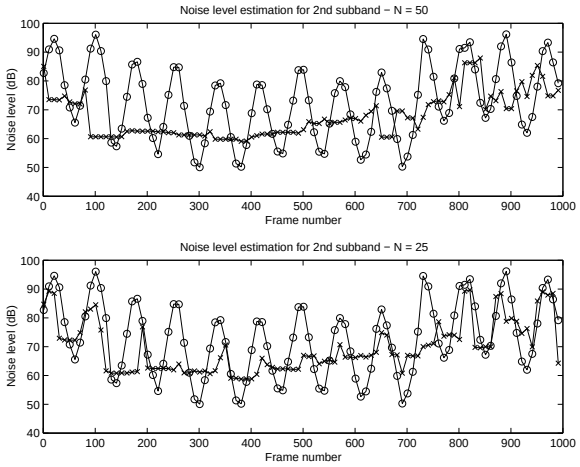


FIG. 4.7 – *Estimation du niveau de bruit (bande de fréquence entre 707 et 1632 Hz) suivant l'algorithme de Hirsch (type de bruit = N2, 0 = niveau de bruit, X = estimation). Figure du haut: N=50, figure du bas: N=25.*

le but d'optimiser la longueur des segments temporels en fonction de la bande de fréquence considérée.

Finalement, notre meilleure méthode d'estimation a été utilisée dans une expérience de reconnaissance vocale utilisant la soustraction spectrale. Les résultats obtenus sont comparables à ceux de la méthode $J-RASTA$. Rappelons que dans le cas de bruits stationnaires, nous avions également obtenu des performances similaires avec ces deux techniques (voir Section 3.10).

Ces techniques d'estimation du spectre de bruit seront également appliquées ultérieurement: (1) dans le cadre de l'approche de décomposition en bandes de fréquence, en vue d'estimer le rapport signal/bruit dans chacune des bandes de fréquence et (2) dans le cadre de la reconnaissance audiovisuelle de la parole, pour l'estimation du rapport signal/bruit global du signal acoustique.

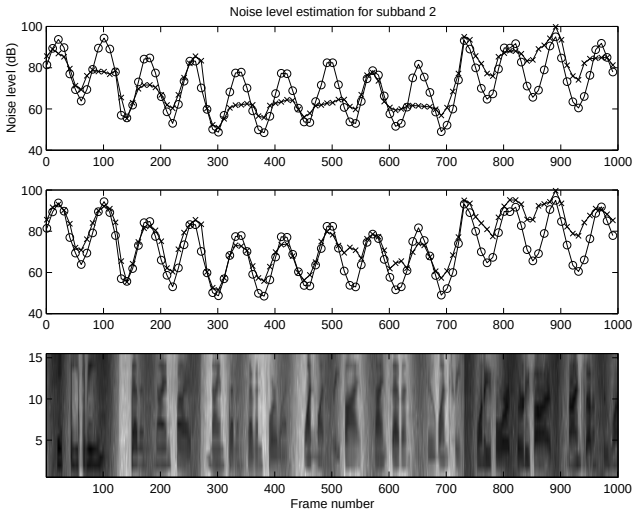


FIG. 4.8 – *Estimation du niveau de bruit (bande de fréquence entre 707 et 1632 Hz) suivant l'algorithme de suivi d'enveloppe utilisant le filtrage des harmoniques (type de bruit = N2, 0 = niveau de bruit, X = estimation). Figure du haut: N=50, figure centrale: N=25, figure du bas: spectrogramme.*

Chapitre 5

Modèle "multi-stream"

5.1 Introduction

Dans les systèmes de reconnaissance vocale actuels, le module d'analyse extrait un vecteur de paramètres représentatifs (vecteurs caractéristiques) toutes les 10 ms environ. On obtient donc une séquence de vecteurs décrivant les caractéristiques locales du signal de parole. Les modèles de Markov cachés, généralement associés à des unités phonétiques de base comme les phonèmes, sont alors caractérisés par des modèles statistiques, par exemple des fonctions de densité de probabilité sur l'espace des paramètres représentatifs. Les mots et les phrases sont ensuite supposés stationnaires par morceaux et représentées par des séquences d'unités de base.

Généralement, chaque vecteur de caractéristiques correspond à la concaténation (approche de **combinaison au niveau des paramètres représentatifs**) de différents types de paramètres comme: les valeurs instantanées de l'énergie, des paramètres spectraux ainsi que les dérivées premières et secondes de ces paramètres. On peut également imaginer l'utilisation de différentes sources d'information comme le mouvement des lèvres, des paramètres représentant la forme du conduit vocal, des paramètres prosodiques... Ceci conduit donc à un espace de dimension élevée sur lequel sont estimés les paramètres des modèles statistiques. Compte tenu de la quantité limitée des données d'entraînement, on suppose parfois que les différents paramètres représentatifs sont indépendants. Une autre solution, basée sur les mêmes hypothèses, est de diviser l'ensemble des paramètres en plusieurs groupes. Chaque groupe est alors considéré comme un vecteur de paramètres indépendant des autres et la **combinaison se fait dans l'espace des probabilités**. Dans les deux cas, on suppose cependant que les différents groupes de paramètres (canaux) sont synchrones. De ce fait:

1. Les segmentations des différents canaux en segments stationnaires (par l'algorithme de Viterbi) doivent être identiques. Ceci signifie que les transitions entre états des modèles de Markov cachés doivent se produire aux mêmes instants pour les différents canaux. Cette contrainte est gênante si les différents canaux sont stationnaires à différents moments. Cela pourrait déjà être le cas

par exemple entre les paramètres classiques et leurs dérivées première ou seconde.

2. La topologie des modèles de Markov cachés est la même pour les différents canaux. Cela implique notamment que le nombre de segments stationnaires est le même pour les différents canaux.

Finalement, la façon dont on combine les canaux, spécialement dans le premier cas, n'est pas adaptative. Elle ne permet donc pas de tenir compte de la fiabilité respective des différentes sources d'information, fiabilité qui pourrait être différente de celle rencontrée lors de l'entraînement (si les conditions de bruit par exemple sont différentes à l'entraînement et à l'utilisation).

Afin de lever ces limitations, nous proposons ici des approches qui n'imposeront pas le synchronisme parfait entre les différentes sources d'information. Une première approche fera appel à une modélisation sous forme de HMMs coopératifs, chaque source d'information ayant son propre jeu de HMMs. Ces modèles de Markov cachés n'interagiront entre eux qu'à des endroits pré-définis du point de vue phonémique. Une deuxième approche fera appel à la construction de HMMs composites (multi-dimensionnels). Chaque état de ces modèles représentera une configuration d'états des modèles coopératifs de l'approche précédente. Les modèles résultants, appelés modèles "multi-stream", seront utilisés ultérieurement dans le cadre de la reconnaissance audiovisuelle de la parole.

Les avantages potentiels des approches proposées ici sont multiples:

- Elles offrent un moyen pour combiner différentes sources d'information,
- Cette combinaison peut être adaptative, certains canaux d'information pouvant facilement être sous-pondérés, voire rejetés s'ils sont identifiés comme très peu fiables.
- La topologie des modèles de Markov cachés peut être adaptée à chaque canal d'observation.
- L'association trame/état est indépendante pour les différents canaux d'observation. Les différents canaux peuvent donc se désynchroniser jusqu'à certains points lexicaux pré-définis. Cette structure est donc destinée à des processus évoluant indépendamment, c'est-à-dire à des canaux d'information dont la dynamique est découplée.

  Avec l'approche de combinaison au niveau des paramètres représentatifs, différents groupes de paramètres sont combinés en un seul vecteur. Si les processus qui ont générés ces groupes de paramètres sont partiellement découplés, cela peut conduire à une augmentation de la variance des modèles et donc à une diminution des performances du système de reconnaissance. Dans ce cas, il serait peut-être préférable de combiner les canaux sur base des probabilités obtenues après décodage. L'approche "multi-stream" opère de la sorte mais permet également de définir des points de resynchronisation phonémique, par exemple entre les mots ou entre les syllabes.
- Le système d'intégration humain semble robuste à de faibles désynchronisations temporelles entre différents canaux d'information. Dans [3], un signal de parole est partitionné en 19 bandes de fréquence d'un quart d'octave. Ces canaux sont ensuite translatés dans le temps suivant une distribution uniforme entre 0 et

un délai maximum D_{max}. Des expériences de perception sur un sous ensemble du corpus *Timit* montrent que le taux de reconnaissance décroît progressivement lorsque D_{max} augmente. Cependant, il reste au dessus de 75% pour un asynchronisme D_{max} aussi élevé que 140 ms, bien que la durée moyenne des segments phonétiques soit de 72 ms.

Dans le domaine de la reconnaissance audiovisuelle de la parole, Massaro [134] introduit des désynchronisations systématiques entre les sources d'information audio et video. Ces expériences d'intelligibilité, basées sur les stimuli /ba/, /da/, /i/ et /u/, indiquent que le processus d'intégration est relativement robuste pour des asynchronismes allant jusqu'à 200 ms. Des résultats de Smeele [183] indiquent que l'intelligibilité audiovisuelle de stimuli CVC ne se dégrade pas pour des désynchronisations allant jusqu'à 80 ms. L'approche "multi-stream" proposée ici pourrait fournir un cadre robuste par rapport à ce type de désynchronisations.

- Alors que les HMMs sont essentiellement utilisés pour modéliser un seul processus, ou plusieurs processus dépendants, les modèles "multi-stream" pourront être utilisés pour des processus qui évoluent indépendamment au sein d'unités lexicales pré-définies. Des points d'ancrage à l'intersection de ces sous-unités permettent la resynchronisation des processus. Cependant, la définition de ces points d'ancrage n'est pas évidente car la dynamique des différents canaux n'est pas connue a priori. De plus, il est fort probable que certains problèmes seront caractérisés par des séquences de vecteurs provenant de processus qui ne sont ni dépendants, ni complètement découplés. Ceci peut conduire à des degrés de synchronisation plus ou moins importants, dépendant probablement des états visités. Finalement, certains processus pourraient également conduire à des motifs de synchronisme/asynchronisme, un des canaux étant systématiquement en avance, ou bien en retard par rapport aux autres canaux. De tels phénomènes pourraient être pris en compte en vue d'améliorer la précision des modèles. Des solutions seront proposées ici sur base des modèles "multi-stream".
- En reconnaissance de mots isolés, il est très aisé de permettre aux différents canaux de se désynchroniser. Il suffit pour cela de combiner les "scores" globaux obtenus à la fin des séquences de vecteurs représentatifs. La reconnaissance de parole continue est cependant moins évidente car nous ne souhaitons pas attendre la fin de la phrase avant de combiner les différents canaux. Cela introduirait un délai important. Cela impliquerait également l'utilisation de listes des N meilleures hypothèses pour chaque canal. Seules des hypothèses identiques peuvent en effet être combinées. Les approches "multi-stream" ne requièrent pas l'utilisation de listes des meilleures hypothèses. Elles permettent le décodage synchrone (trame à trame) de la parole continue.

Ce travail gravitait essentiellement autour de l'idée d'utiliser plusieurs canaux d'information. Dans ce but, nous avons notamment appliqué les méthodes décrites dans ce chapitre. Trois applications ont été envisagées (deux seulement font l'objet de ce rapport de thèse):

- l'approche **multi-bandes** [17, 18, 21, 42] (cf. Chapitre 6) consiste à extraire des paramètres représentatifs de différentes bandes de fréquence, à estimer, pour

chaque bande de fréquence, les probabilités associées aux sous-unités lexicales, et finalement, à recombiner ces probabilités de façon à pénaliser les bandes de fréquence bruitées. Les avantages de cette approche sont multiples. Hormis sa robustesse aux bruits colorés, elle permet d'optimiser indépendamment les systèmes de reconnaissance pour chaque bande de fréquence.

- avec l'approche **multi-échelle** [44, 45, 207, 208], il s'agit de définir plusieurs modèles de Markov cachés coopératifs se focalisant sur différentes propriétés dynamiques du signal de parole pour, par exemple, modéliser les phénomènes dynamiques au niveau de la syllabe (notamment les phénomènes microprosodiques). Un des intérêts de cette approche est de pouvoir intégrer une information temporelle couvrant une durée supérieure à la durée des phonèmes.
- finalement, nous nous sommes intéressés à la reconnaissance de parole **multimodale** [47] (cf. Chapitre 7). Dans toute communication humaine, il est clair que l'image apporte un plus par rapport au son quand des perturbations altèrent la qualité de ce dernier. L'observation des gestes, des expressions et surtout du mouvement des lèvres entraîne une plus grande robustesse face aux perturbations acoustiques. Nous nous sommes intéressés à l'utilisation du mouvement des lèvres comme source d'information complémentaire pour un système de reconnaissance automatique de parole.

5.2 Modèles parallèles

L'approche "multi-stream" [21, 203] proposée ici est une méthode adaptative permettant de combiner différentes sources d'information en utilisant des modèles de Markov cachés coopératifs. Si les sources d'information sont parfaitement synchrones, elles peuvent être combinées facilement. Ces sources peuvent cependant ne pas être synchrones. Il peut également être nécessaire de définir des modèles qui n'ont pas la même topologie pour les différentes sources d'information. Dans ces conditions, nous proposons de traiter les différentes sources indépendamment (sur base de modèles de Markov cachés) jusqu'à certains point d'ancrage où elles sont contraintes à se resynchroniser et à combiner leurs "contributions" partielles. Alors que le niveau de resynchronisation est défini a priori (à l'intersection des mots, ou des syllabes par exemple), l'instant optimal de resynchronisation résultera du processus de décodage.

5.2.1 Formalisme

Soit K le nombre de sources d'information et soit M le modèle correspondant à une transcrition possible de la phrase prononcée. Supposons ce modèle M composé d'une séquence de J modèles qui correspondent à des sous-unités lexicales. La j-ème sous-unité de la séquence est représentée par le modèle M_j. Le choix de ces sous-unités est lié au niveau lexical auquel nous souhaitons resynchroniser les sources d'information: elles correspondent par exemple à des syllabes. Chaque modèle M_j est ensuite composé de K modèles de Markov cachés indépendants, notés M_j^k. Pour une sous-unité lexicale donnée (j fixé), ces différents modèles de Markov cachés n'interagissent pas entre-eux et ont généralement des topologies différentes.

L'interaction entre les différentes sources d'information est réalisée en forçant la transition d'une sous-unité à la suivante à se réaliser au même instant pour les différents canaux. Par exemple, le transition du modèle M_j^k au modèle M_{j+1}^k doit être synchrone avec la transition du modèle M_j^l au modèle M_{j+1}^l. C'est également à cet instant que les "scores" associés aux K modèles M_j^k seront combinés pour fournir un "score" global associé au modèle M_j.

La figure 5.1 illustre un modèle composé d'une séquence de deux sous-unités lexicales, chacune d'entre elles étant représentée par K modèles de Markov cachés correspondant aux K sources d'information disponibles. La contrainte de synchronisation/combinaison dont nous avons parlé est représentée par le symbole $\bigotimes$. Il ne correspond pas à un état d'un modèle de Markov caché. Il indique simplement l'introduction d'un point d'ancrage forçant l'interaction entre les différents modèles, qui ne pourront donc pas évoluer indépendamment les uns des autres. Plus précisément, cette interaction consiste à:

1. combiner les "scores" des différents HMMs (probabilités ou vraisemblances) accumulés depuis le point d'ancrage précédent. Différentes solutions sont envisageables en ce qui concerne la combinaison des "scores" fournis pas les différents canaux. Cela fait l'objet de la Section 6.5.
2. resynchroniser les HMMs, c'est à dire forcer les transitions d'une sous-unité à la suivante à se produire au même instant pour les modèles correspondants aux différentes sources d'information. Cette interaction par resynchronisation nécessite l'utilisation d'algorithmes de décodage particuliers dont nous parlerons ultérieurement.

Au sein de chaque sous-unité lexicale, les différents HMMs n'intéragissent pas entre eux: l'association trames/états est donc indépendante pour les différentes chaînes de Markov. Ce phénomène est illustré à la figure 5.3. Dans cet exemple, les transitions de la chaîne de Markov $\{a,b,c,d,e\}$ sont en retard par rapport aux transitions de la seconde chaîne du modèle. Ainsi, la troisième trame est associée au premier état de la première chaîne (état a) alors qu'elle correspond déjà au deuxième état (état B) de la seconde chaîne. Une structure synchrone classique (figure 5.2) imposera par contre un synchronisme parfait entre les différentes sources d'information.

5.2.2 Reconnaissance

Comme déjà présenté au Chapitre 2, le problème de la reconnaissance vocale consiste à déterminer le modèle M dont la probabilité a posteriori est maximale, étant donnée la séquence de vecteurs d'observation X:

$$M^* = \underset{M}{\operatorname{argmax}}\ P(M|X) \tag{5.1}$$

La loi de Bayes nous donne alors:

$$M^* = \underset{M}{\operatorname{argmax}}\ \frac{P(X|M)P(M)}{P(X)} \tag{5.2}$$

$P(X)$ étant indépendant du modèle M, le problème de la reconnaissance revient à déterminer le modèle M qui maximise le produit de la vraisemblance $P(X|M)$ et

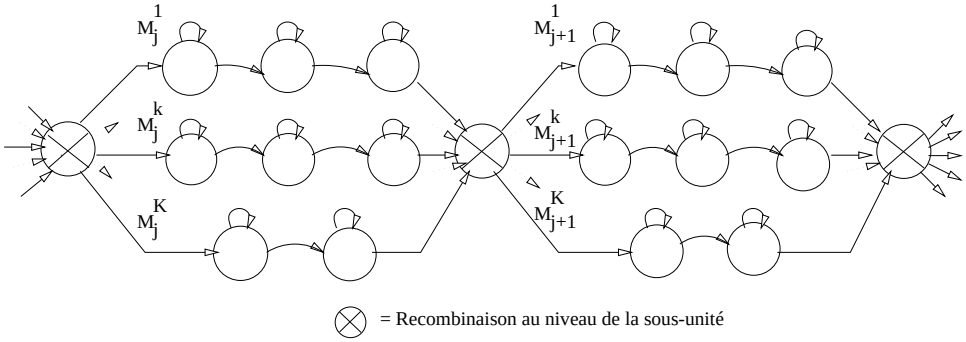


FIG. 5.1 – *Structure générale d'un système de reconnaissance à K canaux. Les points d'ancrage entre les sous-unités de parole permettent l'interaction entre les modèles correspondants aux différents canaux. Notons bien que les topologies des modèles ne sont pas forcément identiques pour tous les canaux.*

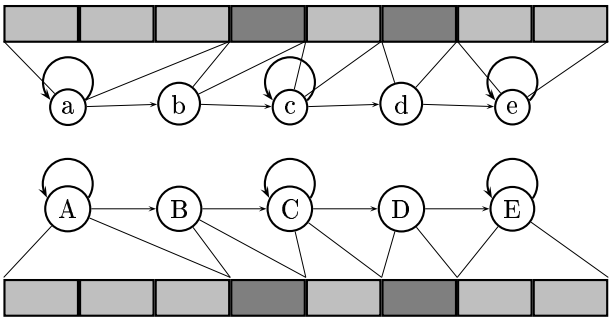


FIG. 5.2 – *Association trame/état sur base d'une structure synchrone.*

de la probabilité $P(M)$, fournie par le modèle de langage (n-gramme par exemple). Dans notre cas, le modèle M est constitué d'une séquence de sous-unités lexicales elles-mêmes composées de modèles de Markov cachés parallèles.

La vraisemblance $P(X|M)$ est calculée de manière exacte comme la somme des vraisemblances associées à chacun des chemins C possibles dans le modèle M. En effet, ces chemins étant **totalement exclusifs**[1], on a:

$$P(X|M) = \sum_{C} P(X,C|M) \tag{5.3}$$

la somme est étendue à l'ensemble des chemins possibles. Constatons que dans notre cas, **ces chemins seront tels que les contraintes imposées par les points d'ancrage $\otimes$ soient respectées**: chaque chemin définit une séquence d'état pour chacun des canaux d'information, ces séquences d'états devant coïncider à chaque transition d'une sous-unité lexicale à la suivante.

1. ils sont mutuellement exculsifs et couvrent tout l'espace

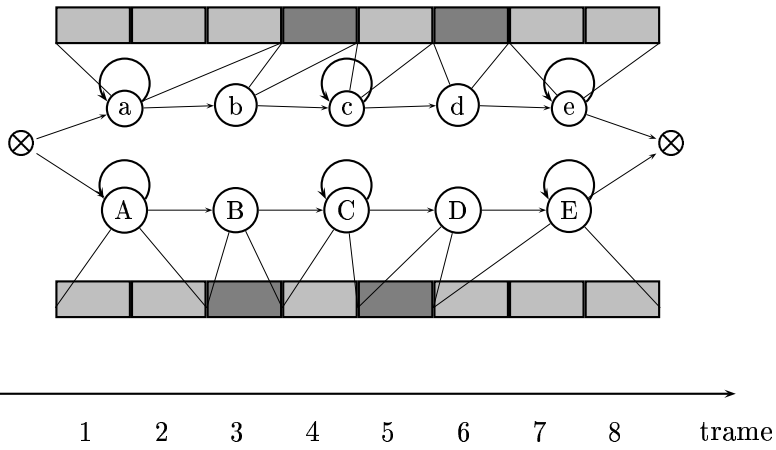


FIG. 5.3 – *Association trame/état sur base de modèles parallèles.*

L'approximation de Viterbi consiste quant à elle à estimer la vraisemblance sur base du meilleur chemin C_{Vit} uniquement:

$$P(X|M) = \max_{C} P(X,C|M) = P(X,C_{Vit}|M) \tag{5.4}$$

Ces calculs peuvent être effectués par énumération exhaustive des chemins possibles. Pour chaque chemin C, on peut alors calculer $P(X,C|M)$ en faisant appel aux hypothèses classiques des modèles de Markov cachés. Tout chemin C défini une segmentation temporelle en sous-unités lexicales. Si le modèle comprend J sous-unités lexicales, cette segmentation conduit à J sous-séquences X_j de vecteurs acoustiques et J sous-chemins d'états notés C_j. On obtient donc:

$$P(X,C|M) = P(X_1,C_1,...,X_j,C_j,...,X_J,C_J|M_1,...,M_j,...,M_J) \tag{5.5}$$

En utilisant deux des hypothèses fondamentales des modèles de Markov cachés (la non-corrélation des vecteurs d'observation et le fait que les probabilités des vecteurs d'observation ne dépendent que de l'état du modèle à l'instant considéré), on obtient:

$$P(X,C|M) = \prod_{j=1}^{J} P(X_j,C_j|M_j) \tag{5.6}$$

Par définition du modèle "multi-stream", **la vraisemblance $P(X_j,C_j|M_j)$ est alors calculée sur base des vraisemblances associées à chacun des canaux d'information**:

$$P(X_j,C_j|M_j) = f(\{P(X_j^k,C_j^k|M_j^k),\forall k\}) \tag{5.7}$$

où X_j^k est la séquence de vecteurs d'observation associés au canal k, C_j^k est la sous-séquence d'états associées au canal k et M_j^k est le modèle de Markov caché pour le canal k. Divers formalismes de combinaison f sont envisageables.

En utilisant les hypothèses classiques suivantes:

– non-corrélation des vecteurs d'observation,

- les probabilités des vecteurs d'observation ne dépendent que de l'état du modèle à l'instant considéré,
- les modèles sont d'ordre 1[2],

on obtient alors:

$$P(X_j^k,C_j^k|M_j^k) = \prod_{n=1}^{N_j} P(x_{j,n}^k|q_{j,n}^k)P(q_{j,n}^k|q_{j,n-1}^k) \tag{5.8}$$

faisant intervenir les probabilités classiques d'émission et de transition, où $q_{j,n}^k$ donne l'état visité par la j-ème sous-unité du canal k à l'instant n et N_j le nombre de trames couvertes par le sous-chemin C_j.

Récurrences

Le calcul des vraisemblances suivant les expressions (5.3) et (5.4) peut également faire appel à des récurrences de formes classiques.

Sur base de l'approximation de Viterbi, et en utilisant les expressions (5.4) et (5.6), on peut écire, pour une segmentation en sous-unités lexicales donnée[3]:

$$P(X|M) = \prod_{j=1}^{J} \max_{C_j} P(X_j,C_j|M_j) \tag{5.9}$$

Pour une fonction de combinaison (voir expression (5.7)) en somme pondérée de vraisemblances, la vraisemblance de la sous-séquence X_j par rapport au modèle de sous-unité M_j est calculée comme:

$$P(X_j|M_j) = \sum_{k=1}^{K} \alpha_k P(X_j^k|M_j^k) \tag{5.10}$$

Maximiser les vraisemblances associées aux M_j^k permet la maximisation de la vraisemblance globale associée à M_j. C'est également le cas pour des fonctions de combinaison en somme pondérée de logarithmes de vraisemblances, de probabilités a posteriori ou de logarithmes de probabilités a posteriori. On a donc:

$$\begin{aligned} P(X|M) &= \textstyle\prod_{j=1}^{J} \sum_{k=1}^{K} \alpha_k \max_{C_j^k} P(X_j^k,C_j^k|M_j^k) \\ &= \textstyle\prod_{j=1}^{J} \sum_{k=1}^{K} \alpha_k \max_{q_{j,1}^k,\ldots,q_{j,N_j}^k} \prod_{n=1}^{N_j} P(x_{j,n}^k|q_{j,n}^k)P(q_{j,n}^k|q_{j,n-1}^k) \end{aligned} \tag{5.11}$$

Le calcul de $\max_{C_j^k} P(X_j^k,C_j^k|M_j^k)$ fait appel à une récurrence Viterbi classique. En laissant tomber les indices k et j:

$$P(q_{l,n},X_{1,n}|M) = \max_m \left(P(q_{m,n-1},X_{1,n-1}|M)P(q_{l,n}|q_{m,n-1},M)\right) P(x_n|q_{l,n}) \tag{5.12}$$

2. Notons qu'on pourrait également envisager des modèles d'ordre > 1 très facilement.

3. c'est-à-dire que les frontières entre les sous-unités de base sont fixées mais pas la segmentation en états au sein des HMMs de ces sous-unités.

où $q_{l,n}$ signifie que l'état q_l est visité à l'instant n, où $X_{1,n}$ représente la sous-séquence de vecteurs d'observation jusqu'à l'instant n (où les indices j et k on été supprimés pour plus de clarté). La recherche du maximum est étendue à tous les prédécesseurs possibles de $q_{l,n}$, soit, dans le pire des cas, à tous les états du modèle de Markov caché associé à M_j^k. De même, cette recherche de maximum est faite pour tous les états $q_{l,n}$ possibles. La complexité de cette récurrence est donc de $L^2 N_j$ où L est le nombre d'états du modèle M_j^k et N_j le nombre de trames dans la sous-séquence considérée. Une approche d'exploration exhaustive sans l'utilisation de cette récurrence aurait une complexité de L^{N_j}.

La recherche du meilleur chemin pour chaque modèle M (et de la vraisemblance associée) consiste donc à factoriser le calcul du maximum de vraisemblance. Pour chaque segmentation en sous-unités, on recherche le chemin conduisant au maximum de vraisemblance. On choisi alors la segmentation en sous-unités pour laquelle le meilleur chemin conduit à la vraisemblance maximale. Dans le cas de parole continue, deux segmentations (en sous-unités) différentes peuvent avoir des portions identiques en termes de trames et de modèles concernés: la première portion des séquences de chiffres "un deux" et "un trois" par exemple. La recherche du chemin maximisant la vraisemblance sur ces portions identiques est bien entendu réalisée une seule fois. La description faite ici est proche de celle de l'algorithme du "Two Level" [172] bien connu en reconnaissance de parole continue. Nous étudierons cet algorithme par la suite.

La reconnaissance de mots isolés sur base de modèles de Markov cachés fait appel à la récurrence Viterbi utilisée précédement ou éventuellement à la récurrence exacte, aussi appellée récurrence de Baum-Welch (voir [163]). Ces récurrences sont utilisées pour estimer la vraisemblance associée à chacun des mots. On choisit alors le mot qui conduit à la vraisemblance maximale, après multiplication éventuelle par une probabilité a priori, comme le veut le formalisme général (voir équation (5.2)). On peut généralement étendre ces récurrences au cas de la parole continue. L'idée à la base de la récurrence de Viterbi (programmation dynamique) par exemple, à savoir qu'un chemin optimal est constitué de sous-chemins optimaux, vaut toujours. Il s'agit donc d'appliquer la récurrence à un modèle de Markov caché représentant toutes les phrases possibles: l'approche récursive permet de ne mémoriser, à chaque instant, que le meilleur chemin aboutissant à chacun des états des différents mots du vocabulaire. La contrainte grammaticale $P(M)$ de l'expression (5.2) doit être intégrée à cette récurrence. Dans le cas d'un bigramme par exemple, la probabilité de la paire de mots sera utilisée en lieu et place de la probabilité de transition $P(q_l|q_m,M)$ dans le cas de transitions entre deux mots ($q_l \in$ à un mot et $q_m \in$ au mot précédent)[4] [5]. La phrase reconnue est un sous-produit de cette récurrence.

4. Remarquons que ces récurrences ne sont valables que dans le cas de modèles de Markov d'ordre 1. Seules les grammaires d'ordre 1 peuvent donc être utilisées (grammaires en paires de mots ou bigrammes). Les grammaires d'odre plus élevé peuvent cependant se ramener à des grammaires d'ordre 1 assez facilement, au prix d'une complexité sensiblement accrue. Pour une grammaire d'ordre 2 par exemple (trigramme), il suffit de considérer chaque paire de mots valide comme un mot indépendant des autres: le but est de conserver le meilleur chemin pour chaque paire de mot, de façon à pouvoir appliquer le trigramme lorsqu'on étend ces paires de mots vers les successeurs possibles.

5. Remarquons aussi que les méthodes non-récurentes présentées précédement (équations (5.3)

Ces récurrences classiques, étendues au cas de la parole continue ne sont cependant pas applicables ici. Les HMMs doivent forcément pouvoir représenter plusieurs chemins d'états car la durée des différents régimes stationnaires n'est pas connue a priori, et dépend de la vitesse d'élocution. Dès lors, à chaque instant lors du décodage Viterbi, il convient de décider du chemin optimal (récurrence de programmation dynamique). Si ce choix se fait indépendamment pour chaque canal, les chemins de programmation dynamique des différents modèles M_j^k n'auront pas forcément le même point de départ et l'objectif de synchronisme n'est pas atteint.

Le problème est donc de forcer une segmentation inter-unités identique pour les différentes sources d'information, sans imposer cependant que les chemins soient identiques au sein même des sous-unités lexicales. Une solution à ce problème est d'effectuer pour chaque sous-unité de chaque canal, autant de récurrences qu'il y de débuts possibles pour la sous-unité en question. On retombe donc sur les méthodes de décodage décrites plus haut dans cette partie et sur l'algorithme "Two-Level", qui peut également être formulé de manière synchrone, comme nous le verrons par la suite.

Two-Level asynchrone

Rappelons brièvement le principe de l'algorithme du "Two Level". Cet algorithme a été introduit pour la première fois par Sakoe [172] comme une extension des techniques du DTW (Dynamic Time Warping) en mots isolés au problème de la reconnaissance de mots connectés.

Le TL est séparé en deux niveaux, imposant deux passes de traîtement:

- un processus au niveau unité (=mot en parole continue, =sous-unité lexicale dans notre cas) qui calcule pour chaque unité du vocabulaire de base, le "score" que celle-ci aurait si elle était localisée entre les trames b et e[6] (b et e compris), avec:

$$\forall b \in [1,N], \forall e \in [1,N], b \leq e \tag{5.13}$$

  N étant le nombre de trames contenues dans la séquence de parole considérée. Ce calcul peut être effectué par énumération exhaustive. L'hypothèse d'ordre 1 (au niveau des états des HMMs) permet cependant d'utiliser des récurrences classiques: soit la récurrence exacte, soit la récurrence de Viterbi.
- un processus au niveau des phrases qui calcule, en se servant des résultats du premier processus, le meilleur empilement possible d'unités pour une longueur d'empilement variant de 1 à L, L étant la durée maximum (de la phrase) en nombre de mots (ou de sous-unités lexicales dans notre cas). Ici également, la recherche du meilleur empilement peut faire appel à une énumération exhaustive ou à une récurrence de Viterbi. Dans le second cas cependant, l'hypothèse d'ordre 1 empêche l'utilisation de grammaires d'ordre plus élevé.

à (5.8)) font appel à l'hypothèse d'ordre 1 localement, c'est-à-dire au niveau des états des HMMs. Ces approches par énumération exhaustive permettent cependant d'envisager des modèles d'ordre plus élevé au niveau local et/ou au niveau des connexions entre mots.

6. Dans notre cas, ce score résulte d'une récurrence exacte ou d'une récurrence sur base de l'approximation de Viterbi.

A la fin de l'algorithme, le résultat est une suite d'unités dont le "score" est connu.

Cet algorithme a deux gros défauts:

- il n'est pas synchrone.
- dans le cadre du modèle "multi-stream", la complexité du premier niveau varie en $O(N_{mod} \times S \times K \times (N!))$ où N_{mod} est le nombre de modèles, S le nombre d'états par modèle, K le nombre de canaux d'information et N le nombre de trames à décoder. Pour N trames, il existe en effet $N!$ paires $\{b,e\}$ différentes. Pour chacune de ces paires, le décodage comprend trois boucles imbriquées: une boucle sur les K canaux d'information, une boucle sur les différents modèles de sous-unités et une boucle sur les différents états de la chaîne de Markov correspondant à la sous-unité considérée par la boucle de niveau supérieur. En pratique, il s'agit donc d'effectuer un décodage Viterbi pour chaque début b possible, et de poursuivre ce décodage jusqu'au bout de la phrase tout en conservant les scores pour chaque fin e possible. Ce algorithme est donc fort lourd. On peut cependant en réduire la complexité pour la rendre proportionelle au nombre de trames en imposant simplement une durée maximum aux modèles de sous-unités lexicales. Soit D cette durée. Alors, pour N suffisament grand, la complexité devient $O(N_{mod} \times S \times K \times D \times N)$, mais le décodage n'est plus optimal. Si D est suffisamment grand cependant (étant donné les durées des modèles dans l'ensemble d'entraînement), on peut espérer obtenir des performances proches de l'optimum. Notons aussi que le décodage optimal dont il est question ici est optimal dans les limites du modèle utilisé. En pratique, imposer une durée maximum aux différentes chaînes de Markov a peut-être un intérêt en termes de reconnaissance car on désactive ainsi certains chemins invraisemblables.

Two-Level synchrone

Nous avons développé une implémentation synchrone de l'algorithme Two-Level. Il s'agit d'une version en une passe. Elle ne nécessite donc pas d'attendre la fin de la phrase avant de commencer le décodage de celle-ci. L'algorithme fait appel à la récurrence Viterbi au niveau local ainsi qu'au niveau de la connexion entre les sous-unités. En pratique, seule une solution basée sur ces hypothèses simplificatrices est acceptable en ce qui concerne la charge de calcul. Nous verrons par ailleurs que même avec ces hypothèses, il sera nécessaire d'avoir recours à des approches de recherche sous-optimales basées sur l'élagage[7].

L'intérêt de cette implémentation synchrone est évident. Ce décodeur reste néanmoins très lourd et inutilisable en temps réel, sa complexité étant toujours de $O(N_{mod}.S.K.(N!))$. Comme pour l'approche asynchrone, une première solution permettant de limiter la charge de calcul consiste à imposer une durée maximale aux modèles. Cette approche manque cependant de souplesse et conduit à un système qui reste toujours fort lourd en mémoire et en ressource de calcul.

7. Notons que ces méthodes d'élagage sont également nécessaires dans le cas de systèmes de reconnaissance classiques à grand vocabulaire.

Comme deuxième approche, nous pouvons avoir recours à des méthodes de recherche en faisceau ("beam search") aussi appelées méthodes d'élagage ("pruning"). Ces méthodes, destinées au décodage Viterbi, opèrent par élimination/désactivation des hypothèses les moins vraisemblables en ne conservant qu'un faisceau plus ou moins large d'hypothèses très vraisemblables. Ceci conduit à classer dynamiquement les états des modèles de Markov cachés en deux catégories: les états actifs, et les états inactifs, ces derniers conduisant à l'abandon des chemins de programmation dynamique qui en résulteraient. Cette approche, relativement simple à implémenter dans le cas d'un système de décodage classique, nécessite d'être revue dans le cas de systèmes "multi-stream". D'une part, lorsqu'un chemin est désactivé (abandonné), il doit l'être pour les différents canaux d'information car un sous-groupe de canaux ne peux pas intervenir dans le "score" global, qui fait toujours appel à l'ensemble des canaux. D'autre part, comme l'occupation mémoire est également un problème critique avec ces modèles parallèles, nous souhaitons développer une approche qui permette en outre de libérer l'espace mémoire occupé par les chemins qui auraient été désactivés.

Plutôt que d'effectuer une désactivation sur base des états des HMMs, nous avons finalement trouvé préférable d'opérer par désactivation complète des chemins correspondants aux paires début/sous-unité lexicale[8] les moins vraisemblables: nous avons en effet vu que le décodage "Two-Level" consiste à effectuer pour chaque sous-unité lexicale, autant de programmations dynamiques qu'il y a de débuts possibles. Le nombre de programmations dynamiques est donc de $N_{mod} \times N$, où N_{mod} est le nombre de modèles de sous-unités et N le nombre de trames dans la phrase considérée. Ces $N_{mod} \times N$ calculs sont considérés comme concurrents dans notre approche d'élagage et seuls les meilleurs en termes de vraisemblance sont conservés. Il est donc tout à fait concevable de conserver plus de débuts possibles pour une sous-unité que pour une autre. Il y est également concevable d'abandonner complètement une sous-unité c'est-à-dire d'abandonner tous ces débuts possibles au profit d'autres sous-unités lexicales.

Ceci permet en outre un gain considérable en espace en ne gardant en mémoire que les structures de données correspondant aux paires début/sous-unité lexicale conservées. Pour éviter des désallocations/réallocations de mémoire intempestives (qui ralentiraient considérablement le décodage), il a été décidé d'effectuer une seule phase d'allocation en début de décodage, et de partager ensuite les structures de données allouées aux différents chemins conservés. Notons finalement que, contrairement au décodeur n'utilisant pas l'élagage, nous n'avons plus besoin d'imposer une durée maximale pour les sous-unités lexicales (pour limiter la charge de calcul), le système décide de lui-même quand il peut abandonner un chemin, éventuellement trop long. Les prononciations particulièrement lentes sont donc correctement traitées.

L'exemple suivant illustre l'importance de cette stratégie d'élagage. Des expériences ont été effectuées sur une base de données de nombres connectés en anglais. Nous avons envisagé un système basé sur ce décodeur Two-Level synchrone utilisant

8. Rappelons en effet que le décodage "two-level" optimal (sans élagage) effectue pour chaque sous-unité lexicale, autant de programmations dynamiques qu'il y a de débuts possibles à cette sous-unité.

deux canaux distincts mais tout à fait identiques (contenant exactement les mêmes vecteurs représentatifs). Cette expérience n'a bien évidement aucun intérêt pratique si ce n'est de permettre la validation du logiciel développé. En effet, les performances du système à deux canaux devraient être identiques à celles obtenues avec un seul canal, à savoir un taux d'erreur de 10.7%. La table 5.1 présente les résultats obtenus en termes de taux de reconnaissance, ressources de calcul et occupation mémoire. On constate qu'un élagage raisonnable permet de diminuer fortement la charge de calcul et l'occupation mémoire sans réduire les performances de reconnaissance.

Configuration	Mémoire (Mg)	CPU (*temps-réel)	Taux d'erreur (%)
100	150	25	10.7
10	13	3	10.7
3	4.5	1	11.6
Décodeur classique	2.5	0.07	10.7

TAB. 5.1 – *Performances du décodeur "multi-stream two-level" pour différentes configurations d'élagage sur station de travail SUN Ultra 1. Le nombre de d.p. (programmations dynamiques) conservées est de x fois le nombre de sous-unités lexicales qui interviennent dans le vocabulaire, x étant le nombre indiqué dans la colonne "Configuration". Toutes ces d.p. sont partagées entre les paires début/sous-unité les plus vraisemblables. Il est donc tout à fait possible de conserver un plus grand nombre de d.p. (débuts possibles) pour une sous-unité lexicale que pour une autre. Il est également possible d'abandonner complètement une sous-unité c'est-à-dire d'abandonner tous ses débuts possibles au profit d'autres sous-unités lexicales.*

Soulignons à nouveau trois points importants concernant cet algorithme et son implémentation:

- **Complexité de l'algorithme**: Conserver le meilleur chemin pour chaque canal d'information ne suffit pas car cela ne permet pas de combiner les "scores" de modèles débutant au même instant. Il est donc nécessaire de conserver le meilleur chemin pour chaque début possible des sous-unités lexicales intervenant dans le vocabulaire, ce qui devient très lourd.
- **Elagage dynamique - Recherche en Faisceau**: Un élagage dynamique basé sur les modèles permet cependant de contrôler efficacement la charge de calcul et l'occupation mémoire.
- **Problèmes d'implémentation**: Cet élagage, pour être efficace, fait appel à quelques particularités d'implémentation, notamment le partage de l'espace mémoire entre différents modèles, sans réallocation nécessaire.

Cet algorithme est présenté sous forme de pseudo-code à la figure 5.4.

Autre algorithme proposé dans la littérature

Un algorithme similaire est proposé dans [29].

Remarque

Le processus de décodage que nous avons proposé consiste à calculer la vraisemblance de la séquence de vecteurs acoustiques étant donnés les différents modèles M_j^k et pour toutes les paires $\{b,e\}$ possibles. Ce calcul de vraisemblance fait appel à l'approximation de Viterbi, et passe donc par une recherche (programmation dynamique) des associations trames/états maximisant la vraisemblance de la séquence de vecteurs. Les vraisemblances ainsi obtenues pour les différents M_j^k d'une même sous-unité (j fixé) sont ensuite combinées dans le but de fournir une estimation de la vraisemblance de la séquence de vecteurs étant donné le modèle M_j.

Si la combinaison des vraisemblances associées aux M_j^k correspond à une somme pondérée, nous obtiendrons le maximum de la vraisemblance associée à M_j. C'est également le cas pour des fonctions de combinaison en somme pondérée de logarithmes de vraisemblances, de probabilités a posteriori ou de logarithmes de probabilités a posteriori. L'approche de décodage proposée permet donc de calculer le maximum de la vraisemblance associée à toute séquence M de sous-unités lexicales et permet donc la reconnaissance vocale suivant l'hypothèse de Viterbi.

5.2.3 Entraînement

Il s'agit ici d'estimer les paramètres des modèles de sous-unités lexicales de façon à maximiser la vraisemblance des séquences de vecteurs de la base d'entraînement par rapport à ces modèles. Nous étudierons ici l'entraînement basé sur l'approximation de Viterbi uniquement. La méthode est basée sur l'algorithme EM [38, 163], les variables cachées étant les états des HMMs à chaque instant. On commence par choisir une segmentation initiale: en d'autres termes, on choisit des valeurs initiales pour les variables cachées. Ensuite, on répète les étapes de maximisation et d'estimation jusqu'à vérification d'un critère de convergence.

L'étape de maximisation consiste à déterminer les paramètres des modèles de façon à maximiser la vraisemblance des observations. L'estimation des paramètres des HMMs correspondant aux modèles M_j^k est donc supervisée et sera résolue grâce aux techniques classiques de maximisation de la vraisemblance [163], appliquées aux modèles de Markov cachés classiques, ou aux modèles de Markov cachés utilisant des réseaux de neurones artificiels comme estimateurs des probabilités a posteriori des classes phonétiques [20].

L'étape d'estimation consiste à déterminer les valeurs des variables cachées qui maximisent la vraisemblance des observations de l'ensemble d'entraînement, étant donnés les modèles et leurs paramètres actuels, obtenus grâce à l'étape de maximisation précédente. Les méthodes de décodage présentées à la section précédente permettent de résoudre ce problème. Il s'agit d'effectuer un décodage Viterbi. L'entraînement étant supervisé, les phrases de l'ensemble d'entraînement sont connues. On applique donc le décodage Viterbi sur base de séquences connues de modèles de Markov cachés. Pour cette raison, on parlera de décodage "forcé", voire d'alignement "forcé". Le but de ce décodage est en effet d'obtenir la meilleure séquence d'états, c'est-à-dire le meilleur alignement entre les états des HMMs et les séquences de vecteurs caractéristiques. Cet alignement est forcé car on impose la séquence de modèles de Markov cachés pour chaque phrase d'entraînement.

```
for (trame=0;trame<nombre_trames;trame++) {

  Un modèle correspond au HMM d'une sous-unité lexicale débutant à une trame donnée. Le
  chemin optimal ne peut être obtenu que s'il existe un modèle pour chaque début possible (chaque
  trame) de chaque sous-unité. En pratique cependant, un élagage basé sur les "scores" sera utilisé
  afin de désactiver les modèles les moins vraisemblables.
    for (modèle=0;modèle<nombre_modèles_actifs;modèle++) {
        // VITERBI INDEPENDANT POUR CHAQUE EXPERT
     Boucle sur les canaux d'information
        for (expert=0;expert<nombre_experts;expert++) {
         Boucle sur les états des HMMs
            for (état=0;état<nombre_états[modèle];état++) {
             Etendre le "score" vers tous les états successifs étant donné les règles intra-unités.
             Ajouter la distance correspondant à la probabilité de transition.
            }
        }
        // COMBINAISON
     Extraire les "scores" accumulés pour chaque expert à la sortie du modèle. Soustraire de ces
     "scores" le "score" au début du modèle (qui est identique pour chaque expert).
     Combiner ces "scores" correspondants aux différents experts. Ajouter le "score" correspon-
     dant au début du modèle.
        // CONNECTION VERS LES SOUS-UNITES
     Etendre ce score combiné vers toutes les sous-unités (modèles) qui suivent étant donnés les
     règles inter-unités (lexique définissant les mots en termes de sous-unités et grammaire entre
     ces mots) ET le fait que chaque modèle a un point de départ. Le score est dont uniquement
     étendu vers les modèle destinés à débuter à la trame suivante, les autres modèles continuant
     à progresser sur base de leurs scores internes. Au score étendu est alors ajouté une distance
     correspondant à une grammaire bigramme.
        // CONTRIBUTION LOCALE
        for (expert=0;expert<nombre_experts;expert++) {
            for (état=0;état<nombre_états[modèle];état++) {
             Ajouter la distance locale correspondant à la probabilité d'émission de l'état
             considéré.
            }
        }
    }
    // ELAGAGE
  Pour chaque modèle, déterminer le meilleur score de programmation dynamique à la trame
  courante.
  Sur base de ces "scores" (logarithmes des vraisemblances jusqu'à la trame courante), ne garder
  que les N meilleurs modèles[a] (les modèles débutant à la trame courante sont toujours gardés
  cependant). Supprimer les autres modèles. En pratique, pas de désallocation/réallocation de
  mémoire (trop coûteuse): les donnés des modèles supprimés sont simplement réinitialisées et
  peuvent alors être réutilisées pour démarrer un nouveau modèle à la trame suivante.

  [a] Notons ici que cette méthode est différente de l'approche classique de recherche en faisceau
  (beam search). Cette dernière consiste à déterminer le meilleur "score" courant et à ne conserver
  que les chemins dont le "score" est compris dans une plage de largeur prédéfinie (faisceau) à
  partir de ce meilleur "score". On conservera donc un nombre important de chemins lorsque leurs
  "scores" sont proches. Par contre, si quelques chemins se démarquent nettement par rapport aux
  autres, l'élagage sera plus important. Dans notre cas, cette méthode ne convient pas car nous
  souhaitons limiter l'espace mémoire requis par l'algorithme. Il est donc préférable de contrôler
  le nombre de modèles actifs plutôt que la largeur d'un faisceau de "scores".
}
```

FIG. 5.4 – *Pseudo-code de l'algorithme Two-Level synchrone avec élagage (recherche en faisceau) au niveau des sous-unités lexicales.*

Pendant l'étape de maximisation, il convient également d'optimiser les paramètres de la fonction de combinaison de l'équation (5.7). Dans le cas d'une combinaison par somme pondérée de logarithmes de vraisemblances (c'est-à-dire par moyenne géométrique pondérée de vraisemblances), l'estimation par maximum de vraisemblance des poids de combinaison échoue cependant. En effet, soit X l'ensemble des vecteurs d'observation utilisé pour l'entraînement, soit C le résultat d'une étape d'estimation de l'algorithme d'entraînement et soit M la séquence de modèles correspondant aux phrases d'entraînement. On a:

$$
\begin{aligned}
P(X,C|M) &= \prod_{j=1}^{J} P(X_j,C_j|M_j) && (5.14)\\
&= \prod_{j=1}^{J}\prod_{k=1}^{K} P(X_j^k,C_j^k|M_j^k)^{\alpha_k} && (5.15)\\
&= \prod_{k=1}^{K}\prod_{j=1}^{J} P(X_j^k,C_j^k|M_j^k)^{\alpha_k} && (5.16)\\
&= \prod_{k=1}^{K} (P_k)^{\alpha_k} && (5.17)
\end{aligned}
$$

où P_k représente donc la vraisemblance globale des données d'entraînement pour le canal k. Nous souhaitons maximiser $P(X,C|M)$ sous la contrainte

$$\sum_k \alpha_k = 1 \tag{5.18}$$

Cela conduit à la solution suivante:

$$
\begin{cases}
\alpha_k = 1 \quad \text{si} \qquad k = \underset{k}{\text{argmax}}\ P_k \\
\alpha_k = 0 \quad \text{sinon}
\end{cases}
\tag{5.19}
$$

Dans [161], l'auteur propose d'utiliser des contraintes additionnelles de la forme:

$$\alpha_k P_k = \alpha_1 P_1, \forall k \tag{5.20}$$

Ces contraintes conduisent à des poids plus élevés pour les canaux dont la vraisemblance est faible. De plus, elles mènent à une solution unique n'impliquant plus aucune maximisation de vraisemblance. Selon les auteurs, elles permettent cependant d'obtenir une solution satisfaisante pour les α_k, du moins comme valeurs initiales de ces poids. Celles-ci peuvent alors être utilisées dans le cadre de méthodes d'optimisation utilisant des critères autres que le maximum de vraisemblance. En utilisant les équations (5.18) et (5.20), on obtient:

$$\alpha_k = \frac{1/P_k}{\sum_s (1/P_s)} \tag{5.21}$$

Alternativement, un entraînement par descente de gradient stochastique généralisée ou GPD [27, 161] utilisant un critère de minimisation de l'erreur de classification peut

également être utilisé (voir la Section 6.5.3 pour plus de détails). Il est également possible d'optimiser les poids de combinaison en utilisant un véritable critère de minimisation de l'erreur de classification au niveau du mot, sur base de données de développement. Cette étape d'estimation est effectuée à chaque itération de l'algorithme de maximisation de la vraisemblance. C'est cette méthode qui sera appliquée dans nos expériences de reconnaissance vocale multimodale.

Finalement, il est important de remarquer que l'estimation des paramètres "multi-stream" proposée ici est différente de l'optimisation indépendante des paramètres des différents canaux, comme proposé dans [161]. Les auteurs proposent d'estimer les paramètres des modèles HMMs de façon tout à fait indépendante pour les différents canaux: les paramètres des modèles M_j^k associés au canal k sont estimés uniquement sur base des vecteurs représentatifs du canal en question. L'entraînement et la reconnaissance utilisent donc deux approches d'estimation de vraisemblances différentes. Les topologies des modèles "multi-stream", particulièrement les points d'ancrage, introduisent des contraintes additionnelles dans les alignements forcés (algorithme de Viterbi) sur les données d'entraînement. Ces contraintes peuvent notamment être importantes pour 'corriger' les mauvais alignements d'un canal d'information peu fiable. Nous avons notamment constaté sur le corpus M2VTS (voir Chapitre 7) que les alignements d'un canal correspondant au mouvement des lèvres sont incorrects pour une part importante des données d'entraînement. Les états de certains mots sont parfois alignés sur des portions de signal appartenant à d'autres mots. Dans d'autres cas, les états recouvrent des portions de signal appartenant à plusieurs mots consécutifs. L'utilisation des contraintes du modèle "multi-stream" nous parait donc importante pendant la phase d'entraînement.

5.3 Modèles composites

Dans les sections précédentes, nous avons décrit une architecture de reconnaissance basée sur la coopération de modèles de Markov cachés. Elle permet la modélisation de processus imparfaitement couplés. Une des particularités de cette structure concerne la possibilité d'associer une trame donnée à un état de façon indépendante pour les multiples HMMs. Les scores fournis par les différents HMMs sont combinés à un niveau lexical pré-défini où l'on force les modèles à se resynchroniser.

Une approche alternative consiste à définir des HMMs composites où chaque état correspond à un K-uple d'états des K modèles coopératifs. La topologie de ces modèles composites possède donc K dimensions et est définie de façon à pouvoir représenter tous les chemins possibles étant données les topologies des modèles coopératifs initiaux. Les contributions locales d'un état composite sont définies comme une combinaison des contributions locales des états intervenant dans cet état composite. La combinaison peut donc se faire à chaque trame et le décodage est tout à fait classique.

5.3.1 Formalisme

A la Section 5.2, nous avons proposé de modéliser toute sous-unité lexicale M_j par K modèles de Markov cachés M_j^k coopératifs, K étant le nombre de sources d'information disponibles. Nous proposons ici d'utiliser des modèles composites dans le même esprit que l'approche de décomposition de modèles de Markov cachés [198][9]. Chaque état du modèle composite M_j représente un K-uple d'états des K modèles coopératifs M_j^k. Les transitions dans le modèle composite permettent quant à elles de représenter toutes les transitions dans les K modèles coopératifs. Ceci permet de remplacer le décodage multidimensionnel imposé par le modèle présenté à la figure 5.1 par un décodage monodimensionnel classique. Il s'agit de construire des modèles composites de sous-unités lexicales sur base des modèles correspondants pour chacun des canaux d'observation. Comme précédement, il sera possible de forcer le synchronisme des transitions d'une sous-unité à l'autre, tout en permettant l'asynchronisme au sein de la sous-unité. Un augmentation du niveau lexical correspondant à la sous-unité augmentera cependant la charge de calcul de façon considérable, car le nombre d'états composites sera plus élevé.

Prenons comme exemple le cas de deux modèles coopératifs de topologies identiques. Ces modèles, représentés à la figure 5.5 conduisent à un HMM composite symétrique représenté à la figure 5.6. Ces figures correspondent à l'exemple déjà présenté à la figure 5.3. L'association trames/états de cette dernière correspond au chemin indiqué en traits gras à la figure 5.3 (en passant deux trames dans l'état $a-A$, 1 trame dans les états $a-B$, $b-C$, $c-D$ et $d-E$ et 2 trames dans l'état $e-E$). On constate en effet à la figure 5.3 que pour les deux premières trames, on se trouve dans l'état a de la première chaîne de Markov et dans l'état A de la seconde. Pour la troisième trame, on se trouve dans l'état a de la première chaîne et dans l'état B de la seconde... Comme second exemple, considérons des modèles coopératifs aux topologies différentes (figures 5.7). Le modèle composite est ici dissymétrique comme illustré à la figure 5.8.

Dans les modèles des figures 5.6 et 5.8, les mouvements haut-bas correspondent à des transitions dans la chaîne $\{a,b,c,d,e\}$, les mouvements gauche-droite correspondent à des transitions dans la chaîne $\{A,B,C,D,E\}$ et les mouvement diagonaux correspondent à des transitions simultanées pour les deux canaux. Un chemin dans ces modèles correspond à des chemins différents dans les deux modèles coopératifs initiaux.

Sur base des mêmes hypothèses que celles conduisant aux expressions (5.6) et (5.8), on obtient, pour un chemin C, une séquence de vecteurs X et un modèle

9. Cette approche correspond à un décodage Viterbi synchrone permettant la décomposition d'une séquence d'observation unique en composantes indépendantes (typiquement la parole et la bruit), chaque composante étant modélisée par son propre jeux de modèles de Markov cachés. Il s'agit de définir des états composites pour chacun des états combinés des différentes composante (HMMs) intervenant dans le modèle. Cela permet d'utiliser un algorithme de décodage classique (Viterbi ou éventuellement Baum-Welch) mais requiert le calcul de probabilités d'émission associées aux différents états composites.

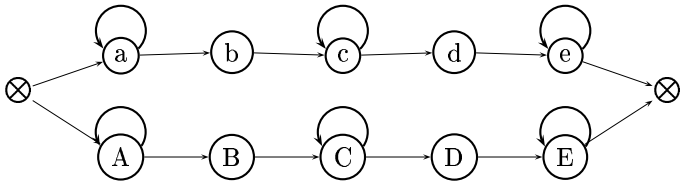


FIG. 5.5 – *Exemple de modèles coopératifs de topologies identiques.*

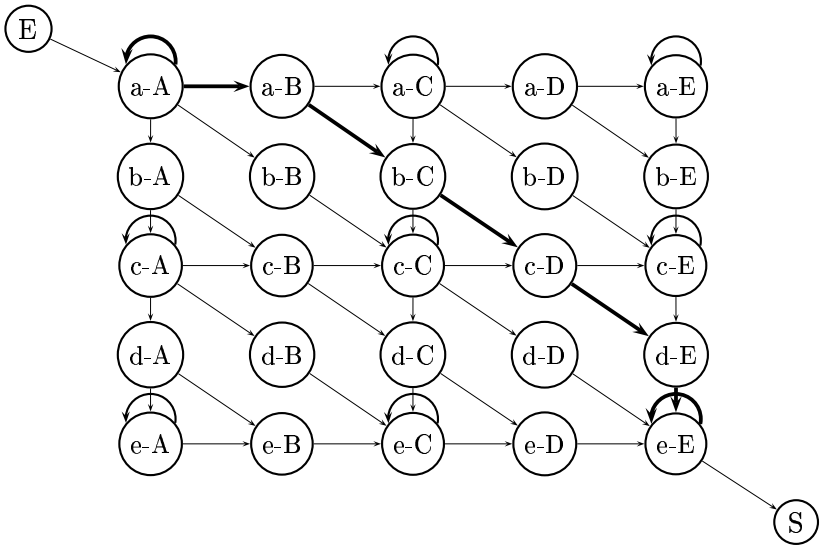


FIG. 5.6 – *Exemple de modèle composite résultant de modèles coopératifs de topologies identiques. Le chemin en traits gras correspond à l'alignement présenté à la figure 5.3. Les mouvements haut-bas correspondent à des transitions dans la chaîne* $\{a,b,c,d,e\}$*, les mouvements gauche-droite correspondent à des transitions dans la chaîne* $\{A,B,C,D,E\}$ *et les mouvement diagonaux correspondent à des transitions simultanées pour les deux canaux.*

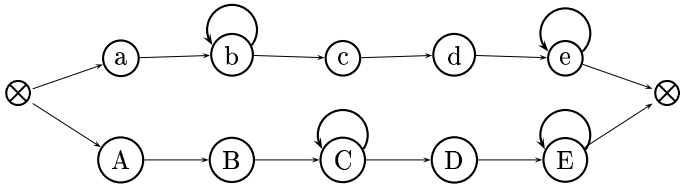


FIG. 5.7 – *Exemple de modèles coopératifs de topologies différentes.*

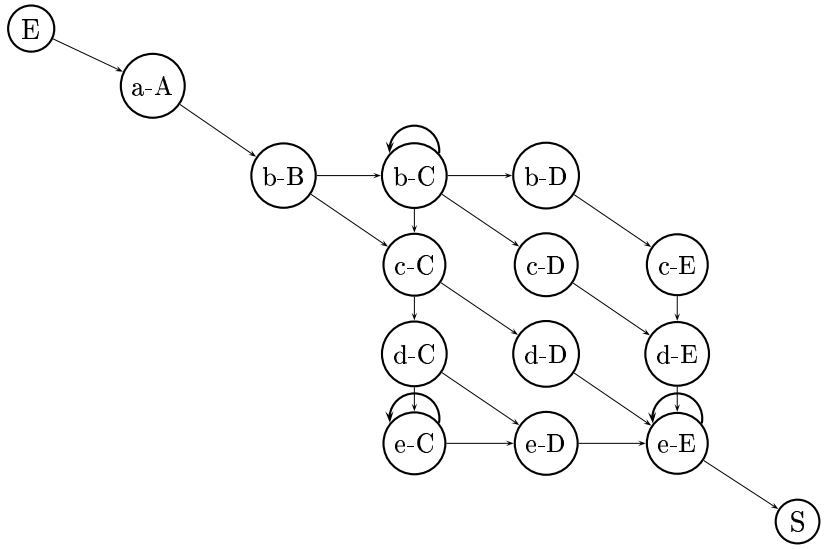


FIG. 5.8 – *Exemple de modèle composite résultant de modèles coopératifs de topologies différentes.*

M donnés:

$$P(X,C|M) = \prod_{j=1}^{J} P(X_j,C_j|M_j) \tag{5.22}$$

$$= \prod_{j=1}^{J} \prod_{n=1}^{N_j} P(x_{j,n}|q_{j,n}) P(q_{j,n}|q_{j,n-1}) \tag{5.23}$$

Rappelons que tout chemin C définit une segmentation temporelle en sous-unités lexicales. Si le modèle M comprend J sous-unités lexicales, cette segmentation conduit à J sous-séquences X_j de vecteurs acoustiques et J sous-chemins d'états notés C_j. Dans les expressions précédentes, N_j est le nombre de trames couvertes par le sous-chemin C_j, $q_{j,n}$ indique l'état visité à l'instant n par le modèle composite associé à M_j, et $x_{j,n}$ est le vecteur d'observation correspondant à la trame n du sous-chemin C_j.

Par **définition du modèle composite**, on a:

$$\begin{cases} P(x_{j,n}|q_{j,n}) = f(P(x^k_{j,n}|q^k_{j,n}),\forall k) \\ P(q_{j,n}|q_{j,n-1}) = f(P(q^k_{j,n}|q^k_{j,n-1}),\forall k) \end{cases} \tag{5.24}$$

où

$$\begin{cases} x_{j,n} = \{x^k_{j,n},\forall k\} \\ q_{j,n} = \{q^k_{j,n},\forall k\} \end{cases} \tag{5.25}$$

où f est une fonction de combinaison. Ces expressions permettent donc l'estimation de $P(X|M)$ suivant l'approche exacte (équation (5.3)) ou suivant l'approche de Viterbi (équation (5.4)).

Récurrences

Sur base de l'approximation de Viterbi, on peut également écrire:

$$P(X|M) = \max_{C} P(X,C|M) \tag{5.26}$$

Pour une segmentation en sous-unités donnée, on obtient alors:

$$\begin{aligned} P(X|M) &= \prod_{j=1}^{J} \max_{C_j} P(X_j,C_j|M_j) && (5.27)\\ &= \prod_{j=1}^{J} \max_{q_1,\ldots,q_{N_j}} \prod_{n=1}^{N_j} P(x_{j,n}|q_{j,n}) P(q_{j,n}|q_{j,n-1}) && (5.28)\\ &= \prod_{j=1}^{J} \max_{q_1,\ldots,q_{N_j}} \prod_{n=1}^{N_j} f\left(P(x^k_{j,n}|q^k_{j,n}),\forall k\right) f\left(P(q^k_{j,n}|q^k_{j,n-1}),\forall k\right) && (5.29)\\ & && (5.30) \end{aligned}$$

Le calcul de chacun des termes du produit sur j fait appel à une récurrence Viterbi classique. Cette récurrence est naturellement étendue vers la reconnaissance de parole continue si la grammaire est d'ordre 1 (bigramme).

La complexité de cette approche par modèles composites est $O(N_{mod}.S^K.N)$ où N_{mod} est le nombre de modèles HMM, S le nombre d'états par modèles, K le nombre de canaux d'information et N le nombre de trames à décoder.

5.3.2 Modélisation de motifs d'asynchronisme

Lorsqu'il s'agit de combiner deux sources d'information différentes, les modèles de Markov cachés conventionnels peuvent être utilisés mais il convient pour cela de concaténer à chaque instant les vecteurs caractéristiques provenant des deux sources d'information. Ces modèles sont donc bien adaptés aux processus qui évoluent de façon parfaitement synchronisée.

Lorsque les processus générateurs évoluent indépendamment les uns des autres, il serait préférable d'utiliser un modèle de Markov caché par processus et de laisser évoluer ces différents modèles indépendamment les uns des autres.

En pratique, les processus intervenant dans la production de parole sont peut-être entre ces deux extrêmes. C'est ce qui nous a conduit à introduire le modèle "multi-stream" qui permet la désynchronisation des différents processus, tout en forçant ceux-ci à se resynchroniser à certains points pré-définis du point de vue lexical.

Une analyse plus fine révèle que le problème n'est pas si simple. Le mouvement des lèvres est lié aux mouvements du conduit vocal et des cordes vocales, mais des

désynchronisations peuvent apparaître à certains moments. Lors de la prononciation d'un mot commençant par une voyelle, les lèvres peuvent commencer à s'ouvrir un peu avant la production du premier son. Dans d'autres cas par contre, les deux processus sont presque parfaitement couplés: lors de la production d'une plosive par exemple. Pour pouvoir exploiter au mieux ces phénomènes, il conviendrait de développer un système qui, sur base de données d'entraînement, détermine les positions lexicales où le couplage entre les différents processus est nécessaire, ainsi que le niveau de couplage requis.

Une première idée est d'utiliser un élagage des états les moins souvent rencontrés dans les modèles composites. Une des conséquences de cet élagage est en effet d'éviter une désynchronisation trop forte des différents modèles là où ils doivent être synchrones, tout en permettant la désynchronisation quand cela est nécessaire. Cette idée a été implémentée. Elle sera testée dans le cadre de la reconnaissance multimodale de la parole.

Une deuxième idée est d'utiliser les probabilités de transition des modèles composites. Les probabilités de transition ne sont en général pas utilisées pour nos systèmes de base. En effet, dans un modèle de Markov caché classique, les probabilités de transition implémentent un modèle de durée très simple (à distribution exponentielle décroissante). Un état pour lequel la probabilité de boucle est élevée correspond à un phonème dont la durée moyenne est plus élevée. Il a cependant été observé qu'en pratique, ces probabilités de transition n'apportent aucune amélioration significative des performances. Cela est peut-être dû au fait que la distribution de durée implémentée est très éloignée de la distribution réelle, qui ressemble plus à une distribution lognormale. Dans le cadre de nos systèmes de base, nous préférons donc utiliser des modèles imposant une durée minimale à chaque phonème (en concaténant plusieurs états identiques sans boucle). Par contre, nous pensons que les probabilités de transition intervenant dans le modèle composite possèdent deux fonctions distinctes: d'une part, elles fournissent un modèle de durée, et d'autre part, elles modélisent le niveau de couplage entre les différents canaux. Considérons par exemple le cas d'un système utilisant deux canaux distincts. Cela conduit à un modèle composite bidimensionnel. Des probabilités de transition élevées vers des états bidimensionnels indicés ij pour lesquels l'écart entre i et j est élevé indiquent un couplage faible entre les deux chaînes. En effet, dans ce cas, le modèle ne défavorise pas la désynchronisation entre les deux canaux. Par contre, si ces probabilités de transition sont faibles, cela indique que le système a tendance à garder le synchronisme entre les deux canaux. Cette idée sera testée dans le cadre du système de reconnaissance multi-modal du Chapitre 7. Notons finalement que l'intérêt des probabilités de transition a également été souligné dans le cadre d'une approche de décomposition en bandes de fréquence [203].

5.4 Equivalence des deux approches

Il est facile de constater que, dans le cadre d'une combinaison par somme pondérée de logarithmes de vraisemblances, si les poids pour chaque canal ont le même signe et sont constants au sein de chaque sous unité lexicale, l'approche Viterbi par modèles

composites proposée ici fournit la même solution que l'approche Viterbi par modèles parallèles de la Section 5.2.

Soit une segmentation en sous-unités lexicales donnée[10], l'estimation de la vraisemblance de la séquence de vecteurs d'observation fait intervenir des termes $P(X_j|M_j)$ où X_j représente le j-ème segment (de longueur N_j) de la séquence de vecteurs d'observation X et M_j le modèle associé à X_j pendant la phase d'alignement temporel. En fonction du niveau de recombinaison, M_j pourrait être le modèle d'un état de HMM, un modèle de phonème ou un modèle représentant n'importe quelle sous-unité lexicale. Pour chaque segment, **supposons que** la recombinaison statistique des canaux obéisse à:

$$\log P(X_j|M_j) = f(\{\log P(X_j^k|M_j^k), \forall k\}) = \sum_{k=1}^{K} \alpha_k \log P(X_j^k|M_j^k) \tag{5.31}$$

où X_j^k est la séquence de paramètres acoustiques associée au canal k, M_j^k est le modèle associé à X_j^k, et α_k sont les paramètres de recombinaison. $P(X_j^k|M_j^k)$ représente donc la vraisemblance d'une séquence partielle X_j^k étant donné un modèle M_j^k. Elle peut être calculée grâce à un HMM standard ou un système hybride HMM/ANN.

Si le segment est une sous-unité lexicale dont le modèle de Markov caché possède plus d'un état, chaque terme de la somme de l'équation (5.31) est le résultat d'un Viterbi classique (approximation du critère du maximum de vraisemblance par le critère de Viterbi, voir les équations (5.9) à (5.11)). On obtient donc:

$$\log P(X_j|M_j) = \sum_{k=1}^{K} \alpha_k \left(\max_{q_1^k, q_2^k, \ldots q_{N_j}^k} \sum_{n=1}^{N_j} \left(\log P(x_{j,n}^k|q_{j,n}^k) + \log P(q_{j,n}^k|q_{j,n-1}^k) \right) \right) \tag{5.32}$$

où $x_{j,n}^k$ est l'observation correspondant au canal k à l'instant n, $q_{j,n}^k$ est l'état du modèle du canal k à l'instant n et $P(q_{j,n}^k|q_{j,n-1}^k)$ est la probabilité de transition de l'état $q_{j,n-1}^k$ à l'état $q_{j,n}^k$. Cela correspond évidemment à chercher le meilleur chemin $q_{j,1}^k, q_{j,2}^k, \ldots q_{j,N_j}^k$ pour chaque canal et à recombiner par somme pondérée les distances accumulées obtenues. La recherche de ce meilleur chemin peut faire appel à une récurrence.

C'est aussi équivalent à chercher le meilleur chemin pour lequel chaque état est cette fois ci défini par K coordonnées, suivant l'expression:

$$P(X_j|M_j) = \max_{\{q_{j,1}^1, \ldots q_{j,1}^K\}, \ldots, \{q_{j,N_j}^1, \ldots q_{j,N_j}^K\}} \prod_{n=1}^{N_j} \prod_{k=1}^{K} \left(P(x_{j,n}^k|q_{j,n}^k) \right)^{\alpha_k} \left(P(q_{j,n}^k|q_{j,n-1}^k) \right)^{\alpha_k} \tag{5.33}$$

C'est ce que fait l'approche par modèles composites utilisant le critère de Viterbi et où les coefficients de pondération α_k (positifs) sont **constants au sein de chaque sous-unité lexicale**. A l'instant n, le modèle multidimensionnel est dans l'état

10. Comme nous l'avons déjà signalé, le calcul global devra considérer toutes les segmentations en sous-unités lexicales possibles, ce qui pourra se faire par programmation dynamique si les grammaires sont d'ordre 1

$q^1_{j,n}$,$q^2_{j,n}$,...$q^K_{j,n}$ et la distance locale[11] associée à cet état est:

$$\sum_{k=1}^{K} \alpha_k \left(\log P(x^k_{j,n} | q^k_{j,n}) \right)$$

De même, la combinaison des probabilités de transition se fait sous forme d'une somme pondérée (avec les mêmes coefficients de pondération) des logarithmes des probabilités de transition.

On peut cependant montrer que l'équivalence n'est plus respectée si $P(X_j|M_j)$ est calculé de manière exacte, c'est-à-dire sans passer par l'approximation de Viterbi.

Comparaison

L'intérêt des modèles parallèles est de présenter une complexité plus faible que la méthode des modèles composites dans le cas d'un nombre important de canaux d'information (supérieur à deux). En effet, le nombre de canaux intervient de façon linéaire dans l'expression de la complexité de la méthode.

5.5 Autres approches similaires

Dans ce paragraphe, nous proposons une brève introduction à d'autres approches similaires généralisant la théorie des modèles de Markov cachés. Les champs de Markov cachés (RFM, random field models) constituent une généralisation des modèles de Markov cachés dans laquelle la variable cachée définit non plus un seul état mais bien un groupe de K états (K étant le nombre de canaux d'information). Il y a donc une similitude avec les modèles "multi-stream" qui sont définis à chaque instant par un K-uple d'états. Ce paragraphe vise à susciter la réflexion à ce sujet.

Soit un processus caché définissant un champ Q. Le champ de Markov caché est défini par un système de voisinages sur un treillis S suivant:

$$P(Q_i = q_i | Q_j = q_j, \forall j \neq i) = P(Q_i = q_i | Q_j = q_j, \forall j \in V_i) \tag{5.34}$$

où V_i indique l'ensemble des points dans le voisinage du point i du treillis, Q_i est la valeur de la variable cachée au point i et Q_j est la valeur de la variable cachée au point j appartenant au voisinage du point i. Ce voisinage peut être défini arbitrairement[12]. A chaque ensemble de points voisins dans le treillis, on peut associer l'équivalent d'une probabilité de transition, appelée fonction de potentiel[13]. C'est une fonction car sa valeur dépend bien entendu des valeurs prises par les variables cachées aux points considérés du treillis.

Comme présenté à la figure 5.9, le champ de Markov caché peut par exemple être défini sur base d'un treillis bidimensionnel $S = \{t,k\}$, où t est l'indice temporel

11. *Distance locale* signifie ici *logarithme de la probabilité d'émission.*

12. Le modèle de Markov caché est donc un champ de Markov caché pour lequel le treillis n'est qu'une simple chaîne indicée dans le temps. Notons bien que cette chaîne ne représente pas la topologie du modèle de Markov caché mais bien la séquence temporelle d'états occupés par le modèle. Le voisinage d'un des points t de cette chaîne est simplement défini comme le point précédent, indicé $t-1$.

13. La valeur de la fonction de potentiel correspond à l'opposé du logarithme de la probabilité de transition.

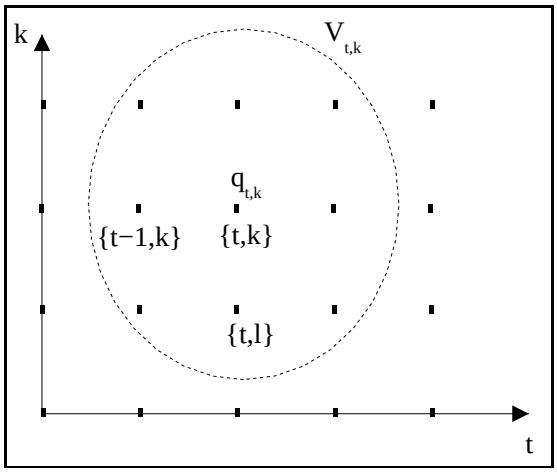


FIG. 5.9 – *Champ de Markov caché (RFM).*

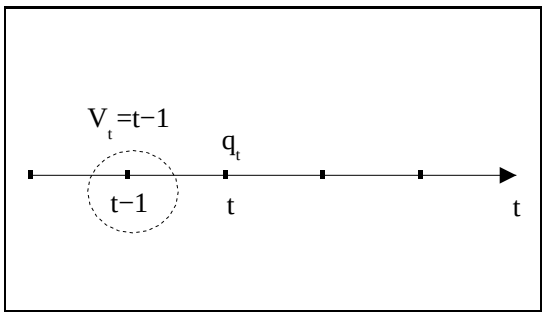


FIG. 5.10 – *Chaîne de Markov cachée (HMM).* $P(q_t|q_j,\forall j \neq t) = P(q_t|q_{t-1})$.

et $1 < k < K$ un indice indiquant le canal d'information. Le processus caché définit donc un champ $Q = \{Q_{t,k}\}$. L'intérêt de la représentation sous forme de champ de Markov caché est de pouvoir introduire des dépendances de l'état au point (t,k) du treillis avec les états aux points (u,l). Lorsque l est différent de k, on introduit donc un couplage entre les différents canaux d'information. Généralement, on aura $u = t-1$, comme c'est le cas pour les modèles de Markov cachés d'ordre un. Dans [76], le voisinage du point (t,k) est définit comme $V_{t,k} = \{(t-1,k),(t,l)\forall l \neq k\}$. Ensuite, on considère que ce voisinage définit deux sortes de cliques. Les cliques $\{(t-1,k),(t,k)\}$ sur lesquelles seront définies des fonctions de potentiel équivalentes aux probabilités de transition des modèles de Markov cachés, et des cliques $\{(t,k),(t,l)\}$ dont les fonctions de potentiel modélisent les interactions entre chaînes.

L'auteur choisit alors de modéliser le couplage entre les chaînes en utilisant, pour la clique $\{(t,k),(t,l)\}$, une fonction potentiel de la forme:

$$U_{t,k,l}(q) = f_{k,l}|q_{t,k} - q_{t,l}| \tag{5.35}$$

où $f_{k,l}$ est un terme de synchronisation car plus il est élevé, plus la différence entre les indices des états $q_{t,k}$ et $q_{t,l}$ (donc la désynchronisation des chaînes) doit être faible pour minimiser la valeur du potentiel et donc maximiser la probabilité (on considère en général que la valeur de la fonction de potentiel équivaut à l'opposé du logarithme d'une probabilité de transition). Une meilleure solution serait de faire dépendre le terme de synchronisation non seulement des chaînes considérées mais également des états $q_{t,k}$ et $q_{t,l}$, certains états conduisant par exemple à une synchronisation plus élevée.

Tout ce qui précède définit l'équivalent des probabilités de transition des modèles de Markov cachés. La distance globale associée à ce modèle est calculée comme la somme des potentiels de chaque point (et son voisinage associé). On ajoutera à cette distance globale un terme calculé comme la somme des distances locales associées aux points du treillis. On pose donc l'hypothèse d'indépendance entre les observations qui correspondent aux différentes chaînes du modèle. Ces distances locales dépendent des états parcourus et peuvent par exemple être calculées sur base de distributions de probabilité multi-gaussiennes ou sur base de réseaux de neurones artificiels.

Des algorithmes existent pour le décodage et pour l'entraînement. Ils sont cependant très lourds. Dans [76], une approche heuristique est utilisée pour estimer les paramètres des modèles. Il s'agit de maximiser la vraisemblance pour chaque canal indépendamment des autres (donc sans tenir compte des interactions entres canaux). Ensuite, les paramètres des fonctions de potentiel associées aux cliques $\{(t,k),(t,l)\}$ sont estimés sur base d'un comptage utilisant le meilleur chemin pour chaque canal.

La figure 5.10 représente un modèle de Markov caché sous forme de champ de Markov caché. Le champ est ici unidimensionnel. Il est important de noter qu'il ne s'agit pas d'une représentation topologique du modèle mais bien d'une représentation sur base d'un axe temporel où chaque point correspond à un instant donné. L'état que peut visiter le modèle à chaque instant est cependant guidé par la topologie du modèle sous-jacent.

D'autres auteurs proposent l'utilisation de modèles de Markov cachés factoriels [68, 122] (FHMMs, factorial hidden Markov models) ou de modèles de Markov

cachés couplés [23, 24] (CHMMs, coupled hidden Markov models). Ceux-ci sont similaires aux champs de Markov cachés bidimensionnels présentés ici. Dans [68, 122] cependant, aucun couplage n'est introduit entre les différentes chaînes. Dans [122], cette approche a été utilisée dans le cadre de la reconnaissance automatique de la parole. Aucun résultat probant n'a été obtenu. Dans [24], l'approche a été utilisée avec succès pour la reconnaissance de mouvement Taï-chi-chuan. Il s'agit d'un art martial chinois consistant en mouvement corporels stylisés et bien définis. Les paramètres utilisés pour la reconnaissance sont les positions spatiales de la tête et des mains. Plusieurs architectures HMM ont été développées, la meilleure conduisant à un taux d'erreur supérieur à 30% pour la reconnaissance sur base d'un 'vocabulaire' composé de trois mouvements différents. Une approche basée sur les modèles de Markov couplés conduit à un taux d'erreur de 5.8% pour la même tâche.

On peut trouver dans la littérature des modèles plus généraux incluant les HMMs ou même les champs de Markov cachés. Ces modèles apparaissent sous différents noms: Bayesian Networks, Graphical (Association) Models, Probabilistic (Independence) Networks, Belief Networks ou Coupled Networks [185, 186, 100, 171]. Zweig [214, 215] en fait une application dans le domaine de la reconnaissance automatique de la parole. Ces modèles ne seront pas discutés ici.

5.6 Conclusions

Nous avons présenté deux approches permettant la modélisation de processus partiellement découplés. La première, appellée approche **"multi-stream" par modèles parallèles**, consiste à modéliser les différents processus par des modèles de Markov cachés coopératifs. Cette coopération réside dans la resynchronisation des chemins d'états à certains endroits prédéfinis dans la topologie des modèles. C'est également à ces endroits que les contributions des différentes chaînes sont combinées.

La seconde approche, appellée approche **"multi-stream" par modèles composites**, consiste à construire des topologies multidimensionnelles. Chacune des directions de ces topologies représentera l'évolution des régimes stationnaires pour un des différents processus. Il existe une équivalence avec l'approche précédente. Il est en effet possible de construire des topologies multidimensionnelles de façon à ce que chacun des états représente un K-uple d'états des modèles parallèles (K étant le nombre de processus modélisés). Les contributions locales des états des modèles composites sont alors estimées comme une combinaison des contributions locales des différentes composantes de ces états. Sous l'hypothèse d'indépendance, c'est-à-dire si la combinaison se fait par produit de vraisemblances, on peut montrer que les deux approches sont tout à fait équivalentes.

Nous avons discuté des problèmes d'entraînement et de reconnaissance (et donc d'estimation) dans le cadre de ces modèles. Ceux-ci étant très lourds, nous avons également proposé un décodeur efficace destiné à l'approche par modèles parallèles. Celui-ci est basé sur l'approximation de Viterbi et sur des techniques de recherche en faisceau. L'approche par modèles composites peut quant à elle utiliser des techniques de décodage et d'élagage classiques.

Finalement, nous proposons une brève introduction à d'autres approches simi-

laires et notamment aux champs de Markov cachés.

Par la suite, l'approche "multi-stream" par modèles composites sera utilisée dans le cadre de la reconnaissance vocale audiovisuelle, les informations acoustiques et les informations concernant le mouvement des lèvres étant considérées comme provenant de processus partiellement découplés. Nous avons suggéré ici que l'approche par modèles composites permet de modéliser le couplage entre les différents processus. Cette possibilté sera également exploitée expérimentalement dans le cadre de la reconnaissance audiovisuelle.

Chapitre 6

Approche multi-bande

6.1 Introduction

Vers 1918, Fletcher entame aux *Bell Labs* des études ayant pour objectif la quantification de la qualité d'un signal vocal. Le but de ces travaux est de tenter d'optimiser l'intelligibilité et le confort d'écoute de signaux transmis par téléphone. Les résultats obtenus, fruits de nombreux tests d'écoute, fournissent un modèle permettant de prédire l'intelligibilité d'un signal vocal sur base du rapport signal/bruit dans 20 bandes de fréquence, chacune couvrant deux bandes critiques. Ce modèle, publié dans [57] (et repris par Allen [2]), a notamment conduit à l'élaboration d'une structure schématisant le processus de reconnaissance vocale humain. Celle-ci suggère que la perception auditive humaine est basée sur des bandes de fréquence analysées indépendamment les unes des autres. Le processus d'analyse consiste à extraire des paramètres représentatifs ainsi qu'une mesure du rapport signal/bruit dans la bande considérée. Les résultats obtenus sont ensuite recombinés à un niveau plus élevé (c'est à dire plus tard) dans le processus de reconnaissance vocale. La recombinaison des résultats fournis par le traitement de ces bandes de fréquence se fait de sorte que le taux d'erreur global au niveau du phonème [1] est égal au produit des taux d'erreur obtenus sur chaque bande indépendamment. Il s'agit d'une des conclusions fondamentales des travaux de Fletcher. Elle signifie que les erreurs commises dans les différentes bandes de fréquence sont indépendantes et qu'une erreur globale ne peut apparaître que si toutes les bandes de fréquence sont en erreur. Si au moins une bande de fréquence fournit un résultat de reconnaissance correct, le résultat sera correct au niveau global également. Le système auditif semble donc capable de déterminer les bandes de fréquence qui conduisent à une reconnaissance vocale correct. Ce résultat fondamental ne fournit cependant aucun moyen de détecter les bandes de fréquence valides.

Les résultats de Fletcher ont inspiré d'autres travaux également basés sur des tests d'écoute. Alors que les expériences de [57] concernent des signaux filtrés soit

1. Dans des conditions idéales (rapport signal/bruit élevé), le taux d'erreur au niveau du phonème est de 1.5%.

passe-haut, soit passe-bas, celles rapportées dans [79] et dans [117] concernent l'intelligibilité sur base de l'utilisation canaux fréquentiels distincts. Dans [117], il est montré que l'intelligibilité de syllabes CVC[2] reste élevée lorsqu'on utilise deux bandes de fréquence distinctes: une bande en basse fréquence (de 0 à 800 Hz) et une bande en haute fréquence (au-delà de 4000 Hz ou même de 8000 Hz). Dans [79], des expériences similaires semblent indiquer que l'intelligibilité humaine est bien supérieure à celle qui peut être prédite par le modèle de Fletcher.

D'autres études psycho-acoustiques suggèrent que des portions fréquentielles différentes sont responsables de la transmission de différentes "qualités" phonétiques (comme le caractère voisé, nasal ou fricatif d'un son). Dans [70], des syllabes CVC sont découpées en 12 blocs sur base de quatre sections temporelles et de trois sections fréquentielles: de 0 à 1000 Hz, de 1000 à 2500 Hz et au-delà de 2500 Hz. L'expérience consiste à intervertir des blocs provenant de syllabes différentes mais positionnés au même endroit dans la découpe en 12 blocs. Les résultats indiquent que le caractère voisé et nasal d'un son est très sensible à l'interversion de blocs correspondant à la première bande de fréquence, alors que le caractère fricatif est très sensible à des interversions dans la troisième bande de fréquence. Dans [141], des expériences basées sur un filtrage passe-bas montrent que le caractère voisé est le mieux préservé, suivi par le caractère nasal et finalement le caractère fricatif. Les conclusions de ces deux travaux sont donc similaires.

Les résultats de [70] suggèrent également que des régions fréquentielles différentes d'un signal vocal ont des caractéristiques dynamiques différentes. Un bon couplage entre les caractéristiques du système de production et du système de perception pourrait donc passer par des options d'analyse optimisées en fonction des caractéristiques dynamiques de chaque bande de fréquence. Par exemple, la largeur du contexte acoustique et la résolution temporelle du processus d'analyse pourrait dépendre de la bande de fréquence traitée.

Malgré certaines contradictions, tous ces résultats suggèrent que le processus de reconnaissance humain est fondamentalement différent des méthodes de classification phonétique utilisées actuellement. Ces dernières sont en effet très sensibles au filtrage et au bruit. Une architecture de reconnaissance originale, que nous qualifierons de multi-bande, est partiellement inspirée et motivée par les travaux cités précédemment. Son principe est le suivant (voir figure 6.1):

- Analyse, estimation de probabilités phonétiques et éventuellement catégorisation phonétiques dans différentes bandes de fréquence, sur base de "sous-reconnaisseurs" dont l'implémentation repose sur des méthodes classiques. Les paramètres de ces méthodes, notamment leurs caractéristiques dynamiques, peuvent dépendre de la bande de fréquence considérée.
- Recombinaison des résultats d'analyse, d'estimation et de catégorisation correspondant aux différentes bandes de fréquence et intégration sur des intervalles de temps plus larges. Le processus de recombinaison vise à obtenir les probabilités associées à des unités lexicales de base. Celui d'intégration temporelle vise à obtenir des hypothèses de reconnaissance complètes (mots ou phrases).

2. Une syllabe CVC est une séquence consonne-voyelle-consonne.

Plusieurs options sont envisageables quant aux traitements effectués dans chacune des bandes de fréquence, avant recombinaison. Il peut s'agir d'une simple analyse dans le but d'obtenir des paramètres représentatifs propres à chacune des bandes de fréquence traitées [155]. Dans ce cas, la recombinaison peut simplement correspondre à une concaténation des paramètres obtenus. Il est également possible de poursuivre jusqu'à l'obtention de paramètres discriminants, voire d'estimations des probabilités associées aux différentes unités phonétiques [17, 21, 86, 144, 155, 192]. Dans ce cas, la recombinaison impliquera vraisemblablement un formalisme statistique permettant la fusion des probabilités fournies pour les différentes bandes de fréquence. On pourra finalement envisager d'aller jusqu'à la pré-catégorisation phonétique dans chacune des bandes de fréquence, comme proposé dans [28, 41]. Les travaux de [70, 141] motivent par ailleurs l'étude de classes phonétiques spécialisées à chacune des bandes de fréquence, comme dans [143].

D'autre part, l'architecture "multi-stream" qui a été étudiée précédemment donne la possibilité de se libérer de la contrainte de synchronisme imposée par l'approche classique (voir Chapitre 5). Comme représenté schématiquement à la figure 6.3, chaque bande pourra évoluer indépendamment des autres. Plus explicitement, dans une bande de fréquence, les transitions d'un état du modèle acoustique à un autre pourront être non synchrones aux transitions dans les autres bandes. L'étude de cette désynchronisation est justifiée par le fait que les événements acoustiques semblent eux-mêmes non synchrones. Il suffit pour cela d'observer des spectrogrammes comme celui des figures 6.2 et 6.3. Le système d'intégration humain semble par ailleurs robuste à de faibles désynchronisations temporelles entre différents canaux d'information [3].

Les figures 6.2 et 6.3 illustrent les différences entre l'approche classique et l'approche multi-bande. Rappelons que l'approche classique procède par extraction d'un vecteur de paramètres représentatifs de la parole dans toute la plage de fréquence. Ce vecteur est calculé périodiquement, toutes les 10 ms par exemple. Il est ensuite utilisé par le système de reconnaissance proprement dit. L'approche multi-bande, quant à elle, extrait un vecteur de paramètres pour chaque bande de fréquence considérée. Chaque vecteur est alors utilisé par un sous-reconnaisseur indépendant des autres, effectuant un traitement plus ou moins avancé, avant de fournir des résultats à un module chargé de la décision globale.

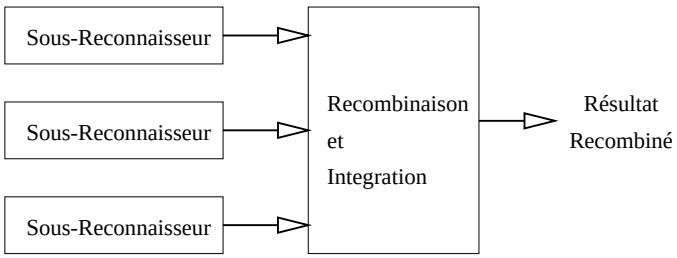


FIG. 6.1 – *Approche multi-bande.*

Outre les motivations psycho-acoustiques résultant des travaux cités au début de

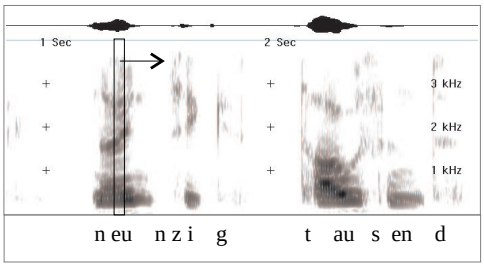


FIG. 6.2 – *Approche classique.*

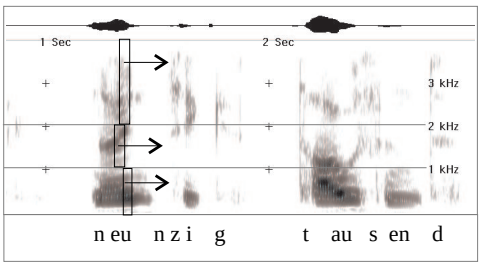


FIG. 6.3 – *Approche multi-bande.*

cette section, une telle architecture offre différents avantages techniques potentiels:

- *Robustesse au bruit* – En effet, un bruit étant très rarement blanc (voir par exemple l'Annexe B qui propose l'analyse de quelques types de bruits), il ne perturbe pas le signal vocal de façon équivalente dans toutes les bandes de fréquence. D'autre part, même si le bruit est blanc, il est possible qu'il n'affecte pas de la même façon l'information linguistique des différentes bandes de fréquence, les différentes qualités phonétiques n'étant pas transmises de la même façon dans les différentes bandes. Avec l'approche classique (figure 6.4), le vecteur acoustique est utilisé comme une seule entité. De ce fait, une composante bruitée de ce vecteur (à cause d'un bruit coloré par exemple) entraînera une dégradation importante des performances du système. Par contre, avec l'approche proposée (figure 6.5), la dégradation ne touchera que les vecteurs acoustiques des bandes de fréquence bruitées. Il est alors possible d'obtenir un meilleur résultat si on parvient à diminuer l'importance, voir à rejeter les décisions des reconnaisseurs utilisant ces bandes.
- *Meilleur usage de l'information temporelle et du compromis temps/fréquence.* Il est en effet possible d'optimiser indépendamment chaque sous-reconnaisseur en ce qui concerne leur résolution temporelle et la largeur du contexte acoustique utilisé. Ce point à déjà été introduit au début de cette section.
- *Optimisation du type de paramètres.* Dans les bandes de fréquence de 700 à 2800 Hz, les performances des sous-reconnaisseurs du système décrit dans [41] sont, selon l'auteur, bien inférieures à celles de l'homme. Cette observation suggère que l'approche de reconnaissance envisagée (HMM) n'utilise pas cor-

rectement l'information présente dans ces bandes de fréquence. Il est probable que les paramètres testés (cepstres, coefficients LPC, fonction d'autocorrélation et leurs dérivées) ne sont pas appropriés à la tâche de reconnaissance dans des bandes étroites. Cette remarque suggère que les différentes bandes de fréquence sont sensibles aux paramètres acoustiques, qui pourraient alors être optimisés indépendamment pour chaque bande de fréquence.

– *Développement de classes phonétiques spécialisées.* L'intuition, et les résultats de [70, 141], suggèrent que les différentes bandes de fréquence contiennent plus ou moins d'information concernant les différents traits phonétiques. Si une bande de fréquence ne transmet que très peu d'information pour certains traits phonétiques, on pourrait fusionner en une seule classe les phonèmes qui se distinguent uniquement par ces traits.

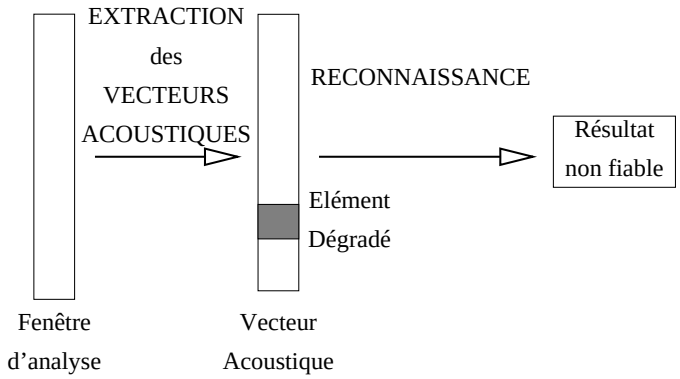


FIG. 6.4 – *Approche classique.*

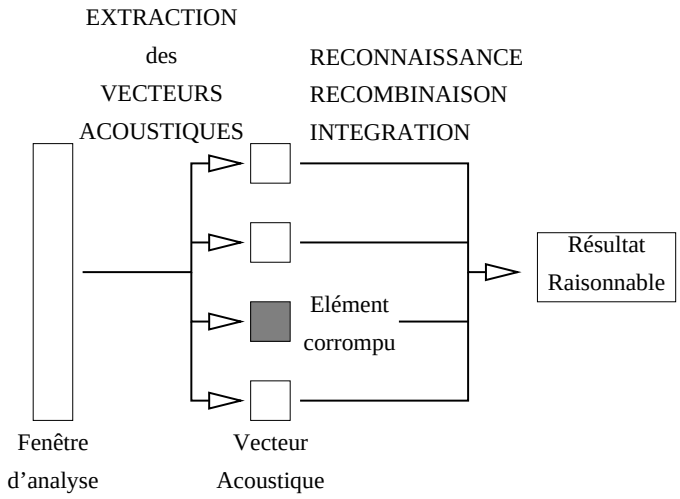


FIG. 6.5 – *Approche de traitement indépendant des bandes de fréquence.*

6.2 Etat de l'art

Des travaux préliminaires exploitant certaines de ces idées sont présentés dans [41] et [147]. Cependant, les expériences ont été faites sans tester les systèmes résultants dans des conditions difficiles de bruit, conditions dans lesquelles ce type de reconnaisseur pourrait révéler quelque intérêt.

Plus récemment, différentes équipes se sont intéressées à l'étude expérimentale du concept multi-bande. Dans le cas de parole perturbée par du bruit coloré, les résultats obtenus (synthétisés ci-dessous), semblent indiquer la supériorité de l'approche multi-bande par rapport à une architecture de reconnaissance classique.

6.2.1 Travaux de Hermansky et Tibrewala

Dans [86], les auteurs envisagent une tâche de reconnaissance de chiffres isolés. Les sous-reconnaisseurs ont pour fonction d'estimer la vraisemblance associée à chacun des mots du vocabulaire[3]. Les paramètres utilisés dans chaque sous-bande sont les énergies de bandes critiques. Le module de recombinaison a pour but d'obtenir un "score" final sur base de ces vraisemblances. Deux approches de combinaison ont été envisagées: une combinaison linéaire des logarithmes des vraisemblances associées aux différentes sous-bandes et finalement une combinaison par perceptron multicouche (MLP) utilisant les logarithmes de vraisemblances. Dans le cas de parole claire, les performances des systèmes de reconnaissance multi-bandes utilisant deux ou sept bandes et une combinaison par MLP sont similaires à celles du système classique utilisant les énergies des bandes critiques comme paramètres représentatifs. La combinaison linéaire conduit à des performances sensiblement moins bonnes.

Un second ensemble d'expériences montre la supériorité du système multi-bande (7 bandes, combinaison par MLP) dans le cas de parole bruitée par un signal sinusoïdal à 900 Hz. A un rapport signal/bruit de 10 dB par exemple, le taux d'erreur passe de 35 % à 20 %. Plutôt que d'utiliser systématiquement les 7 bandes de fréquence, les auteurs proposent ensuite de combiner uniquement les groupes de bandes les moins perturbés. Ceci conduit à développer un ensemble de 127 MLPs de combinaison ainsi qu'une méthode de sélection appropriée. En choisissant pour chaque mot le meilleur des systèmes, on obtient des performances qui restent constantes sur toute la plage de rapport signal/bruit envisagée (de 30 à 0 dB): taux d'erreur inférieur à 5 %. Ceci suggère qu'il existe au moins un groupe de bandes qui conduit à des performances exceptionnelles. Une méthode automatique basée sur un vote de majorité est ensuite proposée: elle conduit à un taux d'erreur de 15 % (à 10 dB). L'inconvénient de cette dernière méthode est d'impliquer un nombre très important de systèmes de reconnaissance (127 groupes de bandes).

Dans [192], les auteurs poursuivent leurs investigations sur la reconnaissance de chiffres isolés. Ils envisagent des systèmes basés sur 2, 4 ou 7 bandes de fréquence et comparent deux types de paramètres représentatifs: les énergies des bandes critiques et les cepstres calculés sur base des énergies de bandes critiques (approche PLP - "Perceptual Linear Prediction" [83]). En parole claire, les performances des systèmes multi-bandes sont similaires à celles des systèmes classiques utilisant le même type

3. Ils constituent donc des systèmes de reconnaissance complets.

de paramètres représentatifs. Globalement, les paramètres PLP conduisent à de meilleurs résultats. En vue de passer à la reconnaissance de parole continue, les auteurs envisagent ensuite une combinaison au niveau de la trame. Dans ce cas donc, les sous-reconnaisseurs fournissent des log-vraisemblances pour chaque trame. Celles-ci sont ensuite combinées par un MLP. Une phase de décodage classique permet finalement la reconnaissance de mots ou de phrases. Les résultats obtenus montrent que cette approche est aussi performante que l'approche précédente de combinaison au niveau du mot, impliquant un décodeur par bande de fréquence.

Des essais sont ensuite effectués sur de la parole perturbée par des bruits réels provenant de la base de données NOISEX. Les résultats sont cette fois-ci plutôt mitigés. Pour certains bruits ("factory", "destroyer-engine", "volvo", "babble" et "pink noise"), le système basé sur une décomposition en sept bandes de fréquence donne de meilleurs résultats que le système de référence: le taux d'erreur moyen passe de 25 % à 10 %. Pour d'autres types de bruits cependant ("white noise", "high-frequency radio channel"), ce système conduit à une dégradation des résultats: le taux d'erreur passe de 25 % à plus de 35 %. Une caractéristique commune de ces derniers bruits est de perturber significativement toutes les bandes de fréquence.

Dans [193] finalement, les auteurs proposent une approche de type multi-bande dans laquelle chaque bande de fréquence (parmi 15) utilise des paramètres représentatifs d'un segment temporel de l'ordre de 1 s. Les scores (log-vraisemblances) fournis par les 15 sous-reconnaisseurs sont ensuite combinés par MLP avant d'effectuer le décodage. Les tests effectués sur une tâche de reconnaissance de nombres connectés conduisent à des résultats inférieurs à ceux du système de référence, en parole claire, comme en parole bruitée (bruit blanc). La combinaison des scores fournis par le système de référence et le système multi-bande conduit cependant à une amélioration significative des performances.

Les conclusions fondamentales de cet ensemble de travaux sont les suivantes:

- l'utilisation de paramètres de type PLP donne de meilleurs résultats que l'utilisation directe des énergies des bandes critiques.
- en parole claire, un système multi-bande utilisant une combinaison par MLP conduit à des performances similaires à celles d'une architecture classique.
- en parole bruitée, l'approche multi-bande conduit à une robustesse inhérente aux bruits colorés, sans nécessité de désigner les bandes les moins perturbées.
- un choix dynamique des bandes de fréquence les moins perturbées conduit à une robustesse accrue, mais peut être fort lourd.
- dans le cas de bruit large bande (bruit blanc par exemple), l'approche multi-bande conduit à une dégradation des résultats.
- dans tous les cas, il semble intéressant de combiner les scores fournis par un système multi-bande et un système classique.

6.2.2 Travaux de Morgan et Mirghafori

Les travaux de ce groupe se sont orientés dans trois directions:

- *la possibilité de désynchronisation des bandes de fréquence* [143, 145]: Des systèmes multi-bandes basés sur deux ou quatre bandes de fréquence ont été

développés[4]. Les sous-reconnaisseurs utilisent des paramètres caractéristiques de type PLP et ont pour but d'estimer les vraisemblances à associer aux unités phonétiques. Ces vraisemblances sont ensuite combinées par simple multiplication (hypothèse d'indépendance) pour conduire aux vraisemblances globales utilisées durant le décodage. Le décodage utilise une approche par modèles composites de mots (voir Section 5.3) et permet donc le choix de chemins de programmation dynamique indépendants pour les différentes bandes de fréquence. Des tests ont été effectués sur une tâche de reconnaissance de nombres connectés, pour de la parole claire et pour de la parole réverbérée. Les résultats indiquent une légère dégradation des performances par rapport à une approche multi-bande utilisant les mêmes sous-reconnaisseurs mais dans le cadre d'une combinaison/intégration forçant le synchronisme entre les bandes de fréquence.

- *la possibilité de construire des classes phonétiques spécialisées pour chaque sous-bande* [143]: Cette étude est motivée par l'observation que certaines bandes de fréquence contiennent plus d'information pour distinguer certaines classes phonétiques que d'autres. L'idée qui a été appliquée est de fusionner les phonèmes les plus proches dans chaque sous-bande. Plusieurs critères de similitude ont été envisagés. Les résultats obtenus sur base de nombres connectés indiquent une très légère amélioration dans le cas de parole réverbérée, par rapport à un système multi-bande utilisant des unités phonétiques classiques. Dans le cas de parole claire ou bruitée, les performances des deux types de systèmes sont similaires.
- *la combinaison des scores fournis par un systèmes multi-bande et un système classique* [144]: cette méthode conduit à une amélioration significative des performances dans le cas de parole claire ainsi que de parole réverbéré. Le taux d'erreur de 4.7 % en parole claire sur la base de données NUMBERS'95 est le plus faible publié jusqu'à présent.

Les conclusions de ces travaux sont donc les suivantes:

- quelles que soient les conditions, l'utilisation d'une approche de combinaison/intégration permettant l'asynchronisme des bandes de fréquence n'apporte aucune amélioration des performances. Une étude plus approfondie pourrait cependant être effectuée dans le but d'imposer des contraintes de synchronisme dont les paramètres seraient estimés sur base de données d'entraînement.
- l'utilisation de classes phonétiques spécialisées aux différentes bandes de fréquence n'apporte généralement pas d'amélioration, sauf dans le cas de parole réverbérée, où une très légère amélioration a été observée. Cette approche est donc un sujet potentiellement intéressant pour une étude beaucoup plus approfondie.
- dans tous les cas, il semble intéressant de combiner les scores fournis par un système multi-bande et un système classique.

4. Notons que pour les expériences de désynchronisation, l'auteur utilise l'approche par modèles composites ainsi que l'approche par modèles parallèles. Dans le premier cas, il se focalise uniquement sur le système à 2 bandes de fréquence, celui à 4 bandes étant fort lourd.

6.2.3 Travaux de Haton et Cerisara

Les travaux de cette équipe [27, 29] se sont d'abord orientés vers une approche multi-bande inspirée des travaux de Duchnowsky [41]. Il s'agit d'effectuer une catégorisation phonétique par reconnaissance complète dans chaque bande de fréquence. Il convient ensuite de combiner les étiquettes phonétiques ainsi obtenues.

Par la suite, cette méthode a été abandonnée au profit d'approches n'effectuant pas de pré-catégorisation phonétique au sein des bandes de fréquence. Le travail s'est alors orienté vers quatre thèmes:

- *Le formalisme de combinaison*: Cerisara propose de comparer une approche de combinaison par somme pondérée des vraisemblances associées aux différentes bandes et une approche de combinaison des vraisemblances par réseau de neurones artificiels. Comme précédemment (voir Section 6.2.1), l'approche non-linéaire donne de meilleurs résultats, même si l'approche linéaire fait ici appel à une optimisation des poids par minimisation de l'erreur de classification (critère MCE). Finalement, l'approche multi-bande conduit à une amélioration par rapport à un système classique pour la reconnaissance de parole continue non bruitée en français (corpus BREF-80 [67]).
- *L'asynchronisme des bandes de fréquence*: Un algorithme de décodage inspiré du "Two-Level" est proposé. Les résultats ne sont cependant pas très clairs et ne confirment en tout cas pas l'intérêt de cette approche.
- *La définition de classes phonétiques spécialisées*: on trouvera une étude préliminaire concernant la responsabilité de différentes bandes de fréquence pour la classification des différents types de phonèmes.
- *L'estimation des paramètres du modèle multi-bande*: les auteurs proposent un algorithme d'entraînement "global" sur base du critère de minimisation de l'erreur de classification. Une légère amélioration est obtenue par rapport au système multi-bande basé sur une optimisation indépendante pour chaque bande de fréquence.

Une des conclusions de ces travaux est à nouveau l'intérêt d'une combinaison par réseau de neurones artificiels. Les résultats concernant l'asynchronisme et la définition de classes phonétiques ne sont pas concluants.

6.2.4 Travaux de Bourlard et Dupont

Ces travaux [17, 18, 21], concernant le choix des paramètres représentatifs, la reconnaissance de parole claire ou bruitée et la désynchronisation des bandes de fréquence ont conduit aux conclusions suivantes:

- l'utilisation de paramètres cepstraux de type PLP donne de meilleurs résultats que l'utilisation directe des énergies des bandes critiques.
- en parole claire, un système multi-bande utilisant une combinaison par MLP conduit à des performances similaires à celles d'une architecture classique.
- en parole bruitée, l'approche multi-bande conduit à une robustesse inhérente aux bruits colorés, sans nécessité de désigner les bandes les moins perturbées.

– l'utilisation d'une approche de combinaison/intégration permettant l'asynchronisme des bandes de fréquence n'apporte aucune amélioration des performances.

6.2.5 Autres contributions

Dans [138], les auteurs combinent les vraisemblances obtenues sur base de paramètres représentatifs de différentes bandes de fréquence. Ils utilisent une simple somme des logarithmes de vraisemblances (hypothèse d'indépendance). Ils obtiennent une amélioration des performances en parole claire comme en parole bruitée (bruit blanc) en combinant le système de référence avec un système basé sur deux ou quatre bandes de fréquence. La combinaison se fait ici aussi suivant l'hypothèse d'indépendance.

Dans [150], les auteurs développent une approche où, comme dans [86], chaque groupe de bande possède son modèle. Les probabilités a posteriori fournies par les différents groupes de bandes de fréquence peuvent alors être combinées suivant l'approche de mixture d'experts (voir Section 6.5.3), la difficulté étant d'estimer les coefficients de pondération à associer à chaque groupe de bandes. L'approche étant très lourde, les auteurs proposent également une approximation sur base de l'hypothèse d'indépendance des bandes de fréquence. Les résultats obtenus sur une tâche de reconnaissance de nombres connectés montrent une légère amélioration par rapport au système de référence dans le cas d'un bruit de voiture (NOISEX-92). De plus, l'hypothèse d'indépendance ne semble pas affecter fortement le système.

Dans [72], Glotin propose également d'utiliser un modèle par groupe de bandes. Cette fois cependant, il s'agira par la suite de sélectionner un des groupes de bandes, comme dans [86]. Le critère de sélection fait intervenir une approche simple d'analyse de scène auditive. Une importante réduction du taux d'erreur est observée dans le cadre de la reconnaissance de nombres connectés affectés par un bruit ne perturbant qu'une seule des quatre bandes de fréquence intervenant dans le système.

Dans [155], les auteurs comparent deux approches de combinaison sur base d'un système à trois bandes de fréquence. L'approche de combinaison des paramètres consiste simplement à concaténer les paramètres représentatifs des différentes bandes de fréquence. La différence avec un système classique est minime: il s'agit simplement de calculer des cepstres MFCC par bandes de fréquence plutôt que sur base du spectre large-bande. La deuxième approche envisagée est plus proche des travaux cités précédemment. Il s'agit d'utiliser des sous-reconnaisseurs estimant les vraisemblances pour chacune des bandes de fréquence. Ces vraisemblances sont ensuite combinées par simple produit (hypothèse d'indépendance) faisant éventuellement intervenir des coefficients de pondération en exposant. Les vraisemblances combinées sont ensuite utilisées dans une phase de décodage classique. Des résultats sont rapportés sur base d'une tâche de reconnaissance de parole continue indépendante du locuteur (corpus ATIS - "Air Travel Information Service"). Pour certains types de bruits ("babble", "destroyer engine" et "machine gun" de la base de données NOISEX-92), les deux approches multi-bandes conduisent à une robustesse accrue par rapport à une approche classique. Pour d'autres types de bruits cependant (bruit blanc par exemple), les systèmes multi-bandes sont moins performants. Finalement,

la combinaison des paramètres donne de meilleurs résultats que la combinaison des vraisemblances. Rappelons cependant que la seconde approche fait ici appel à l'hypothèse d'indépendance alors que d'autres travaux (voir Section 6.2.1) ont montré la supériorité d'une recombinaison sur base d'un réseau de neurones artificiels.

Dans [195], les auteurs s'intéressent à la possibilité de désynchronisation des bandes de fréquence. Ils ont développé un système basé sur deux bandes de fréquence: une bande de 0 à 4000 Hz et une bande de 4000 à 8000 Hz. Les états des modèles de Markov cachés sont représentés par des modèles gaussiens à matrices de covariance diagonales et les observations correspondant aux deux bandes de fréquence sont par conséquent supposées indépendantes. Le décodage utilise une approche par modèles composites de phonèmes (voir Section 5.3), les phonèmes étant constitués d'une séquence de trois états différents. Ce décodage permet donc le choix de chemins de programmation dynamique indépendants au sein des phonèmes. Des tests ont été effectués sur une tâche de reconnaissance de parole continue (500 mots) dépendante du locuteur. Les résultats rapportés montrent une réduction importante du taux d'erreur. Des essais ont finalement été effectués sur base d'un modèle à trois bandes de fréquence, sans succès cette fois ci. De même, les essais d'extension de la plage d'asynchronisme au delà du phonème n'ont pas été concluant. Bien qu'il s'agisse là du premier travail dont les résultats expérimentaux concernant l'asynchronisme sont encourageants, la généralisation des conclusions est encore incertaine.

Dans [123], on suggère l'utilisation de modèles de Markov cachés factoriels pour permettre la désynchronisation des bandes de fréquence. Les résultats de reconnaissance phonétiques sur base du corpus TIMIT sont cependant non concluants.

Dans [6, 182], on trouvera quelques travaux dans le domaine de l'identification/vérification du locuteur.

6.2.6 Critique des travaux précédents

Une faiblesse importante de tous les travaux de recherche résumés ici concerne les essais dans le cas de parole bruitée. La robustesse au bruit est présentée comme une des potentialités de l'approche mais les méthodes développées sont très rarement comparées avec des approches de reconnaissance robustes classiques, faisant maintenant partie de l'état de l'art: la soustraction spectrale par exemple. Dans ce chapitre, nous tenterons d'éclaircir quelque peu ce problème en présentant des résultats expérimentaux couvrant une gamme importante de techniques, de paramètres représentatifs et de bruits.

6.3 Développements

L'étude théorique et expérimentale qui suivra vise à clarifier le concept et l'intérêt de l'approche multi-bande. Elle portera essentiellement sur les points suivants:

- *paramètres représentatifs*: nous utiliserons des paramètres représentatifs de type standard et comparerons les performances obtenues sur base de différentes formes de paramètres et de différents types de traitement robuste.
- *combinaison*: nous présenterons différentes méthodes d'ensembles permettant la combinaison des "décisions" fournies par les différents sous-reconnaisseurs.

Nous envisagerons la possibilité de donner plus ou moins de responsabilité aux différentes bandes de fréquence, ainsi que la possibilité de traiter le bruit grâce à des coefficients de pondération adaptatifs.

- *tests*: les tests effectués seront plus complets que dans les travaux antérieurs. Nous comparerons/couplerons notamment l'approche multi-bande avec des méthodes classiques de reconnaissance robuste.

La *désynchronisation* des chemins de programmation dynamique associés aux différentes bandes de fréquence sera abordée très brièvement. Nous présenterons une discussion concernant des résultats obtenus sur base d'architectures "multi-stream".

Les points suivants sortent cependant du cadre de ce travail:

- *optimisation de la découpe en bandes de fréquence*: l'optimisation du nombres de bandes et des fréquence de coupures n'a pas été envisagée ici. Les bandes de fréquence utilisées par nos sous-reconnaisseurs seront cependant relativement larges (donc relativement peu nombreuses). La raison est que si ces bandes sont trop étroites, l'information présentée au reconnaisseur est insuffisante pour effectuer une reconnaissance fiable dans chaque bande de fréquence et l'on risque finalement d'aboutir à une ineptie. Nous travaillerons par la suite (dans ce chapitre et dans le chapitre 8) avec un nombre de bandes inférieur à 10.
- *optimisation du compromis temps/fréquence*: l'optimisation des caractéristiques dynamiques des différents sous-reconnaisseurs n'a pas été envisagée, ni dans ce travail, ni dans les travaux cités précédemment.
- *classes phonétiques spécialisées*: le lecteur trouvera dans [143] une étude préliminaire concernant la définition de classes phonétiques dépendantes des bandes de fréquence.
- *tests dans des conditions de réverbération*: nous nous sommes seulement focalisés sur le problème de la robustesse au bruit additif. Quelques résultats concernant la réverbération peuvent être trouvés dans des publications parallèles [143].

6.4 Paramètres acoustiques

Comme nous le verrons dans cette Section, les méthodes générales présentées à la Section 2.7 permettront l'extraction de paramètres représentatifs des bandes de fréquence intervenant dans l'approche multi-bande.

6.4.1 Approche par banc de filtres

Pour pouvoir mettre en oeuvre l'approche multi-bande, le premier niveau de reconnaissance doit extraire plusieurs ensembles de paramètres acoustiques (premier bloc de la figure 2.1), chaque ensemble représentant une bande de fréquence. Nous pourrions simplement utiliser une transformée de Fourier discrète et ne garder pour chaque bande de fréquence que les valeurs représentant effectivement cette bande.

Il est cependant bien connu que les performances obtenues avec ce type de paramètres ne sont pas très bonnes, contrairement aux performances que l'on peut

espérer des approches axées sur les méthodes d'analyse standard présentées à la figure 2.7. Nous avons choisi d'utiliser des méthodes basées sur une analyse en banc de filtres. En l'occurence, nous avons utilisé l'analyse à la base de la méthode PLP ("Perceptual Linear Prediction") [83] résumée à la section 2.7.4.

Pour chaque bande de fréquence, nous pouvons isoler les énergies des bandes critiques représentatives de cette bande et les utiliser comme paramètres représentatifs. Ce sous-groupe d'énergies peut également servir de point de départ à une analyse plus adaptée au problème de la reconnaissance de la parole, comme le calcul de cepstres par exemple.

Pour ce travail, nous avons envisagé et testé quatre types de paramètres acoustiques:

- Les énergies des bandes critiques: chaque bande de fréquence est représentée par un sous-groupe de bandes critiques, les énergies de ces bandes critiques ainsi que leurs dérivées premières et secondes sont utilisées comme paramètres représentatifs. L'inconvénient de cette approche est que les paramètres obtenus dépendent du niveau de signal. Les trois autres types de paramètres corrigent ce défaut.
- Les énergies des bandes critiques normalisées par rapport à l'énergie de la bande de fréquence considérée à la trame courante. Comme précédemment, les dérivées premières et secondes de ces paramètres, sont également utilisées.
- Les cepstres calculés sur base des énergies des bandes critiques.
- Les cepstres du modèle autorégressif dont les paramètres sont calculés sur base des énergies des bandes critiques.

Ces différentes options d'analyse n'empêchent pas l'utilisation de techniques de reconnaissance robuste de la parole. Elles peuvent être appliquées au niveau de représentation intermédiaire le plus approprié. Nous envisagerons dans ce travail les approches log-RASTA, J-RASTA ainsi que la soustraction spectrale (voir Chapitre 3). Celles-ci opèrent sur la représentation en banc de filtres. Les résultats des expériences effectuées sont donnés à la Section 6.6.

6.4.2 Approche par analyse LPC

Nous avons rappelé à la Section 2.7.3 le principe de la modélisation autorégressive du signal de parole. Les paramètres de ce modèle peuvent être obtenus par récursion sur base des $p+1$ premiers coefficients de la fonction d'autocorrélation du signal. Dans le cadre de l'approche multi-bande, nous pourrions envisager la modélisation autorégressive de différentes bandes de fréquence. Cette approche consiste à utiliser un filtre passe-bande qui isolerait les composantes fréquentielles de la bande considérée. L'approche classique de modélisation autorégressive serait alors utilisée sur base du signal filtré. Cette approche implique cependant l'observation d'une précaution supplémentaire. En effet, l'utilisation brute de la modélisation autorégressive conduirait à un modèle représentant également les caractéristiques du filtre passe-bande. Il convient donc, avant d'estimer les paramètres du modèle autorégressif, de ramener le signal en bande de base (de 0 Hz jusqu'à la fréquence de Nyquist), par multiplication avec un signal sinusoïdal adéquat et sous-échantillonnage, comme proposé dans [41].

Les paramètres qui peuvent être obtenus sur base d'une étape initiale d'analyse LPC sont fondamentalement similaires à ceux de la section précédente et n'ont pas été envisagées dans le cadre de ce travail.

6.5 Méthodes d'ensemble

Si le calcul des paramètres représentatifs est un aspect important de l'approche multi-bande, la combinaison des "décisions" fournies par les différents sous-reconnaisseurs est tout aussi fondamentale. Nous envisagerons ici différents formalismes ainsi que la possibilité de traiter le bruit grâce à des coefficients de pondération adaptatifs.

De nombreux travaux, dans le domaine de la reconnaissance de la parole, de la reconnaissance de formes, ou de la classification en général, se sont penchés sur l'idée de combiner plusieurs classificateurs ou plusieurs estimateurs de probabilité dans le but d'améliorer les capacités de classification ou de généralisation de l'ensemble. Avant de passer à des exemples spécifiques, citons quelques articles généraux sur le sujet: [81, 91, 107, 205, 206, 211].

Dans [80], les auteurs proposent l'utilisation de méthodes d'ensemble pour contourner la difficulté posée par l'augmentation de dimensionalité provenant de l'utilisation de mesures hétérogènes (par exemple l'utilisation de vecteurs d'observation présentant des résolutions spectro-temporelles différentes). Chaque expert prend finalement ses décisions sur base d'un ensemble réduit de mesures et le problème revient à déterminer la combinaison optimale de ces experts. Trois méthodes sont proposées: le vote, la combinaison linéaire pondérée de probabilités a posteriori et l'hypothèse d'indépendance (produit des vraisemblances). Bien qu'elle soit fausse, l'hypothèse d'indépendance fournit généralement les meilleurs résultats. Les auteurs proposent également l'utilisation de méthodes hiérarchiques. Dans ce cas, on ne joue plus sur le partitionnement du vecteur d'entrée mais bien sur le partitionnement des classes de décision (phonèmes). Le vecteur de "scores" est construit de façon hiérarchique en passant par la définition de groupes de phonèmes, certains experts étant chargés de la classification inter-groupes, d'autres de la classification intra-groupes.

Dans le cadre de l'approche de reconnaissance multi-bande également, l'utilisation de structures hiérarchiques a été proposée dans [143].

Dans [105], deux groupes de paramètres représentatifs sont utilisés. L'un correspondant aux caractéristiques acoustiques classiques, et l'autre correspondant à des paramètres articulatoires (c'est à dire à des traits phonétiques comme le voisement, le lieu d'articulation...). Les deux groupes de caractéristiques sont utilisés par deux experts dont les décisions sont combinées soit par somme pondérée des probabilités a posteriori, soit par l'hypothèse d'indépendance. Ces deux méthodes apparaissent généralement dans la littérature comme *règle de la somme* et *règle du produit*. Bien que plus robuste au niveau de la trame, la règle de la somme fournit de moins bons résultats au niveau du mot.

Dans [107], ces deux types d'approches ont été envisagées dans le cadre d'une tâche de vérification visuelle d'identité, en vue de combiner les décisions fournies sur

bases de 6 images différentes de la même personne.

Dans le cadre de la reconnaissance de caractères manuscrits, l'utilisation de plusieurs jeux de paramètres représentatifs différents est envisagée dans [75].

Il est également possible d'utiliser des experts obtenus sur base de conditions d'apprentissage différentes. Dans [109], des experts différents sont utilisés pour modéliser des groupes de locuteurs différents, par exemple les locuteurs masculins et féminins. Une approche similaire est utilisée dans [94].

Plus généralement, la technique de *"boosting"* [34, 39] consiste à partitionner l'ensemble d'entraînement de façon à pouvoir développer des experts généralisant différemment. La procédure est séquentielle. Un premier expert est entraîné. Ensuite, cet expert est utilisé sur des données différentes de façon à déterminer un ensemble d'entraînement pour le second expert. Cette sélection de données d'entraînement peut faire appel à différents critères. On pourra par exemple utiliser les données mal classifiées par le premier expert. Plusieurs experts peuvent par la suite être entraînés successivement en filtrant les données sur base d'une combinaison des estimations (par exemple par somme pondérée) fournies par les experts précédents. La technique de *"bootstrapping"* [26] est similaire et consiste également à obtenir des ensembles d'entraînement différents conduisant à des experts généralisant différemment. Une extension de cette approche consiste à développer des ensembles d'entraînement indépendants basés sur une modélisation du bruit inhérent aux données d'apprentissage [165]. Les techniques citées dans ce paragraphe fournissent donc un moyen de définir différents experts (et non pas une approche de combinaison des décisions des différents estimateurs.)

Dans la technique de *"mixture of experts"*, les experts sont entraînés sur des données différentes et l'espace est partionné par un expert supplémentaire appelé "gating network", dont le seul but est (dans le cadre d'une combinaison sous forme de somme pondéré des estimations) d'associer un coefficient de pondération à chacun des classificateurs du système.

Dans [107], les estimations fournies par différentes types de classificateurs (classificateur gaussien, réseau de neurones artificiels...) sont combinées avec succès pour la reconnaissance automatique de caractères manuscrits. Plus simplement, différents experts du même type mais aux architectures légèrement différentes pourraient également être utilisés. On peut trouver une application au cas de la reconnaissance de la parole dans [129].

Utilisations possibles

Comme on peut le constater, il existe essentiellement trois utilisations différentes des méthodes d'ensemble. La première consiste à utiliser des *classificateurs d'architecture ou de type différents.* Les performances d'un classificateur se jugent à sa capacité à généraliser sur des données différentes des données d'entraînement. Des classificateurs différents n'auront généralement pas la même capacité à généraliser, d'où l'intérêt d'utiliser des méthodes d'ensemble. L'idée la plus simple consiste à estimer les performances de généralisation des différents experts par validation croisée sur des données différentes des données d'entraînement. On choisit alors l'expert présentant les meilleures performances. Une méthode plus sophistiquée consiste à combiner les différents experts plutôt que de choisir le meilleur. De la sorte, on peut espérer obtenir un système qui soit meilleur que chacun des experts pris indépendamment.

Comme les exemples précédents le suggèrent, différentes méthodes de combinaison peuvent être envisagées. Nous y reviendrons dans les sections suivantes.

Une deuxième utilisation des méthodes d'ensemble consiste à utiliser des *experts spécialisés sur des conditions ou sur des régions de l'espace différentes.* La troisième application, utilisée dans le cadre du multi-bande, consiste à avoir des *experts travaillant sur des espaces d'entrée différents.* Ces deux types d'applications sont justifiées par la flexibilité accrue et le caractère adaptatif de telles architectures.

Modes d'intégration

Dans le cadre de ce travail, les méthodes d'ensemble ont été utilisées dans le but de combiner plusieurs jeux de paramètres différents: notamment les paramètres représentatifs de différentes bandes de fréquence. Notre travail s'est donc orienté vers une comparaison des différentes approches possibles. La *règle de la somme* (où l'estimation de la probabilité a posteriori d'une classe est calculée comme une somme pondérée des estimations fournies par les différents éléments de l'ensemble) et la *règle du produit* (découlant de l'hypothèse d'indépendance de différentes sources d'information), ainsi que certaines extensions de celles-ci ont été envisagées. Les techniques de votes ont également été considérées. Ces techniques consistent à utiliser les décisions de classification et non plus les vecteurs de probabilités a posteriori. La classe qui reçoit le nombre de votes le plus important est finalement choisie. Cette façon de faire est trop brutale. En effet, dans le cadre de la reconnaissance automatique de la parole, il ne s'agit pas de prendre une décision tranchée à chaque instant mais bien d'améliorer la qualité globale du vecteur de "scores". La décision finale sera alors fournie par la coopération entre ces scores acoustiques et les contraintes imposées par les modèles de Markov et par la grammaire. Ces techniques de vote ont donc été adaptées pour pouvoir être utilisées dans le cadre de notre problème. Nous avons choisi d'associer une probabilité P_0 (de l'ordre de 0.9) à toutes les classes recevant au moins un vote, P_0 étant réparti entre les différentes classes en fonction du nombre de votes. Les classes restantes se répartissent alors une une probabilité de $1 - P_0$.

Nous avons également envisagé la *reclassification* (utilisation d'un système de classification) sur base des estimations fournies par les différents experts. Celles-ci peuvent en effet être considérées comme des paramètres représentatifs, au même titre que les paramètres issus de l'analyse d'une trame du signal de parole. On peut même envisager de compléter ce vecteur d'estimation par des paramètres représentatifs issus du signal, comme ceux utilisés à l'entrée des différents experts. Comme le système de combinaison est un classificateur, les différents experts ne doivent pas forcément fournir des estimations des probabilités a posteriori des classes phonétiques. Plus simplement, ils peuvent se contenter d'effectuer une étape de *pré-classification* par extraction de paramètres représentatifs.

Finalement, nous avons considéré d'autres règles de combinaison simples: la règle du *maximum*, la règle du *minimum* et la *règle médiane.* Toutes les approches envisagées ici peuvent être vues comme des cas particuliers de la *généralisation en pile* exposée par Wolpert [205].

Les sections qui suivent exposent ces différentes méthodes de combinaison ainsi qu'une discussion concernant la décomposition en bandes de fréquence dans le cadre de ces approches.

6.5.1 Règle de la somme et extensions

Supposons que pour résoudre un problème de classification donné, nous disposions de K estimateurs de probabilités différents. La spécificité de chacun de ces estimateurs peut résider soit dans la choix de paramètres représentatifs, soit dans l'approche d'estimation choisie; l'idée sous-jacente étant qu'il n'existe pas d'estimateur qui soit toujours le meilleur et que l'un ou l'autre de ces estimateurs se démarquera en fonction de la position dans l'espace ou des conditions d'utilisation. Associons alors pour chaque estimateur/expert k un événement E_k signifiant que cet expert est un *meilleur estimateur* de probabilité a posteriori que les autres. Les événement E_k sont donc *totalement exclusifs*:

$$\sum_{k=1}^{K} P(E_k) = 1 \tag{6.1}$$

où $P(E_k)$ représente la probabilité que l'expert/estimateur k soit meilleur que les autres estimateurs. Soit x un vecteur de paramètres représentatifs et q une des classes (ou états) du problème de classification. Il vient donc:

$$P(q|x) = \sum_{k=1}^{K} P(q,E_k|x) = \sum_{k=1}^{K} P(q|x,E_k)P(E_k|x) = \sum_{k=1}^{K} P_k(q|x)P(E_k|x) \tag{6.2}$$

où $P_k(q|x)$ est l'estimation de probabilité fournie par l'expert k. En effet, lorsque l'expert k est le meilleur, la probabilité a posteriori est idéalement estimée par celui-ci. Le *terme de fiabilité* $P(E_k|x)$ désigne la probabilité que k soit le meilleur étant donné le vecteur de paramètres représentatifs. Notons que la somme sur k de ces "coefficients de pondération" doit valoir l'unité (équation (6.1)).

Sous cette formulation générale, cette approche ne semble souffrir d'aucune hypothèse contraignante. Remarquons cependant qu'elle suppose que chacun des experts fournit une bonne estimation de la probabilité a posteriori. De plus, la méthode ne dit pas comment estimer le terme de fiabilité, qui dépend de l'expert mais également du vecteur d'observation.

Hypothèses simplificatrices, extensions et alternatives

Une première hypothèse consisterait à supposer que le choix du meilleur expert est indépendant du vecteur d'observation:

$$P(q|x) = \sum_{k=1}^{K} \alpha_k P_k(q|x) \tag{6.3}$$

Une extension (partant de l'expression (6.2)) serait d'ajouter une dépendance avec l'état considéré:

$$P(q_d|x) = \sum_{k=1}^{K} P_k(q_d|x)P(E_k|x,q_d) \tag{6.4}$$

Ici également, on pourrait supposer l'indépendance par rapport au vecteur d'observation:

$$P(q_d|x) = \sum_{k=1}^{K} P_k(q_d|x)P(E_k|q_d) \tag{6.5}$$

L'optimisation/estimation des *termes de fiabilité* dans ces différentes formulations sera abordé à la Section 6.5.3. Les expressions (6.2) et (6.4) correspondent à une approche de type mixture d'experts. Quant aux expressions (6.3) et (6.5), nous verrons qu'elles peuvent aisément être optimisées sur base du critère des moindres carrés. Ces approches semblent attractives dans le cas de conditions d'utilisation particulières, identiques aux conditions d'entraînement. Cependant, lorsque les conditions d'utilisation changent, les probabilités associées à chacun des experts ($P(E_k|q_d)$) pourraient évoluer, traduisant le fait qu'un expert particulier peut être le meilleur sous certaines conditions d'utilisation seulement. Imaginons par exemple le cas d'un système composé d'un expert développé sur base de voix masculines et d'un autre développé sur base de voix féminines; ou aussi d'un système constitué d'un estimateur efficace dans le cas de parole claire et d'un autre destiné à la parole bruitée. Il conviendrait alors d'utiliser des termes de fiabilité reflétant les conditions d'utilisation $P(E_k|\text{condition d'utilisation, ...})$. Cependant, il est parfois difficile de couvrir l'ensemble des conditions d'utilisation et d'optimiser les probabilités des différents événements E_k pour chacune de ces conditions, par exemple si l'on désire développer un système robuste à différents types de bruits. Une approche heuristique pourrait alors être envisagée en pondérant les termes de fiabilité sur base d'estimations du rapport signal/bruit. Nous y reviendrons par la suite dans le cas de l'approche multi-bande.

Les formulations issues de la règle de la somme correspondent à des moyennes pondérées des différentes estimations des probabilités a posteriori ou des vraisemblances (suivant les hypothèses faites, les coefficients de pondération dépendent des observations et/ou des états considérés). Une alternative serait d'effectuer une moyenne géométrique pondérée des estimations:

$$P(q|x) = \prod_{k=1}^{K} P_k(q|x)^{\alpha_k} \tag{6.6}$$

C'est équivalent à une somme pondérée de logarithmes de vraisemblances, c'est à dire une moyenne géométrique de vraisemblances:

$$P(x|q) = \prod_{k=1}^{K} P_k(x|q)^{\alpha_k} \tag{6.7}$$

Cette approche est donc relativement proche de l'hypothèse d'indépendance.

6.5.2 Hypothèse d'indépendance - Règle du produit

Dans le cas où l'on dispose de différents estimateurs utilisant des paramètres représentatifs distincts et *statistiquement indépendants*, la probabilité a posteriori de chacune des classes du problème est alors aisément calculée sur base des probabilités a priori des différentes classes et des probabilités a posteriori fournies par

les différents experts. Soit $x = (x_1,...,x_K)$, l'observation constituée des K vecteurs d'observation indépendants. La loi de Bayes donne:

$$P(q|x) = \frac{P(x|q)P(q)}{P(x)} \tag{6.8}$$

En supposant que les différents éléments du vecteur d'observation sont statistiquement indépendants, on a successivement:

$$\begin{aligned} P(q|x) &= \frac{P(q)}{P(x)} \prod_{k=1}^{K} P(x_k|q) \\ &= \frac{P(q)}{P(x)} \prod_{k=1}^{K} \frac{P(q|x_k)P(x_k)}{P(q)} \\ &= \left[\frac{\prod_{k=1}^{K} P(x_k)}{P(x)}\right] \left[\frac{\prod_{k=1}^{K} P(q|x_k)}{(P(q))^{(K-1)}}\right] \end{aligned} \tag{6.9}$$

Le premier terme de l'expression précédente est indépendant de q. Si l'hypothèse d'indépendance est correcte, il vaudra l'unité. En pratique cependant, la probabilité a posteriori sera estimée sur base du second terme normalisé de façon à ce que la somme des estimations sur toutes les classes du problème donne l'unité.

Une extension de cette approche, incluant des coefficients de pondération heuristiques, sera proposée et justifiée à la Section 6.5.5.

6.5.3 Règle de la somme - Optimisation de la pondération

L'optimisation des paramètres de la combinaison linéaire proposée à la Section 6.5.1 peut se faire sur base de différents critères. Un premier consiste à minimiser l'erreur quadratique moyenne (MMSE: minimum mean square error) entre les estimations calculées par la règle de combinaison et des valeurs désirées (par exemple 1 pour la bonne classe phonétique et 0 pour les autres classes). Cette approche, relativement simple, trouve souvent une solution analytique par l'utilisation de techniques d'estimation linéaire ou d'optimisation sous contraintes. Pour des fonctions de combinaison plus complexes (perceptron multicouche..., voir plus loin), une approche itérative est possible par rétro-propagation et descente du gradient de l'erreur [96]. Bien qu'on puisse utiliser le critère MMSE dans ce cas, on préfère souvent utiliser le critère du minimum de l'entropie relative (voir [167]). On peut montrer que ce dernier conduit à une convergence plus rapide.

Il n'est pas garanti que ces méthodes d'optimisation conduiront à une séparation parfaite de classes linéairement séparables ([167], p.116 propose un exemple illustrant ce fait). Cette situation se présente lorsque de nombreux exemples sont localisés au voisinage de la frontière de décision et de façon dissymétrique. Ils introduisent donc un coût important qui peut être réduit en déplaçant la frontière de décision, quitte à mal classifier certains exemples. L'utilisation de règles de combinaison non-linéaires diminue cependant l'importance de ce phénomène. De plus, nous souhaitons surtout que le système final **généralise** correctement. On dit qu'un système généralise lorsque la relation entre ses entrées et ses sorties est correcte, même pour des exemples qui n'ont pas été utilisés lors de l'apprentissage. Dans cette optique, la frontière de décision obtenue par un des critères précités pourrait être meilleure que

la frontière de décision "optimale", sachant que l'ensemble servant à l'apprentissage est nécessairement fini.

Une deuxième approche consisterait à estimer les paramètres de combinaison de façon à minimiser l'erreur de classification. Cette approche n'a cependant pas de solution analytique et est plus compliquée à mettre en oeuvre car le nombre d'erreurs n'est pas une fonction de coût dérivable. La solution proposée dans la littérature, appelée critère du *minimum de l'erreur de classification* (MCE: Minimum Classification Error) consiste alors à définir une mesure de mauvaise classification. Celle-ci est généralement exprimée sous forme d'une différence entre le score associé à la bonne classe et les scores associés aux mauvaises classes. Cette mesure sert alors à définir une fonction de coût dérivable qui sera minimisée de façon itérative en adaptant les paramètres de combinaison par une technique de descente de gradient. La méthode est appelée *descente de gradient stochastique généralisée* (GPD - Generalized Probabilitic Descent). Elle permet l'optimisation de paramètres dans le cadre d'une combinaison par règle de la somme [27] (bien que l'auteur propose une somme pondérée de vraisemblance et non pas de probabilités a posteriori, voir équation (6.5)), ainsi que dans le cas d'une combinaison par somme pondérée des logarithmes de vraisemblances [12, 138, 155] (voir équation (6.7))[5] L'approche MCE peut également être utilisée pour optimiser une fonction de combinaison par perceptron multicouche [29]

Etant donné que les critères MMSE et MCE sont fondamentalement assez similaires, et que le critère MCE ne conduit pas forcément à une minimisation de l'erreur de classification (même si on parvient à obtenir le minimum de la fonction de coût), seul le critère MMSE a été utilisé pour la combinaison linéaire. Dans le cadre de fonctions de combinaison par perceptron multicouche par contre, nous avons utilisé le critère de minimum de l'entropie relative.

Pour être complet, signalons finalement un avantage du critère MCE: il peut être utilisé quel que soit le type de scores fournit par les experts (probabilités a posteriori, vraisemblances ou autres). Dès qu'un expert fournit des valeurs qui peuvent être interprétées comme des scores pour différentes classes, il peut intervenir dans un ensemble optimisé par le critère MCE. Les critères MMSE et de minimum de l'entropie relative sont généralement associés à une interprétation statistique des scores fournis par les experts. Ils demandent l'utilisation d'estimations cibles (par exemple des probabilités a posteriori valant 1 pour la classe phonétique correcte et 0 pour les autres classes) en vue de calculer les erreurs quadratiques ou les entropies relatives. Le MCE par contre ne demande pas ce genre de cibles car la fonction à optimiser est généralement exprimée sous forme d'une différence entre le score de la bonne classe et ceux des autres classes. L'approche heuristique de type MCE est donc plus souple. Pour modérer quelque peut cet avantage, signalons aussi que quel que soit le classificateur utilisé, il est toujours possible d'en extraire des probabilités a posteriori sur base de sa matrice de confusion [210] et donc d'utiliser une approche de type MMSE ou de minimisation de l'entropie relative. En effet, considérons la classe re-

5. Remarquons qu'il s'agit là d'une généralisation de la règle du produit qui au sens strict correspond à une simple somme des logarithmes des vraisemblances. Soulignons également que le résultat obtenu ne représente pas en général une fonction de densité de probabilité (intégrale différente de l'unité). Il peut cependant être interprété comme une estimation statistique correspondant a un score pouvant être utilisé en classification.

connue (quelle que soit la méthode de classification) comme une variable aléatoire. Pour chaque classe du problème, il est alors possible d'estimer la distribution de probabilité de la variable discrète indiquant la classe reconnue. Ces distributions conduisent à la matrice de confusion $[n_{de}]$ où n_{de} indique le nombre de fois où la classe q_e est reconnue alors que la classe q_d a généré l'observation. Lorsque la classe q_e est reconnue, la probabilité a posteriori d'une classe q_d peut être calculé suivant:

$$P(q_d|q_e) = \frac{n_{de}}{\sum_i n_{ie}} \tag{6.10}$$

Recombinaison de Bishop

Nous souhaitons ici combiner de façon linéaire et optimale les estimations fournies par différents experts. Nous nous limitons ici à des coefficients de pondération indépendants des vecteurs d'observation. L'expression (6.3) peut être écrite:

$$y(x) = \sum_{k=1}^{K} \alpha_k y_k(x) \tag{6.11}$$

où $y_k(x)$ est le vecteur de sortie de l'expert k. Nous souhaitons également respecter la contrainte suivante:

$$\sum_{k=1}^{K} \alpha_k = 1 \tag{6.12}$$

Nous allons estimer les paramètres α_k de façon à minimiser l'erreur quadratique moyenne à la sortie du système complet, tout en respectant la contrainte 6.12.

Etant donné $t(x)$ la sortie désirée du système pour l'observation x, le vecteur d'erreur de l'expert k est:

$$e_k(x) = y_k(x) - t(x) \tag{6.13}$$

Le vecteur d'erreur du système complet est donc:

$$e(x) = \sum_{k=1}^{K} \alpha_k e_k(x) \tag{6.14}$$

et l'erreur quadratique est:

$$e(x)^2 = \sum_{d=1}^{D} \left(\sum_{k=1}^{K} \alpha_k e_{kd}(x) \right) \left(\sum_{l=1}^{K} \alpha_l e_{ld}(x) \right) \tag{6.15}$$

où $e_{kd}(x)$ est l'élément d du vecteur d'erreur $e_k(x)$. L'erreur quadratique moyenne est finalement estimée par l'expression suivante:

$$E = E\{e(x)^2\} = \sum_k \sum_l \alpha_k \alpha_l C_{kl} = \alpha^T C \alpha \tag{6.16}$$

où C est la matrice de covariance des erreurs qui peut être estimée sur base des N vecteurs d'entraînement:

$$C_{kl} = \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} e_{kd}(x_n) e_{ld}(x_n) \tag{6.17}$$

La méthode des multiplicateurs de Lagrange permet la minimisation de E sous la contrainte 6.12. La fonction à minimiser est:

$$\alpha^T C\alpha - \lambda(U\alpha - 1) \tag{6.18}$$

où α est le vecteur colonne des poids α_k, U un vecteur ligne dont tous les éléments valent l'unité, et T l'opérateur de transposition. La matrice C étant symétrique, l'annulation de la dérivée par rapport à α conduit à:

$$2C\alpha - \lambda U^T = 0 \tag{6.19}$$

soit:

$$\alpha = \frac{\lambda}{2} C^{-1} U^T \tag{6.20}$$

et l'utilisation de la contrainte 6.12 permet finalement d'obtenir:

$$\alpha = \frac{C^{-1}U^T}{UC^{-1}U^T} \tag{6.21}$$

Recombinaison avec poids dépendant du phonème Il s'agit ici d'étendre l'approche décrite dans la section précédente en estimant les K (nombre d'experts) poids optimaux indépendamment pour chacune des D classes intervenant dans le problème. L'expression (6.11) devient donc:

$$y_d(x) = \sum_{k=1}^{K} \alpha_{kd} y_{kd}(x) \tag{6.22}$$

où $y_{kd}(x)$ est l'élément d vecteur de sortie de l'expert k. La solution consiste a résoudre D fois un problème d'optimisation similaire à celui décrit dans la section précédente, chacun de ces problèmes étant constitué de K experts ne comportant qu'un seule sortie.

L'intérêt de cette approche est de permettre de favoriser tel ou tel expert, suivant la classe considérée. Dans le cadre de l'approche multi-bande, on peut en effet supposer que chacune des bandes de fréquence n'apporte pas la même quantité d'information vis-à-vis des différentes classes de sons.

Mixture d'experts

Bien que la loi de combinaison 6.11 semble attractive, nous souhaiterions également pouvoir l'utiliser localement dans l'espace des observations et donc permettre aux poids α_k d'évoluer avec x. Ce point de vue est par ailleurs justifié par la théorie (expression (6.2)).

Le formalisme des mixtures d'experts (MofE - Mixture of Experts) a initialement été introduit comme extension à la théorie des réseaux de neurones artificiels pour leur permettre de modéliser des distributions de probabilité complexes par coopération de plusieurs distributions relativement simples [101]. Ils pouvaient donc être utilisés dans des tâches de régression. Récemment, il a été montré que ce formalisme permet aussi d'estimer des probabilités a posteriori et peut donc être utilisé

pour une tâche de classification, ou dans le cadre des modèles de Markov cachés pour la reconnaissance vocale [146]. Ce formalisme est basé sur l'utilisation de plusieurs 'experts' statistiques (réseaux de neurones artificiels) se focalisant sur des régions différentes de l'espace des paramètres, les probabilités fournies par ces différents experts étant recombinées linéairement avec des coefficients de pondération estimés par un réseau de neurones artificiels appelé réseau de décision ("gating network"). Dans notre cas, les experts statistiques utiliseront les paramètres représentatifs des différentes bandes de fréquence, et le réseau de décision estimera les coefficients de pondération à associer aux experts sur base de l'ensemble des paramètres précédents ou de tout autre paramètre représentant le signal large bande. Deux méthodes sont envisageables pour entraîner un système MofE complet. La première consiste à appliquer une méthode de descente de gradient sur le système complet (algorithme de rétro-propagation de l'erreur). La deuxième méthode met en oeuvre un algorithme itératif EM ("Estimation-Maximisation") et présente parfois de meilleures propriétés de convergence [146]. Seule la première méthode sera utilisée ici.

L'entraînement consistera à

- estimer les paramètres des experts (sous-reconnaisseurs) de façon classique.
- estimer les paramètres du réseau de décision par descente de gradient. L'algorithme de rétro-propagation du gradient de l'erreur (à la sortie du système complet) peut en effet être appliqué simplement pour optimiser les paramètres du réseau de décision seul.

Dans nos expériences de reconnaissance multi-bande, chaque expert utilisera les paramètres représentatifs de la bande de fréquence dont il est responsable. Le réseau de décision quant à lui utilisera un vecteur de paramètres correspondant à la concaténation des paramètres représentatifs des diverses bandes de fréquence.

Dans l'expression (6.2), les termes $P_k(q|x)$ et $P(E_k|x)$ sont respectivement estimés par les experts et par le réseau de décision.

L'intérêt des MofE dans le cadre de la reconnaissance multi-bande est que chacune des bandes (plus précisément, chacun des sous-reconnaisseurs) peut se spécialiser sur une région de l'espace, c'est-à-dire sur certains phonèmes et pas sur d'autres. Il parait en effet clair que certaines bandes de fréquence sont peu utiles à la discrimination entre certains sons (par exemple les basses fréquences pour les fricatives). Cette idée de particulariser chacune des bandes à certains sons de la langue rejoint alors l'idée d'un travail parallèle [142] explorant également les potentialités de l'approche multi-bande.

6.5.4 Autres approches simples

Cette section comprend trois règles de combinaison très simples proposées par certains auteurs [107, 196]. La règle du maximum consiste à approximer la probabilité a posteriori d'une classe par le maximum des probabilités a posteriori de cette classe pour les différents experts:

$$P(q_d|x) = \max_{k=1}^{K} P_k(q_d|x) \tag{6.23}$$

La règle du minimum consiste quant à elle à choisir le minimum:

$$P(q_d|x) = \min_{k=1}^{K} P_k(q_d|x) \tag{6.24}$$

Finalement, la règle médiane consiste à utiliser comme estimation la valeur médiane des estimations fournies par les différents experts:

$$P(q_d|x) = med_{k=1}^{K} P_k(q_d|x) \tag{6.25}$$

Cette règle peut être vue comme une approximation robuste de la règle de la somme (équation (6.2)) qui utilise une valeur moyenne pondérée comme estimation. Si l'estimation fournie par l'un des experts est incorrecte, la valeur moyenne sera fortement affectée. La valeur médiane, quant à elle, est relativement robuste à ce type d'erreurs d'estimation.

6.5.5 Discussion

Il est légitime de se questionner sur l'intérêt d'étudier expérimentalement tant d'approches différentes. En fait, ces approches étant basées sur des hypothèses différentes, elles conduiront à des systèmes dont les performances seront différentes. De plus, une étude théorique des caractères propres à chacune de ces approches ne semble pas évidente. Certains aspects théoriques ont été envisagés dans [106, 107]. Sur base de ce travail, cette section a pour but de proposer un embryon de discussion et de préciser les idées concernant certains aspects des méthodes de combinaison qui font l'objet des sections précédentes.

L'étude proposée dans [106, 107] est basée sur la sensibilité des méthodes de combinaison aux imprécisions des différents experts. Le fait est que ces experts ne fournissent qu'une **estimation** des probabilités a posteriori. Ces estimations sont entachées d'une erreur qui provient des limitations propres des systèmes d'estimation et qui peut également devenir d'autant plus grande que les conditions d'utilisation sont différentes des conditions d'entraînement. Les auteurs comparent alors l'effet de ces erreurs sur l'estimation fournie par l'ensemble d'experts. Par des développements simples, ils montrent que l'erreur sur cette estimation finale est plus importante pour la règle du produit que pour la règle de la somme. Cette conclusion suggère une robustesse (aux erreurs d'estimation) plus importante des méthodes par somme d'estimations. Donc, de telles approches, pourtant basées sur des hypothèses fortement restrictives, peuvent éventuellement conduire à des performances supérieures à la règle du produit lorsque l'hypothèse que nous avons de bonnes estimations des probabilités a posteriori est mise en défaut. Des résultats expérimentaux pour une tâche de reconnaissance visuelle du locuteur sur base de plusieurs images différentes confirment cette conclusion.

Cependant, cette sensibilité aux erreurs d'estimation n'indique que l'effet des erreurs sur les différentes règles d'intégration. Or, ces règles d'intégration/combinaison découlent d'hypothèses concernant l'interaction entre les différents experts (voir sections précédentes). Les erreurs découlant de ces hypothèses ont également un effet sur les estimations finales, effet qui pourrait éventuellement devenir plus important

que celui des erreurs introduites par les différents experts. L'analyse ne semble donc pas aussi évidente que celle proposée dans [106, 107].

Le fait que les experts peuvent fournir des estimations entachées d'erreurs justifie également l'utilisation de poids adaptatifs dans les règles d'intégration, même si l'utilisation de tels poids ne découle pas de l'analyse théorique. Ainsi, l'hypothèse d'indépendance pourrait conduire à une règle d'intégration de la forme:

$$P(q_d|x) = P(q_d) \prod_{k=1}^{K} \left(\frac{P_k(q_d|x)}{P(q_d)} \right)^{\alpha_k} \tag{6.26}$$

où $\alpha_k = 1$ si l'information fournie à l'expert k est consistante avec ce qui a été vu lors de l'apprentissage, α_k pouvant également tendre vers 0 si l'information fournie à l'expert k est fortement dégradée, impliquant d'importantes erreurs d'estimation. Rappelons cependant que l'estimation de ces poids α_k n'est pas évidente et fera généralement appel à des heuristiques. Par exemple, nous utiliserons des estimations de rapport signal/bruit pour pondérer les différents experts d'un système de reconnaissance multi-bande.

Une extension des méthodes précédentes consiste à développer un système de classification fonctionnant sur base des estimations fournies par les différents experts. Ces estimations tiennent donc lieu de paramètres représentatifs. L'avantage de cette approche est de ne pas devoir recourir:

- à l'hypothèse que chacun des experts fournit une estimation de la probabilité a posteriori étant donné le vecteur d'observation.
- aux hypothèses simplificatrices permettant d'obtenir ces formulations de combinaison simples.

Un des inconvénients de cette approche est la difficulté d'adapter le système ainsi obtenu à des conditions d'utilisation parfois changeantes. Les deux sections suivantes traitent respectivement de la combinaison linéaire sur base de la règle du perceptron (avec solution analytique) et de la combinaison non-linéaire sur base d'un perceptron multicouche.

6.5.6 Règle du perceptron

Nous envisageons ici la combinaison des estimations par transformation affine, ce qui correspond à utiliser la règle du perceptron comme classificateur basé sur les estimations fournies par les différents experts intervenant dans le système. La règle de combinaison est donc la suivante:

$$Z = AY \tag{6.27}$$

où Y est la matrice des estimations fournies par les experts, chaque colonne correspondant à un vecteur d'estimations à un instant donné, augmenté d'un 1 de façon à pouvoir introduire des biais pour réaliser la transformation affine. A est la matrice de transformation composée de C lignes (C étant le nombre de classes intervenant dans le problème). Finalement, Z est la matrice des estimations finales, où chaque colonne est un vecteur comprenant C éléments.

Soit X, une matrice de sorties idéales, où chaque colonne est un vecteur de C sorties désirées (par exemple 1 pour la bonne classe et 0 pour toutes les autres classes). Comme à la Section 6.5.3, nous pouvons utiliser le critère de minimisation de l'erreur quadratique moyenne pour estimer les valeurs des éléments de la matrice A. L'erreur quadratique pour la trame n est donnée par:

$$E(n) = (AY_n - X_n)^T (AY_n - X_n) \tag{6.28}$$

où T représente la transposition, où Y_n est la colonne n de Y et X_n est la colonne n de X. L'erreur quadratique moyenne pour toute la base de donnée est estimée par:

$$E = \frac{1}{N} \sum_n (AY_n - X_n)^T (AY_n - X_n) \tag{6.29}$$

La dérivée de cette erreur par rapport à A vaut:

$$\frac{\delta E}{\delta A} = \frac{1}{N} \sum_n 2(AY_n - X_n)Y_n = 2(AYY^T - XY^T) \tag{6.30}$$

On obtient donc directement l'estimation de la matrice A par (voir également [139]):

$$A = XY^T (YY^T)^{-1} \tag{6.31}$$

Dans [149], on propose une interprétation sur base de la matrice de confusion:

$$(XY^T)_{ij} = \sum_n P(i|x_n)P(j|x_n) \tag{6.32}$$

où $P(q_i|x_n)$ est la probabilité a posteriori de la classe q_i étant donné l'observation x_n à l'instant n et $P(j|x_n)$ est l'estimation de cette probabilité par notre système de classification.

6.5.7 Combinaison non-linéaire utilisant un perceptron multicouche

Il s'agit ici d'utiliser un perceptron multicouche (MLP - Multilayer Perceptron) dont les entrées correspondent aux estimations fournies par les différents experts. Ce système est entraîné de façon classique pour résoudre une tâche de classification. Ses sorties seront donc des estimations des probabilités a posteriori des différentes classes intervenant dans le problème. Cette approche, et dans une moindre mesure celle décrite à la section précédente permettent aisément plusieurs extensions potentiellement intéressantes:

- il est possible d'utiliser à l'entrée du MLP plusieurs trames successives représentant un contexte temporel et pouvant éventuellement améliorer la qualité de l'estimation,
- les entrées du MLP ne seront pas nécessairement des estimations de probabilités a posteriori car le réseau de neurones effectuera une opération de classification, indépendamment du type de paramètres placés à ses entrées. Cela permet notamment d'appliquer une approche d'analyse discriminante linéaire [60], voire même non-linéaire (NLDA) [58]. En effet, chacun des experts peut être

assimilé à une procédure d'analyse discriminante dont le seul but est d'obtenir un vecteur de paramètres discriminant de dimension réduite. Le vecteur de probabilités a posteriori en est déjà un. Cependant, les méthodes d'analyse discriminante citée ci-dessus permettent plus de souplesse quant à la construction de ces vecteurs discriminants, notamment en ce qui concerne la taille de ces vecteurs. Du point de vue strictement théorique cependant, ce type de paramètres n'a pas d'intérêt. En effet, les probabilités a posteriori sont par définition les meilleurs paramètres discriminants (ceux qui permettent la minimisation du taux d'erreur) pour une tâche de classification. Dans le cas de l'analyse discriminante non-linéaire, la procédure est la suivante. Chaque expert est un MLP comprenant deux couches cachées et entraîné comme classificateur. Les sorties des deuxièmes couches cachées des différents experts sont alors concaténées et fournies comme entrées du perceptron multicouche effectuant la combinaison des décisions des différents experts. Donc, pour chaque bande de fréquence k, une fonction non-linéaire est appliquée aux paramètres représentatifs x_k de façon à obtenir des paramètres représentatifs x'_k qui sont utilisés en entrée d'un MLP réalisant l'estimation des probabilités a posteriori des états des modèles de Markov cachés:

$$x'_k = NLDA_k(x_k) \tag{6.33}$$

– par extension, il est également possible d'utiliser plusieurs types de paramètres différents à l'entrée du système effectuant la combinaison des décisions. Par exemple, nous utiliserons conjointement des paramètres discriminants et des paramètres acoustiques (classiques) provenant d'une analyse du signal de parole (voir table 6.1, méthode H).

Dans le cas de conditions d'utilisation similaires aux conditions d'entraînement, nous montrerons que les approches de ce type sont de loin les plus efficaces. Elle pêchent cependant sur un point: leur caractère non-adaptatif. En effet, le gros avantage du perceptron multicouche, à savoir sa capacité à approximer n'importe quelle fonction multivariable est aussi un inconvénient: l'espace des paramètres étant très vaste, ce système est plus difficile à entraîner et il est difficile à adapter rapidement. Sous certaines conditions, des approches plus simples, telles que proposées dans les sections précédentes, pourraient alors se révéler plus intéressantes. Une discussion plus complète peut être trouvée au paragraphe suivant. Au niveau de l'évaluation expérimentale dans le cadre de la stratégie multi-bande, nous verrons que les approches simples conduisent dans certains cas à de meilleurs résultats.

6.5.8 Généralisation en pile - Discussion

Dans [205], Wolpert introduit une méthode de classification composée de plusieurs classificateurs de niveau 0. Les sorties de ces classificateurs, et éventuellement d'autres paramètres représentatifs, sont ensuite utilisées comme entrée à un processus de classification de niveau 1. Ce schéma général recouvre donc toutes méthodes d'ensemble rappelées précédemment. Dans la présentation de Wolpert, le système de niveau 1 est généralement entraîné sur des données différentes de celles utilisées pour entraîner les experts de niveau 0. L'idée est d'améliorer la capacité de généralisation

du système complet. Si les données d'entraînement sont identiques pour les deux niveaux, l'approche permet éventuellement d'améliorer l'apprentissage mais pas la qualité de la généralisation.

Les estimations fournies par les experts de niveau 0 peuvent donc être combinées par un expert de niveau 1 suivant plusieurs approches différentes. Certaines approches peuvent par exemple être indépendantes de l'ensemble d'entraînement. C'est le cas d'un système effectuant la moyenne des différentes estimations ou utilisant l'hypothèse d'indépendance. D'autres approches dépendent de l'ensemble d'entraînement mais restent très simples, comme une moyenne pondérée par exemple. Comme nous l'avons vu, ces approches simples découlent d'hypothèses de combinaison simples (par exemple l'hypothèse d'indépendance), et également du fait que les sorties des experts de niveau 0 sont des estimations des probabilités a posteriori. Rien ne nous empêche cependant de voir le problème de combinaison comme un problème de classification. Ceci conduit à des experts de niveau 1 plus complexes mais potentiellement plus efficaces, car ne reposant pas sur des hypothèses inexactes. C'est le cas par exemple pour une combinaison utilisant un réseau de neurones artificiels. De plus, ces méthodes peuvent être envisagées quels que soient les experts de niveau 0, qui ne sont donc plus tenus de fournir des estimations des probabilités a posteriori.

Un des inconvénients de cette dernière approche (classificateur de niveau 1 complexe) est le manque d'adaptabilité de l'ensemble. Alors qu'il est possible, lorsqu'on utilise la règle de la somme par exemple, de diminuer l'influence d'un des experts si l'on sait qu'il est inefficace dans des conditions particulières, cela ne sera généralement pas possible si l'on utilise un classificateur complexe comme un réseau de neurones artificiels par exemple. Un réentraînement sur les nouvelles conditions est possible mais reste cependant plus lourd.

Dans les expériences multi-bandes qui vont suivre, les paramètres de combinaison seront **optimisés sur base de parole claire**. Pour les méthodes de combinaison simples, une **adaptation heuristique** basée sur une mesure du rapport signal/bruit sera également utilisée dans le cas de conditions d'utilisation bruitées. Pour les méthodes de combinaison complexes cependant, aucune adaptation ne sera réalisée. Nous constaterons malgré tout que ces méthodes conduisent aux meilleurs résultats dans le cadre de l'approche multi-bande.

6.5.9 Tableau récapitulatif

La table 6.1 donne une liste des approches qui ont été utilisées dans ce travail.

6.5.10 Cas de l'approche multi-bande

Il s'agit ici d'envisager les différentes méthodes d'ensembles proposées ci-dessus dans le cadre d'un système de reconnaissance de parole utilisant une décomposition en bandes de fréquence et des experts opérant sur ces différentes bandes de fréquence.

Nous espérons notamment pouvoir rassembler les informations disparates provenant des différentes bandes de fréquence pour:

1. Obtenir des performances au moins équivalentes aux reconnaisseurs classiques

A	linéaire	$P(q_d\|x) = \sum_{k=1}^{K} \alpha_k P_k(q_d\|x)$	$\sum_{k=1}^{K} \alpha_k = 1$
B	mixture d'experts	$P(q_d\|x) = \sum_{k=1}^{K} \alpha_k(x) P_k(q_d\|x)$	$\sum_{k=1}^{K} \alpha_k(x) = 1$
C	extension	$P(q_d\|x) = \sum_{k=1}^{K} \alpha_{kd} P_k(q_d\|x)$	$\sum_{k=1}^{K} \alpha_{kd} = 1$
D	moyenne géométrique	$P(q_d\|x) = \prod_{k=1}^{K} P_k(q_d\|x)^{\alpha_k}$	$\sum_{k=1}^{K} \alpha_k = 1$
E	perceptron	$P(q_d\|x) = \sum_{k=1}^{K} \sum_{l=1}^{D} \alpha_{kdl} P_k(q_l\|x) + \beta_d$	β = terme de biais
F	MLP	$P(q_d\|x) = f(W, P_k(q_l\|x), \forall k, \forall l)$	W = paramètres du MLP
G	(N)LDA & MLP	$P(q_d\|x) = f(W, (param.discr.)_k)$	W = paramètres du MLP
H	(N)LDA+x & MLP	$P(q_d\|x) = f(W, (param.discr.)_k, x)$	W = paramètres du MLP x = paramètres acoustiques
I	indépendence	$P(q_d\|x) = P(q_d) \prod_{k=1}^{K} (\frac{P_k(q_d\|x)}{P(q_d)})^{\alpha_k}$ $normalisation \sum_d P(q_d\|x) = 1$	$\alpha_k = 1$ si parole claire
J	min	$P(q_d\|x) = \min_{k=1}^{K} P_k(q_d\|x)$	– – –
K	max	$P(q_d\|x) = \max_{k=1}^{K} P_k(q_d\|x)$	– – –
L	median	$P(q_d\|x) = med_{k=1}^{K} P_k(q_d\|x)$	– – –
M	vote	$P(q_d\|x) = P_0 \frac{N_d}{K}$ si $N_d > 0$ $P(q_d\|x) = \epsilon$ si $N_d = 0$	N_d = nbre. votes pour q_d $P_0 = 0.9$

TAB. 6.1 – *Méthodes de combinaison d'experts. Estimation de* $P(q|x)$.

dans le cas de parole claire.

2. Obtenir des performances supérieures aux reconnaisseurs classiques dans le cas de parole avec du bruit additif.

Décomposition en bandes de fréquence

Le signal est donc divisé en bandes de fréquence. A chaque bande de fréquence est alors associé un estimateur/expert développé pour effectuer une opération de classification (estimation de probabilités a posteriori) sur base de l'information contenue dans la bande de fréquence. Les estimations fournies par ces différents experts sont alors combinées suivant l'un ou l'autre des critères proposés aux sections précédentes, et rappelés à la table 6.1.

Il est important ici de rediscuter ces différents critères dans le cadre de l'application multi-bande. Les approches A, B, C, et D du tableau 6.1 (règle de la somme et approches dérivées) supposent qu'au moins un des experts fournit une 'bonne' estimation des probabilités a posteriori des classes phonétiques. Dans le cas d'une

décomposition en bandes de fréquence, aucun des experts ne pourra satisfaire à cette condition. En effet, chacun des experts n'utilise qu'un fraction de l'information disponible sur tout le spectre[6]. Par exemple, sur base du formalisme de la Section 6.5.1, $P_k(q|x)$ est estimé par $P_k(q|x_k)$ où x_k est le vecteur acoustique représentant la bande de fréquence k. Une solution à ce problème consiste à faire intervenir des experts dont les estimations sont basées sur l'information contenue dans des groupes de bandes de fréquence plutôt que dans une seule bande de fréquence. Cette approche est décrite à la section suivante.

L'approche I repose sur l'hypothèse d'indépendance des bandes de fréquence. Or, chacun sait que l'information présente dans une bande de fréquence est fortement corrélée avec celle des autres bandes, même si ces bandes ont des fréquences centrales relativement différentes. Le lecteur pourra trouver dans [92] des exemples d'analyse de la corrélation entre bandes de fréquence. Brièvement, les auteurs calculent la corrélation entre différentes bandes de fréquence en se basant sur des séquences de parole libre d'une durée de 10 à 15 secondes. En prenant 18 bandes de fréquence dans la plage 0-6000 Hz, la corrélation entre bandes adjacentes oscille entre 0.8 et 1.0. Si on considère les deux bandes extrêmes du spectre, la corrélation peut tomber à une valeur proche de 0. Cependant, pour fixer les idées, la corrélation entre une bande centrée à 666 Hz et une bande centrée à 1666 Hz vaut toujours 0.8 approximativement. Nous pouvons donc conclure que les vraisemblances fournies par les différents reconnaisseurs seront certainement corrélées. Ce formalisme, somme toute relativement simple, n'est peut-être pas le plus adapté. En pratique cependant, cette approche donne des résultats acceptables et meilleurs que ceux obtenus sur base de la règle de la somme (A).

Les approches E, F, G et H quant a elles ne reposent sur aucune hypothèse particulière. Nos résultats expérimentaux montrerons que ces approches conduisent aux meilleurs résultats lorsque les conditions d'utilisation sont similaires aux conditions d'entraînement. Cependant, le caractère non-adaptatif de ces approches semble incompatible avec le concept multi-bande permettant l'adaptation à des conditions changeantes de bruit coloré. Nous verrons malgré cela que ces méthodes confèrent un gain en robustesse non négligeable, résultat qui confirme des travaux similaires [155]. Nous en discuterons. Dans de très rares cas cependant, les approches simples (l'hypothèse d'indépendance par exemple) conduisent à de meilleurs résultats.

Parmi les différentes méthodes envisagées, nous verrons que les approches heuristiques J, K, L et M conduisent à des résultats qui sont parmi les moins bons.

Approche de combinaison complète et approximation

Il s'agit ici de diviser le signal en bandes de fréquence et de considérer par la suite toutes les combinaisons possibles de bandes de fréquence, chaque combinaison disposant d'un estimateur spécialisé [150]. La règle de la somme peut alors être utilisée pour combiner les estimations. Cette fois ci, l'hypothèse qu'il existe des experts estimant correctement les probabilités a posteriori des différentes classes est

6. Notons que même si elles ne sont pas totalement justifiées par la théorie, l'ensemble des approches proposées au tableau 6.1 ont néanmoins été envisagées expérimentalement.

justifiée. En effet, ces experts travaillent sur base de groupes de bandes de fréquence et non plus sur base de bandes isolées.

Cette approche est donc basée sur le formalisme de la Section 6.5.1. A partir de K bandes de fréquence, il est possible de définir $L = 2^K$ événements notés E_l, chaque événement indiquant qu'un sous groupe de bandes de fréquence (sous-groupe l) est non bruité, alors que les autres bandes sont bruitées. Ces événements étant totalement exclusifs, nous pouvons écrire:

$$P(q|x) = \sum_{l=1}^{L} P_l(q|x)P(E_l|x) \tag{6.34}$$

où $p_l(q|x)$ est l'estimation de la probabilité a posteriori de l´état q étant donnée l'observation x et le fait que le groupe de bandes de fréquence défini par l est non bruité alors que les autres bandes sont bruitées.

Comme l'on montré des études récentes [35, 119, 43], il est possible d'améliorer les performances d'un système de reconnaissance en ignorant simplement les composantes perturbées des vecteurs d'observation. De ce fait, il sera préférable d'estimer la probabilité $p_l(q|x)$ en ne tenant compte que de l'information fournie par les bandes définies par l. On obtient donc:

$$P(q|x) = \sum_{l=1}^{L} P(q|x_l)P(E_l|x) \tag{6.35}$$

Bien entendu, il reste à estimer le terme $P(E_l|x)$ qui représente la probabilité que le groupe de bandes l conduise à un meilleur estimateur de probabilités phonétiques que les autres groupes de bandes. Comme dans [150], une approche heuristique a été utilisée ici pour estimer ce terme sur base de mesures du rapport signal/bruit.

L'inconvénient majeur de ce type d'approche est de nécessiter un grand nombre d'estimateurs [86]. Une solution proposée dans [150] consiste à supposer l'indépendance des bandes de fréquence et à estimer ainsi les probabilités a posteriori associées à chacun des experts sur base d'estimations fournies par des experts utilisant chacun une seule bande de fréquence. Suivant l'équation (6.9), on a:

$$P(q|x_l) = \frac{P(q)\prod_{o\in l}\frac{P(q|x_o)}{P(q)}}{\sum_m P(q_m|x_l)} \tag{6.36}$$

Le dénominateur permet de normaliser les estimations.

Adaptation de la pondération - Erreurs d'estimation

Nous avons vu précédemment comment optimiser les paramètres des différentes approches de combinaison rappelées à la table 6.1. Cependant, ces méthodes ne valent que si les conditions d'utilisation sont similaires aux conditions d'entraînement. Dans le cas contraire, il conviendrait d'estimer les paramètres de combinaison de façon adaptative. L'idée mise en oeuvre dans ce travail consiste à utiliser le fait que plus on s'écarte des conditions d'entraînement, plus le taux de reconnaissance est

faible. Dans notre cas, les experts sont développés sur base d'un signal clair[7]. Nous considérerons donc que plus le signal est bruité, plus le système de reconnaissance doit être pénalisé. Des mesures de rapport signal/bruit dans chacune des bandes de fréquence peuvent donc être utilisées comme base d'un procédé de pondération adaptatif. Cette pondération heuristique est complémentaire à la pondération issue de l'optimisation en parole claire 6.5.3. Dans nos expériences, les coefficients de pondération issus des deux procédés sont simplement multipliés de façon à traduire le fait que l'utilité d'une bande de fréquence vis-à-vis des autres bandes dépend non seulement de son importance intrinsèque en parole claire, mais également des perturbations affectant cette bande. Rappelons cependant que pour certaines méthodes de combinaison, notamment celles utilisant des réseaux de neurones artificiels, une telle adaptation heuristique est impossible. Dans ce cas, le système de combinaison reste inchangé, quelles que soient les conditions d'utilisation.

Décrivons plus en détails les approches d'adaptation heuristique qui ont été utilisées ici dans le cadre des différentes méthodes de combinaison envisagées. Dans l'équation (6.26) (extension de l'hypothèse d'indépendance, ligne I du tableau récapitulatif), l'exposant α_k peut être interprété comme la probabilité que la bande k soit non bruitée ($P(k_{claire})$). Pour estimer cette probabilité, nous avons utilisé une mesure du rapport signal/bruit dans chaque bande de fréquence. Cette mesure est normalisée de façon à obtenir une estimation de la probabilité recherchée:

$$\alpha_k = P(k_{claire}) = \frac{\max\left(\min\left(\widehat{SNR}_k, SNR_{max}\right), SNR_{min}\right) - SNR_{min}}{SNR_{max} - SNR_{min}} \tag{6.37}$$

où $\widehat{SNR}_k$ est l'estimation du rapport signal/bruit dans la bande k. SNR_{min} est le seuil en dessous duquel on considère la bande de fréquence comme inutile. De même, SNR_{max} est le seuil au dessus duquel la bande de fréquence est considérée comme parfaitement claire. Comme dans [150], nous fixerons ces seuils à 0dB et 30dB. La méthode d'estimation du rapport signal/bruit qui sera utilisée consiste à mesurer le niveau de bruit sur base des 100 ms initiales de chaque phrase de test. Le niveau de parole est estimé (pour chaque phrase également) comme le niveau d'énergie moyen sur toute la phrase auquel on soustrait le niveau de bruit moyen de la phrase. Cette méthode peut être appliquée car les bruits utilisés dans ce chapitre sont stationnaires.

Dans la règle de la somme (ligne A du tableau récapitulatif), le coefficient α_k représente la probabilité que la bande de fréquence k fournisse le meilleur estimateur de probabilités phonétiques. Elle sera interprétée comme la probabilité que la bande k soit claire et que les autres bandes soient bruitées et sera donc (sous l'hypothèse d'indépendance) calculée comme suit:

$$\alpha_k = P(k_{claire}) \prod_{j \neq k} (1 - P(j_{claire})) \tag{6.38}$$

suivi d'une normalisation pour vérifier $\sum_{k=1}^{K} \alpha_k = 1$. Les différentes probabilités intervenant dans cette expression sont estimées par l´équation 6.37.

7. car nous ne pouvons pas prévoir les conditions de bruit.

Pour l'approche de combinaison complète finalement (équation (6.35)), le terme $P(E_l|x)$ représente la probabilité que le groupe de bande l conduise à un meilleur estimateur de probabilités phonétiques que les autres groupes de bandes. Cette probabilité sera interprétée de façon similaire au cas précédent:

$$P(E_l|x) = \prod_{k \in l} P(k_{claire}) \prod_{j \notin l} (1 - P(j_{claire})) \tag{6.39}$$

On multiplie donc la probabilité que les bandes intervenant dans le groupe l soient claires par la probabilité que les bandes n'appartenant par au groupe l soient bruitées.

Si nous ne disposons pas d'estimations des rapports signal/bruit, ces coefficients de pondération heuristiques sont supposés égaux pour les différentes bandes de fréquence.

6.6 Evaluation en parole bruitée

Des expériences ont été effectuées en vue d'évaluer l'intérêt de l'approche multi-bande et de comparer les paramètres représentatifs ainsi que les approches de combinaison/adaptation proposées précédemment.

Toutes les expériences faisant l'objet de cette section concernent des systèmes multi-bandes forçant le synchronisme entre les différentes bandes de fréquence. Nous n'utilisons donc pas l'architecture "multi-stream" présentée au chapitre 5. Une discussion traitant de l'asynchronisme entre bandes de fréquences sera cependant présentée à la section 6.7.

Pour rappel, on associe un expert à chaque bande de fréquence. Chaque expert est chargé d'estimer un vecteur de probabilités a posteriori (ou plus généralement de paramètres discriminants). Ces vecteurs sont ensuite utilisés par un système de combinaison fournissant un vecteur de probabilités a posteriori qui peut alors être utilisé par un moteur de reconnaissance basé sur des modèles de Markov cachés.

Dans un premier temps, nous avons travaillé sur base du corpus Numbers93. Cependant, nous avons rencontré des problèmes avec l'approche d'analyse discriminante non-linéaire. Pour rappel, cette approche consiste à utiliser les sorties de la seconde couche cachée d'un perceptron multicouche comme paramètres discriminants pour la classification. Nous avons constaté que l'entraînement de ce réseau à deux couches cachées ne convergeait pas correctement. Plusieurs choix différents pour les paramètres de l'algorithme d'entraînement ont été envisagés, sans succès. Il est généralement admis que l'entraînement de ce type de réseau de neurones est plus difficile et implique parfois l'utilisation d'une quantité plus importante de données d'apprentissage. Les entraînements ont finalement été menés à bien sur base du corpus Numbers95. Les expériences de reconnaissance, quant a elles, ont toujours été effectués sur la fraction de test du corpus Numbers93. Ce choix a en outre permis de comparer les performances de systèmes de reconnaissances entraînés sur des corpus de tailles différentes: 225855 trames pour Numbers93 et 376399 trames pour Numbers95 [8]. Nous verrons que les performances des systèmes basés sur NUMBERS'95 sont généralement légèrement supérieures à celles des systèmes basés sur NUMBERS'93, dont certains résultats ont été présentés à la table 3.1.

8. Notons que les réseaux de neurones utilisés ici comprennent au moins 150000 paramètres.

6.6.1 Systèmes de reconnaissance

Les systèmes de reconnaissance envisagés ici sont basés sur des unités phonétiques. Les transcriptions phonétiques des mots du vocabulaire ont été obtenues à partir du dictionnaire CMU 0.4, basé sur 46 phonèmes (un sous-ensemble du vocabulaire phonétique du corpus TIMIT) et contenant 110000 mots. Les phonèmes intervenant dans le vocabulaire considéré sont au nombre de 33. Pour chaque phonème, les topologies HMM ont été construites de façon à imposer une durée minimale égale à la moitié de la durée moyenne du phonème. Une grammaire de type "paire de mots" est également utilisée, mais son influence est mineure.

L'estimation des probabilités a posteriori des états des HMMs est fournie par des réseaux de neurones artificiels. Quel que soit le système (système de référence ou sous-reconnaisseurs pour chaque bande de fréquence), neuf trames adjacentes sont utilisées aux entrées du réseau de neurones artificiels. L'analyse utilise des trames de 25 ms décalée de 12.5 ms. Pour permettre l'application d'une approche discriminante non linéaire (NLDA), nous avons considéré des perceptrons multicouches composés de deux couches cachées: la première comportant 400 neurones et la seconde 30 neurones. Aucune amélioration significative des performances n'est obtenue lorsqu'on augmente la dimension des couches cachées au delà de ces valeurs, probablement à cause de la quantité relativement limitée de données d'entraînement. De plus, les performances obtenues sont comparables à celles fournies par un perceptron composé d'une seule couche cachée comprenant 400 neurones.

L'étiquetage phonétique de la base de données est obtenu comme suit: dans un premier temps, un système de reconnaissance de base est développé à partir d'un étiquetage manuel (basé sur les phonèmes TIMIT et fourni avec la base de donnée). Ensuite, cet étiquetage est adapté de façon itérative sur base de l'approximation de Viterbi de la procédure d'entraînement EM des modèles de Markov cachés. L'étiquetage obtenu avec ce système de base est utilisé pour développer tous les autres systèmes.

6.6.2 Paramètres représentatifs

Sur base des étapes initiales de la méthode d'analyse PLP [83], quatre types de paramètres représentatifs ont été évalués:

- **CBE ("Critical Band Energies")**: les énergies des bandes critiques (pt. 5, figure 2.7): chaque bande de fréquence est représentée par un sous-groupe de bandes critiques, les énergies de ces bandes critiques (après compression par une racine cubique) ainsi que leurs dérivées premières et secondes sont utilisées comme paramètres représentatifs. L'inconvénient de cette approche est que les paramètres obtenus dépendent du niveau de signal. Les trois autres types de paramètres corrigent ce défaut.
- **CBE normalisés**: les énergies des bandes critiques normalisées par rapport à l'énergie de la bande de fréquence considérée à la trame courante (suivi par une compression par racine cubique). Comme précédemment, les dérivées premières et secondes de ces paramètres, sont également utilisées.

- **cepstres**: les cepstres calculés sur base des énergies (compressées par racine cubique) des bandes critiques (pt. 6, figure 2.7).
- **cepstres-PLP ("Perceptual Linear Prediction")**: les cepstres du modèle autorégressif dont les paramètres sont calculés sur base des énergies (compressées par racine cubique) des bandes critiques (pt. 4, figure 2.7).

De plus, l'énergie de la bande de fréquence considérée est calculée comme la somme des énergies des bandes critiques correspondantes. Les dérivées premières et secondes de cette énergie sont adjointes au vecteur de paramètres représentatifs. Notons également qu'un filtrage passe-bande log-RASTA est appliqué aux énergies des bandes critiques.

L'objectif de ces expériences est d'envisager un système utilisant une décomposition en quatre bandes de fréquence. Cette décomposition est basée sur la définition des bandes critiques de l'algorithme d'analyse utilisé. A partir d'une analyse en 15 bandes critiques, les quatre bandes de fréquence sont obtenues comme suit[9] (les fréquences données ici correspondent aux fréquences de coupure à 3 dB):

- bande 1: bandes critiques de 1 à 6, fréquences de coupure à 3 dB: 0 et 778 Hz,
- bande 2: bandes critiques de 7 à 10, fréquences de coupure à 3 dB: 707 et 1632 Hz,
- bande 3: bandes critiques de 11 à 13, fréquences de coupure à 3 dB: 1506 et 2709 Hz,
- bande 4: bandes critiques de 13 à 15, fréquences de coupure à 3 dB: 2122 et 4000 Hz.

Ce choix provient d'expériences préliminaires [17] basées sur une décomposition en 3, 4 ou 6 bandes de fréquence, les meilleurs résultats ayant été obtenus avec 4 bandes. D'autres auteurs proposent des découpages différents (voir [86, 88, 144, 155, 192]). L'optimisation du découpage en bandes de fréquence, bien qu'il semble important, n'a pas fait l'objet de plus d'attention dans ce travail. Le but est en effet d'étudier une architecture originale, pour en souligner les avantages et pour en détecter les faiblesses éventuelles. Une optimisation plus fine est envisageable mais n'a à notre avis que peu d'intérêt dans un travail de recherche comme celui-ci.

Pour le quatrième type de paramètres, les ordres des modèles autorégressif et le nombre de cepstres conservés valent:

- bande 1: ordre 5, 6 cepstres,
- bande 2: ordre 3, 4 cepstres,
- bande 3: ordre 2, 3 cepstres,
- bande 4: ordre 2, 3 cepstres.

Donc, quel que soit le type de paramètres, les vecteurs représentatifs obtenus ont la même dimension.

Les figures 6.6, 6.7, 6.8 et 6.9 donnent les taux d'erreur obtenus pour les différentes bandes de fréquence. Pour chaque bande, les quatre types de paramètres sont envisagés dans différentes conditions de bruit (bruit blanc gaussien à différents niveaux). Comme on peut le constater, les différents types de paramètres conduisent à des performances similaires. Pour la première bande de fréquence cependant (la plus

9. Cette décomposition correspond approximativement à celle proposée par Shannon dans [175].

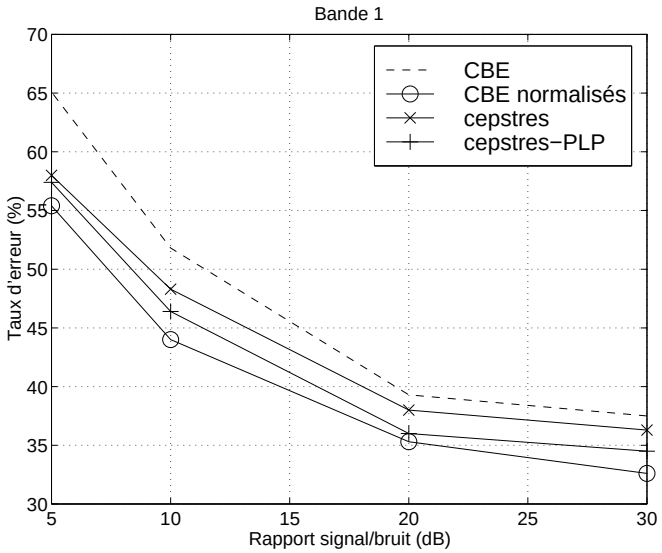


FIG. 6.6 – *Taux d'erreur au niveau du mot pour la première bande de fréquence.*

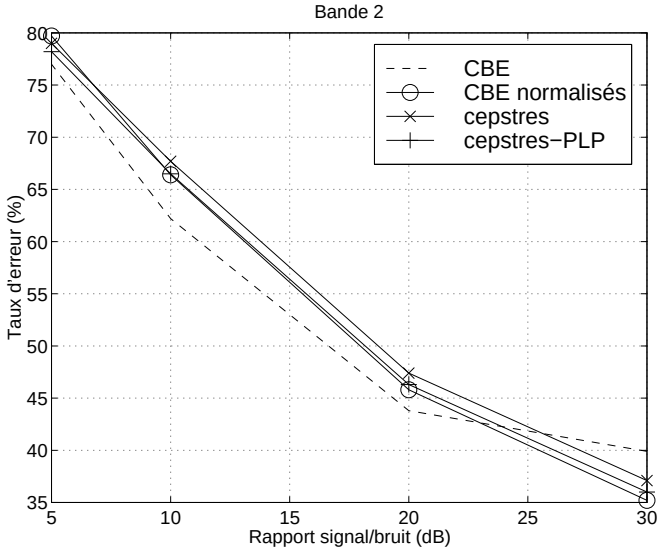


FIG. 6.7 – *Taux d'erreur au niveau du mot pour la deuxième bande de fréquence.*

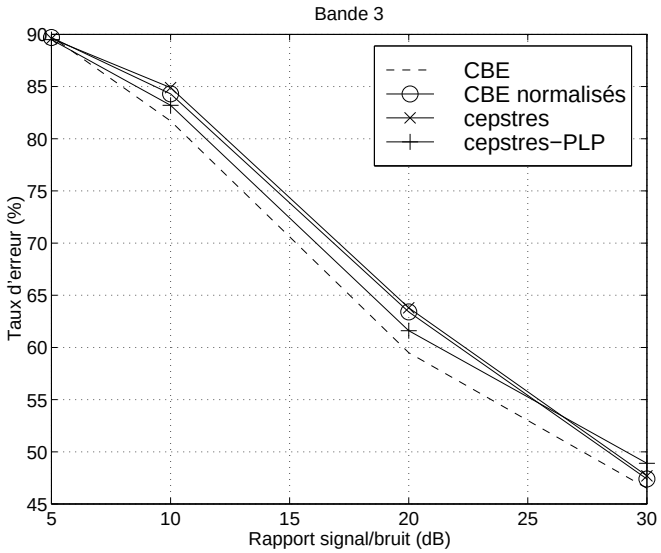


FIG. 6.8 – *Taux d'erreur au niveau du mot pour la troisième bande de fréquence.*

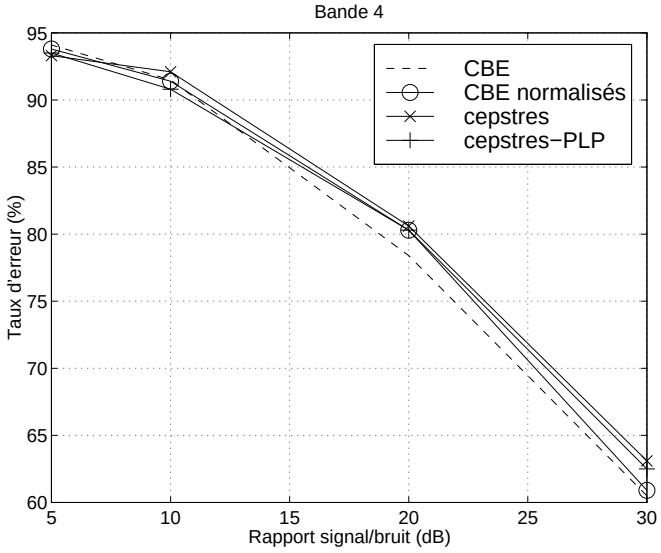


FIG. 6.9 – *Taux d'erreur au niveau du mot pour la quatrième bande de fréquence.*

importante), les meilleurs résultats sont obtenus sur base des énergies normalisées. C'est ce type de paramètres qui sera utilisé par la suite. Bien entendu, il ne s'agit là que d'un choix concernant la forme des paramètres. Les énergies des bandes critiques qui sont à la base du calcul de ces paramètres peuvent être traitées de diverses façon, notamment dans le but d'augmenter la robustesse du système. Dans les expériences qui vont suivre, nous envisagerons trois types de traitement: le log-RASTA (déjà utilisé pour les expériences précédentes), la soustraction spectrale et le J-RASTA. Nous confirmerons l'intérêt des deux dernières approches dont le but est de diminuer la variabilité des paramètres représentatifs et par conséquent d'augmenter la robustesse du système de reconnaissance aux bruits additifs.

6.6.3 Approches de combinaison et pré-traitement robuste

Sur base des experts introduits à la section précédente, nous avons envisagé l'utilisation des différents types d'approches de combinaison rappelées au tableau 6.1.

Les résultats qui vont être présentés concernent trois types de pré-traitement robuste: la soustraction spectrale, le log-RASTA et le J-RASTA. Pour chaque type de pré-traitement, les différents formalismes de combinaison sont testés et comparés. Ces expériences sont répétés à des niveaux de bruits différents et pour trois types de bruits: un bruit blanc gaussien, un bruit blanc gaussien filtré par un résonateur du second ordre et finalement un bruit très fortement coloré; tous trois sont stationnaires. Les deux bruits colorés affectent essentiellement la première bande de fréquence.

Les systèmes multi-bandes sont comparés à des systèmes de références utilisant des paramètres de type PLP, éventuellement pré-traités soit par soustraction spectrale, soit par log-RASTA ou J-RASTA.

Les méthodes de combinaison suivantes sont comparées:

- **Référence**: système de référence utilisant le même type de pré-traitement que les système multi-bandes envisagés (par exemple la soustraction spectrale). Les paramètres sont du type PLP lorsque le pré-traitement est du type log-RASTA ou J-RASTA. Pour le pré-traitement par soustraction spectrale, nous avons directement utilisé les énergies des bandes critiques: ce type de paramètres donne en effet de meilleurs résultats dans le cadre de la soustraction spectrale (voir table 3.1). Comme toujours, nous utilisons également les dérivées premières et secondes de ces paramètres ainsi que les dérivées première et seconde de l'énergie de la trame.
- **A**: règle de la somme.
- **B**: mixture d'experts.
- **I**: hypothèse d'indépendance.
- **D**: moyenne géométrique.
- **Full**: approximation de la méthode de combinaison complète (expression (6.35)).
- **A-bishop**: règle de la somme avec optimisation des α_k suivant le critère MMSE.
- **C-bishop**: règle de la somme avec coefficients de pondération dépendant du phonème et optimisés suivant le critère MMSE.

- **E**: règle du perceptron.
- **G**: combinaison par perceptron multicouche utilisant des paramètres représentatifs obtenus par analyse discriminante non-linéaire. Notons que cette approche nous permet d'utiliser les paramètres discriminants obtenus sur base de plusieurs trames consécutives. Nous en utiliserons trois.
- **H**: combinaison par perceptron multicouche utilisant des paramètres représentatifs obtenus par analyse discriminante non-linéaire concaténés aux paramètres représentatifs classiques (à savoir les énergies des bandes critiques compressées par racine cubique). Comme précédemment, nous utilisons les paramètres discriminants ainsi que les énergies des bandes critiques obtenus sur base de trois trames consécutives.
- **A-adapt**: règle de la somme avec estimation heuristique des coefficients α_k sur base de mesures du rapport signal/bruit dans chaque bande.
- **I-adapt**: hypothèse d'indépendance avec estimation heuristique des coefficients α_k.
- **Full-adapt**: approximation de la méthode de combinaison complète et estimation heuristique des coefficients de pondération sur base de mesures du rapport signal/bruit dans chaque bande.
- **A-bishop-adapt**: les coefficients de pondération sont calculés comme le produit des coefficients optimaux et des coefficients heuristiques.
- **C-bishop-adapt**: les coefficients de pondération sont calculés comme le produit des coefficients optimaux et des coefficients heuristiques.
- **J**: règle du minimum.
- **K**: règle du maximum.
- **L**: règle médiane.
- **M**: vote.
- **oracle**: Cette méthode permet d'avoir une idée des performances que l'on pourrait obtenir avec la méthode **A-adapt** si l'on disposait d'un oracle capable de détecter à chaque instant la meilleure des quatre bandes de fréquence. Un alignement forcé des phrases de test est utilisé pour décider de la meilleure bande de fréquence. A chaque trame, on choisi la bande de fréquence qui fournit la probabilité la plus élevée pour la bonne classe phonétique. On utilise ensuite les sorties du MLP correspondant à cette bande de fréquence comme estimation des probabilités a posteriori. Cette méthode s'apparente à la recombinaison optimale du modèle de Fletcher (voir l'introduction de ce chapitre): celui-ci est en effet capable de déterminer les bandes de fréquences qui fournissent un résultat de reconnaissance correct [10]. Notons cependant que pour le modèle de Fletcher, la détermination des bandes valides se fait pour chaque mot et non pour chaque trame.

10. Le modèle de Fletcher ne donne cependant aucun moyen de détecter les bandes de fréquence valides.

6.6.4 Résultats

La définition du rapport signal/bruit est identique à celle donnée à la section 3.10.1.

Les résultats complets sont présentés à l'annexe C car ils sont assez nombreux. Ceux concernant le pré-traitement log-RASTA font l'objet des figures C.1 à C.3: la première[11] concerne le bruit blanc, la suivante concerne le bruit coloré et la dernière concerne le bruit fortement coloré. Les figures C.7 à C.9 présentent les résultats obtenus avec un soustraction spectrale en phase de pré-traitement et les figures C.4 à C.6 concernent le pré-traitement J-RASTA. Pour chaque type de pré-traitement, les performances du système de base ainsi que celles du meilleur système multi-bande sont également rappelées à chaque figure. Les résultats sont résumés aux tables 6.2, 6.3 et 6.4.

Rapport S/B (dB)	5	10	20	moyenne
Référence	49.7	33.0	16.3	**33.0**
A	72.9	64.1	40.9	59.3
B	60.4	50.8	35.6	48.9
I	42.5	29.2	16.3	29.3
D	73.6	62.9	37.6	58.0
Full	48.7	34.1	19.2	34.0
A-bishop	70.5	54.9	32.2	52.5
C-bishop	70.3	54.4	32.3	52.3
E	62.8	46.5	25.0	44.8
G	46.7	30.1	13.5	30.1
H	42.4	27.0	12.3	**27.2**
A-adapt	65.0	54.4	34.9	51.4
I-adapt	49.3	32.0	16.5	32.6
Full-adapt	46.7	31.3	17.7	31.9
A-bishop-adapt	58.4	46.2	30.5	45.0
C-bishop-adapt	71.5	56.5	33.5	53.8
J	75.2	61.5	36.9	57.9
K	75.6	66.5	44.6	62.2
L	71.3	61.6	38.0	57.0
M	69.2	61.2	40.9	57.1
oracle	33.5	23.3	11.9	22.9

TAB. 6.2 – *Taux d'erreur au niveau du mot: moyenne pour les trois types de bruits, filtrage log-RASTA.*

Les conclusions majeures sont les suivantes:

– les meilleures performances sont obtenues avec des paramètres utilisant la technique de soustraction spectrale en phase de pré-traitement robuste.

11. Les différents types de combinaisons sont répartis en trois figures pour plus de clarté.

Rapport S/B (dB)	5	10	20	moyenne
Référence	32.3	21.8	12.3	**22.1**
A	46.5	36.4	27.2	36.7
B	42.6	35.3	28.4	35.4
I	28.6	22.0	17.1	22.6
D	48.4	35.9	23.7	36.0
Full	28.6	22.8	18.6	23.3
A-bishop	46.3	37.4	25.9	36.5
C-bishop	46.8	37.7	26.4	37.0
E	35.8	28.8	19.2	27.9
G	26.8	18.9	11.4	19.0
H	26.0	18.3	10.6	**18.3**
A-adapt	39.5	32.2	26.3	32.7
I-adapt	28.9	21.6	17.2	22.6
Full-adapt	28.3	22.7	18.5	23.1
A-bishop-adapt	40.6	33.6	26.0	33.4
C-bishop-adapt	52.1	41.7	27.8	40.5
J	52.8	39.8	25.3	39.3
K	48.0	38.5	30.6	39.0
L	48.7	37.6	26.1	37.5
M	47.3	39.0	31.2	39.1
oracle	16.9	12.0	8.9	12.6

TAB. 6.3 – *Taux d'erreur au niveau du mot: moyenne pour les trois types de bruits, filtrage J-RASTA.*

Rapport S/B (dB)	5	10	20	moyenne
Référence	17.8	12.8	9.6	**13.4**
A	69.3	57.6	32.0	52.9
B	57.7	48.7	33.0	46.5
I	39.4	25.6	16.8	27.3
D	69.6	56.9	30.2	52.2
Full	45.2	29.2	18.2	30.9
A-bishop	66.6	46.6	26.4	46.6
C-bishop	67.0	47.0	26.9	47.0
E	58.3	36.4	20.1	38.3
G	40.9	22.8	11.2	25.0
H	19.9	13.1	8.3	**13.8**
A-adapt	62.1	49.0	29.6	46.9
I-adapt	48.2	28.4	16.8	31.1
Full-adapt	44.2	27.1	18.0	29.8
A-bishop-adapt	57.1	42.6	26.8	42.2
C-bishop-adapt	66.8	48.2	28.5	47.9
J	71.7	56.2	31.1	53.0
K	72.4	60.0	34.4	55.6
L	67.4	54.7	32.3	51.5
M	65.8	51.8	31.8	49.8
oracle	11.0	8.7	6.9	8.9

TAB. 6.4 – *Taux d'erreur au niveau du mot: moyenne pour les trois types de bruits, soustraction spectrale.*

- dans le cas de paramètres utilisant les approches log-RASTA et J-RASTA, le système multi-bande fournit généralement de meilleurs résultats que le système de référence. L'amélioration est très marquée dans le cas du bruit fortement coloré mais est également significative dans les autres cas et même lorsque le rapport signal/bruit est élevé.
- dans le cas de paramètres utilisant la soustraction spectrale cependant, le multi-bande améliore légèrement la robustesse dans le cas de bruits colorés ou fortement colorés et à faible niveau de bruit. Dans les autres cas, on observera une légère dégradation des performances.
- généralement, les meilleures approches de combinaison sont celles utilisant des réseaux de neurones artificiels. Viennent ensuite les approches basées sur l'hypothèse d'indépendance des bandes de fréquence, l'approche utilisant la règle du perceptron et les approches basées sur la règle de la somme. Finalement, on trouvera les méthodes basées sur les règles médiane, du minimum et du maximum, ainsi que la technique de vote.
- dans certains cas cependant (voir figure C.3), les techniques simples basées sur l'hypothèse d'indépendance surpassent les techniques basées sur des réseaux de neurones artificiels.
- généralement, les techniques basées sur une adaptation heuristique des poids de combinaison permettent d'augmenter très sensiblement les taux de reconnaissance,
- cependant, ces techniques ne permettent généralement pas de surpasser les performances obtenues sur base d'une combinaison par réseau de neurones artificiels.
- comme nous pouvions nous y attendre, l'oracle fournit les meilleurs résultats.

En résumé, les meilleures performances sont obtenues avec le système de référence utilisant la soustraction spectrale (le système multi-bande basé sur des paramètres utilisant la soustraction spectrale conduit à des performances comparables). Viennent ensuite les systèmes multi-bandes utilisant la technique J-RASTA. Les autres systèmes conduisent à des performances nettement inférieures. Vu sous cet angle, l'intérêt de l'approche multi-bande semble donc limité. Rappelons cependant que ces expériences concernent des bruits parfaitement stationnaires et que l'approche de soustraction spectrale, basée sur une mesure du niveau de bruit pendant les trames silencieuses au début de chaque phrase, n'est donc pas très réaliste. En pratique, l'hypothèse de bruit stationnaire ne tient pas et il convient d'utiliser une méthode adaptative pour la mesure de niveau de bruit. Toute méthode adaptative conduira à des estimations plus approximatives, et vraisemblablement d'autant moins bonnes que le bruit est non stationnaire. En vue d'évaluer le danger d'une méthode basée sur la soustraction spectrale, nous avons appliqué l'algorithme de soustraction sur base d'une estimation du niveau de bruit erronée. Cette estimation provient d'un signal bruité à 20 dB et est utilisée pour le débruitage dans les autres conditions de bruit (10 dB et 5 dB). Dans ces conditions, le taux d'erreur du système de référence utilisant la soustraction spectrale grimpe à 29.1% (globalement pour les différentes conditions et niveaux de bruit) alors qu'il était de 13.4% avec l'estimation correcte du niveau de bruit (voir table 6.4). Le système multi-bande basé sur les paramètres

pré-traités par J-RASTA est plus performant avec un taux d'erreur moyen de 18.3%. Au Chapitre 8, nous expérimenterons sur base de bruits réels.

6.6.5 Discussion

Comme nous avons pu le constater, la combinaison par réseau de neurones artificiels fournit généralement les meilleurs résultats, que ce soit en parole très faiblement bruitée, ou en parole fortement bruitée. De plus, lorsque le bruit est fortement coloré, et de niveau élevé, les performances obtenues avec ce type d'architecture sont nettement meilleures que celles obtenues à partir d'une architecture mono-bande classique.

En parole faiblement bruitée, ces conclusions se justifient aisément par le fait que l'approche de combinaison par réseau de neurones artificiels ne fait appel a aucune hypothèse particulière concernant les experts et leurs paramètres représentatifs. La justification des performances en parole plus fortement bruitée est moins évidente. En effet, le MLP opérant la combinaison des décisions, tout comme les MLPs correspondant aux bandes de fréquence, est entraîné sur base d'un signal de parole clair [12]. Pourquoi conduit-il à un gain significatif, qui peut même être très important dans le cas de bruit fortement coloré? Il est possible que la robustesse de cette approche provienne du fait que naturellement, dans l'ensemble d'entraînement même non bruité, il existe des cas où certaines bandes de fréquence ne sont pas en accord avec les autres [13]. Le MLP de combinaison pourrait donc "apprendre" à distinguer ces bandes de fréquence des autres. Dans le cas de bruit blanc, certaines bandes pourrait également être plus efficaces que d'autres en fonction du type de phonème (voir notamment [29]).

Les approches de combinaison simples, quant à elles, peuvent être parmi les plus efficaces dans certains cas de bruit coloré et à faible rapport signal/bruit. Il s'agit de situations où le gain d'une décomposition en bandes de fréquence surpasse la dégradation due aux hypothèses simples, comme l'hypothèse d'indépendance par exemple.

Concluons finalement en constatant qu'en moyenne, sur toutes les conditions de bruit envisagées, l'approche multi-bande conduit à une réduction du taux d'erreur de 18% pour les paramètres pré-traités par un filtre log-RASTA et de 17% pour les paramètres pré-traités par un filtre J-RASTA. On passe respectivement de 33% à 27.2% et de 22.1% à 18.3%. Pour les paramètres pré-traités par soustraction spectrale cependant, on remarque une légère dégradation des performances: on passe de 13.4% à 13.8%. L'inconvénient de cette méthode de soustraction spectrale est de nécessiter une bonne estimation du niveau de bruit, ce qui est le cas dans les expériences réalisées ici, le bruit étant stationnaire et le niveau de bruit étant mesuré sur base des trames initiales de chaque phrase. Nous avons cependant montré qu'une estimation incorrecte du niveau de bruit entraîne une forte dégradation des performances de cette approche, qui peut devenir moins efficace que l'approche de filtrage J-RASTA.

12. Il est en effet difficilement envisageable de considérer tous les types de bruits susceptibles d'affecter le système de reconnaissance durant son utilisation

13. Ou que le vecteur de probabilité estimé par certaines bandes présente une entropie relativement élevée par rapport aux estimations des autres bandes.

Des expériences sur base de bruits réels et d'un estimateur adaptatif du niveau de bruit sont donc nécessaires pour évaluer l'intérêt de l'approche multi-bande (avec pré-traitement J-RASTA ou soustraction spectrale éventuelle) par rapport à la technique classique de soustraction spectrale.

L'approche multi-bande, ainsi que d'autres approches, standard et originales, seront évaluées sur des signaux perturbés par des bruits réels. Cette évaluation fait l'objet du Chapitre 8.

6.6.6 Tâches de reconnaissance complexes

Au cours de la réunion de travail sur SWITCHBOARD qui s'est tenu en 1996 à Baltimore, l'approche multi-bande à été testée sur une base de donnée de conversations téléphoniques (corpus SWITCHBOARD). Les données d'entraînement consistaient en 4 heures de parole et l'ensemble de test était constitué de 240 phrases. Nous avons développé un système de reconnaissance basé sur 4 bandes de fréquence (identiques aux quatre bandes définies dans les sections qui précèdent) et utilisant des paramètres de type PLP sous-bandes. Les transcriptions phonétiques des mots étaient basées sur des phonèmes indépendants du contexte et les probabilités a posteriori de ces phonèmes étaient estimées par des perceptrons multicouches. Chacun des 4 réseaux de neurones comportait 500 neurones cachés alors que le système de référence en comportait 2000. Un perceptron sans couche cachée a été utilisé pour combiner les décisions. Les performances du système de référence (63.6% d'erreur au niveau du mot) étaient inférieures à celles du système multi-bande (61.4% d'erreur). Finalement, un gain plus important a pu être obtenu en combinant les décisions provenant des sous-bandes avec celles du système de référence, toujours sur base d'un perceptron sans couche cachée (59.7% d'erreur) [21].

Dans [98], les auteurs envisagent l'utilisation d'une approche multi-bande pour la reconnaissance automatique de journaux parlés (Broadcast News). Les différentes variantes proposées ont toutes conduit à une dégradation des performances; les raisons avancées dans cet article portant sur la difficulté de la tâche. Pour une tâche compliquée[14], il est en effet possible que la réduction d'information opérée par la décomposition en bandes de fréquence conduise a un problème de classification devenant trop difficile à résoudre par des opérations de combinaison ultérieures.

Au vu de ces derniers résultats, il paraît nécessaire de poursuivre l'investigation de l'approche multi-bande dans le cas de tâches plus complexes que la reconnaissance de nombres connectés. Dans certains cas, l'approche conduit à des dégradations significatives. D'autre part, aucune expérience concernant le bruit n'a été envisagée dans les travaux cités dans ce paragraphe.

6.7 Modèles "multi-stream"

Toutes les expériences faisant l'objet des sections précédentes concernent des systèmes multi-bandes forçant le synchronisme entre les différentes bandes de fréquence. On associe un expert à chaque bande de fréquence. Pour chaque trame

14. Le taux d'erreur du système de référence est proche de 30%.

d'analyse, les décisions fournies par chacun de ces experts sont utilisées par un système de combinaison qui estime les probabilités des classes phonétiques.

Les approches "multi-stream" qui ont été présentées au Chapitre 5 peuvent cependant être utilisées pour développer des systèmes qui permettent aux bandes de fréquence de se désynchroniser au sein d'unités linguistiques telles que la syllabe par exemple. On évite ainsi de modéliser la variabilité du signal vocal due à l'indépendance plus ou moins importante des sons présents dans les différentes bandes de fréquence et de leurs processus générateurs: la frication, la vibration des cordes vocales et les différentes constrictions du conduit vocal. Il semble d'autre part que le système de reconnaissance humain soit peu sensible à la désynchronisation de bandes de fréquence [3] ou, dans le cadre de la reconnaissance multimodale de la parole, à la désynchronisation d'un canal audio par rapport à un canal video [134, 183]. Cette partie présente très brièvement les résultats majeurs de la communauté dans ce domaine de recherche.

Des expériences préliminaires sont rapportées dans [17] sur base d'une tâche de reconnaissance de 108 mots isolés sur ligne téléphonique en langue allemande (corpus HER). Trois configurations différentes sont développées à partir d'une décomposition utilisant trois bandes de fréquence: (1) un système "multi-stream" avec points d'ancrage à l'intersection des syllabes, (2) un système "multi-stream" avec points d'ancrage à l'intersection des phonèmes, sachant que ceux-ci sont modélisés par des HMM comportant trois états différents, et (3) un système avec combinaison des probabilités à chaque trame. Dans les trois cas, la combinaison est basée sur l'hypothèse d'indépendance des bandes de fréquence. Le taux d'erreur au niveau du mot est plus faible pour la première configuration (2.3%) que pour les deux autres (2.6%). Cependant, ces taux d'erreur ne sont pas significativement différents. Le lecteur intéressé trouvera une description plus complète des configurations utilisées (définition des bandes de fréquence, paramètres acoustiques utilisés...) dans l'article cité.

Comme déjà discuté dans l'état de l'art présenté plus haut (revoir la Partie 6.2 qui donne plus de détails sur les résultats rappelés ici), les résultats concernant la désynchronisation possible des bandes de fréquence sont mitigés. Dans [195], les résultats expérimentaux concernant un système basé sur deux bandes de fréquence et une désynchronisation au sein des phonèmes sont concluants. L'article montre cependant que la généralisation des conclusions, en ce qui concerne le nombre de bandes et le niveau d'asynchronisme, est encore incertaine.

Dans [143], les résultats ne vont pas en faveur de l'approche de désynchronisation (avec désynchronisation au sein de mots) qui conduit généralement à une légère dégradation des performances, que se soit en parole claire ou en parole réverbérée. Le taux d'erreur a même tendance à augmenter avec l'augmentation de l'asynchronisme maximal entre les bandes de fréquence.

Dans nos travaux ([17] et divers résultats récents non encore publiés), l'utilisation d'une approche de combinaison/intégration permettant l'asynchronisme des bandes de fréquence n'apporte aucune amélioration des performances. Elle conduit même à une légère dégradation dans certains cas. Les résultats de [203] vont également dans ce sens.

On constate donc qu'un relâchement de la contrainte de synchronisme entre différentes bandes de fréquence conduit généralement à une dégradation des résultats.

De plus, il implique une charge de calcul et un espace mémoire importants, dûs à l'utilisation de modèles parallèles ou de modèles composites.

Il est possible que la contrainte de synchronisme soit nécessaire et que se libérer de cette contrainte entraîne une augmentation du taux d'erreur suite à l'augmentation du nombre d'alternatives possibles dans les chemins d'états. Si des désynchronisations existent, elles apparaissent peut-être à certains moments alors qu'elle sont quasi inexistantes à d'autres moments: lors de la production d'une plosive par exemple. Il nous semble donc que dans une étape de recherche ultérieure, nous pourrions tenter d'imposer des contraintes de synchronisme apprises sur les données d'entraînement: on pourrait imposer le synchronisme parfait dans certains cas et permettre le déphasage dans d'autres cas. Les méthodes introduites à la Section 5.3.2 et utilisées dans le cadre de la reconnaissance vocale audiovisuelle (Chapitre 7) pourrait être utilisées. Nous n'avons malheureusement pas eu le loisir d'approfondir le problème dans le cadre de cette thèse.

6.8 Conclusions

Ce chapitre s'est d'abord attaché à présenter un bref état de l'art des travaux touchant au domaine du multi-bande. Les conclusions fondamentales de ces travaux antérieurs sont multiples. Premièrement, l'utilisation de cepstres sous-bandes donne généralement de meilleurs résultats de reconnaissance que l'utilisation d'une représentation de type spectrale: cette amélioration provient vraisemblablement du fait que les paramètres cepstraux sont normalisés par rapport à l'énergie de la trame courante. Ensuite, une combinaison par réseau de neurones artificiels est préférable à une combinaison simple de type somme de probabilités ou produit de probabilités: le réseau de neurones n'est en effet basé sur aucune hypothèse restrictive. Les essais rapportés dans la littérature indiquent que les approches de type multi-bande conduisent à une robustesse accrue dans le cas de bruits colorés. Dans le cas de bruits large-bande cependant, aucune amélioration n'est observée.

Nous nous sommes ensuite attaché à présenter nos derniers résultats de recherche. Ceux-ci portent sur les paramètres représentatifs des différentes bandes de fréquence ainsi que sur les méthodes de combinaison des résultats provenant de l'utilisation indépendante des différentes bandes de fréquence. Nous avons notamment proposé un état de l'art des méthodes d'ensemble. Celles-ci permettent de combiner plusieurs experts (par exemple des systèmes de classification) dans le but d'améliorer les capacités de l'ensemble. Elles sont bien entendu d'un intérêt particulier dans le cadre de l'approche multi-bande, ou plus généralement de l'approche "multi-stream". Les conclusions obtenues sont similaires à celles des travaux précédents. La normalisation des paramètres par rapport à l'énergie de la bande de fréquence conduit à une légère amélioration, l'utilisation d'une combinaison par réseau de neurones artificiels est généralement la meilleure des diverses méthodes qui ont été envisagées ici et, finalement, l'approche multi-bande conduit à une robustesse nettement accrue dans le cas de bruit coloré. Pour un bruit blanc cependant, les performances des systèmes multi-bandes et des systèmes de référence sont similaires. Nous avons également envisagé l'utilisation de techniques de reconnaissance robuste en vue d'accroître les perfor-

mances des systèmes de référence et des systèmes multi-bandes. Pour la meilleure de ces techniques (la soustraction spectrale), l'approche multi-bande ne conduit à aucun gain en robustesse, même dans le cas de bruits colorés. Notons cependant que les tests rapportés ici correspondent à des bruits stationnaires, cas idéal pour l'approche de soustraction spectrale. Dans le cas de bruits réels non-stationnaires, l'estimation du niveau de bruit nécessaire à la soustraction spectrale est toujours entachée d'erreur et peut donc conduire à des performances inférieures. Des tests plus réalistes sur base de bruits réels et d'une méthode automatique d'estimation du niveau de bruit sont donc nécessaires et seront envisagés au Chapitre 8.

Nous nous sommes finalement intéressés à l'utilisation de modèles "multi-stream" pour la reconnaissance en bandes de fréquence. Ils permettent de relâcher la contrainte de synchronisme entre les différentes bandes. Nos résultats, peu concluants, nous laissent à penser que la contrainte de synchronisme devrait peut-être être assouplie seulement pour certaines classes de sons, et pas pour tous les sons comme cela a été fait dans ce travail ainsi que par d'autres. A ce propos, les suggestions des Sections 5.3.2 et 5.5 ouvrent de nouvelles perspectives de recherche.

L'optimisation du découpage en bandes de fréquence, bien qu'il semble important, n'a pas fait l'objet de plus d'attention dans ce travail. Le but est en effet d'étudier une architecture originale, pour en souligner les avantages et pour en détecter les faiblesses éventuelles. Une optimisation plus fine est envisageable ultérieurement. De même, il serait intéressant dans une phase ultérieure d'analyser dans quelle mesure l'utilisation de résolutions temporelles et de contextes temporels différentes pour chacune des bandes de fréquence permettrait d'améliorer les performances globales.

Chapitre 7

Reconnaissance multi-modale de la parole

7.1 Introduction

La reconnaissance automatique de la parole est un domaine de recherche actif depuis plusieurs décénies. Malgré les efforts fournis, les performances des systèmes actuels sont encore loin de celles des humains. Ces systèmes sont notamment très sensibles aux bruits acoustiques. Une part importante des efforts de recherche s'oriente désormais vers la robustesse aux environnements bruités, problème qui a été identifié par plusieurs comme un des grands défis pour les années à venir [31]

La perception vocale humaine implique essentiellement l'analyse du signal acoustique. En présence de bruit acoustique, le mouvement des lèvres est également utilisé dans un processus appelé "lecture labiale". Il est bien connu que l'intelligibilité de la parole augmente lorsque l'on peut voir le visage du locuteur. Par exemple, une expérience de perception [169] sur un vocabulaire de sept voyelles montre qu'à un rapport signal/bruit de 0 dB, l'audiovisuel permet d'obtenir un taux de reconnaissance supérieur à 90% (100% en parole claire) alors que le signal acoustique seul conduit à un taux de reconnaissance inférieur à 80% (100% en parole claire).

Quelques travaux récents utilisent l'information visuelle dans le but d'augmenter les performances et surtout la robustesse des systèmes automatiques, comme c'est le cas pour la reconnaissance humaine. Deux types d'approches sont généralement proposées en ce qui concerne l'intégration des informations acoustiques et visuelles. L'approche d'intégration avant classification (EI - "Early Integration") [25, 194] construit des vecteurs de paramètres représentatifs sur base des paramètres acoustiques et visuels. Ces vecteurs composites sont utilisés par un module de classification unique. Cette approche découle de résultats [22, 55, 77, 191] montrant que la reconnaissance humaine de syllabes qui diffèrent par la plosive initiale (/bi/ et /pi/ par exemple) repose sur l'information de bas niveau des deux modalités: elle passe par la détermination du délai de voisement (VOT - "Voice Onset Time" - délai entre le relâchement de la constriction du conduit vocal et le début de vibration des cordes

vocales)

L'approche d'intégration après classification (LI - "Late Integration") [158, 177, 191] consiste à effectuer une phase de pré-catégorisation phonétique avant la phase d'intégration. Cette dernière s'effectue alors sur base des probabilités de chacune des classes phonétiques calculées par deux modules de classification indépendants. Cette approche est en accord avec des résultats [2, 57] montrant que la perception humaine sur base d'un signal acoustique opère par pré-classification phonétique dans différentes bandes de fréquence.

Alors que la plupart des systèmes testés au chapitre précédent (dans le cadre de la décomposition en bandes de fréquence) effectuaient la combinaison des décisions au niveau local, l'approche qui va être décrite ici repose sur le modèle "multi-stream" présenté au Chapitre 5. Elle correspond par ailleurs à un modèle d'intégration après classification (modèle LI). Des modèles composites sont définis à partir des modèles audio et video, l'intégration étant basée sur l'hypothèse d'indépendance. Ce travail, lié à [194], propose en plus différentes stratégies pour la modélisation du comportement asynchrone des deux modalités.

La plupart des travaux en reconnaissance vocale audiovisuelle (AVSR - "Audio-Visual Speech Recognition") s'intéressent aux mots isolés. Ils se focalisent essentiellement sur la recherche de méthodes de pondération appropriées garantissant de bonnes performances dans une large gamme de rapports signal/bruit. La reconnaissance de parole continue est cependant plus délicate que la reconnaissance de mots isolés. Nous ne souhaitons en effet pas attendre la fin de la phrase avant d'amorcer l'intégration des modalités. Cela introduirait un délai important et cela nécessiterait de générer une liste des N meilleures hypothèses pour chacune des modalités. Seules les hypothèses identiques peuvent en effet être mises en correspondance pour intégrer les "scores" des deux modalités.

L'approche "multi-stream" ne fait pas appel à des listes des meilleures hypothèses. Il s'agit d'un candidat intéressant pour la reconnaissance multimodale de parole continue. Il permet en effet: (1) le décodage synchrone de parole continue, (2) l'asynchronisme des canaux acoustique et visuel avec la possibilité de définir des points de resynchronisation phonologiques, (3) la définition de modèles HMMs spécifiques pour les canaux acoustique et visuel, (4) la modélisation de motifs d'asynchronisme.

Ce chapitre est essentiellement expérimental. Il repose sur les approches décrites au Chapitre 5 et utilise également les résultats du Chapitre 4.

7.2 Expériences de reconnaissance vocale

Nous avons utilisé la base de données M2VTS, décrite à l'Annexe A. Durant la phase de reconnaissance, seules les hypothèses comportant le nombre correct de mots (10 mots) ont été considérées. Ceci permet d'éviter l'optimisation d'un paramètre de reconnaissance pénalisant l'ajout de mots ("word entrance penalty").

Système	Video	Audio	Audio-Visuel
Taux d'erreur	40.3%	1.4%	1.2%

TAB. 7.1 – *Taux d'erreur au niveau du mot pour les systèmes acoustique (paramètres PLP), visuel et audio-visuel (*MODÈLE 1*), en parole claire.*

7.2.1 Reconnaissance vocale acoustique

Dans un premier temps, le signal audio (48 kHz) est sous-échantillonné à 8 kHz. Nous calculons ensuite des paramètres cepstraux de type PLP ("Perceptual Linear Prediction") [83] sur base de trames de 30 ms, analysées toutes les 10 ms. Le vecteur complet comporte 25 paramètres: 12 coefficients PLP, les dérivées premières de ces coefficients (12 ΔPLP) et la dérivée première de l'énergie (Δénergie).

Sur base de la durée moyenne des chiffres, nous définissons des modèles HMMs de mots[1] comprenant de 3 à 9 états indépendants, conduisant à un total de 52 états différents (incluant un état représentant le silence).

Les séquences du corpus sont d'abord segmentées en chiffres par alignement Viterbi sur base de modèles de Markov cachés entraînés à partir du corpus POLYPHONE SUISSE FRANCOPHONE [30]. Chaque chiffre est ensuite segmenté linéairement sur base du nombre d'états du HMM correspondant. Cette segmentation initiale est utilisée pour développer des modèles de Markov cachés utilisant des réseaux de neurones artificiels pour l'estimation des probabilités associées aux états. Nous utilisons des Perceptrons Multicouches (MLP - "Multi-Layer Perceptron") entraînés avec neuf trames acoustiques adjacentes en entrée. Cela permet de modéliser la corrélation temporelle locale et cela accroît le taux de classification [20]. L'algorithme de rétropropagation est utilisé pour adapter les poids des réseaux de neurones par descente de gradient. Le système de référence comporte 150 neurones cachés, nombre au delà duquel les performances n'augmentent plus.

Les entraînements et les tests sont effectués suivant le partitionnement des données décrit à la Section A.4. Les tables 7.1 et 7.2 synthétisent les résultats obtenus pour de la parole claire et pour de la parole bruitée (bruit blanc gaussien), à différents rapports signal/bruit[2]. On peut observer l'effet néfaste du bruit, même à des niveaux très modérés.

7.2.2 Reconnaissance vocale visuelle

Les paramètres visuels sont issus des travaux de Juergen Luettin [127]: 12 paramètres représentent les contours interne et externe des lèvres et 12 autres représentent la luminosité au voisinage de ces contours. Les dérivées premières de ces paramètres sont également utilisées, conduisant à des vecteurs de 48 paramètres. Le principe de cette méthode d'analyse est le suivant: dans un premier temps, les

1. Des phonèmes identiques présents dans des mots différents sont modélisés par des états différents: par exemple /t/ dans "trois" et dans "quatre".

2. Pour ces expériences, le rapport signal/bruit est calculé pour chaque séquence sans éliminer les portions de silence. Le rapport signal/bruit est donc quelque peu sous-estimé (voir également la section 3.10.1.

Rapport signal/bruit (dB)	clair	20	15	10	5	moyenne
Acoustique	1.4	24.5	56.3	76.2	83.4	48.4%
Modèle0 (EI)	1.6	10.6	25.8	47.5	63.0	29.6%
Modèle1	1.2	9.4	19.4	27.3	36.7	18.8%
Modèle2	1.2	9.9	17.3	26.8	34.3	17.9%
Modèle2+Elagage	1.1	8.0	16.4	22.5	32.3	16.1%
Acoustique+Durée	1.1	21.4	52.5	74.7	83.5	46.7%
Modèle1+Durée	1.1	6.7	14.8	22.9	30.7	15.2%
Modèle2+Durée	1.0	7.6	14.5	22.0	30.5	15.1%
Modèle2+Elagage+Durée	1.0	7.4	14.3	21.5	30.2	14.9%

TAB. 7.2 – *Taux d'erreur au niveau du mot pour la reconnaissance de séquences de chiffres, le nombre correct de chiffres étant connu a priori. Les paramètres acoustiques sont calculés par l'algorithme PLP. Cinq conditions de bruit sont considérées: parole claire, 20 dB, 15 dB, 10 dB et 5 dB. Il s'agit d'un bruit blanc gaussien stationnaire. Pour chaque condition, le poids de combinaison de l'équation (7.1) est optimisé sur un ensemble de développement sujet au même niveau de bruit que l'ensemble de test.*

Rapport signal/bruit (dB)	clair	20	15	10	5	moyenne
Acoustique	1.1	3.6	7.2	17.3	31.4	12.1%
Modèle0 (EI)	1.2	3.1	4.8	8.5	15.9	6.7%
Modèle1	0.8	2.1	2.8	5.3	10.4	4.3%
Modèle2	1.0	2.1	3.1	6.3	12.7	5.1%
Modèle2+Elagage	1.0	2.0	2.8	5.9	12.5	4.8%
Acoustique+Durée	0.7	3.1	6.4	15.4	29.8	11.1%
Modèle1+Durée	0.8	1.8	2.5	5.4	10.6	4.2%
Modèle2+Durée	0.5	2.2	2.8	5.9	11.6	4.6%
Modèle2+Elagage+Durée	0.7	2.1	2.5	5.9	11.4	4.5%

TAB. 7.3 – *Idem table 7.2. Paramètres acoustiques J-RASTA-PLP.*

contours des lèvres sont tracés manuellement pour un nombre important d'images d'entraînement. Ces contours sont ensuite discrétisés en un nombre limité de points définis par leurs coordonnées dans le plan. Ces contours discrétisés sont ensuite normalisés de façon à minimiser tout effet de translation, de rotation ou de facteur d'échelle. Pour chaque image, l'ensemble des points obtenus constitue un vecteur. Une analyse en composantes principales (ACP) est effectuée sur base de l'ensemble des vecteurs, de façon à isoler un nombre limité de directions principales (12 dans notre cas) sur lesquelles il sera possible de projeter chacun des vecteurs: ces directions principales constituent en quelque sorte un modèle des contours des lèvres. Pendant l'utilisation, une phase d'analyse appellée *suivi de contour* permet d'extraire la projection des contours suivant ces directions principales: elle consiste à minimiser un critère d'erreur en utilisant une approche itérative. Ce critère d'erreur est notamment basé sur un modèle de luminosité dans le voisinage des contours

intérieur et extérieur des lèvres. Ce modèle de luminosité est également obtenu par ACP sur base de vecteurs issus du corpus d'entraînement.

Nous utilisons les mêmes topologies et la même segmentation initiale que pour le système acoustique décrit précédement. Cette fois, le MLP comprend 70 neurones cachés.

Le taux d'erreur moyen pour les cinq partitions de la base de données est de 40.3%. Comme le signal visuel ne fournit qu'une information partielle, ce taux d'erreur est nettement plus élevé que celui du système acoustique. Cela est essentiellement dû à la similitude visuelle des chiffres "quatre", "cinq", "six", et "sept".

7.2.3 Reconnaissance vocale audio-visuelle

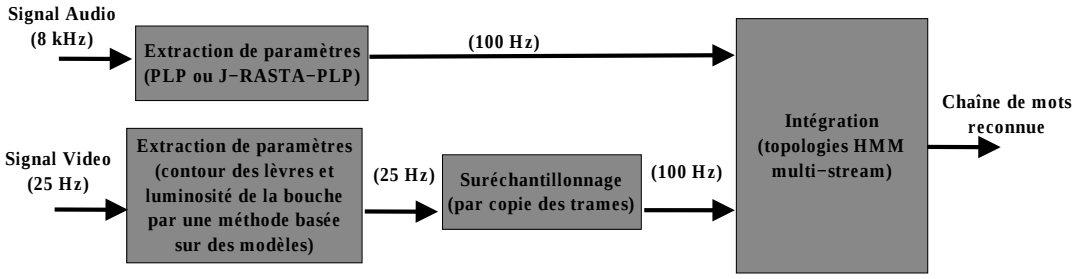


FIG. 7.1 – *Architecture du système de reconnaissance vocale audio-visuelle.*

Des essais de reconnaissance audio-visuelle ont ensuite été effectués. L'architecture du système est présentée à la figure 7.1. Trois types de modèles ont été comparés. Ces modèles, basés sur les topologies HMMs utilisées précédement, diffèrent essentiellement par la désynchronisation possible des deux canaux d'information.

Le premier modèle (MODÈLE 0) implémente l'approche EI (voir l'introduction de ce chapitre). Les paramètres acoustiques et visuels constituent les entrées d'un MLP unique comportant 150 neurones cachés.

Le second modèle (MODÈLE 1) est un modèle "multi-stream" avec fusion au niveau de l'état. Son avantage par rapport à l'approche EI est de permettre la pondération des deux canaux d'information. Cependant, il ne permet toujours pas l'asynchronisme des canaux, comme le modèle LI.

Le troisième modèle (MODÈLE 2) est un modèle "multi-stream" avec fusion au niveau du mot (figure 7.2). Il permet donc aux chemins de programmation dynamique d'être indépendants du début à la fin des mots.

De façon similaire à l'approche LI, l'approche "multi-stream" fait appel à un formalisme de combinaison des informations des différentes modalités. Dans nos expériences, nous effectuons une combinaison par multiplication des vraisemblances, supposant donc l'indépendance des canaux acoustique et visuel. Le formalisme est le suivant

$$P(X^a,X^v|M) = P(X^a|M^a)^{\alpha} P(X^v|M^v)^{(1-\alpha)}, \tag{7.1}$$

Le coefficient de pondération α ($0 \leq \alpha \leq 1$) permet de représenter la fiabilité des deux modalités. Il dépendra de leurs performances intrinsèques et de la présence de bruit acoustique ou visuel. Nous optimiserons α sur un ensemble de développement

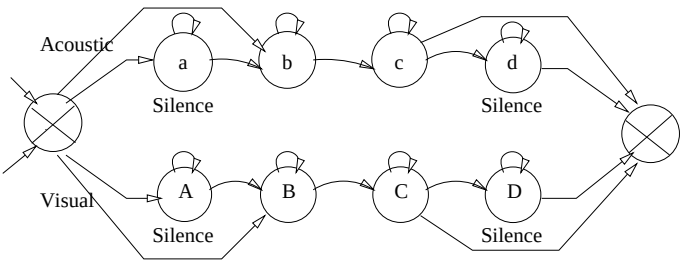


FIG. 7.2 – *Modèle "multi-stream" d'un mot pour la reconnaissance audio-visuelle de la parole. Les états de silence sont optionnels (voit texte).*

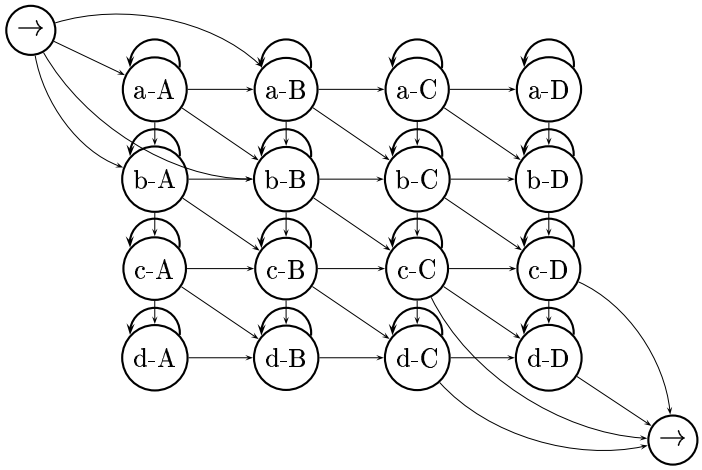


FIG. 7.3 – *Topologie HMM du modèle composite construit sur base du modèle "multi-stream" de la figure 7.2.*

perturbé de la même façon que les données de test. Dans nos expériences finales cependant, une estimation automatique du rapport signal/bruit permettra d'ajuster α de façon dynamique.

Nous avons observé que la valeur optimale du coefficient de pondération peut être obtenue par régression linéaire sur le rapport singal/bruit (RSB). Pour notre meilleur système, le coefficient de corrélation entre le RSB et le poids optimal est de 0.99. La loi linéaire est la suivante:

$$\alpha = 0.009 * RSB(dB) + 0.512 \tag{7.2}$$

les valeurs de α étant en outre limitées à 0 et 1.

Le MODÈLE 2 permet notamment aux transitions silence/parole et parole/silence de se produire à des instants différents pour les deux canaux, les lèvres pouvant commencer à bouger avant ou après la production de sons. La figure 7.4 représente en parallèle un spectrogramme de parole et l'évolution du premier paramètre visuel, représentant esentiellement les déplacements du contour labial inférieur [125]. A partir de cette figure, ainsi que de l'étude sur base d'histogrammes d'asynchronisme (voir la Section suivante et la figure 7.7), on pourra se convaincre que les deux signaux sont partiellement synchrones sur certaines portions et partiellement asynchrones sur d'autres. Idéalement, nous souhaiterions utiliser un modèle forçant les deux canaux à être synchrones pendant les portions naturellement synchrones, et à être asynchrones où les signaux sont typiquement asynchrones. Ceci justifie l'utilisation du MODÈLE 2. Ce modèle est implémenté suivant l'approche en "modèles composites" de la Section 5.3. Le modèle de la figure 7.2 conduit ainsi au modèle composite de la figure 7.3. Dans les expériences effectuées ici, l'asynchronisme entre les deux modalités est limité à un état. Sur la figure 7.3, les états suivants ne sont donc pas permis: $a-C$, $a-D$, $b-D$, $c-A$, $d-A$ et $d-B$. Le coefficient de pondération de l'équation (7.1) est optimisé par minimisation du taux d'erreur sur les données de développement, perturbées au même rapport signal/bruit que les données de test.

Nous avons utilisé les mêmes paramètres caractéristiques qu'aux paragraphes précédents. Le taux de raffraichissement video (25 Hz) étant différent du taux audio (100 Hz), les vecteurs visuels ont été copiés pour aboutir à un taux de 100 Hz.

Les résultats sont résumés dans les tables 7.1 et 7.2. Les intervalles de confiance pour des taux d'erreur de 10%, 30% et 50% sont $\pm 1.4\%$, $\pm 2.1\%$ et $\pm 2.3\%$ respectivement ($\alpha = 0.05$). Pour la parole claire, utiliser l'information visuelle n'apporte pas d'amélioration significative. Pour la parole perturbée par du bruit additif, par contre, le gain est très net. La faible amélioration due au MODÈLE 2 (par rapport au MODÈLE 1) n'est cependant pas significative[3]. Constatons finalement que l'approche EI conduit à des résultats inférieurs. Pour cette tâche, le gain d'un coefficient de pondération adaptatif semble donc surpasser les possibles dégradations dues à l'hypothèse d'indépendance.

3. Ceci est en contradiction avec les résultats de [194] qui vont en faveur d'une approche permettant l'asynchronisme des canaux d'information même si celle-ci est limitée au phonème (les modèles de phonèmes étant constitués de 3 états différents).

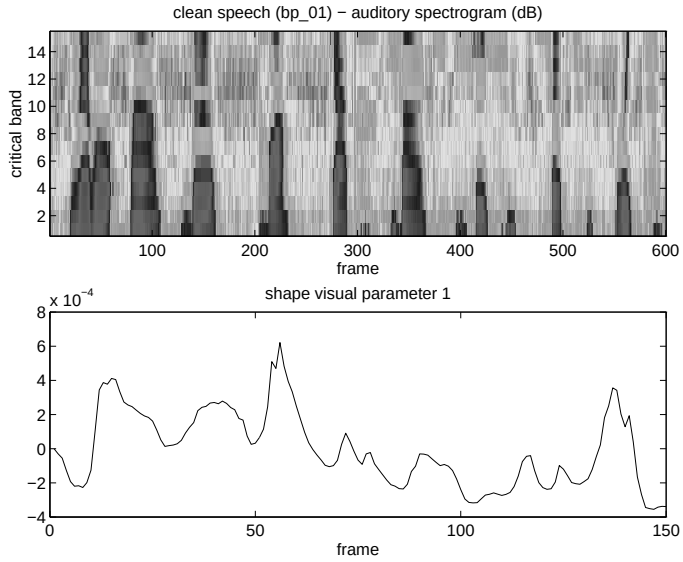


FIG. 7.4 – *Spectrogramme en bandes critiques et évolution du premier paramètre visuel pour une portion (de '0' à '8') d'une des phrases de la base de données* M2VTS.

7.2.4 Modélisation de l'asynchronisme des canaux

Nous avons appliqué les approches proposées à la Section 5.3.2 en vue de modéliser l'asynchronisme entre les deux sources d'information. Un alignement Viterbi forcé de l'ensemble d'entraînement a été obtenu sur base du MODÈLE 2. Les probabilités a priori des états des modèles composites sont calculées sur base de ces alignements et les états les moins probables sont éliminés des modèles (les états synchronisés sont cependant toujours conservés). Nous appellerons cette approche **élagage statique**. Le nombre d'états conservés est optimisé de façon à obtenir les meilleures performances sur l'ensemble de développement. On constatera avec intérêt que les meilleurs modèles comportent 25 états hors-diagonale (états désynchronisés) en plus des 52 états synchronisés (voir figure 7.5). Les données de test sont ensuite utilisées pour effectuer des expériences de reconnaissance sur base de ces modèles composites "simplifiés". Les résultats sont significativement meilleurs que ceux obtenus à partir des modèles initiaux (voir table 7.2, 5ème ligne).

Comme proposé à la Section 5.3.2, nous avons également développé des **modèles de durée** basés sur les modèles composites multidimensionnels. A cette fin, les états des modèles multidimensionnels (figure 7.3) sont remplacés par des topologies HMMs particulières (figure 7.6) visant à modéliser la durée des états. La longueur de ces HMMs est fixée à 20 états, permettant une modélisation précise de la durée jusqu'à 200 ms. Une boucle sur le dernier état permet des durées supérieures, avec une loi de probabilité en exponentielle décroissante. Les paramètres de ces modèles sont estimés de façon classique.

Les délais de transition (d'un état HMM à un autre) entre les deux canaux ont ensuite été mesurés sur base de l'alignement forcé obtenu à partir du MODÈLE 2. L'observation des histogrammes (voir les figures 7.7, 7.8, 7.9 et 7.10) de ces délais

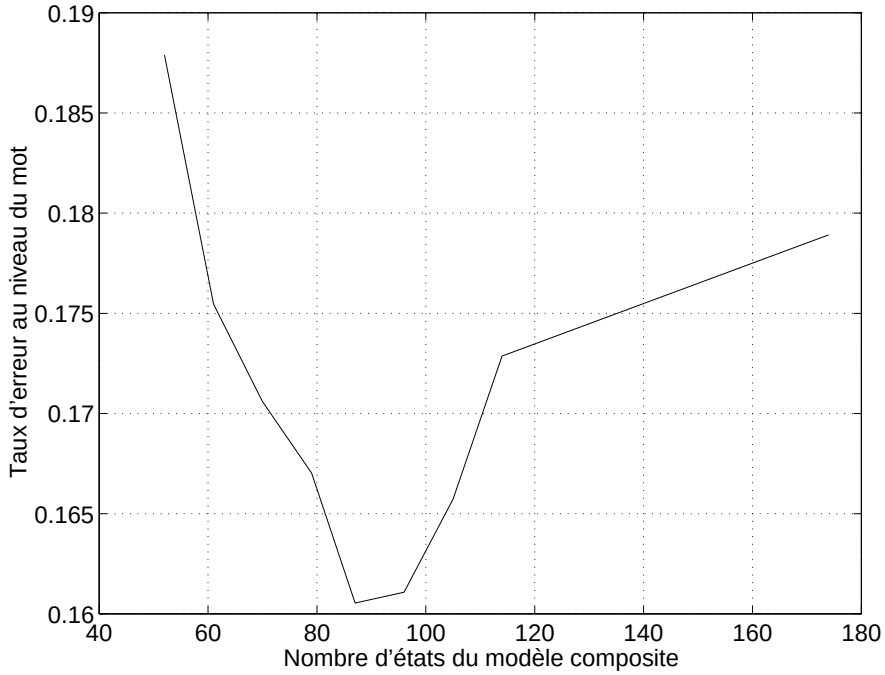


FIG. 7.5 – *Taux d'erreur au niveau du mot pour différents degrés d'élagage statique. Une configuration sans élagage (174 états) conduit à un taux d'erreur de 17.9%. En ne conservant que les états synchronisés (52 états), le taux d'erreur est de 18.8%. En conservant 25 états supplémentaires, le taux d'erreur est de 16.1%. Ces taux d'erreur sont une moyenne sur 5 conditions de bruit: parole claire, 20 dB, 15 dB, 10 dB et 5 dB. Voir également la table 7.2.*

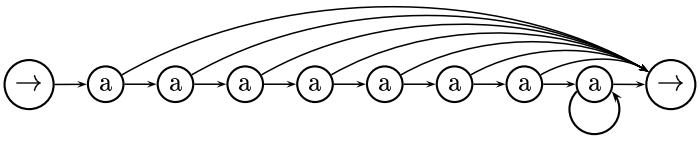


FIG. 7.6 – *Exemple de topologie HMM pour la modélisation de la durée. Les probabilités d'émission sont identiques pour les différents états. Le modèle de durée est encodé dans les probabilités de transition.*

nous montre que certains sont relativement étroits, alors que d'autres ont une variance élevée (voir la figure 7.10). De plus, le délai de transition moyen n'est pas toujours proche de 0. Une observation plus approfondie nous montre même que certains histogrammes sont significativement décalés vers les valeurs positives, indiquant un motif où la transition visuelle est généralement en retard sur la transition acoustique; d'autres par contre sont décalés vers les valeurs négatives (voir la figure 7.9). Ces observations tendent à indiquer que les délais de transition ne sont pas simplement le produit d'un bruit d'alignement (comme suggéré dans [143] pour des canaux correspondant à différentes bandes de fréquence), mais reflètent également une structure au niveau des motifs d'asynchronisme audio-visuels, structure qui pourrait être utile pour la reconnaissance vocale.

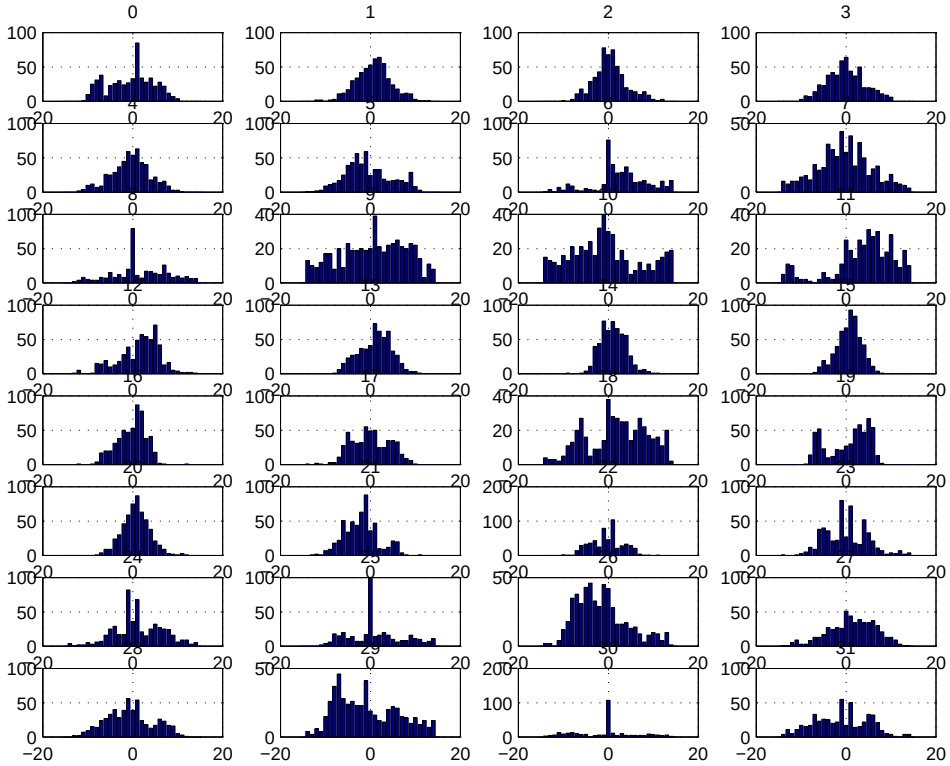


FIG. 7.7 – *Distributions de probabilité des délais de transition (délai du canal visuel sur le canal acoustique, de -14 à 14 trames de 10 ms).*

Avec cette approche, les modèles de durée, introduits initialement pour modéliser les motifs d'asynchronisme, modélisent également la durée des états synchronisés. Cette approche ne peut donc être comparée à un HMM standard sans modélisation de durée. Le même type de modèle de durée a donc été utilisé pour des expériences additionnelles avec les topologies purement synchronisées (HMMs standards). Il a également été appliqué aux meilleurs modèles composites obtenus par "élagage statique". Nous avons observé que cette modélisation de durée accroît significativement la robustesse au bruit de tous les systèmes envisagés ici.

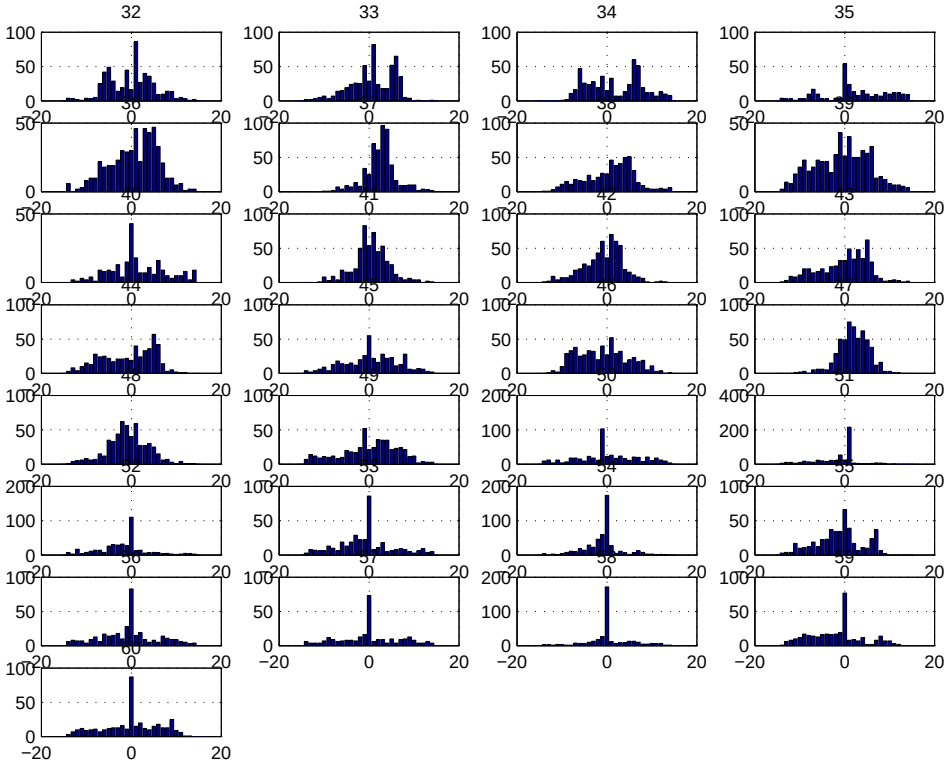


FIG. 7.8 – *Distributions de probabilité des délais de transition (suite).*

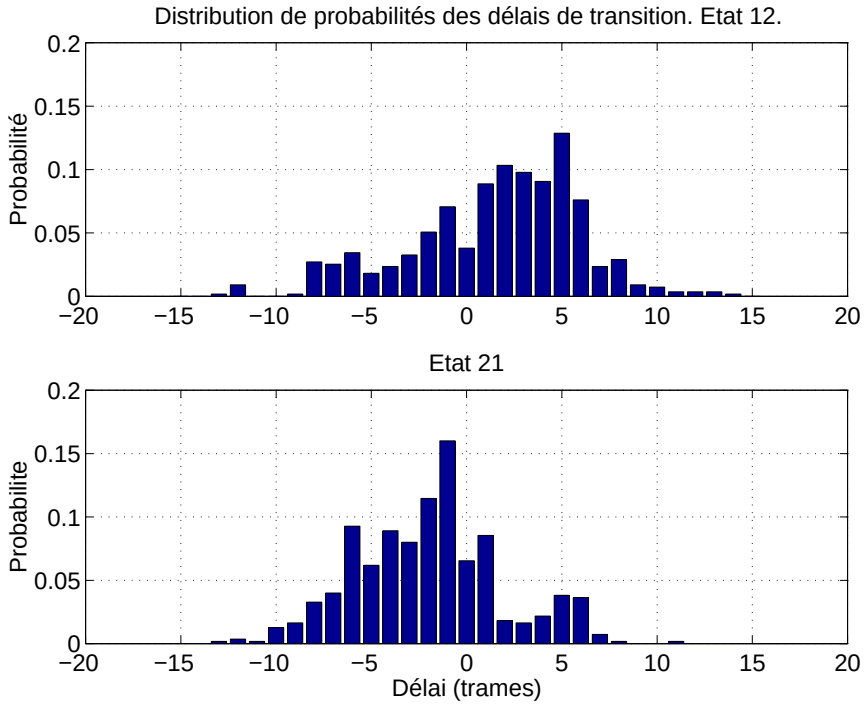


FIG. 7.9 – *Distribution de probabilité du délai de transition. La figure du haut correspond à la transition du premier au deuxième état du mot "trois". La figure du bas correspond à la transition du troisième au quatrième état du mot "quatre".*

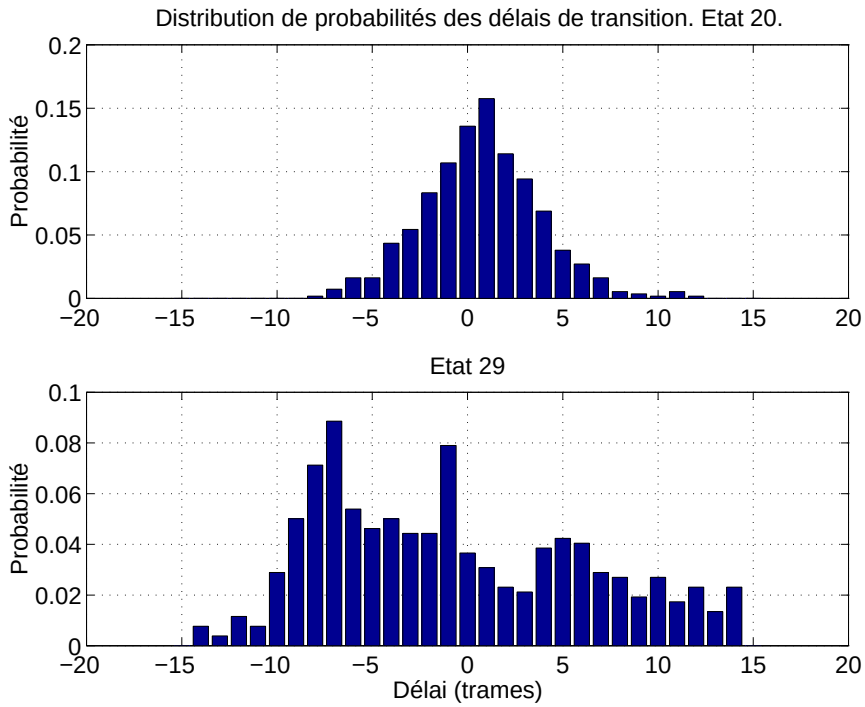


FIG. 7.10 – *Distribution de probabilité du délai de transition. La figure du haut correspond à la transition du deuxième au troisième état (sur 7 états) du mot "quatre". La figure du bas correspond à la transition du quatrième au cinquième état (sur 5) du mot "cinq".*

Comme on peut le constater à la table 7.2, le MODÈLE 2 (utilisant l'élagage statique) est significativement meilleur que les autres modèles. Ce résultat suggère que l'asynchronisme des deux modalités, modélisé sous forme de topologie multidimensionnelle, apporte un gain significatif en terme de robustesse. Pour les systèmes avec modélisation de durée cependant, ce modèle n'est pas significativement meilleur.

7.2.5 Paramètres acoustiques robustes

Pour permettre une comparaison avec les techniques de reconnaissance acoustique robuste, les expériences décrites précédemment ont été reproduites sur base de paramètres acoustiques J-RASTA-PLP ("Relative Spectra") [84]. L'intérêt de cette technique a été illustré au Chapitre 3 où nous avons notamment montré qu'elle supporte la comparaison avec la technique de soustraction spectrale. Les résultats (voir tables 7.2 et 7.3) obtenus ici confirment l'intérêt de ce type de paramétrisation. Comme précédemment, l'utilisation des deux modalités conduit à un pas en avant très important en terme de robustesse. Par contre, la modélisation des motifs d'asynchronisme et l'utilisation d'approches de décodage permettant la désynchronisation des canaux n'apportent aucune amélioration.

Les résultats concernant l'utilisation du modèle "multi-stream" et la modélisation de l'asynchronisme des deux modalités sont donc mitigés. Soulignons cependant un des résultats marquants de ces expériences. A un rapport signal/bruit de 15 dB, les paramètres PLP et J-RASTA-PLP conduisent respectivement à des taux d'erreur de 56.3% et 7.2%. L'utilisation additionnelle des paramètres visuels conduit à un taux d'erreur final de 2.5% (voir figure 7.11).

7.2.6 Bruit réel

Finalement, nous avons perturbé les données de test par un bruit non-stationnaire de la base de données *Madras*: bruit de voitures enregistré le long d'une chaussée (voir Annexe A). Ce bruit a été utilisé car il est plus réaliste que le bruit blanc stationnaire utilisé précédemment. Contrairement aux expériences précédentes, le niveau de bruit sera estimé de manière automatique par la méthode décrite au Chapitre 4. Il s'agit d'une méthode de suivi d'enveloppe avec filtrage des harmoniques. Sur base de cette estimation et des poids optimaux résultant des expériences précédentes, une régression linéaire est utilisée pour adapter le poids de combinaison (équation (7.1)) de façon dynamique. Les résultats de reconnaissance sont présentés à la table 7.4. Ici encore, la conclusion principale est que l'intégration audio-visuelle apporte un gain en robustesse important et que le gain d'un coefficient de pondération adaptatif semble surpasser les dégradations dues à l'hypothèse d'indépendance.

7.2.7 Entraînement sur une base de données plus importante

Comme nous l'avons déjà mentionné, la base de données M2VTS est relativement limitée. Elle ne permet pas de développer des systèmes comportant un grand nombre de paramètres. Il existe cependant de nombreuses bases de données acoustiques qui pourraient être utilisées pour développer des systèmes de reconnaissance acoustique plus performants.

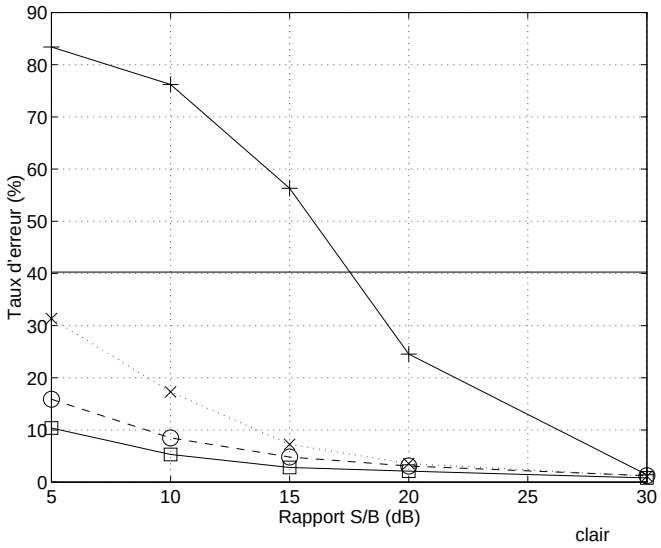


FIG. 7.11 – *Taux d'erreur au niveau du mot pour la reconnaissance de séquences de chiffres, le nombre de chiffres étant connu a priori. La courbe du haut correspond au système utilisant des paramètres PLP. La courbe en pointillés correspond au système basé sur des J-RASTA-PLP. La troisième courbe donne les performances du système multimodal EI basé sur des J-RASTA-PLP et la dernière courbe correspond à notre meilleur système multimodal basé sur les mêmes paramètres représentatifs.*

Système	Visuel	Acoustique	Modèle0 (EI)	Modèle1	Modèle2
Taux d'erreur	40.3%	10%	6.9%	3.7%	4.2%

TAB. 7.4 – *Taux d'erreur au niveau du mot pour l'acoustique, le visuel et plusieurs systèmes audio-visuels. Le signal est perturbé par un bruit de voiture non stationnaire de la base de données* Madras *(rapport signal/bruit moyen = 10 dB). Des paramètres de type J-RASTA-PLP sont utilisés pour le système acoustique.*

Nous avons notamment utilisé 2850 phrases (à comparer avec les 185 phrases de chacun des ensembles d'entraînement de M2VTS), prononcées par 32 locuteurs différents, provenant de la base de données de chiffres connectés BDSONS. Des systèmes basés sur les paramètres PLP et J-RASTA-PLP ont été d'eveloppés. Nous nous attendions à ce que ces systèmes conduisent à des performances supérieures à celles des systèmes acoustiques développés sur M2VTS. Nous avons au contraire observé une dégradation que nous pouvons expliquer par:

- des différences dans les conditions d'enregistrement. Les séquences de BDSONS sont de très bonne qualité alors que le niveau de bruit est relativement élevé pour M2VTS.
- le fait que les locuteurs utilisés pour les tests sont les mêmes que ceux utilisés pour l'entraînement du système acoustique M2VTS, alors que les locuteurs à la base de l'entraînement du système BDSONS sont bien entendu différents.

Finalement, comme les performances du système acoustique basé sur M2VTS (taux d'erreur de 1.1%) sont proches de l'état de l'art, nous n'avons pas poursuivi cette investigation.

7.3 Conclusions

Nous avons décrit un système complet pour la reconnaissance vocale audio-visuelle. Ce système correspond à une tâche de reconnaissance multi-locuteurs de chiffres connectés. Il a été testé dans différentes conditions de bruit.

Nous avons proposé différentes méthodes de fusion des données audio-visuelles sur base de l'approche "multi-stream". Celle-ci permet la collaboration de plusieurs sources d'information sur base de différents modèles de Markov caché. Plusieurs avancées importantes sont rapportées ici. Tout d'abord, la méthode permet le décodage synchrone de parole continue. De plus, la fiabilité des deux modalités peut être introduite sous forme d'un poids adaptatif dans le formalisme d'intégration. Nous avons montré que le gain dû à l'adaptation du poids de combinaison surpasse l'effet défavorable de l'hypothèse d'indépendance sur laquelle repose notre critère d'intégration. Finalement, l'approche permet la désynchronisation des deux canaux.

L'observation de statistiques obtenues sur base d'alignements Viterbi audio et video semble indiquer que les délais de transition entre les deux canaux ne sont pas simplement le produit d'un bruit d'alignement mais reflètent également une structure dont la modélisation pourrait être bénéfique à la reconnaissance vocale. Les résultats expérimentaux sont cependant très mitigés. La modélisation de l'asynchronisme conduit à une amélioration significative en terme de robustesse pour un système utilisant des paramètres représentatifs standard. Lorsqu'on utilise des paramètres robustes cependant, le gain est nul. Ce point important pourrait faire l'objet de recherches plus approfondies. Insistons également sur la nécessité d'obtenir des corpus audiovisuels de plus grande taille (M2VTS ne contient que 185 séquences de chiffres) et d'étendre les recherches vers des tâches plus complexes.

Finalement, par rapport à un système de reconnaissance acoustique (même si l'on utilise des paramètres acoustiques robustes), le système audio-visuel conduit à une réduction importante du taux d'erreur en présence de bruit additif: il est en effet divisé par trois dans certaines conditions. Le système obtenu est totalement automatique, la seule contrainte étant de lui fournir des séquences d'images de la région entourant les lèvres du locuteur. Des techniques de suivi de visage ("face tracking") [10, 170] pourraient être utilisées à cette fin.

Chapitre 8

Vers un système robuste

Le but de ce chapitre est d'introduire une nouvelle approche de reconnaissance automatique de la parole et de la comparer aux approches proposées dans la littérature ainsi qu'aux techniques originales développées dans les chapitres précédents. Cette comparaison portera sur de nombreux tests à partir de données de parole perturbées par des bruits réels provenant de bases de données de bruit.

Nous rappelons également quelques résultats de recherche récents dans des domaines autres que la reconnaissance vocale en milieu bruité. Ces résultats pourraient bénéficier à tout système de reconnaissance vocale.

8.1 Introduction

Différentes sources de variabilité rendent difficile la tâche de reconnaissance, comme par exemple les différences de voix d'une personne à l'autre, les mauvaises prononciations, les accents locaux, les conditions d'enregistrement de la parole et le bruit ambiant. Ainsi, si l'utilisation de systèmes de reconnaissance vocale dans des conditions bien contrôlées donne généralement satisfaction, le taux d'erreur de tels systèmes augmentent sensiblement en présence de bruit.

L'effet du bruit est de conduire à un "désaccord" entre les modèles de reconnaissance et les vecteurs représentatifs issus de l'analyse du signal de parole: les signaux de parole rencontrés lors de l'utilisation des systèmes ne correspondent plus aux signaux rencontrés à l'entraînement et donc aux modèles. Ce "désaccord" entraîne une dégradation des performances des systèmes de reconnaissance: leurs taux d'erreurs augmentent. Comme nous l'avons déjà illustré au chapitre 3, cette dégradation est d'autant plus importante que le niveau de bruit est élevé.

De nombreuses techniques ont été développées en vue de diminuer la sensibilité de ces systèmes au bruit, notamment au bruit additif. Ces techniques de **reconnaissance vocale robuste** ont fait l'objet du chapitre 3. Rappelons ici les principales familles de méthodes.

Une première famille vise à effectuer un traitement dont le but est d'obtenir soit une version sensiblement débruitée d'un signal bruité, soit une version sensiblement débruitée (compensée) de paramètres représentatifs [115]. Bien que très efficaces

dans certains cas, ces techniques présentent néanmoins l'inconvénient d'introduire des distorsions (bruit musical) et sont généralement insuffisantes pour permettre l'application de la reconnaissance vocale à différents environnements acoustiques, et particulièrement dans le cas de niveaux de bruit élevés.

Une autre famille de méthodes concerne l'obtention de paramètres représentatifs intrinsèquement moins sensibles au bruit que les paramètres classiques [201, 85] introduits à la section 2.7.

Au lieu de chercher à transformer les paramètres représentatifs, on peut également tenter de transformer les paramètres des modèles intervenant dans les systèmes de reconnaissance vocale de façon à les adapter aux conditions d'utilisation courantes [199, 112]. Ces techniques d'adaptation sont en fait des techniques d'apprentissage rapide et ne sont efficaces que si les conditions de bruit varient lentement. En effet, elles nécessitent plusieurs secondes de signaux de parole bruités pour adapter les modèles de reconnaissance. Si suite à cette adaptation, les conditions de bruit changent à nouveau, le modèle adapté ne conviendra plus.

Une quatrième famille de techniques consiste à contaminer, par bruitage à plusieurs niveaux de bruits différents[1], la totalité ou une partie du corpus d'apprentissage et à estimer les paramètres des modèles intervenant dans les systèmes de reconnaissance sur base de ce corpus bruité [148]. L'intérêt de ces méthodes est de présenter des performances quasi-optimales lorsque le bruit caractérisant les conditions d'utilisation est similaire au bruit utilisé pour contaminer le corpus d'apprentissage. En revanche, lorsque les deux bruits sont différents, ces méthodes ont peu d'intérêt. Leur domaine d'application est donc malheureusement limité, dans la mesure où il n'est pas envisageable d'effectuer la contamination sur base de bruits diversifiés qui couvriraient tous les bruits pouvant être rencontrés lors de l'utilisation (entraînement multi-styles).

Une cinquième famille de méthodes consiste à mener une analyse permettant d'obtenir des paramètres représentatifs de bandes de fréquence. Un modèle peut alors être développé pour chacune de ces bandes; l'ensemble des bandes devant idéalement couvrir tout le spectre utile[2]. L'intérêt de ces techniques "multi-bandes", étudiées au chapitre 6, est de pouvoir minimiser, dans une phase de décision ultérieure, l'importance de bandes de fréquence fortement bruitées. Cependant, comme nous l'avons déjà observé, ces techniques sont peu efficaces lorsque le bruit couvre une portion importante du spectre fréquentiel utile.

Les limitations de ces deux dernières classes de techniques peuvent être résumées comme suit. Les techniques basées sur la contamination du corpus d'apprentissage ont peu d'intérêt dans le cas de conditions d'utilisation présentant des bruits aux caractéristiques très diverses[3]. Ce problème a notamment été identifié dans [164]. Les techniques basées sur le multi-bande quant à elles n'ont pas d'intérêt lorsque le bruit présente une caractéristique large bande.

L'approche présentée ici propose une variation importante des techniques opérant

1. Niveaux susceptibles d'être rencontré en pratique

2. Bande spectrale utilisée par le système de reconnaissance, qui peut-être inférieure à la bande spectrale du signal de parole (de 100Hz à 12000Hz environ)

3. Par exemple, un bruit essentiellement basse fréquence suivi par un autre essentiellement haute fréquence

par contamination. Cette variation utilise une architecture de type multi-bande. Elle est basée sur l'observation que, si l'on considère une bande de fréquence relativement étroite, les bruits ne différent essentiellement que par leur niveau. De ce fait, des modèles associés à chacune des bandes de fréquence du système peuvent être entraînés après contamination du corpus d'apprentissage par un bruit quelconque; ces modèles demeurent relativement insensibles à d'autres types de bruits. Une phase de décision ultérieure utilise alors ces modèles robustes en vue de la reconnaissance automatique de la parole. Nous montrerons que l'approche peut également trouver des applications dans d'autres domaines des technologies vocales: en débruitage et en codage de la parole par exemple.

La méthode proposée implique l'obtention de paramètres représentatifs pour différentes bandes de fréquence. Ces paramètres sont similaires aux paramètres classiques. Ils seront par exemple calculés par prédiction linéaire et/ou sur base d'une analyse en banc de filtres (voir section 6.4).

Pour chaque bande de fréquence, nous développons ensuite un système dont le but est d'estimer des paramètres représentatifs insensibles au bruit. Ces paramètres sont calculés sur base des paramètres représentatifs propres à la bande considérée. Pour atteindre cet objectif, ce système est entraîné sur base d'un corpus de parole contaminé par du bruit couvrant une plage raisonnable de niveaux de bruits différents, étant donné les niveaux qu'on est susceptible de recontrer en pratique, lors de l'utilisation: rapport signal/bruit au dessus de 0 dB en général. Nous avons trouvé approprié d'utiliser un réseau de neurones artificiels calculant des paramètres discriminants suivant l'approche d'analyse discriminante non-linéaire [58]. Cependant, d'autres approches pourraient également convenir: par exemple des techniques d'analyse discriminante linéaire [60], d'analyse en composantes principales [60], ou des techniques de régression permettant l'estimation d'une version débruitée des paramètres représentatifs.

La méthode porte donc essentiellement sur l'utilisation d'une procédure d'apprentissage particulière (basée sur une contamination des données) dans le cadre d'une architecture particulière (basée sur une décomposition en bandes de fréquence). Comme on peut s'en convaincre en analysant la littérature touchant au domaine concerné, et comme il a également été rappelé dans les sections précédentes, ces deux aspects de cette nouvelle technique sont d'un intérêt relativement limité lorsqu'ils sont utilisés de façon disjointe. C'est la combinaison des deux qui confère à l'ensemble des propriétés particulièrement intéressantes.

Dans la mesure où elles sont complémentaires, les différents types de techniques robustes, notamment celles rappelées dans l'introduction de ce chapitre, peuvent être combinées dans le cadre du système proposé ici. Nous donnerons des illustrations à la section suivante, qui décrit le système que nous avons développé en vue d'évaluer la méthode proposée.

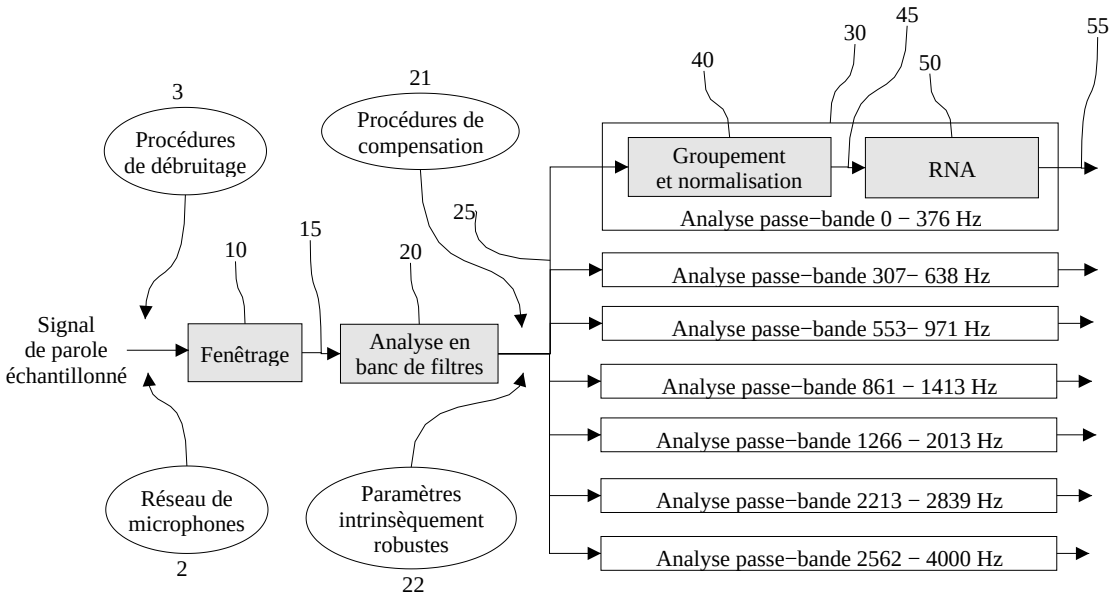


FIG. 8.1 – *Schéma des premières étapes de traitement, jusqu'à l'obtention de paramètres représentatifs relativement insensibles au bruit, associés à sept bandes de fréquence.*

8.2 Description

Tout d'abord, le signal échantillonné à une fréquence de 8 kHz (5) [4] passe par un module de fenêtrage (10) qui divise le signal en une succession de trames temporelles (15) de 30 ms (240 échantillons). Deux trames successives se recouvrent de 20 ms. Les éléments de chaque trame sont pondérés par une fenêtre de Hamming.

Dans une première étape de traitement, on effectue une analyse (20) en bandes critiques sur chaque trame de signal échantillonné. Cette analyse est représentative de l'échelle de résolution fréquentielle de l'oreille humaine. L'approche utilisée est inspirée de la première phase d'analyse de la technique PLP [83]; elle opère dans le domaine fréquentiel. Les filtres utilisés sont trapézoïdaux et l'écart entre les fréquences centrales de deux filtres successifs est fixé à 0.5 Bark dans notre cas. D'autres valeurs pourraient être envisagées. Pour un signal échantillonné à 8 kHz, cette analyse conduit à un vecteur comprenant les énergies de 30 bandes de fréquence (25). La procédure inclu également une accentuation des hautes fréquences.

Ce vecteur de 30 éléments est ensuite dissocié en différents sous-vecteurs représentatifs de l'enveloppe spectrale dans diverses bandes de fréquence. Parmi les configurations testées, une dissociation en sept bandes de fréquence donne les meilleurs résultats (voir figure 8.1). La décomposition suivante est utilisée: 1-4 (les filtres indicées de 1 à 4 constituent la première bande de fréquence), 5-8, 9-12, 13-16, 17-20, 21-24 et 25-30 (les fréquences concernées par ces sept systèmes sont données à la

4. Les nombres entre parenthèses font référence aux figures 8.1, 8.2 et 8.3.

figure 8.1). Chaque sous-vecteur est normalisé en divisant les valeurs de ses éléments par la somme de tous les éléments du sous-vecteur, c'est-à-dire par une estimation de l'énergie du signal dans la bande de fréquence considérée. Cette normalisation confère au sous-vecteur une insensibilité vis-à-vis du niveau d'énergie du signal. Pour chaque bande de fréquence, les paramètres représentatifs sont finalement constitués du sous-vecteur normalisé correspondant à la bande ainsi que de l'estimation de l'énergie du signal dans cette bande. Pour chacune des sept bandes de fréquence, le traitement décrit dans ce paragraphe est réalisé par le module (40) qui fournit un vecteur de paramètres représentatifs (45) de la bande considérée.

Les modules (10), (20) et (40) pourraient être remplacés par toute autre approche permettant d'obtenir des paramètres représentatifs de bandes de fréquence différentes.

Pour chaque bande de fréquence, les paramètres représentatifs correspondants sont ensuite utilisés par un système (50) dont le but est d'estimer un vecteur de paramètres (55) relativement insensibles au bruit présent dans le signal de parole échantillonné. Les vecteurs de paramètres insensibles au bruit associés à chacune des bandes de fréquence sont ensuite concaténés. Le vecteur obtenu (56) est finalement utilisé comme vecteur de paramètres représentatifs de la trame considérée. Il pourra servir à la reconnaissance automatique de la parole (60) dont le but est de fournir la séquence d'unités de parole qui ont été prononcées.

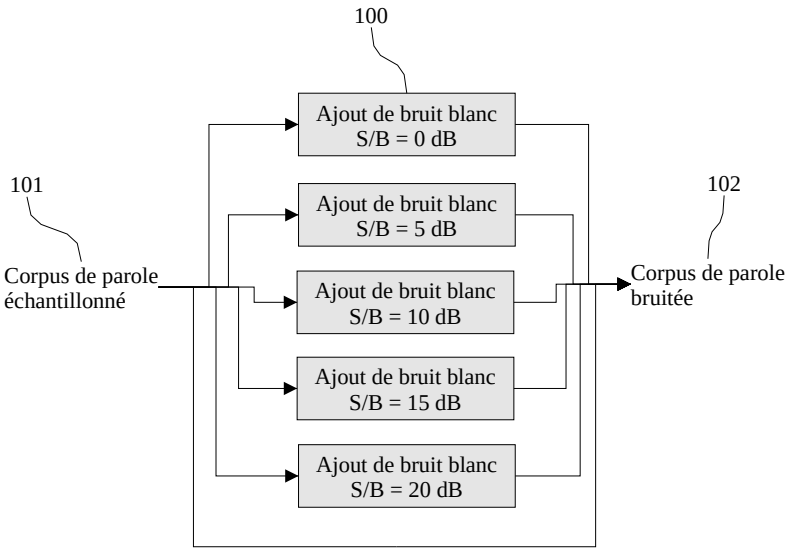


FIG. 8.2 – *Principe de contamination du corpus d'apprentissage par du bruit blanc.*

Pour réaliser la fonctionnalité souhaitée, nous avons utilisé un réseau de neurones artificiels (RNA) [167] comme implémentation du système (50). Comme nous l'avons déjà signalé, d'autres approches pourraient également convenir: par exemple des techniques d'analyse discriminante linéaire [60], d'analyse en composantes principales [60], ou des techniques de régression permettant l'estimation d'une version

débruitée des paramètres représentatifs. Le réseau de neurones utilisé ici est un perceptron multicouche comprenant deux couches de neurones cachés. Les fonctions de sorties de ce perceptron sont des sigmoïdes. Ce réseau de neurones artificiels est entraîné par l'algorithme de rétro-propagation sur base d'un critère de minimisation de l'entropie relative. L'entraînement est supervisé et fait appel à des cibles correspondant aux classes phonétiques des exemples d'entraînement présentés. A l'utilisation, les sorties de ce réseau de neurones estimeront donc, pour chaque trame de paramètres représentatifs (45), les probabilités à posteriori des différentes classes phonétiques. Comme spécifié dans l'introduction, les paramètres de ce RNA sont estimés sur base d'un corpus d'apprentissage (101) contaminé par du bruit (102). De manière à couvrir une majorité des niveaux de bruits susceptibles d'être rencontrés en pratique, six versions du corpus d'apprentissage sont utilisées ici. Une des versions est utilisée telle quelle, c'est-à-dire sans bruit ajouté. Les autres versions sont bruitées (100) à des rapports signal/bruit différents: 0 dB, 5 dB, 10 dB, 15 dB et 20 dB. Ces six versions sont utilisées pour entraîner le RNA. Ces données d'entraînement sont utilisées en entrée du système présenté à la figure 8.1. Ce système permet d'obtenir des paramètres (45) représentatifs des différentes bandes de fréquence envisagées. Ce sont ces paramètres qui alimentent ces réseaux de neurones artificiels et permettent l'entraînement par rétro-propagation de l'erreur [167]. Remarquons également que toutes les techniques qui sont généralement mises en oeuvre lorsqu'on utilise des réseaux de neurones en traitement de la parole peuvent être appliquées ici. Nous avons notamment utilisé ici la possibilité d'appliquer en entrée du RNA plusieurs vecteurs représentatifs de trames successives de signal (9 trames dans notre cas) afin de modéliser la corrélation temporelle du signal de parole.

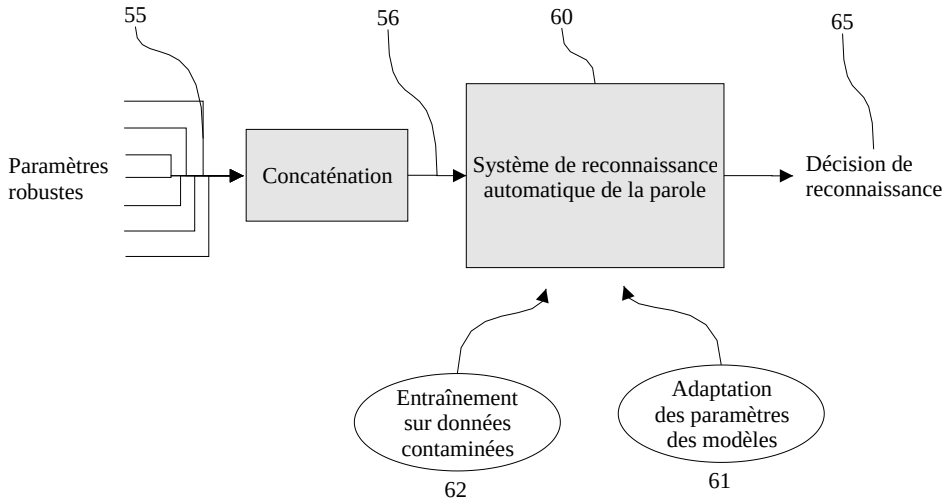


FIG. 8.3 – *Application en reconnaissance automatique de la parole.*

Lors de l'utilisation du RNA, les sorties de la seconde couche cachée, au nombre de 30, sont utilisées comme paramètres (55) insensibles au bruit pour la bande de fréquence associée. Pour chaque bande de fréquence k donc, une fonction non-linéaire

est appliquée aux vecteurs de paramètres représentatifs x_k de façon à obtenir des vecteurs de paramètres représentatifs x'_k:

$$x'_k = NLDA_k(x_k) \tag{8.1}$$

Les vecteurs de paramètres x'_k associés à chacune des sept bandes de fréquence sont ensuite concaténés pour conduire à un vecteur (56) de 210 paramètres. A chaque trame de signal, ce vecteur est alors utilisé comme entrée d'un système de reconnaissance automatique de la parole (60). Ce système est bien évidement entraîné sur base de paramètres représentatifs calculés par la technique décrite ci-dessus (système illustré à la figure 8.1) à partir d'un corpus de parole (bruité ou non bruité) en adéquation avec la tâche de reconnaissance qui nous intéresse. Notons tout de suite que le corpus de données permettant le développement des systèmes (50) associés à chaque bande de fréquence n'est pas forcément le même que celui servant à l'entraînement du système de reconnaissance vocale (60).

Comme nous l'avons déjà signalé, tous les types de techniques robustes, notamment celles rappelées dans l'introduction, peuvent intervenir librement dans le cadre du système proposé ici. Les techniques d'acquisition robustes par exemple, notamment celles basées sur les réseaux de microphones peuvent être utiles pour obtenir un signal de parole relativement débruité (figure 8.1, Num. 2). De même, les techniques de débruitage (par exemple la soustraction spectrale [11]) peuvent être envisagées (figure 8.1, Num. 3). Toute technique de calcul de paramètres intrinsèquement robustes (22) ou de compensation des paramètres représentatifs peut également être utilisée (figure 8.1, Num. 21). Ainsi, les blocs (10), (20) et (40) peuvent être remplacés par toute autre technique permettant d'obtenir des paramètres représentatifs de différentes bandes de fréquence. Plus ces paramètres sont insensibles au bruit ambiant, mieux le système global se comportera. Dans le cadre de l'application à la reconnaissance vocale, des techniques d'adaptation des modèles peuvent être utilisées. L'entraînement du système de reconnaissance sur base d'un corpus de parole contaminé par du bruit est également possible (voir la figure 8.1).

8.3 Evaluation de l'approche en parole bruitée

Des expériences ont été effectuées en vue d'évaluer l'intérêt de l'approche proposée ici et de la comparer à la méthode multi-bande (chapitre 6) ainsi qu'à d'autres approches classiques (section 3.10).

Les systèmes multi-bandes envisagés ici sont basés sur les conclusions du chapitre 6. Ainsi, ils utiliseront une combinaison par réseau de neurones. Les entrées de ce réseau de neurones comprennent:

- 3 trames successives de paramètres discriminants calculés par la méthode d'analyse discriminante non-linéaire dans chacune des différentes bandes de fréquence considérées,
- 3 trames successives de paramètres représentatifs standards.

Les systèmes utilisés sont donc identiques aux meilleurs des systèmes utilisés à la Section 6.6. De même, les systèmes de référence basés sur des approches de débruitage classiques sont identiques à ceux utilisés à la Section 6.6.

Comme pour la Section 6.6, les systèmes sont développés sur les données d'entraînement du corpus NUMBERS'95. Les tests, par contre, sont effectués sur les données de test du corpus NUMBERS'93.

8.3.1 Systèmes de reconnaissance

Rappellons que les systèmes de reconnaissance envisagés ici sont basés sur des unités phonétiques. Les transcriptions phonétiques des mots du vocabulaire ont été obtenues à partir du dictionnaire CMU 0.4 contenant 110000 mots. Les phonèmes intervenant dans le vocabulaire considéré sont au nombre de 33 (un sous-ensemble du vocabulaire phonétique du corpus TIMIT). Pour chaque phonème, les topologies HMM ont été construites de façon à imposer une durée minimale égale à la moitié de la durée moyenne du phonème. Une grammaire de type "paire de mots" est également utilisée, mais son influence est mineure.

L'estimation des probabilités a posteriori des états des HMMs est fournie par des réseaux de neurones artificiels. Quel que soit le système (système de référence ou sous-reconnaisseurs pour chaque bande de fréquence), neuf trames adjacentes sont utilisées aux entrées des réseaux de neurones artificiels. Pour permettre l'application d'une approche discriminante non linéaire (NLDA), nous avons considéré des perceptrons multicouches composés de deux couches cachées: la première comportant 400 neurones et la seconde comportant de 30 à 200 neurones, suivant la largeur des bandes utilisées.

L'étiquetage phonétique de la base de données est obtenu comme suit: dans un premier temps, un système de reconnaissance de base est développé à partir d'un étiquetage manuel (fourni avec la base de données et basé sur les phonèmes TIMIT). Ensuite, cet étiquetage est adapté de façon itérative sur base de l'approximation de Viterbi de la procédure d'entraînement EM des modèles de Markov cachés. L'étiquetage obtenu avec ce système de référence est ensuite utilisé pour développer tous les autres systèmes.

8.3.2 Evaluation sur bruits réels

Les méthodes suivantes sont comparées:

- **J-RASTA ("Relative Spectra")**: système de référence utilisant des paramètres cepstraux de type PLP obtenus après un pré-traitement J-RASTA ($J = 1*10^{-6}$ conduit ici au meilleurs résultats, quel que soit le niveau de bruit). Comme toujours, nous utilisons également les dérivées premières et secondes de ces paramètres ainsi que les dérivées première et seconde de l'énergie de la trame.
- **SPS ("Spectral Subtraction")**: système de référence utilisant un pré-traitement par soustraction spectrale. Ici, nous avons directement utilisé les énergies des bandes critiques (avec compression par racine cubique): ce type de paramètres donne en effet de meilleurs résultats que les PLP dans le cadre de la soustraction spectrale (voir table 3.1).
- **J-RASTA-multi-bande**: Il s'agit d'un système multi-bande basé sur des bandes critiques traitées par la technique J-RASTA. Le système utilise 4

bandes de frequence (voir Section 6.6). Les paramètres représentatifs utilisés en entrée des sous-reconnaisseurs sont les énergies des bandes critiques normalisées par rapport à l'énergie de la bande de fréquence considérée à la trame courante (suivi par une compression par racine cubique). La combinaison est réalisée par un perceptron multicouche utilisant des paramètres représentatifs obtenus par analyse discriminante non-linéaire. Ceux-ci sont concaténés aux paramètres représentatifs classiques (à savoir les énergies des bandes critiques compressées par racine cubique). Nous utilisons trois trames consécutives de paramètres en entrée du réseau de combinaison.

- **SPS-multibande**: système multi-bande du même type que le précédent. Les énergies des bandes critiques sont ici traitées par soustraction spectrale.
- **nouveau**: il s'agit d'une architecture multi-bande utilisant la méthode décrite dans ce chapitre. La combinaison des différentes bandes est réalisée grâce à un perceptron multicouche utilisant les paramètres obtenus par analyse discriminante non-linéaire. Les paramètres sont calculés sur base de 30 bandes critiques d'un demi Bark de largeur couvrant la plage de frequence de 0 à 4000 Hz. Un pré-traitement J-RASTA est appliqué aux énergies des bandes critiques. Quatre configurations ont été envisagées (les frequences données ici correspondent aux frequences de coupure à 3 dB):
  1. 7 bandes:
     - bande 1: bandes critiques de 1 à 4, fréquences de coupure à 3 dB: 0 et 264 Hz,
     - bande 2: bandes critiques de 5 à 8, fréquences de coupure à 3 dB: 217 et 494 Hz,
     - bande 3: bandes critiques de 9 à 12, fréquences de coupure à 3 dB: 439 et777 Hz,
     - bande 4: bandes critiques de 13 à 16, fréquences de coupure à 3 dB: 708 et 1142 Hz,
     - bande 5: bandes critiques de 17 à 20, fréquences de coupure à 3 dB: 1052 et 1629 Hz,
     - bande 6: bandes critiques de 21 à 24, fréquences de coupure à 3 dB: 1507 et 2288 Hz,
     - bande 7: bandes critiques de 25 à 30, fréquences de coupure à 3 dB: 2123 et 4000 Hz.

     Dans ce cas, la seconde couche cachée des MLP sous-bandes compte 30 neurones.
  2. 4 bandes:
     - bande 1: bandes critiques de 1 à 12, fréquences de coupure à 3 dB: 0 et 777 Hz,
     - bande 2: bandes critiques de 13 à 20, fréquences de coupure à 3 dB: 708 et 1629 Hz,
     - bande 3: bandes critiques de 21 à 26, fréquences de coupure à 3 dB: 1507 et 2704 Hz,

- bande 4: bandes critiques de 25 à 30, fréquences de coupure à 3 dB: 2123 et 4000 Hz.

Il y a ici un recouvrement plus important entre les bandes 3 et 4, comme pour les expériences multi-bandes du Chapitre 6. La seconde couche cachée des MLP sous-bandes compte ici 50 neurones.

3. 2 bandes:
   - bande 1: bandes critiques de 1 à 15, fréquences de coupure à 3 dB: 0 et 1041 Hz,
   - bande 2: bandes critiques de 16 à 30, fréquences de coupure à 3 dB: 957 et 4000 Hz.

   La seconde couche cachée des MLP sous-bandes compte ici 100 neurones.
4. 1 bande: par rapport aux systèmes de référence, les architectures multi-bandes envisagées ici ont quatres particularités (outre la décomposition en bandes de fréquence): (1) elles sont entraînées sur base de données bruitées à différents rapport signal/bruit, (2) elles ont un nombre très important de paramètres, (3) les paramètres représentatifs sont basés sur une analyse en 30 bandes de fréquence et (4), les architectures possèdent deux niveaux, le premier niveau étant chargé d'extraire des paramètres discriminants et le second étant chargé de l'estimation des probabilités a posteriori des classes phonétiques. En outre, le deuxième niveau utilise trois trames consécutives de paramètres discriminants, portant le contexte utilisé par le système à 11 trames. Nous avons donc développé un système de référence utilisant ces particularités. Il est entraîné sur les mêmes données bruitées que nos systèmes mutli-bandes. Il comprend un nombre de paramètres équivalent à ceux des systèmes multi-bandes (65000 paramètres contre 673000 pour le système à deux bandes, 710000 pour le système à 4 bandes, 727000 pour le système à 7 bandes et seulement 150000 pour le système de référence J-RASTA). Finalement, il possède une architecture à deux niveaux. Le premier niveau utilise les énergies des 30 bandes critiques (plus leurs dérivées premières et secondes) filtrées par la technique J-RASTA et normalisées par rapport à l'énergie moyenne de la trame. Les dérivées premières et secondes de l'énergie de la trame sont également utilisées. Le second niveau utilise trois trames de contexte. Nos résultats montreront qu'il est peu vraisemblable que le gain obtenu par l'approche multi-bande provienne de l'une ou l'autre de ces particularités.

Estimation du niveau de bruit

Contrairement aux expériences de la Section 6.6, la soustraction spectrale[5] repose ici sur une estimation automatique du niveau de bruit. Nous avons utilisé une des méthodes décrites au Chapitre 4. Il s'agit d'une méthode de suivi d'enveloppe avec filtrage des harmoniques. Nous avons montré qu'elle conduit à une erreur quadratique moyenne inférieure à 10 dB^2 pour des bruits ne variant pas trop rapidement.

5. Comme à la Section 3.10 et au Chapitre 6, une approche de soustraction spectrale généralisée est appliquée aux énergies des bandes critiques avec un facteur de surestimation α égal à 2 et un facteur de seuil β égal à 0.001.

Rapport S/B (dB)	5	10	20
log-RASTA	35.8	22.3	13.9
J-RASTA	23.0	15.4	10.0
SPS	22.0	15.0	10.6
J-RASTA-multi-bande	23.1	14.1	8.7
SPS-multi-bande	23.8	14.9	9.3
nouveau-7bandes	16.9	10.9	7.5

TAB. 8.1 – *Taux d'erreur au niveau du mot: moyenne sur les six types de bruits envisagés.*

Dans le cas de bruit blanc stationnaire (table 8.3), cette approche conduit à de moins bons résultats qu'une simple estimation du niveau de bruit sur base des 100 ms initiales de chaque phrase. Pour d'autres types de bruits par contre (ex: bruit le long d'une chaussée - base de donnée MADRAS - table 8.4), l'approche adaptative donne de meilleurs résultats. Pour les bruits qui ont été envisagés ici, l'approche adapative donne globalement de meilleurs résultats que l'approche simple.

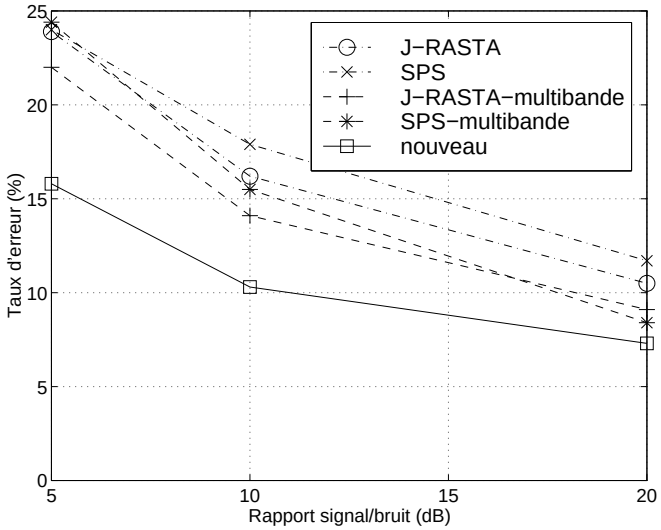


FIG. 8.4 – *Taux d'erreur au niveau du mot: bruit blanc. Le niveau de bruit est estimé automatiquement (voir texte).*

Nos résultats sont présentés aux figures 8.4, 8.5, 8.6, 8.7, 8.8, 8.9, 8.10 et 8.11, et aux tables 8.1 et 8.2.

8.3.3 Discussion

La définition du rapport signal/bruit est identique à celle donnée à la section 3.10.1.

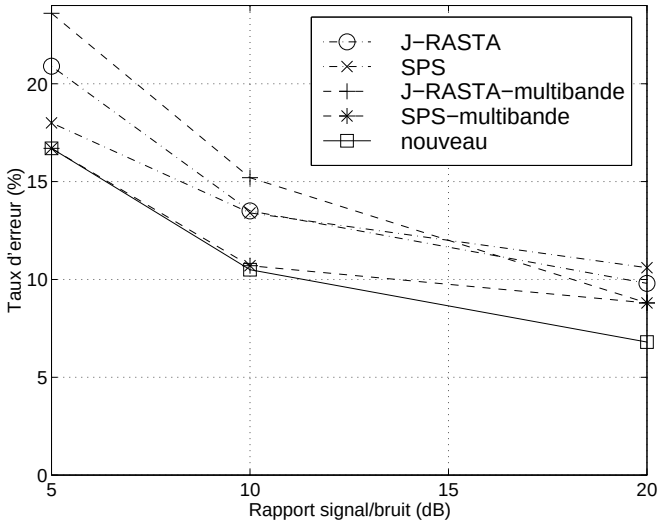


FIG. 8.5 – *Taux d'erreur au niveau du mot: bruit d'hélicoptère* (NOISEX).

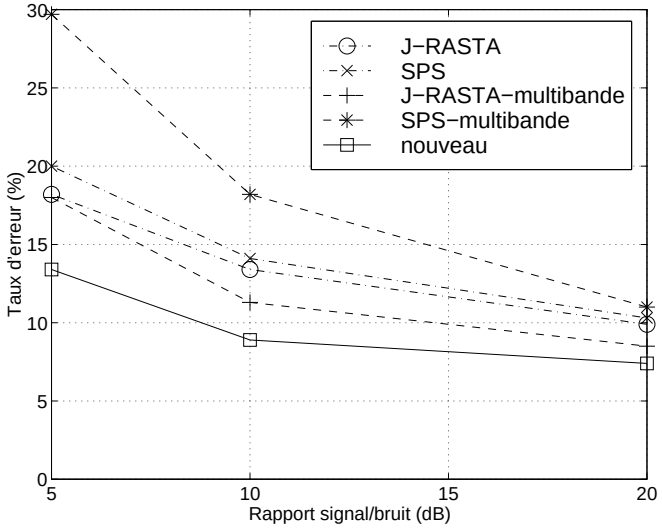


FIG. 8.6 – *Taux d'erreur au niveau du mot: bruit le long d'une chaussée* (MADRAS).

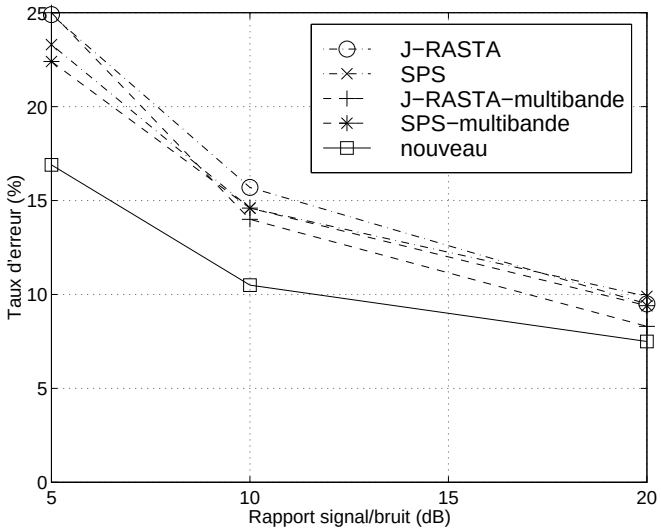


FIG. 8.7 – *Taux d'erreur au niveau du mot: bruit à l'intérieur d'une voiture* (DAIMLER).

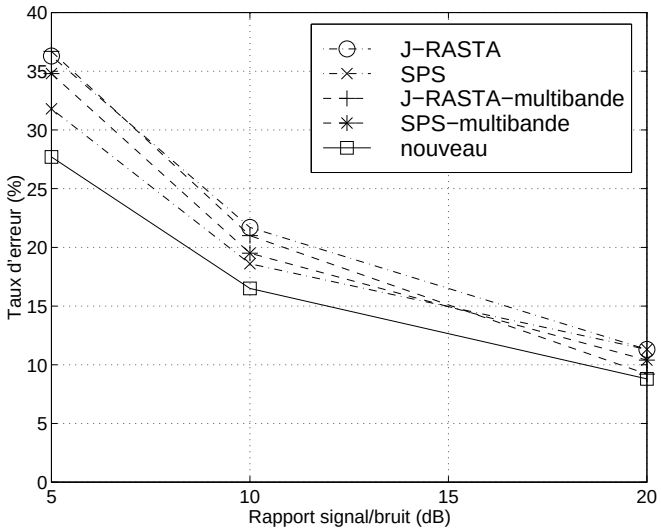


FIG. 8.8 – *Taux d'erreur au niveau du mot: bruit de hall public.*

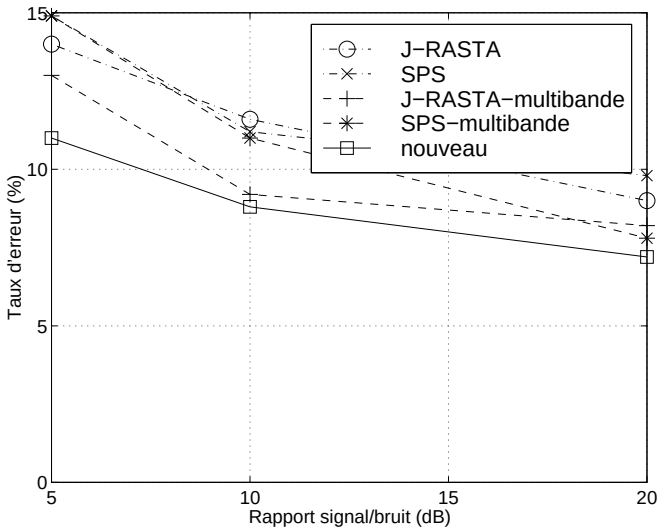


FIG. 8.9 – *Taux d'erreur au niveau du mot: bruit de galerie commerciale.*

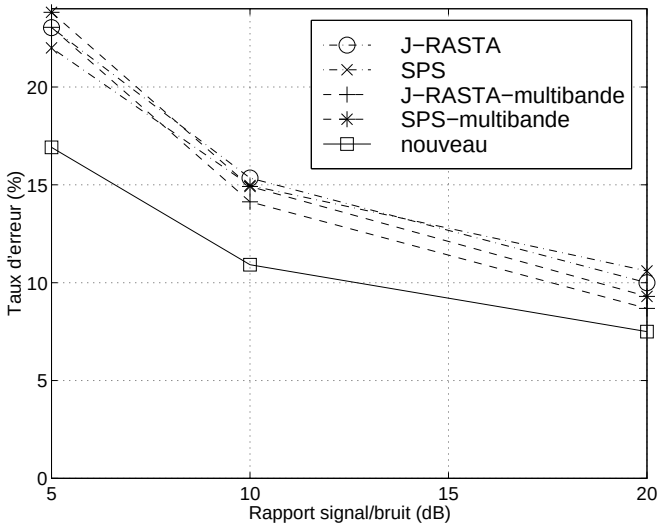


FIG. 8.10 – *Taux d'erreur au niveau du mot: moyenne sur les six types de bruits envisagés (voir table 8.1).*

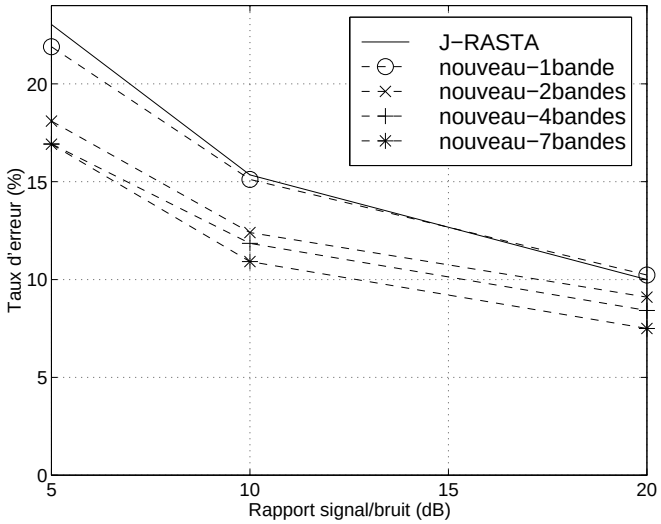


FIG. 8.11 – *Taux d'erreur au niveau du mot: moyenne sur les différents types de bruits envisagés (voir table 8.2).*

Rapport S/B (dB)	5	10	20	moyenne
J-RASTA	23.0	15.4	10.0	16.1%
nouveau-1bande	21.9	15.1	10.2	15.9%
nouveau-2bande	18.1	12.4	9.1	13.2%
nouveau-4bande	16.9	11.9	8.4	12.4%
nouveau-7bande	16.9	10.9	7.5	11.8%

TAB. 8.2 – *Taux d'erreur au niveau du mot: moyenne sur les six types de bruits envisagés.*

Les résultats obtenus suggèrent que les systèmes multi-bandes classiques (et basés sur la soustraction spectrale ou le filtrage J-RASTA) sont globalement meilleurs que les systèmes de référence basés sur la soustraction spectrale ou le filtrage J-RASTA. L'amélioration concerne surtout les faibles niveaux de bruit et reste relativement faible en moyenne.

La nouvelle approche multi-bande conduit quant à elle aux meilleurs résultats, quel que soit le type et le niveau de bruit. Par rapport au système de référence utilisant la soustraction spectrale, on observe globalement une réduction du taux d'erreur proche de 30%. Nos résultats vont en faveur d'une augmentation du nombre de bandes de fréquence. Nous nous sommes cependant limités à 7 bandes. Rappelons que la différence fondamentale entre cette nouvelle approche et les systèmes multi-bandes classiques réside dans l'entraînement sur base d'un corpus de parole contaminé par du bruit blanc.

Rapport S/B (dB)	5	10	20
100 ms	21.2	14.7	9.5
100 ms (NUMBERS'93)	23.6	16.3	12.7
adaptatif	24.0	17.9	11.7

TAB. 8.3 – *Taux d'erreur au niveau du mot (%) pour une tâche de reconnaissance de nombres connectés. Comparaison entre deux approches d'estimation du niveau de bruit dans le cadre de la soustraction spectrale (soit les 100 ms initiales de chaque phrase, soit une méthode adaptative). Influence d'un bruit additif: bruit blanc gaussien à différents niveaux. Nous rappellons également les résultats obtenus avec un système développé sur base du corpus* NUMBERS'93*, de plus petite taille que le corpus* NUMBERS'95.

Rapport S/B (dB)	5	10	20
100 ms	28.5	19.1	11.3
adaptatif	20.0	14.1	10.3

TAB. 8.4 – *Taux d'erreur au niveau du mot (%) pour une tâche de reconnaissance de nombres connectés. Comparaison entre deux approches d'estimation du niveau de bruit dans le cadre de la soustraction spectrale. Influence d'un bruit additif: bruit le long d'une chaussée (corpus* MADRAS*)*

8.3.4 Tâches de reconnaissance complexes

Nous avons ensuite travaillé sur le corpus RESOURCE MANAGEMENT qui correspond à une tâche de reconnaissance de parole continue, indépendante du locuteur, avec un vocabulaire de près de 1000 mots. L'architecture des systèmes envisagés ainsi que les méthodes de calcul des paramètres caractéristiques sont tout à fait identiques à celles envisagées dans les expériences de reconnaissance de nombres connectés décrites ci-dessus. Comme précédemment, les tests effectués correspondent donc à l'approche de reconnaissance hybride HMM/ANN basée sur des perceptrons multicouches analysant neuf trames acoustiques consécutives. Les MLP de référence comprennent une seule couche cachée de 1000 noeuds. Les sous-reconnaisseurs associés à chaque bande comprennent deux couches cachées: une première de 500 noeuds et une seconde de 30 noeuds. Les sorties de la deuxième couche cachée seront utilisées comme paramètres discriminants (approche d'analyse discriminante non linéaire).

Nous avons utilisé la définition officielle des ensembles d'entraînement et de test (voir Annexe A). Le dictionnaire utilisé comprend une seule transcription phonétique par mot et a été développé au laboratoire: ces transcriptions phonétiques sont basées sur 68 classes phonétiques (silence inclu) issues de TIMIT. Pour chaque phonème, la chaine de Markov est construite de façon à imposer une durée minimale égale à la moitié de la durée moyenne du phonème. Nous avons en outre utilisé la grammaire en "paire de mots" de perplexité 60 fournie avec le corpus.

L'étiquetage phonétique des données d'entraînement est obtenu une fois pour toutes sur base du développement complet d'un système de référence utilisant des paramètres PLP. La méthode utilisée consiste à segmenter chaque phrase d'entraînement en autant de portions (de durées identiques) qu'il y a de phonèmes dans la phrase. Cette première segmentation permet de développer un modèle statistique (en l'occurence un MLP) initial. L'entraînement est alors poursuivi de façon classique.

Les méthodes suivantes ont été comparées:

- **PLP**: système de référence utilisant des paramètres PLP.
- **PLP-Log-RASTA**: système de référence utilisant des paramètres PLP et un pré-traitement log-RASTA.
- **PLP-J-RASTA**: système de référence utilisant des paramètres PLP obtenus après un pré-traitement J-RASTA ($J = 1 * 10^{-7}$ conduit ici aux meilleurs résultats). Comme toujours, nous utilisons également les dérivées premières et secondes de ces paramètres ainsi que les dérivées première et seconde de l'énergie de la trame.
- **CBE-SPS**: système de référence utilisant un pré-traitement par soustraction spectrale. Ici, nous avons directement utilisé les énergies des bandes critiques (avec compression par racine cubique): ce type de paramètres donne en effet de meilleurs résultats que les PLP dans le cadre de la soustraction spectrale (voir table 3.1). Le bruit étant relativement stationnaire, nous estimons son niveau sur base des 100 ms initiales de chaque phrase. Pour ce type de bruit, une estimation adaptative conduit d'ailleurs à de moins bons résultats en termes de reconnaissance vocale.
- **Référence B.L.**: il s'agit des résultats du système de référence présenté dans la thèse de Beth Logan [121]. Ce système est basé sur une modélisation multi-gaussienne de triphones. Chaque triphone comprend trois états. Chaque état est modélisé par une distribution multi-gausienne à 5 composantes. Les paramètres caractéristiques utilisés sont des MFCC.
- **nouveau**: il s'agit d'une architecture à sept bandes de frequence utilisant la méthode décrite dans ce chapitre. La fréquence d'échantillonnage étant ici de 16 kHz, ces bandes de fréquence sont différentes de celles utilisées à la section précédente. Cependant, elles suivent toujours une échelle non-linéaire en Bark. La combinaison des différentes bandes est réalisée grâce à un perceptron multicouche utilisant les paramètres obtenus par analyse discriminante non-linéaire.
- **Débruitage B.L.**: il s'agit des résultats du système robuste présenté dans la thèse de Beth Logan [121]. Le système de reconnaissance est identique au système **Référence B.L.**. Les paramètres utilisés sont des MFCC calculés sur base d'une version débruitée du signal. Ce débruitage fait appel à des filtres de Wiener (soustraction spectrale) basés sur des modèles: 512 filtres de Wiener différents sont développés sur base des données d'entraînement. Le choix du filtre de Wiener fait appel à une optimisation par maximum de vraisemblance. Ce type d'approche a déjà été brièvement présenté à la Section 3.4.3.
- **PLP-J-RASTA contaminé**: ce système est du même type que le système

Rapport S/B (dB)	clair	20	12
PLP	6.4	71.3	96.2
PLP-Log-RASTA	7.8	49.5	87.3
PLP-J-RASTA	8.4	28.2	63.0
CBE-SPS	7.5	27.1	63.8
Référence B.L.	6.6	38.9	80.4
nouveau	**6.1**	**10.9**	**35.3**
Débruitage B.L.	6.6	18.1	42.8
PLP-J-RASTA contaminé	8.4	14.1	21.0

TAB. 8.5 – *Taux d'erreur au niveau du mot (%) pour* RESOURCE MANAGEMENT *(test set de février 89). Comparaison entre différentes techniques de reconnaissance robuste pour de la parole bruitée par ajout d'un bruit d'hélicoptère (*NOISEX*) à différents niveaux.*

J-RASTA. Il est cependant entraîné sur base de données contaminées de façon à ce que les conditions d'entraînement et de test soient identiques. Les résultats présentés correspondent donc à trois systèmes différents: un système entraîné sur des données claires, un deuxième entraîné sur des données bruitées à 20 dB et un dernier entraîné sur base de données bruitées à 12 dB. Le bruit considéré étant relativement stationnaire, ces systèmes devraient présenter des performances quasi-optimales.

Pour permettre la comparaison avec les résultats de [121], nous avons appliqué la même procédure d'ajout de bruit: le bruit est d'abord atténué (de 20 dB ou de 12 dB) et ensuite ajouté au signal de parole du corpus de test. La seule différence réside dans la portion de bruit qui est effectivement ajouté: Logan choisi aléatoirement un segment dans l'ensemble du fichier de bruit; nous utilisons l'entièreté du fichier de bruit en boucle de façon à couvrir l'ensemble des données de test. Le bruit étant relativement stationnaire, ces deux approches ne devraient pas conduire à une différence trop importante. Pour vérifier cette hypothèse, nous avons utilisé l'utilitaire NIST "*wavmd*" qui permet l'estimation du rapport singal/bruit. Sur l'ensemble de test *feb89*, Logan obtient des rapports signal/bruit de 17.7 dB et 11.1 dB pour les deux conditions. Nous obtenons 17.8 dB et 11.3 dB.

Le taux d'erreur est calculé de façon classique comme le rapport du nombre d'erreurs (d'insertion, de suppression et de substitution) au nombre de mots de la base de test. Les intervalles de confiance pour des taux d'erreur de 10%, 30% et 50% sont $\pm 1.2\%$, $\pm 1.8\%$ et $\pm 1.9\%$ respectivement ($\alpha = 0.05$). Comme on peut le constater à la table 8.5, l'approche proposée confère un gain en robustesse exceptionnel. Ainsi, pour un rapport signal/bruit de 20 dB, le taux d'erreur est de 10.9% alors qu'il vaut 18.1% avec la méthode proposée dans [121], et 27.1% avec une de nos méthodes de référence robustes. A ce niveau de bruit, notre système donne même de meilleurs résultats que la contamination des données d'entraînement. Ceci est vraisemblablement dû au fait que le bruit n'est pas parfaitement stationnaire: le recouvrement entre les différentes classes phonétiques est donc relativement important

et le système basé sur la contamination n'est donc pas optimal.

8.4 Améliorations possibles

Les systèmes de reconnaissance vocale présentés ici et dans les chapitres précédents sont conformes à l'état de l'art actuel en termes de technologies utilisées. Qu'il s'agisse des systèmes de références ou bien des systèmes implémentant les approches originales proposées dans cette thèse, ils pourraient cependant bénéficier de résultats de recherche récents dans des domaines actifs autres que la reconnaissance vocale robuste. En voici quelques-uns en vrac:

- **Dictionnaire contenant des prononciations multiples** [163]: les dictionnaires utilisés en reconnaissance vocale comprennent généralement une seule transcription phonétique par mot. Les modèles de Markov cachés des mots sont alors construits par concaténation des modèles de Markov cachés des phonèmes. Les prononciations des mots peuvent cependant présenter des variantes liées à l'origine géographique, à l'accent, aux liaisons entre mots[6] ou au phénomène d'élision de phonèmes pouvant apparaître en parole continue. Une solution à ce problème consiste à utiliser plusieurs prononciations différentes pour certains mots.
- **Combiner les probabilités fournies par plusieurs systèmes différents**: Dans [143], l'auteur propose de combiner les probabilités phonétiques fournies par un estimateur classique et par un estimateur multi-bande. La combinaison fait appel à l'hypothèse d'indépendance. Pour une tâche de reconnaissance de nombres connectés, les taux d'erreur au niveau du mot du système classique et du système multi-bande sont respectivement de 6.6% et de 6.3%. Le taux d'erreur rapporté pour le système combiné est de 4.7%. Cette approche de combinaison tombe dans le cadre des méthodes d'ensemble qui, comme nous l'avons déjà signalé au Chapitre 6, permettent d'envisager l'utilisation de classificateurs d'architectures ou de types différents, éventuellement spécialisés sur des conditions ou sur des régions de l'espace différentes ou finalement travaillant sur des espaces d'entrée différents.
- **Boosting [34, 39]**: Il s'agit ici encore d'une méthode d'ensemble consistant à éclater les données d'entraînement en plusieurs parties pour entraîner plusieurs estimateurs qui seront combinés lors de la phase de reconnaissance. Différentes méthodes d'éclatement sont envisageables. Elles font généralement appel à une classification des exemples d'entaînement et à l'utilisation des exemples incorrectement classifiés pour développer un autre estimateur. Dans [39], l'auteur montre une réduction significative du taux d'erreur dans le cadre de la reconnaissance de mots isolés. Pour un vocabulaire de 600 mots, le taux d'erreur passe de 7.0% pour un système basé sur un seul classificateur (en l'occurence un perceptron multicouche) à 5.8% par simple combinaison linéaire des sorties de 3 classificateurs entraînés par la procédure d'éclatement.

6. En français par exemple, certains types de liaisons ne sont pas obligatoires et la réalisation de la fin du mot peut donc présenter deux variantes.

- **Utiliser une méthode d'adaptation au locuteur [112, 46]**: pour un dispositif d'acquisition donné, une des sources de variabilité les plus importantes dans la réalisation des sons de parole provient des différences entre locuteurs. Un domaine de recherche très actif pour le moment concerne donc l'adaptation des modèles de reconnaissance à la voix du locuteur. Cette adaptation peut-être supervisée: le système propose à la personne de lire un texte. Si cette méthode est envisageable dans le cas de systèmes destinés à la bureautique par exemple, d'autres applications, comme les serveurs vocaux ou les bornes d'information interactives peuvent nécessiter l'utilisation de méthodes non supervisées. Pour ce type d'applications, l'adaptation doit également pouvoir apporter un gain significatif sur base de très peu de données d'adaptation: quelques phrases, voire quelques mots.
- **Problème des voix féminines**: dans [94], l'auteur constate que les performances de reconnaissance, sur base de corpus enregistrés dans l'habitacle de voitures, sont inférieures pour les voix féminines que pour les voix masculines. Ni le décalage des formants (dont les fréquences sont de 10 à 15% supérieures pour les femmes que pour les hommes), ni le niveau de la voix ne semblent expliquer ce phénomène. L'auteur rapporte par contre que le taux d'erreur est corrélé avec la fréquence fondamentale des sons voisés. L'explication donnée est la suivante: les harmoniques des voix féminines peuvent être séparés par une distance plus importante que la largeur des filtres intervenant dans le banc de filtres (échelle Mel) utilisé pour le calcul de paramètres MFCC (ça serait également le cas pour un banc de filtres Bark comme celui utilisé pour le calcul de paramètres PLP et dans les systèmes proposés ici). Cela conduit donc à des canaux fréquentiels ne contenant que du bruit et également à une variance inter-locuteur accrue pour ces canaux. Ce problème peut être résolu en adaptant quelque peu le banc de filtres intervenant dans la méthode d'analyse.
- **Modèles de durée**: dans un modèle de Markov caché classique, les probabilités de transition implémentent un modèle de durée très simple (à distribution exponentielle négative). Un état pour lequel la probabilité de boucle est élevée correspond à un phonème dont la durée a une moyenne plus élevée. Il a cependant été montré qu'en pratique, ces probabilités de transition n'apportent aucune amélioration significative des performances. Cela est vraisemblablement dû au fait que la distribution de durée implémentée est très éloignée de la distribution réelle, qui resemble plus à une distribution log-normale. L'utilisation de modèles de durée plus précis peut cependant améliorer les performances dans le cas de parole bruitée, comme montré dans [212].
- **Modèles dépendant du contexte [163, 40]**: le phénomène de coarticulation est une source de variabilité correspondant à l'influence des sons voisins dans la réalisation d'un phonème particulier. L'intérêt des modèles phonétiques dépendant du contexte est de modéliser explicitement ce phénomène. Dans [39], l'auteur montre une réduction significative du taux d'erreur dans le cadre de la reconnaissance de mots isolés. Pour un vocabulaire de 600 mots, le taux d'erreur passe de 7.0% pour un système basé sur un classificateur modélisant des phonèmes indépendants du contexte (CI) à 5.0% par utilisation d'un classificateur basé sur un ensemble d'états comprenant des phonèmes CI ainsi que

64 états représentant des classes de transitions entre phonèmes. L'utilisation de diphones, de triphones, ou plus généralement d'allophones est aussi envisageable [163].

- **Entraînement sur un nombre plus important de trames, sans demander de données d'entraînement additionnelles**: il est montré dans [8] qu'une translation de la fenêtre d'analyse de quelques échantillons modifie sensiblement les paramètres représentatifs issus de cette fenêtre. Cette source de variabilité entraîne des performances sous-optimales. Deux solutions à ce problème sont proposées dans l'article. La première consiste à effectuer une moyenne des paramètres représentatifs en utilisant deux ou trois fenêtres decalées de quelques millisecondes. Cette méthode implique une modification du système de reconnaissance mais est assez facile à implémenter. La deuxième consiste à entraîner le système sur base de plusieurs versions des paramètres représentatifs, décalés de quelques millisecondes. De cette façon, on augmente artificiellement la taille de la base d'entraînement. Cette méthode implique seulement de modifier l'entraînement. Sur une tâche de reconnaissance de mots isoles (vocabulaire de 20K mots, base d'entraînement comprenant 7K phrases), cette méthode diminue sensiblement le taux d'erreur: de 21.26% à 19.85%. L'article donne des résultats pour la première méthode proposée mais sur une autre tâche seulement. L'amélioration est moins importante mais la tâche est aussi beaucoup plus facile.

8.5 Conclusions

Les approches de débruitage font généralement appel à des méthodes d'estimation du niveau de bruit. Pour effectuer cette estimation, on a souvent recours à des techniques heuristiques comme celles présentées au Chapitre 4. Ces méthodes impliquent une découpe en bandes de fréquence de façon à obtenir une estimation du spectre de bruit. Pour chaque trame acoustique, le spectre débruité est finalement estimé par soustraction du spectre de bruit au spectre du signal bruité.

Dans ce chapitre, nous proposons plutôt d'estimer de façon optimale un vecteur de paramètres représentatifs d'une version débruitée du spectre vocal. Cette estimation fait appel à des réseaux de neurones artificiels. Ceux-ci sont entraînés à partir de corpus de parole contaminés par du bruit. Ce genre d'approche demande un corpus d'entraînement couvrant correctement toutes les situations pouvant se présenter en pratique, ce qui est quasiment impossible vu la diversité des formes de bruits. Une découpe en bandes de fréquence relativement étroites, comme celle proposée ici, justifie cependant cette approche. Nous supposons que le bruit est quasi-blanc dans chacune des bandes de fréquence. Nous pouvons ainsi justifier l'entraînement de réseaux de neurones artificiels (un par bande de fréquence) sur base d'un corpus perturbé par du bruit blanc à différents niveaux. Ces réseaux de neurones sont entraînés par rétro-propagation de l'erreur de classification et permettent l'estimation de paramètres représentatifs suivant l'approche d'analyse discriminante nonlinéaire (NLDA). Les paramètres NLDA fournis par les différents réseaux sont alors concaténés et utilisés comme paramètres robustes pour la reconnaissance vocale.

Nous avons tout d'abord travaillé sur base des corpus NUMBERS'93 et NUMBERS'95. Les résultats obtenus suggèrent que les systèmes multi-bandes classiques sont globalement meilleurs que les systèmes de référence basés sur la soustraction spectrale ou le filtrage J-RASTA. L'amélioration concerne surtout les faibles niveaux de bruit et reste relativement faible cependant.

La nouvelle approche multi-bande conduit quant à elle aux meilleurs résultats, quel que soit le type et le niveau de bruit. Par rapport au système de référence utilisant la soustraction spectrale, on observe globalement une réduction du taux d'erreur proche de 30%. Nos résultats vont en faveur d'une augmentation du nombre de bandes de fréquence. Nous nous sommes cependant limités à 7 bandes.

Nous avons ensuite travaillé sur le corpus RESOURCE MANAGEMENT. L'approche proposée ici confère un gain en robustesse exceptionnel. Ainsi, pour un rapport signal/bruit de 20 dB par exemple, le taux d'erreur est de 10.9% alors qu'il vaut 18.1% avec la méthode proposée dans [121], et de 27.1% avec une de nos méthodes de référence robuste utilisant la soustraction spectrale.

Ce gain important est obtenu sans avoir recours à une estimation du spectre de bruit. En parole non bruitée finalement, les performances obtenues sont similaires à celles des systèmes de référence.

La stratégie multi-bande envisagée ici implique cependant l'utilisation de plusieurs réseaux de neurones en parallèle. Elle implique donc un nombre plus important de paramètres et entraîne une charge de calcul accrue par rapport aux approches classiques utilisant un seul réseau de neurones. Elle reste malgré tout utilisable en temps réel sur base de microprocesseurs récents.

Chapitre 9

Conclusions

Au cours de ce travail de thèse, nous nous sommes essentiellement intéressés au problème de la reconnaissance de la parole en milieu bruité. Un premier chapitre s'attache à proposer un état de l'art sommaire des technologies actuelles dans le domaine de la reconnaissance de la parole par modèles de Markov cachés et grammaires stochastiques. Le chapitre suivant vise à donner un aperçu relativement complet mais synthétique des différentes méthodes de reconnaissance vocale robustes au bruit. Les méthodes de reconnaissance vocale envisagées ici sont basées sur l'utilisation de bases de données. Les corpus qui ont été utilisés font l'objet de l'Annexe A. Bien que le chapitre concernant la reconnaissance robuste contienne déjà quelques résultats expérimentaux, la majorité du travail personnel est présentée à partir du Chapitre 4. Nous nous intéressons d'abord au problème de l'estimation du niveau de bruit sur base de méthodes ad hoc. Une méthode originale est présentée, comparée à d'autres méthodes, et enfin utilisée dans le cadre du débruitage de la parole. Le chapitre suivant introduit le modèle "multi-stream". Celui-ci permet la coopération de différents modèles de Markov cachés. Ce chapitre propose notamment des algorithmes de décodage originaux. Au Chapitre 6, nous présentons un état de l'art ainsi que nos travaux récents dans le cadre de l'approche de décomposition en bandes de fréquences ou multi-bande. Cette approche consiste à effectuer une analyse et éventuellement une estimation de probabilités phonétiques, dans différentes bandes de fréquences, sur base de méthodes classiques. On combine ensuite les résultats d'analyse ou d'estimation sur base de méthodes d'ensemble. Ce rapport comprend une description des différents types de méthodes d'analyse ainsi que de méthodes d'ensemble qui ont été utilisées dans le cadre du multi-bande. Le chapitre 6 se termine par une importante section décrivant nos résultats expérimentaux. Le chapitre suivant est consacré au problème de la reconnaissance audiovisuelle de la parole. Après un bref état de l'art, l'approche "multi-stream" est appliquée et comparée à d'autres méthodes plus classiques. Finalement, le dernier chapitre est dédié à une nouvelle méthode de reconnaissance robuste. Celle-ci est basée sur une architecture multi-bande. De nombreux résultats expérimentaux sont rapportés à la fin du chapitre. Contrairement à ceux du chapitre consacré au multi-bande, les résultats concernent ici des bruits réels.

9.1 Résumé des résultats

Les travaux réalisés et les résultats obtenus au cours de ce travail sont résumés ici. Nous avons étudié différentes méthodes classiques d'estimation adaptative du spectre de bruit. Celles-ci sont basées sur l'observation que tout segment de parole suffisamment long contient des portions exemptes de parole qui peuvent être utilisées pour mettre à jour l'estimation du spectre de bruit. Sur cette base, nous avons développé une méthode originale qui met à profit les portions de parole voisées et permet donc une réactualisation plus fréquente de l'estimation. Nous avons pu montrer que pour certains bruits fortement non stationnaires, cette méthode conduit à de meilleures estimations qu'une approche basée sur une détection parole/silence idéale. Finalement, toujours dans le cas de bruits non stationnaires, l'estimation obtenue a été utilisée pour un débruitage par soustraction spectrale, conduisant à des résultats comparables à ceux de la technique de débruitage J-RASTA.

Nous avons étudié la possibilité de diviser l'information disponible en plusieurs canaux, chaque canal étant traité indépendamment des autres sur base de son propre modèle de Markov caché. Nous nous sommes proposés d'étudier cette idée dans le cadre de la reconnaissance vocale sur base d'un canal acoustique et d'un canal visuel axé sur le mouvement des lèvres ainsi que dans le cadre de l'approche de décomposition en bandes de fréquence. Notre étude a conduit à la formulation de deux méthodes de coopération de modèles de Markov cachés: une méthode basée sur des modèles parallèles et une méthode basée sur des modèles composites (multi-dimensionnels). Sous certaines conditions, ces deux méthodes sont identiques. Alors que les HMMs peuvent essentiellement être utilisés pour modéliser un seul processus, ou plusieurs processus dépendants, les modèles que nous proposons ici, appelés modèles "multi-stream", pourront être utilisés pour des processus imparfaitement couplés, qui évoluent indépendamment au sein d'unités linguistiques pré-définies. Ces modèles permettent l'indépendance des régimes stationnaires des différents canaux. Ceux-ci peuvent s'aligner sur des portions temporelles différentes et donc se désynchroniser les uns par rapport aux autres, la transition d'un état (régime stationnaire) au suivant ne se produisant pas nécessairement au même instant pour les différents canaux. Nos travaux ont conduit au développement d'un algorithme de décodage optimal pour la méthode des modèles parallèles. Celui-ci est cependant très lourd. Nous proposons donc également un algorithme sous-optimal efficace utilisant une technique de recherche en faisceau. L'implémentation réalisée permet de contrôler la charge de calcul ainsi que l'occupation mémoire. Dans le cadre de l'approche par modèles composites, nous proposons des méthodes permettant de modéliser l'asynchronisme (et donc le couplage) entre les différents canaux d'information.

Dans le cadre de l'approche multi-bande, nos conclusions sont similaires à celles des travaux précédents:

- l'utilisation de cepstres comme paramètres représentatifs de chaque bande de fréquence donne généralement de meilleurs résultats que l'utilisation de paramètres spectraux, en l'occurence, les énergies dans différentes bandes critiques. Nous avons par ailleurs montré que cette amélioration provient vraisemblablement du caractère normalisé (par rapport à l'énergie) des cepstres.

Ainsi, normaliser les énergies des différentes bandes critiques conduit à des performances similaires à celles des cepstres.

- parmi les différentes méthodes d'ensemble qui ont été envisagées, les réseaux de neurones artificiels (RNA) conduisent généralement aux meilleures performances. La supériorité de cette approche provient du fait qu'on ne pose aucune hypothèse forte concernant les résultats fournis par les différentes bandes de fréquences et concernant la façon dont ces résultats peuvent être combinés. Notons que même dans le cas où les conditions d'utilisation sont différentes des conditions d'entraînement, les RNAs restent supérieurs.
- l'approche conduit à une robustesse nettement accrue dans le cas de bruits colorés. Dans le cas de bruits blancs cependant, le gain est négligeable, voir nul.

Notons cependant que ces résultats sont obtenus sur base de techniques d'analyse et de reconnaissance qui ne font appel à aucune stratégie robuste: les systèmes de référence et les systèmes multi-bandes, outre leur architecture particulière, sont tout à fait classiques. Les expériences effectuées ont donc été répétées à partir de paramètres représentatifs robustes obtenus soit par soustraction spectrale, soit par filtrage J-RASTA. Les résultats sont alors beaucoup moins clairs. L'intérêt du multi-bande est moindre, même dans le cas de bruits colorés. Lorsqu'on utilise la soustraction spectrale, l'approche multi-bande conduit même à une légère dégradation des performances lorsque le niveau de bruit est élevé. Les résultats précédents concernent des bruits artificiels (bruit blanc et bruit blanc filtré) stationnaires. Dans le cas de bruits réels, nos résultats ont malgré tout montré que l'approche multi-bande conduit globalement à une légère amélioration par rapport à la technique J-RASTA ou à la soustraction spectrale.

Nous nous sommes aussi intéressés à l'utilisation de modèles "multi-stream" pour la reconnaissance en bandes de fréquence. Ils permettent de relâcher la contrainte de synchronisme entre les différentes bandes. Nos résultats sont peu concluants à ce stade. On constate qu'un relâchement de la contrainte de synchronisme entre différentes bandes de fréquence conduit généralement à une dégradation des résultats. De plus, suite à l'utilisation de modèles parallèles ou de modèles composites, cela implique une charge de calcul élevée et un espace mémoire important.

Nous nous sommes également intéressés à la reconnaissance audiovisuelle de la parole. Des paramètres représentatifs du contour et de la luminosité des lèvres, fruits d'un travail de thèse [124], sont utilisés en plus de l'acoustique. Nous avons proposé différentes méthodes d'intégration des données audiovisuelles sur base de l'approche "multi-stream". La fiabilité des deux modalités peut être introduite sous forme d'un poids adaptatif dans le formalisme d'intégration. Nous avons ainsi montré que le gain dû à l'adaptation du poids de combinaison surpasse l'effet défavorable de l'hypothèse d'indépendance sur laquelle repose notre critère d'intégration.

Toujours dans le cadre de la reconnaissance vocale audiovisuelle, nous avons observé, sur base d'alignements Viterbi audio et video, les délais entre les transitions d'un état à un autre pour une des modalités et les transitions correspondantes pour l'autre modalité. Les statistiques calculées semblent indiquer que ces délais de transition ne sont pas simplement le produit d'un bruit d'alignement mais reflètent également une structure dont la modélisation pourrait être bénéfique à la reconnais-

sance vocale. Les résultats expérimentaux résultant de l'approche "multi-stream" sont cependant très mitigés. La modélisation de l'asynchronisme ne conduit pas toujours à une amélioration en terme de robustesse. Notons cependant que par rapport à un système de reconnaissance acoustique, même si l'on utilise des paramètres acoustiques robustes, le système audiovisuel conduit à une réduction importante du taux d'erreur en présence de bruit additif: il est en effet divisé par trois pour un rapport signal/bruit acoustique de 10 dB.

Finalement, les faiblesses de l'approche multi-bande "standard" ont conduit au développement d'une méthode originale basée sur le même type d'architecture, et qui a fait l'objet d'un dépôt de brevet. Cette méthode est basée sur l'observation que, si l'on considère une bande de fréquence relativement étroite, les bruits large-bandes ne diffèrent essentiellement que par leur niveau. De ce fait, des modèles associés aux bandes de fréquence du système peuvent être entraînés après contamination du corpus d'apprentissage par un bruit quelconque, à différents niveaux; ces modèles demeureront relativement insensibles à d'autres types de bruits. Cette méthode est testée dans le cas de bruits réels, provenant de bases de données de bruit, et ajoutés aux corpus de test. Dans le cadre d'une tâche de reconnaissance de nombres connectés sur ligne téléphonique, nous avons observé une réduction du taux d'erreur proche de 30% par rapport aux techniques de soustraction spectrale ou de filtrage J-RASTA. Dans le cadre du corpus RESOURCE MANAGEMENT, la réduction du taux d'erreur va jusqu'à 40% par rapport à des résultats récents proposés dans la littérature concernant la reconnaissance robuste de la parole. Ce gain important est obtenu sans avoir recours à une estimation du spectre de bruit.

En parole non bruitée, les performances obtenues sont similaires (voire meilleures dans le cas de la reconnaissance de nombres connectés sur ligne téléphonique) à celles des systèmes de référence.

Les systèmes de reconnaissance vocale qui ont été développés ici font appel à la librairie de classes et de programmes STRUT, conçue au sein du laboratoire TCTS. Les nouveaux développements décrits dans ce rapport de thèse s'intègrent dans cet outil déjà très complet.

9.2 Perspectives

Nous avons présenté une approche, appelée "multi-stream", qui permet la modélisation de processus imparfaitement couplés. Les techniques par champs de Markov cachés et par modèles de Markov couplés méritent également toute notre attention. Elles permettent aussi de relâcher la contrainte de synchronisme entre les différentes sources d'information. Dans le cadre de la reconnaissance en bandes de fréquence, l'intérêt pratique de ces méthodes n'est toujours pas démontré. Nos résultats négatifs avec la stratégie "multi-stream" nous suggèrent que la contrainte de synchronisme devrait peut-être être assouplie seulement pour certaines classes de sons, et pas pour tous les sons comme cela a été fait dans ce travail ainsi que par d'autres. A ce propos, les suggestions des Sections 5.3.2 et 5.5 ouvrent certaines perspectives de recherche.

Les techniques "multi-stream" ont également été appliquées à la reconnaissance vocale audio-visuelle. Ici encore, les résultats sont peu concluants malgré les tenta-

tives de modélisation des motifs d'asynchnonisme. L'analyse de statistiques concernant les délais de transition entre les deux modalités suggère cependant l'existence d'une structure: la moyenne de ces délais est nettement positive ou négative pour certaines classes de sons. A notre sens, ces résultats encouragent donc la poursuite de recherches concernant l'utilisation de techniques de modélisation de processus imparfaitement couplés. Insistons également sur la nécessité d'obtenir des corpus audiovisuels de plus grande taille (M2VTS ne contient que 185 séquences de chiffres) et d'étendre les recherches vers des tâches plus complexes.

De manière plus générale, en ce qui concerne la reconnaissance de la parole en milieu bruité, nos travaux ont conduit à deux résultats importants. Nous avons tout d'abord confirmé l'intérêt de sources d'information alternatives: l'utilisation d'une séquence video des lèvres du locuteur conduit à une réduction importante du taux d'erreur. Nous avons ensuite développé une nouvelle méthode robuste de reconnaissance multi-bande sur base d'un seul signal acoustique. Cette méthode surpasse les approches simples de soustraction spectrale et de filtrage J-RASTA ainsi qu'une approche plus complexe de débruitage basé sur des modèles statistiques.

Les techniques faisant intervenir des sources d'information alternatives et les méthodes robustes appliquées à un signal unique n'excluent pas la possibilité de mettre tout en oeuvre pour obtenir le signal le moins bruité possible. Il va de soi que les performances de reconnaissance vocale seront d'autant meilleures que le signal utilisé est moins bruité. On peut par exemple faire appel aux réseaux de microphones, qui permettent de "focaliser" la prise de son dans une direction précise et de façon adaptative. Ces techniques conduisent à un gain en rapport signal/bruit de l'ordre de 6 à 12 dB lorsque le bruit est diffus. Dans le cadre de certaines applications (la borne multimédia interactive par exemple), ces trois familles de techniques, relativement au point, pourraient être combinées afin d'accumuler leurs gains respectifs.

Outre les perspectives de recherche et de développement que peut susciter la lecture de ce rapport de thèse, d'autres aspects nous semblent importants. Insistons sur une classe émergente de méthodes qui semblent promises à un bel avenir dans le cadre de la reconnaissance robuste de la parole: l'adaptation rapide. Les méthodes d'adaptation sont en fait des méthodes d'apprentissage. Elle utilisent des algorithmes d'entraînement/estimation classiques pour adapter directement ou indirectement les paramètres des modèles statistiques ou les paramètres représentatifs du signal. L'objectif de ces techniques est de réduire les différences entre les conditions d'entraînement et les conditions d'utilisation. Elles peuvent donc être envisagées pour l'adaptation au locuteur mais également pour l'adaptation au canal de transmission et aux bruits perturbateurs. La difficulté dans ce cas réside dans le caractère non-stationnaire et imprévisible du bruit. Les recherches actuelles portent donc une attention particulière aux méthodes permettant une adaptation très rapide c'est-à-dire sur base d'une très petite quantité de données d'adaptation. On parle ainsi d'adaptation sur base d'une seule phrase, voire d'un seul mot. L'idée à la base de ces méthodes est souvent d'utiliser une information a priori concernant les propriétés statistiques des paramètres d'adaptation [111, 181, 46], cette information a priori pouvant provenir d'une analyse des données d'entraînement. Le lecteur trouvera de plus amples informations dans [153].

Finalement, nous ne nous sommes pas intéressés à la modélisation du langage.

Nous savons cependant que les performances humaines en termes de perplexité sont jusqu'à 3 fois supérieures à celles des modèles tri-grammes conventionnels. Cet aspect est donc fondamental dans des conditions de bruit et également dans le cas de la parole spontanée, pour laquelle les systèmes actuels présentent encore un taux d'erreur supérieur à 30%.

Annexe A

Bases de données

Ce chapitre présente les différentes bases de données qui ont été utilisées dans le cadre de cette thèse. D'une part, nous avons des corpus de parole permettant le développement et l'évaluation de systèmes. D'autre part, on trouvera des corpus contenant des enregistrements de bruits divers. Ces bruits ont été ajoutés artificiellement aux signaux de parole pour évaluer la robustesse des méthodes développées au bruit additif.

Nous avons également utilisé une base de données audiovisuelle. Ce corpus contient des séquences vidéo du visage du locuteur, synchronisées avec le signal audio.

Toutes ces bases de données concernent la reconnaissance de parole continue et indépendante du locuteur. On constatera rapidement qu'elles correspondent à des tâches de reconnaissance relativement simples. Nous sommes loin des systèmes à grand vocabulaire, ou des tâches basées sur de la parole spontanée qui intéressent certains laboratoires de recherche. Cependant, même ces tâches simples n'ont pas encore trouvé de solution suffisamment efficace dans des situations où le bruit additif est important. Ainsi, la plupart des recherches dans le domaine de la reconnaissance robuste de la parole se focalisent sur ce genre de tâches.

A.1 OGI Numbers'93

Cette base de données contient des séquences de nombres prononcées de façon naturelle sur les lignes téléphoniques du réseau fixe américain [32]. Cette base de données a été collectée par OGI-CLSU. La base de données Numbers'93 est une portion de la base de données Numbers. Elle contient 2167 séquences de nombres prononcées par 1132 locuteurs. Dans nos expériences, 1534 phrases ont été utilisées pour l'entraînement[1] et 384 phrases (1332 mots) pour le test. Cela correspond à la décomposition qui avait été choisie initialement par plusieurs laboratoires américains. Les phrases inutilisées contiennent soit des prononciations incomplètes pour certains nombres, soit des nombres ordinaux. L'utilisation de cette décomposition permet la comparaison des résultats de différents laboratoires.

1. Dans le cas de l'entraînement de réseaux de neurones artificiels, 1400 phrases ont été utilisées pour l'estimation des poids et 134 pour la validation croisée

A.2 OGI Numbers'95

Il s'agit d'une extension de la base de données précédente. La version distribuée contient pour le moment 15000 séquences de nombres. Pour nos expériences, 3590 séquences ont été utilisées pour l'entraînement et 1227 (4753 mots) pour le test. Ces ensembles correspondent à ceux choisis par d'autres institutions [144].

Les deux bases de données précédentes permettent le développement de systèmes de reconnaissance de nombres connectés indépendants du locuteur, sur ligne téléphonique et en anglais américain.

A.3 DARPA Resource Management (RM)

Il s'agit d'une base de données de parole continue à vocabulaire moyen en anglais américain. Elle a été développée par DARPA [162] et contient des phrases basées sur un vocabulaire lié au problème de la gestion de ressources dans le domaine naval. L'enregistrement est de très bonne qualité et a été effectué à une fréquence d'échantillonage de 16 kHz. Le vocabulaire utilisé est de 992 mots. Nous avons uniquement utilisé la partie destinée à la reconnaissance indépendante du locuteur (RM1). La décomposition officielle entre données d'entraînement et données de tests est la suivante:

- ensemble d'entraînement: 3990 phrases (sur bases de 100 locuteurs prononçant chacun 40 phrases).
- ensemble de test: il existe quatre ensembles de tests comportant chacun 300 phrases (10 locuteurs prononçant chacun 30 phrases). Ces quatres ensembles (feb89 -2561 mots-, oct89 -2684 mots-, feb91 -2484 mots- et sep92 -2559 mots-) correspondent à différentes campagnes d'enregistrement au cours de plusieurs années.

A partir de l'ensemble d'entraînement, 3591 phrases ont été utilisées pour l'entraînement des réseaux de neurones et 999 pour la validation croisée.

Une grammaire de type 'paire de mots' de perplexité 60 est fournie avec le corpus. Elle a été utilisée pour nos expériences.

Cette base de données permet le développement de systèmes de reconnaissance indépendants du locuteur, en anglais américain, à partir d'un signal de qualité et pour un vocabulaire et une grammaire bien définis. Elle n'a donc pas grand intérêt pratique. Elle a cependant été très utilisée pour comparer différentes méthodes. Les algorithmes ayant progressé, cette base de données est devenue trop facile et très peu utilisée actuellement. Dans le cas de parole bruitée cependant, ce corpus a toujours son intérêt.

A.4 M2VTS

La base de données M2VTS [160] (projet européen M2VTS - "Multi Modal Verification for Teleservices and Security applications") a été utilisée pour nos expériences de reconnaissance audiovisuelle de la parole. Cette base de données

contient 185 enregistrements de 37 sujets (12 femmes et 25 hommes). Chaque enregistrement contient les signaux audio et vidéo de la séquence de chiffre (de '0' à '9') en français. Pour chaque locuteur, on dispose de cinq enregistrements effectués chacun à une semaine d'intervalle. Les séquences vidéo sont composées d'images en couleurs de 286*360 pixels, à une fréquence de rafraîchissement de 25 Hz. Le signal audio est enregistré à la fréquence de 48 kHz. La base de données contient plus de 27,000 images couleurs qui ont été converties en images avec 256 niveaux de gris.

Bien que ce corpus soit un des plus volumineux de sa catégorie, il est relativement petit en comparaison aux bases de données audio utilisées dans le domaine de la reconnaissance vocale. Pour accroître le niveau de signification de nos expériences, nous avons utilisé une approche "jack-knife": cinq répartitions différentes du corpus ont été définies, chacune étant constituée de:

- 3 prononciations des 37 locuteurs comme données d'entraînement.
- 1 prononciation des 37 locuteurs comme données de développement. Ces prononciations sont utilisées pour optimiser les coefficients de pondération permettant la combinaison des décisions audio et vidéo.
- 1 prononciation des 37 locuteurs comme ensemble de test.

Cette procédure permet d'utiliser l'ensemble du corpus (185 phrases, soit 1850 mots) pour la phase de test, en développant cinq systèmes de reconnaissance indépendants pour chacune des approches envisagées. Ces systèmes pourraient être qualifiés de multi-locuteurs: il ne sont en effet pas indépendants des locuteurs. Notons pour terminer que la séquence de chiffres à reconnaître est toujours la même (chiffres de '0' à '9').

A.5 Madras

La base de données de bruits environnementaux MADRAS est un produit du projet européen MADRAS [197] ("Methods for Automatic Detection and Recognition of Acoustic Sources"). Le corpus contient des enregistrements de haute qualité de bruits provenant de différentes sources: voitures, trains, camions, avions, usines...

Dans nos expériences, un seul bruit a été utilisé. Il s'agit d'un bruit enregistré le long d'une chaussée et caractérisé par le passage de deux voitures. Ce bruit a été choisi pour son caractère non-stationnaire, contrairement aux autres bruits de voiture dont nous disposions qui correspondent à des enregistrements effectués à l'intérieur de l'habitacle.

A.6 Noisex-92

La base de données NOISEX-92 [152] contient des enregistrements de haute qualité de chiffres isolés ainsi que de séquences de trois chiffres. Elle contient également des enregistrements de bruits environnementaux divers ainsi que des versions bruitées des données de parole.

Dans nos expériences, nous avons uniquement utilisé le bruit d'hélicoptère Lynx (bruit 12).

A.7 Autres bruits

Additionnellement, nous avons utilisé trois autres bruits:

- un bruit enregistré à l'intérieur de l'habitacle d'une voiture roulant à 80 km/h. Cet enregistrement nous a été fourni par la société Daimler dans le cadre du projet européen RESPITE.
- un bruit enregistré dans une galerie commerciale, fourni par la société BABEL TECHNOLOGIES.
- un bruit de hall public simulé constitué d'un bruit de fond stationnaire (bruit de parole dans la base de données Noisex-92) et de 15 personnes parlant en même temps (phrases issues de la base de données Resource Management), le tout fortement réverbéré.

A.8 Types de bruits

Comme indiqué aux sections précédentes, divers types de bruits on été utilisés pour nos expérimentations. En voici un résumé:

- un bruit blanc gaussien stationnaire, peu réaliste mais souvent utilisé dans ce genre de recherches.
- le bruit d'hélicoptère de Noisex car il contient des composantes fortement colorées superposées à un bruit large bande quasi-stationnaire.
- un bruit de passage de voiture (Madras) pour son caractère fortement non stationnaire.
- un bruit provenant de l'habitacle d'une voiture (fourni par Daimler) car l'application semble importante.
- un bruit de galerie commerciale (fourni par Babel Technologies) car l'application de borne interactive est importante.
- un bruit de hall public simulé.

Une analyse spectrale des bruits utilisés peut être trouvée en Annexe B.

Annexe B

Analyse de quelques types de bruits

Les figures qui suivent présentent quelques signaux de bruit qui ont été utilisés dans nos expériences. Elles illustrent la diversité des structures spectro-temporelles des bruits réels.

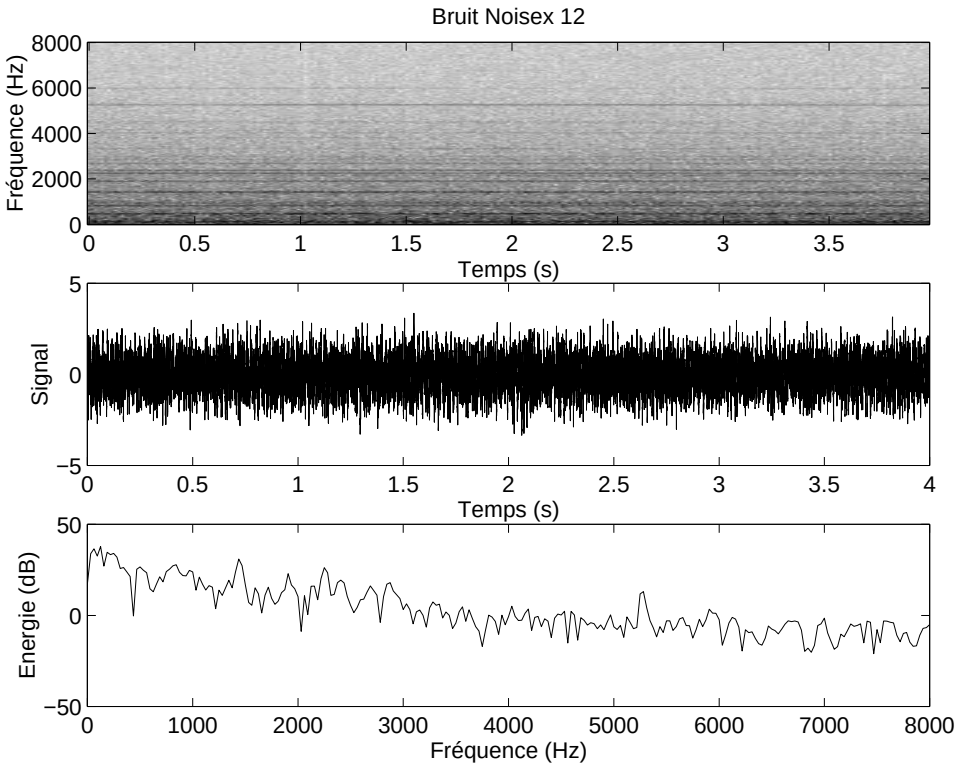


FIG. B.1 – *Spectrogramme, signal temporel et spectre pour 4 secondes d'un bruit d'helicoptère Lynx (base de données* NOISEX, *bruit numéro 12).*

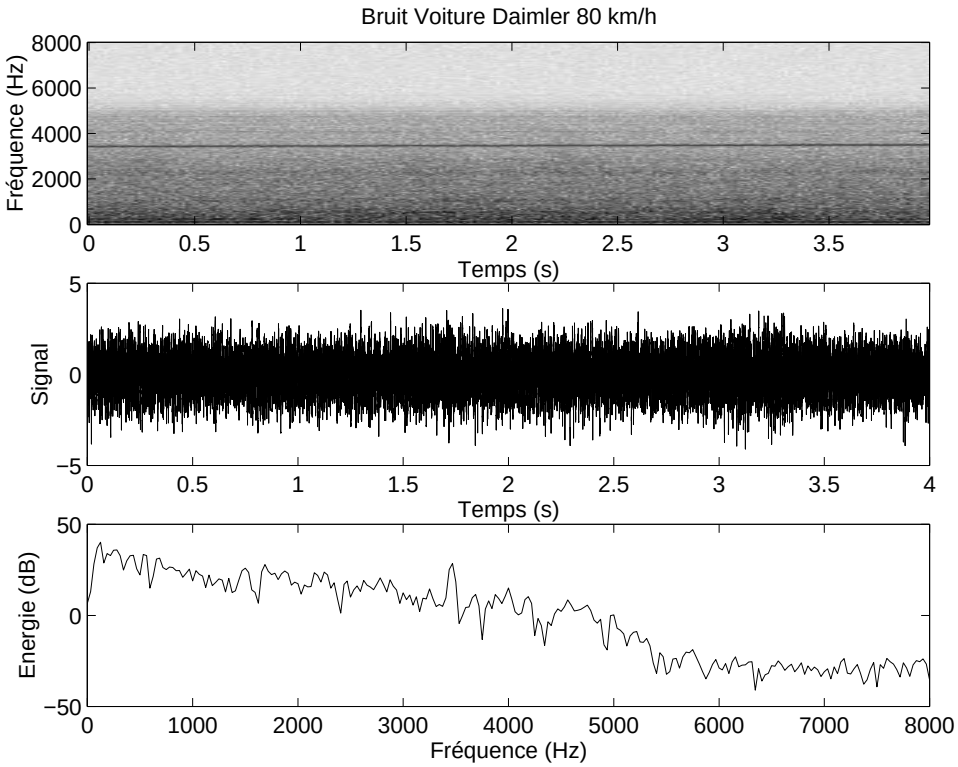


FIG. B.2 – *Spectrogramme, signal temporel et spectre pour 4 secondes d'un bruit dans l'habitacle d'une voiture roulant à 80 km/h (Daimler-Benz).*

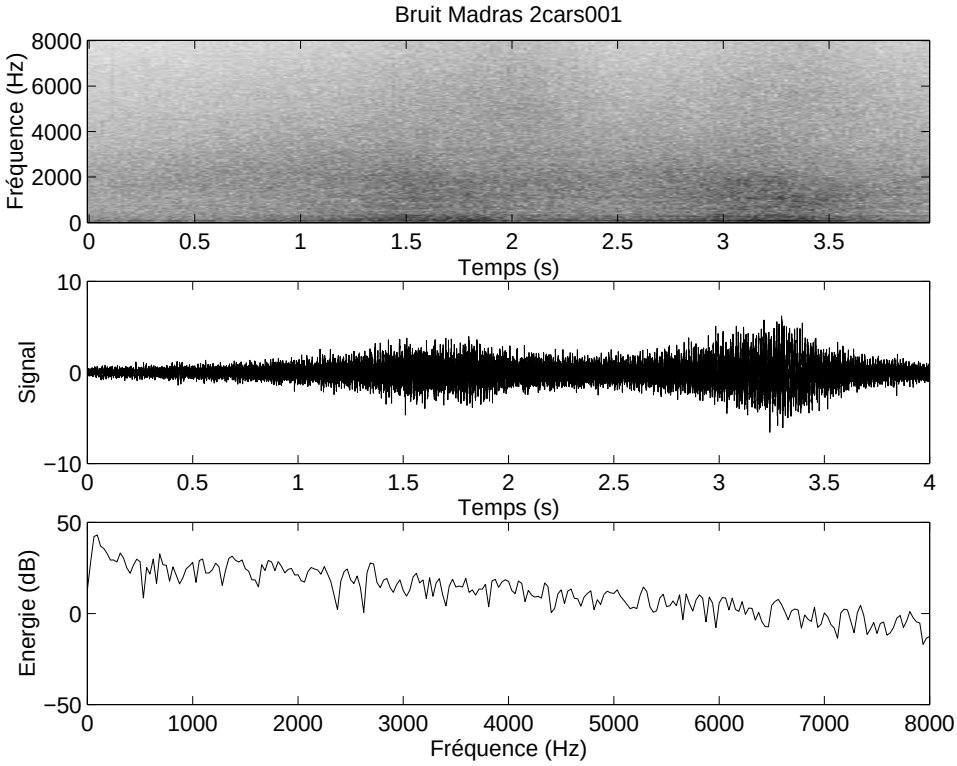


FIG. B.3 – *Spectrogramme, signal temporel et spectre pour 4 secondes d'un bruit créé par le passage de deux voitures sur une chaussée (base de données* MADRAS, *bruit '2cars001').*

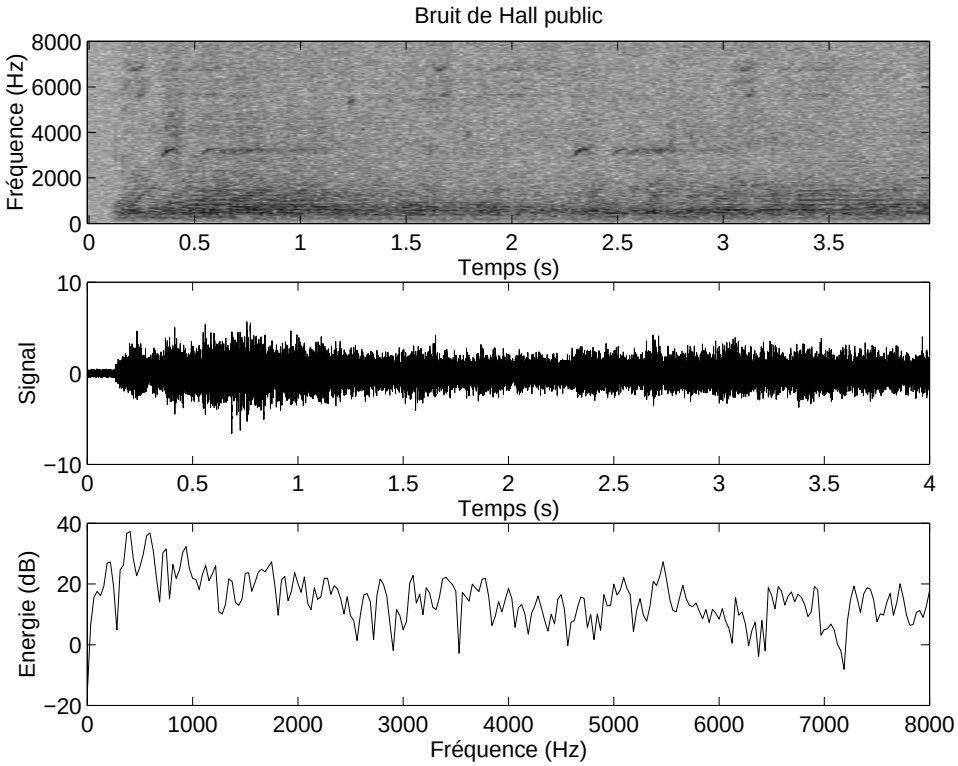


FIG. B.4 – *Spectrogramme, signal temporel et spectre pour 4 secondes d'un bruit de hall public.*

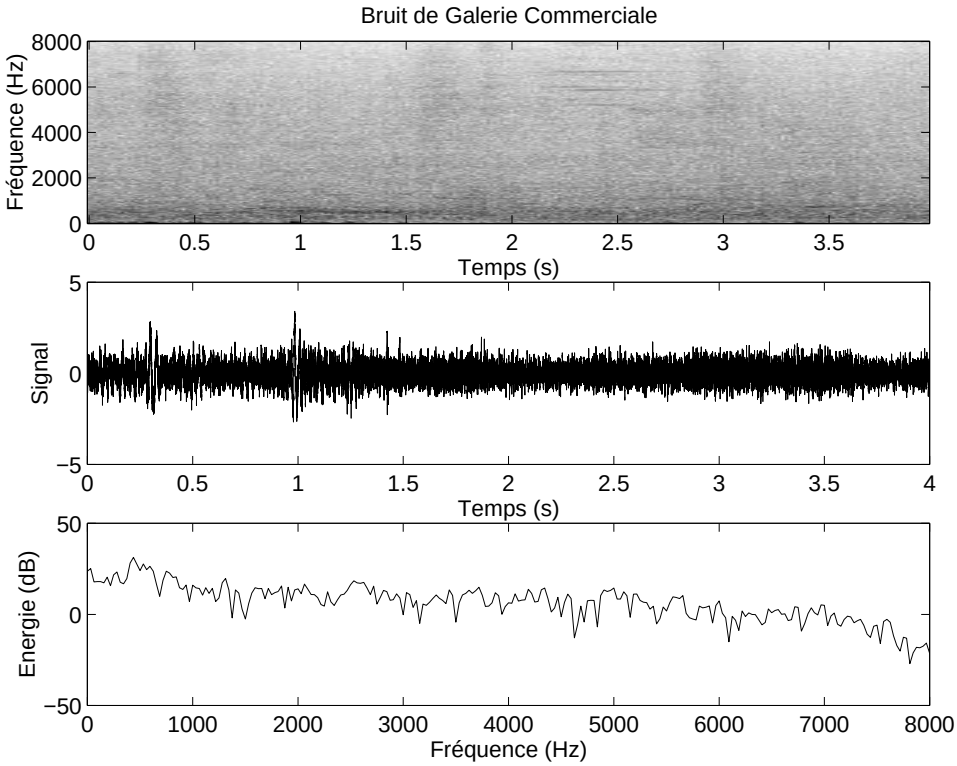


FIG. B.5 – *Spectrogramme, signal temporel et spectre pour 4 secondes d'un bruit de galerie commerciale.*

Annexe C

Résultats pour l'approche multi-bande

Voir le chapitre 6.

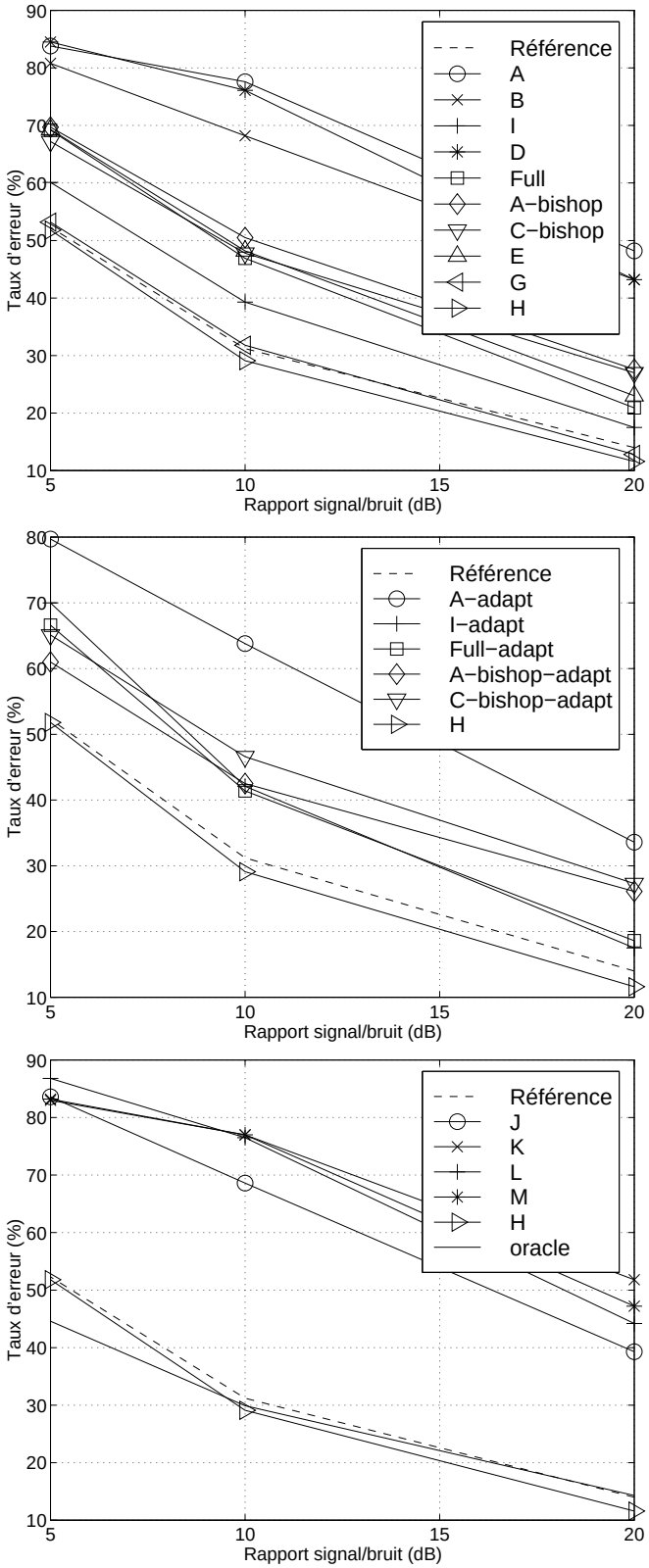


FIG. C.1 – *Taux d'erreur au niveau du mot: bruit blanc, filtrage log-RASTA ("Relative Spectra").*

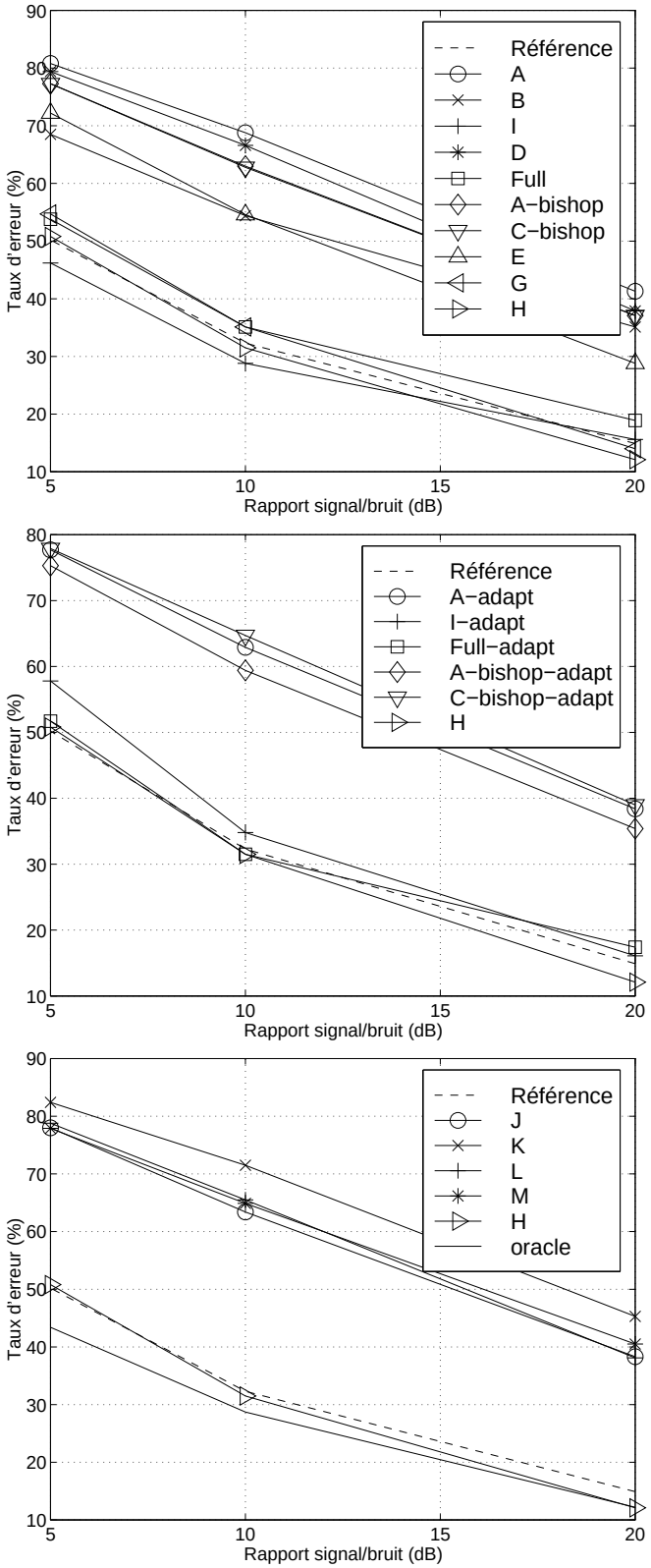


FIG. C.2 – *Taux d'erreur au niveau du mot: bruit coloré, filtrage log-RASTA.*

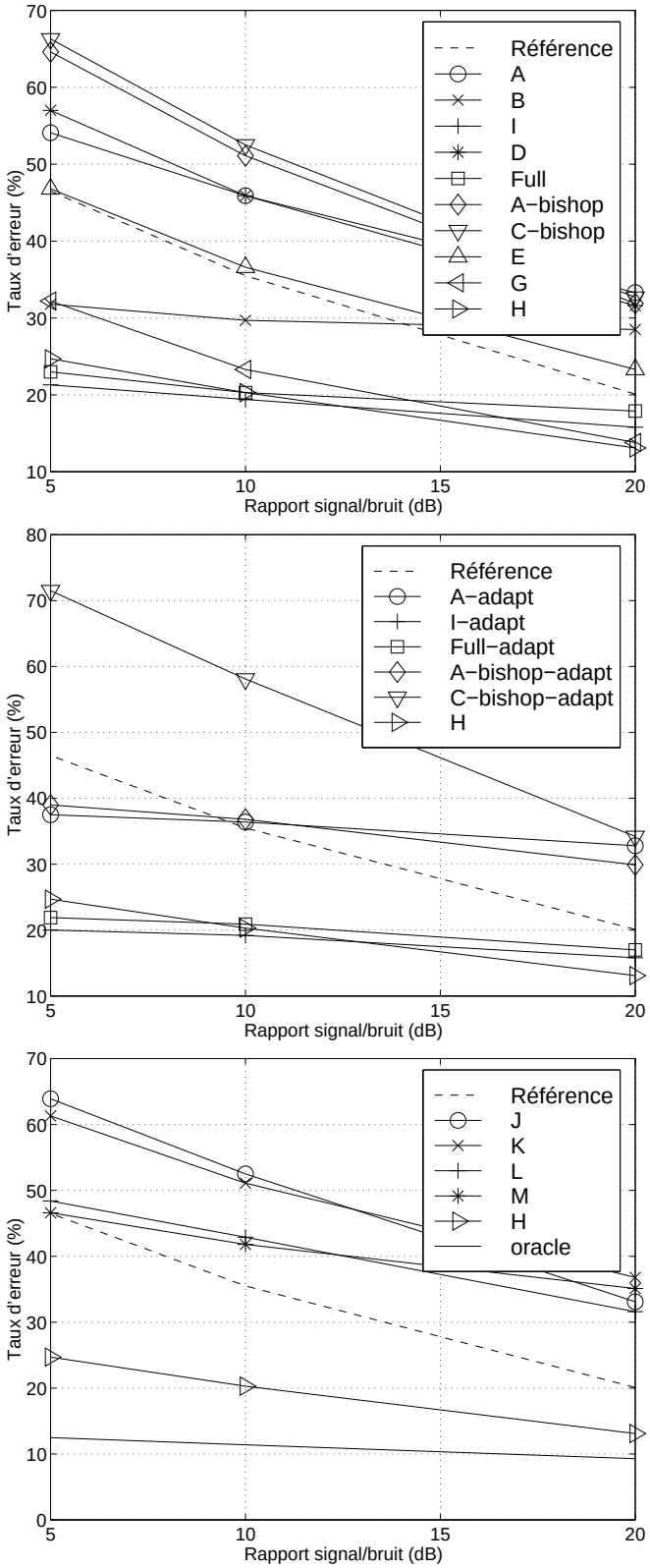


FIG. C.3 – *Taux d'erreur au niveau du mot: bruit fortement coloré, filtrage log-RASTA.*

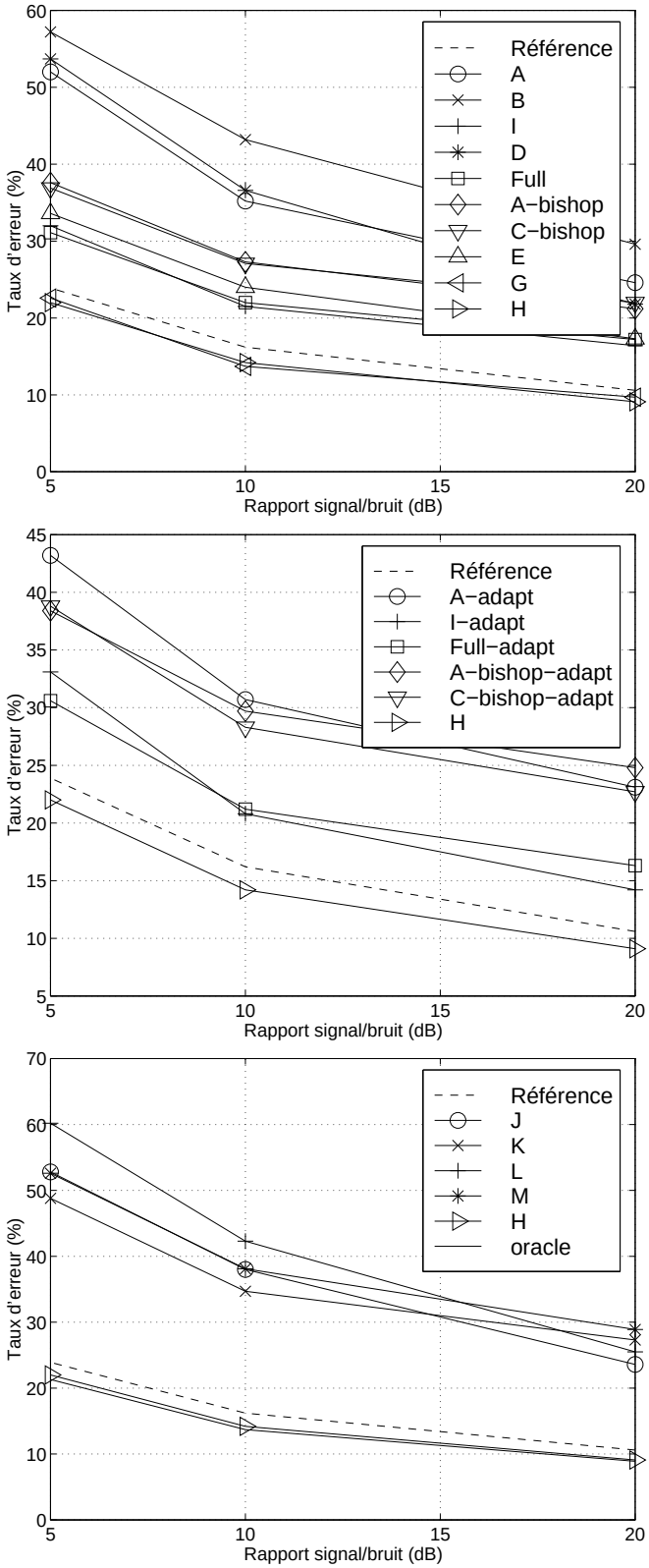


FIG. C.4 – *Taux d'erreur au niveau du mot: bruit blanc, filtrage J-RASTA ("Relative Spectra").*

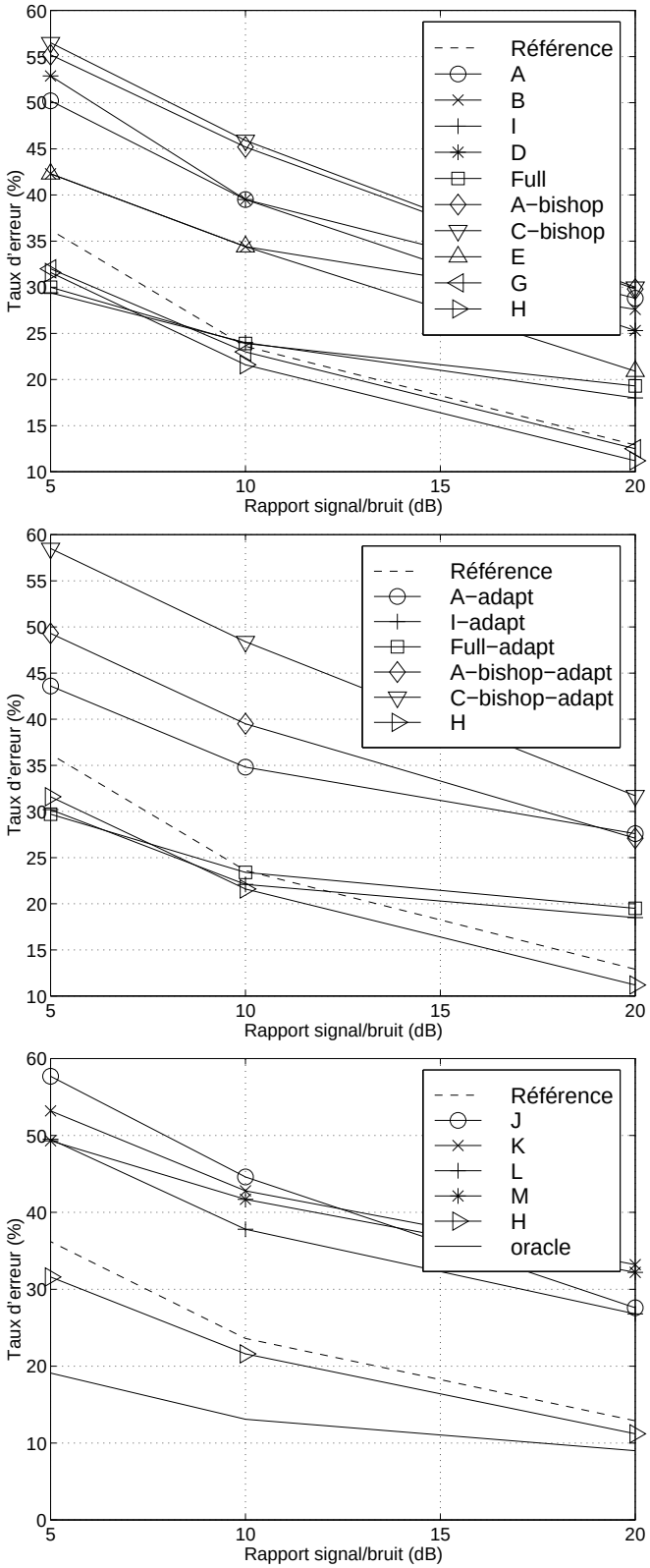


FIG. C.5 – *Taux d'erreur au niveau du mot: bruit coloré, filtrage J-RASTA.*

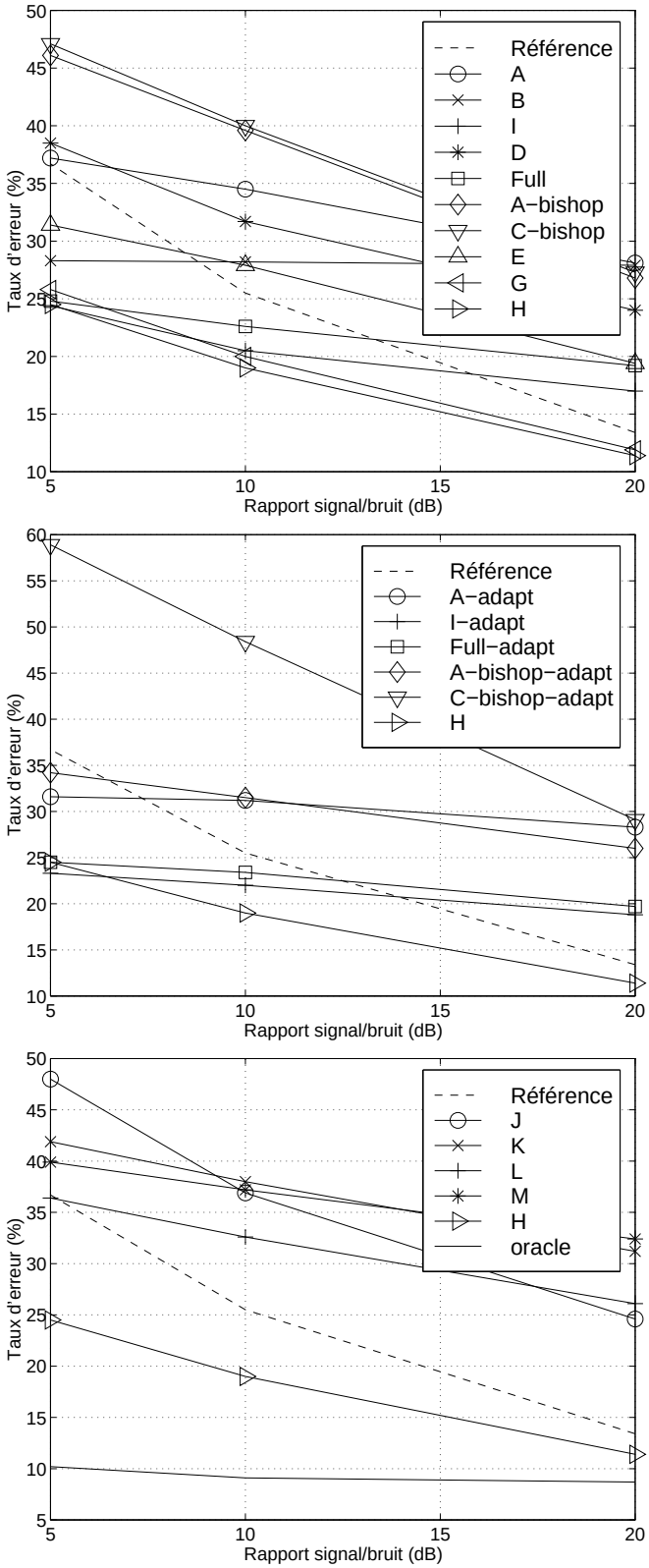


FIG. C.6 – *Taux d'erreur au niveau du mot: bruit fortement coloré, filtrage J-RASTA.*

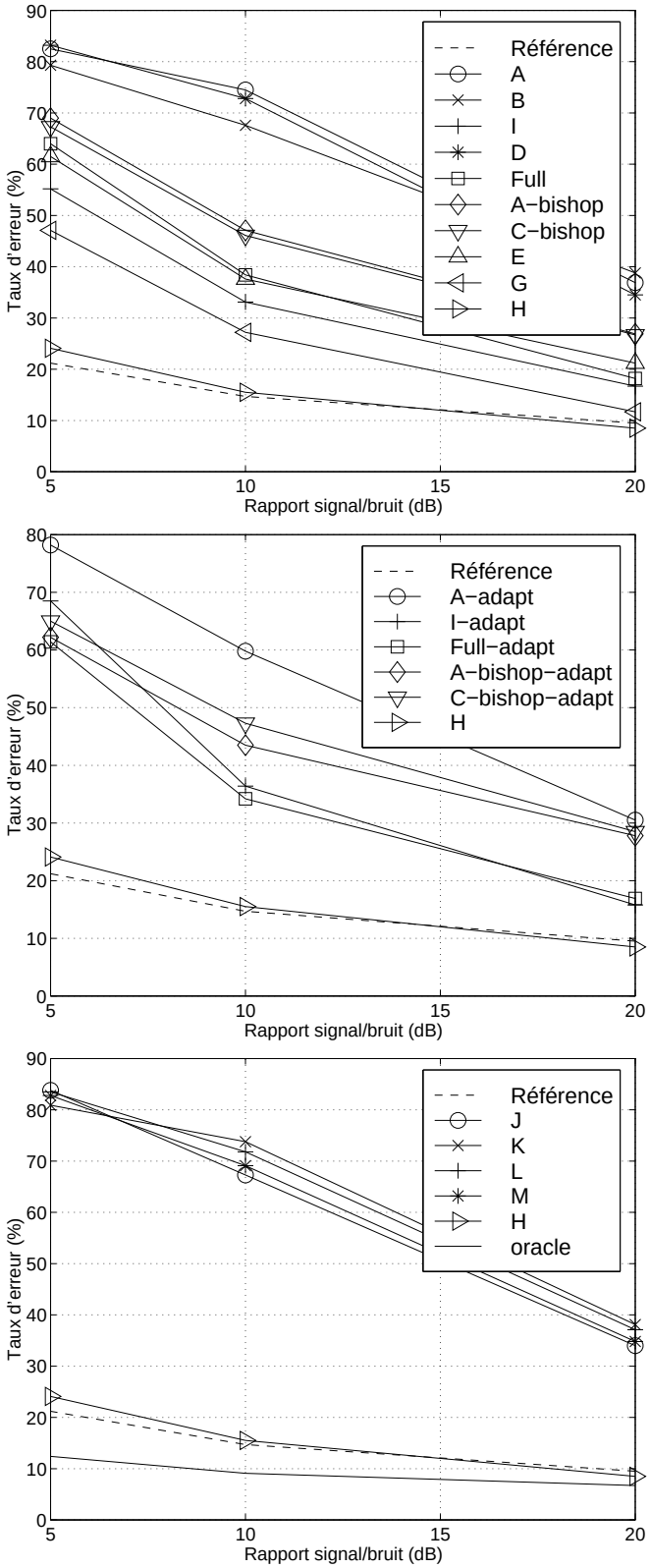


FIG. C.7 – *Taux d'erreur au niveau du mot: bruit blanc, soustraction spectrale.*

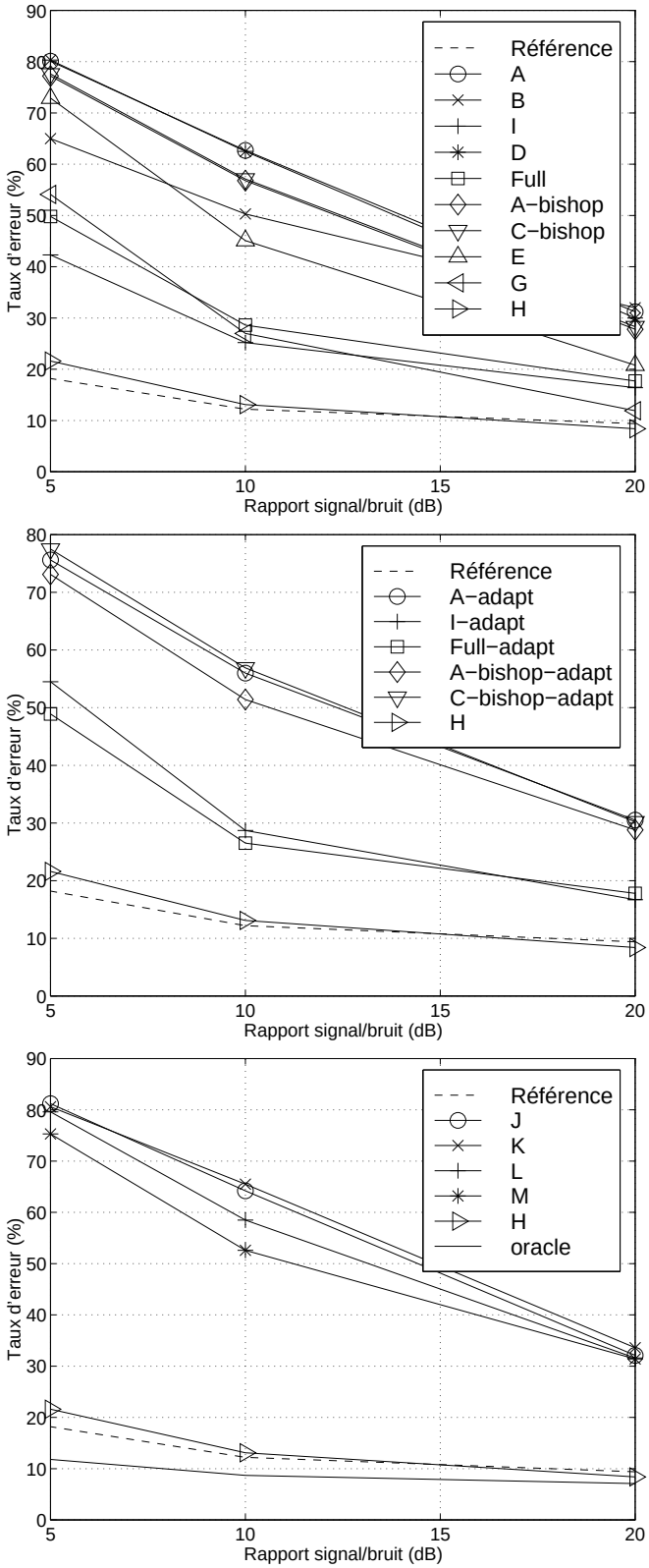


FIG. C.8 – *Taux d'erreur au niveau du mot: bruit coloré, soustraction spectrale.*

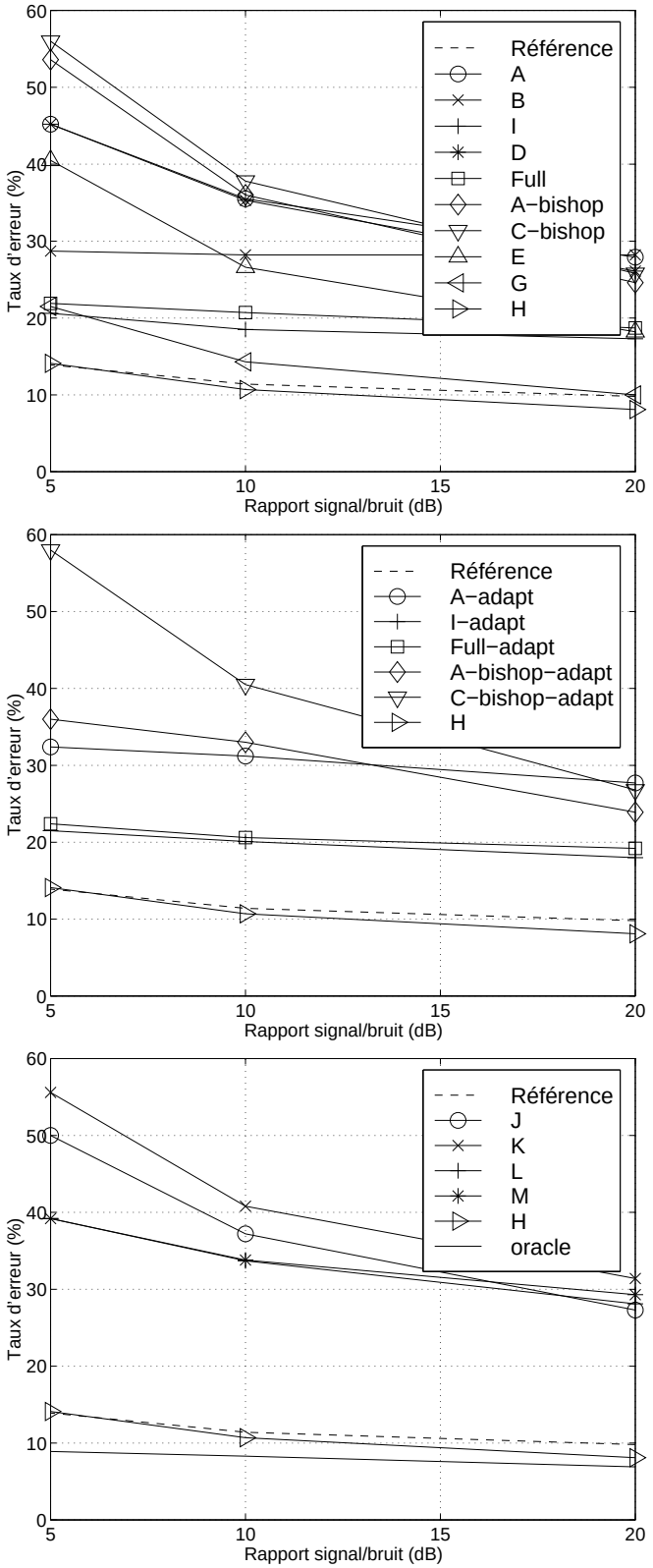


FIG. C.9 – *Taux d'erreur au niveau du mot: bruit fortement coloré, soustraction spectrale.*

Bibliographie

[1] A. Adjoudani and C. Benoît. On the integration of auditory and visual parameters in an HMM-based ASR. In *Speechreading by Humans and Machines: Models, Systems and Applications*. Springer-Verlag, 1996.

[2] J.B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Processing*, 2(4):567–577, 1994.

[3] Takayuki Arai and Steven Greenberg. Speech intelligibility in the presence of cross-channel spectral asynchrony. In *Proc. of ICASSP'98*, pages 933–936, 1998.

[4] ARTIST: Articulatory Representation To Improve Speech Technologies. `http://www.idiap.ch/vision/artist.html/`.

[5] Roland Aubauer, Ralk Kern, and Dieter Leckschat. Optimized second order gradient microphone for hands-free speech recordings in cars. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions. 1999*, pages 191–194, Tampere, Finland, 1999.

[6] Roland Auckenthaler and John S. Mason. Equalizing sub-band error rates in speaker recognition. In *Proc. of EUROSPEECH*, pages 2303–2306, Rhodes, Greece, 1997.

[7] Personal communication. Babel Technologies, 1998.

[8] S. Basu, A. Ittycheriah, and S. Maes. Time shift invariant speech recognition. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[9] V.L. Beattie and S. J. Young. Noisy speech recognition using hidden Markov model state-based filtering. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 917–920, 1991.

[10] S Ben-Yacoub, B. Fasel, and J. Luettin. Fast face detection using mlp and fft. In *Proc. Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, 1999.

[11] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. of ICASSP'79*, pages 208–211, April 1979.

[12] Peter Beyerlein. Discriminative model combination. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 481–484, Seattle, WA, 1998.

[13] René Boite, Hervé Bourlard, Thierry Dutoit, Joël Hancq, and Henri Leich. *Traitement de la parole*. Presses Polytechniques et Universitaires Romandes, 1999.

[14] René Boite and Murat Kunt. *Traitement de la parole.* Presses Polytechniques Romandes, 1987.

[15] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE ASSP*, 2(27), 1979.

[16] S. E. Bou-Ghazale and J. H. L. Hansen. Duration and spectral based stress token generation for HMM speech recognition under stress. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 413–416, Adelaide, Australia, April 1994.

[17] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proc. of Intl. Conf. on Spoken Language Processing*, pages 422–425, Philadelphia, October 1996.

[18] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan. Towards sub-band-based speech recognition. In *Proc. of European Signal Processing Conference*, pages 1579–1582, Trieste, Italy, September 1996.

[19] H. Bourlard, Y. Konig, and N. Morgan. Remap: Recursive estimation and maximization of a posteriori probabilities – application to transition-based connectionist speech recognition. Technical Report TR-94-064, Intl. Computer Science Institute, Berkeley, CA, 1994.

[20] H. Bourlard and N. Morgan. *Connectionist Speech Recognition – A Hybrid Approach.* Kluwer Academic Publishers, ISBN 0-7923-9396-1, 1994.

[21] Hervé Bourlard and Stéphane Dupont. Sub-band-based speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 1251–1254, Munich, April 1997.

[22] L. Braida. Crossmodal integration in the identification of consonants. *Quarterly Journal of Experimental Psychology*, 43A(3):647–677, 1991.

[23] Matthew Brand. Coupled hidden Markov models for modeling interaction processes. Technical Report 405, MIT Media Lab, 1997.

[24] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden Markov models for complex action recognition. Technical Report 407, MIT Media Lab, 1997.

[25] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *IEEE International Conference on Computer Vision*, pages 494–499, Piscataway, NJ, USA, 1995.

[26] Leo Breiman. Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkeley, California, September 1994.

[27] C. Cerisara, J.P. Haton, and D. Fohr. A recombination model for multi-band speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 717–718, Seattle, WA, 1998.

[28] C. Cerisara, J.P. Haton, Jean-François Mari, and D. Fohr. Multi-band continuous speech recognition. In *Proc. of EUROSPEECH*, pages 1235–1238, Rhodes, Greece, 1997.

[29] Christophe Cerisara. *Contribution de l'approche Multi-Bandes à la reconnaissance automatique de la parole.* PhD thesis, LORIA, September 1999.

[30] G. Chollet, J. L. Cochard, Cédric Jaboulet A. Constantinescu, and P. Langlais. Swiss French PolyPhone and PolyVar: Telephone speech databases to model inter and intra-speaker variability. Technical Report IDIAP-RR 1, IDIAP, Martigny, Switzerland, 1996.

[31] R. Cole, L. Hirschmann, L Atlas, and et al. The challenge of spoken language processing: research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3(1):1–20, 1995.

[32] R.A. Cole, M. Fanty, and T. Lander. Telephone speech corpus at CSLU. In *Proc. of Intl. Spoken Language Processing*, Yokohama, Japan, September 1994.

[33] R.A. Cole, J. Mariani, H. Uszkoreit, A. Zuenen, and V. Zue, editors. *Survey of state-of-the-art in Human Language Technology.* Center for Spoken Language Understanding, Oregon Graduate Institute, November 1995.

[34] Gary Cook and Tony Robinson. Boosting the performance of connectionist large vocabulary speech recognition. In *Proc. of the Intl. Conf. on Spoken Language Processing.*, Philadelphia, PA, 1996.

[35] Martin Cooke, Andrew Morris, and Phil Green. Missing data techniques for robust speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, Munich, April 1997.

[36] Thomas Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley and Sons, Inc., 1991.

[37] S. Das, A. Nadas, D. Nahamoo, and M. Picheny. Adaptation techniques for ambience and microphone compensation in the IBM tangora speech recognition system. In *Proc. of ICASSP'94*, pages 21–23, Adelaide, Australia, 1994.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B*, 34:1–38, 1977.

[39] Olivier Deroo. *Modèles Dépendant du Contexte et Fusion de Données Appliqués à la Reconnaissance de la Parole par Modèle Hybride HMM/ANN.* PhD thesis, Faculté Polytechnique de Mons, December 1998.

[40] Olivier Deroo, Christophe Ris, and Stéphane Dupont. Context dependent hybrid HMM/ANN systems for large vocabulary continuous speech recognition system. In *EUROSPEECH'99*, Budapest, Hungary, September 1999.

[41] P. Duchnowski. *A new structure for automatic speech recognition.* PhD thesis, MIT, September 1993.

[42] S. Dupont and H. Bourlard. Multiband approach for speech recognition. In *Proc. of ProRISC/IEEE Workshop on Circuits, Systems and Signal Proc.*, pages 113–118, Mierlo, The Netherlands, November 1996.

[43] Stéphane Dupont. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[44] Stéphane Dupont and Hervé Bourlard. Using multiple time scales in a multi-stream speech recognition system. In *Proc. of EUROSPEECH'97.*, volume 1, pages 3–6, Rhodes, Greece, September 1997.

[45] Stéphane Dupont, Hervé Bourlard, and Christophe Ris. Robust speech recognition based on multi-stream features. In *Proc. of ESCA/NATO Workshop*

on Robust Speech Recognition for Unknown Communication Channels, pages 95–98, Pont-à-Mousson, France, April 1997.

[46] Stéphane Dupont and Leila Cheboub. Fast speaker adaptation of artificial neural networks for automatic speech recognition. In *Proceeding of ICASSP'2000*, Istanbul, Turkey, June 2000.

[47] Stéphane Dupont and Juergen Luettin. Using the multi-stream approach for continuous audio-visual speech recognition: Experiments on the M2VTS database. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[48] Stéphane Dupont and Christophe Ris. Assessing local noise level estimation methods. In *Workshop on Robust Methods For Speech Recognition in Adverse Conditions (Nokia, COST249, IEEE)*, pages 115–118, Tampere, Finland, May 1999.

[49] Mounir El-Maliki, Philippe Renevey, and Andrzej Drygajlo. Rehaussement par soustraction spectrale et compensation des paramètres manquants pour la reconnaissance robuste du locuteur et de la parole. In *Proc. XXIIèmes Journées d'Etude sur la Parole*, pages 409–412, Martigny, Switzerland, 1998.

[50] D.P.W. Ellis. *Prediction-driven computational auditory scene analysis.* PhD thesis, Dept. of Elec. Eng & Comp. Sci., M.I.T., 1996.

[51] The ELRA Home Page.
`http://www.icp.inpg.fr/ELRA/`.

[52] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10), October 1992.

[53] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, December 1984.

[54] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square log-spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 33(2):443–445, 1985.

[55] N. P. Erber and C. L. De Filippo. Voice-mouth synthesis of tactual/visual perception of /pa, ba, ma/. *Journal of the Acoustical Society of America*, 64:1015–1019, 1978.

[56] S. Fisher and K. U. Simmer. Beamforming microphone arrays for speech acquisition in noisy environment. *Speech Communication*, 20(3-4):215–227, 1996.

[57] H. Fletcher. *Speech and Hearing in Communication.* New York – Krieger, 1953.

[58] Vincent Fontaine, Christophe Ris, and Jean-Marc Boite. Nonlinear discriminant analysis for improved speech recognition. In *Proc. of EUROSPEECH'97.*, Rhodes, Greece, 1997.

[59] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd. The voice activity detector for the pan-european digital cellular mobile telephone service. In *Proc. of ESCA/NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 369–372, Glasgow, Scotland, 1989.

[60] Fukunaga. *Introduction to Statistical Pattern Analysis.* Academic Press, 1990.

[61] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech Signal Process.*, 29(2):254–272, 1981.

[62] S. Furui. Speaker independant isolated word recognizer using dynamic features of speech spectrum. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34(1):52–59, 1986.

[63] M.J.F. Gales. Nice model-based compensation schemes for robust speech recognition. In *Proc. of ESCA/NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 55–64, Pont-à-Mousson, France, April 1997.

[64] M.J.F. Gales and S. Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 233–236, San Francisco, California, 1992.

[65] Yuqing Gao, Taiyin Huang, and Jean-Paul Haton. Central auditory model for spectral processing. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 704–707, 1993.

[66] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and AUdio Processing*, 2:291–298, April 1994.

[67] J.L. Gauvain, L.F. Lamel, and M. Eskénazi. Bref, a large vocabulary spoken corpus for french. In *EUROSPEECH'91*, pages 505–508, Genova, Italy, 1991.

[68] Zoubin Gharhamani and Michael I. Jordan. Factorial hidden Markov models. Technical Report Computational Cognitive Science Technical Report 9502, University of Toronto / Massachusetts Institute of Technology, 1996.

[69] O. Ghitza. Auditory nerve representation as front-end for speech recognition in a noisy environment. *Computer Speech and Language*, 1(2):109–130, 1986.

[70] Oded Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Speech and AUdio Processing*, 2(1):115–132, January 1994.

[71] Franck Giron, Yasuhiro Ninami, Masashi Tanaka, and Ken'ichi Furuya. Compensation of speaker directivity in speech recognition using hmm composition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 253–256, Seattle, WA, 1998.

[72] Hervé Glotin, Emmanuel Tessier, Hervé Bourlard, and Frédéric Berthomier. Reconnaissance multi-bandes de la parole bruitée par couplage entre niveaux primitif et d'identification. In *Journées d'étude sur la Parole*, pages 375–378, 1998.

[73] Pedro Gomez, Agustin Alvarez, Rafael Martinez, Victor Nieto, and Victoria Rodellar. A hybrid signal enhancement method for robust speech recognition. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions. 1999*, pages 203–206, Tampere, Finland, 1999.

[74] Yifan Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291, 1995.

[75] Bernard Gosselin. *Application des Réseaux de Neurones Artificiels à la Reconnaissance Automatique des Caractères Manuscrits.* PhD thesis, Faculté Polytechnique de Mons, June 1996.

[76] G. Gravier, M. Sigelle, and G. Chollet. Toward Markov random field modeling of speech. In *Proc. of the Intl. Conf. on Spoken Language Processing.*, Sydney, Australia, December 1998.

[77] K. P. Green and J. L. Miller. On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, 38(3):269–276, 1985.

[78] S. Greenberg and B.E.D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *Proc. of ICASSP'97*, pages 1647–1650, Munich, 1997.

[79] Steven Greenberg, Takayuki Arai, and Rosaria Silipo. Speech intelligibility derived from exceedingly sparse spectral information. In *Proc. of the Intl. Conf. on Spoken Language Processing.*, Sydney, Australia, December 1998.

[80] Andrew K. Halberstadt and James R. Glass. Heterogeneos measuremnts and multiple classifiers for speech recognition. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[81] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001, 10 1990.

[82] S. Haykin. *Adaptive Filter Theory.* Prentice-Hall, Englewood Cliffs, N.J., 1996.

[83] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[84] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):578–589, 1994.

[85] H. Hermansky, N. Morgan, and H.G. Hirsch. Recognition of speech in additive and convolutional noise based on rasta spectral processing. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 83–86, 1993.

[86] H. Hermansky, M. Pavel, and S. Tibrewala. Towards ASR using partially corrupted speech. In *Proc. of Intl. Conf. on Spoken Language Processing*, pages 458–461, Philadelphia, October 1996.

[87] H. Hermansky, M. Pavel, and S. Tibrewala. Down-sampling speech representation in ASR. In *Proc. of EUROSPEECH'99.*, pages 73–76, Budapest, Hungary, September 1999.

[88] Hynek Hermansky. TRAPS - classifiers of temporal patterns. In *Proc. of the Intl. Conf. on Spoken Language Processing.*, Sydney, Australia, November 1998.

[89] H. G. Hirsch. Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. Technical Report TR-93-012, Intl. Comp. Science Institute, Berkeley, CA, 1993.

[90] H.G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *Proc. ICASSP'95*, pages 153–156, 1995.

[91] Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, January 1994.

[92] Tammo Houtgast and Jan A. Verhave. A physical approach to speech quality assessment: Correlation patterns in the speech spectrogram. In *Proc. of EUROSPEECH'91.*, Genova, Italy, 1991.

[93] Hunt and Melvyn. System for separating speech from background noise. U.S. Patent 5319736, June 1994.

[94] Melvyn Hunt. Some experience in in-car speech recognition. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions. 1999*, pages 25–32, 1999.

[95] M.J. Hunt and C. Lefebvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, Signal Processing*, pages 262–265, 1989.

[96] Don R. Hush and B. Horne. Progress in supervised neural networks. *IEEE Signal Processing Magazine*, 78(9), September 1990.

[97] International Computer Science Institute. `http://www.icsi.berkeley.edu`.

[98] Adam Janin, Dan Ellis, and Nelson Morgan. Multi-stream speech recognition: Ready for prime time. In *Proc. of EUROSPEECH*, pages 591–594, 1999.

[99] Frederick Jelinek, Robert L. Mercer, and Salim Roukos. Principles of lexical language modeling for speech recognition. In Sadaoki Furui and M. Mohan Sondhi, editors, *Advances in Speech Signal Processing*, pages 651–699. Marcel Dekker, Inc., 1992.

[100] M. I. Jordan, Z. Ghahramani, and L. K. Saul. Hidden Markov decision trees. In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge,MA, 1997.

[101] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[102] B.H. Juang, S.E. Levinson, and M.M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Trans. Information Theory*, IT-32(2):307–309, March 1986.

[103] Y. Kaneda and J. Ohga. Adaptative microphone array system for noise reduction. *IEEE Transaction on ASSP*, 34(6):1391–1400, 1986.

[104] G. S. Kang and L. J. Fransen. Experimentation with an adaptive noise-cancellation filter. *IEEE Trans. on Circuits and Systems*, 34:753–758, 1987.

[105] Katrin Kirchhoff. Combining articulatory and acoustic information for speech recognition in noisy and reverbarant environments. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[106] J Kittler, M Hatef, R P W Duin, and J Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 3 1998.

[107] J Kittler, Y. P. Li, J Matas, and M. U. Ramos Sanchez. Combining evidence in multimodal personal identity recognition systems. In *Proceedings of Int. Conf. on Audio- and Viseo-based Biometric Person Authentication*, Crans Montana, Switzerland, September 1997.

[108] Hdefumi Kobatake, Kaoru Gyoutoku, and Sheng Lim. Enhancement of noisy speech by maximum likelihood estimation. In *Proc. IEEE Internat. Conf.*

Acoust. Speech and Signal Process., pages 973–976, Toronto, Canada, May 1991.

[109] Y Konig and N. Morgan. Supervised and unsupervised clustering of the speaker space for connectionist speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process. 1993*, Minneapolis, Minnesota, 1993.

[110] A. Korthauer. Robust estimation of the snr of noisy speech signals for the quality evaluation of speech databases. In *Proc. ROBUST'99 workshop*, pages 123–126, Tampere, 1999.

[111] R. Kuhn, P. Nguyen, J-C. Junqua, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Fast speaker adaptation using a priori knowledge. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, Phoenix, Arizona, 1999.

[112] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation. *Computer Speech and Language*, 9:171–185, 1995.

[113] Jae S. Lim and Alan V. Oppenheim. All-pole modeling of degraded speech. *IEEE Trans. Acoust. Speech Signal Process.*, 26(3):101–114, June 1978.

[114] Jae S. Lim and Alan V. Oppenheim. Evaluation of an adaptative comb filtering method for enhancing speech degraded by white noise addition. *IEEE Trans. Acoust. Speech Signal Process.*, 26(4):354–358, 1978.

[115] Jae S. Lim and Alan V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, December 1979.

[116] Klaus Linhard and Heinz Klemm. Noise reduction with spectral subtraction and median filtering for suppression of musical tones. In *Proc. of ESCA-NATO WOrkshop on Robust Speech Recognition for Unknown Communication Channels*, pages 159–162, Pont-à-Mousson, France, April 1997.

[117] Richard P. Lippmann. Accurate consonant perception without mid-frequency speech energy. *IEEE Transactions on Speech and AUdio Processing*, 4(1):66–69, 1996.

[118] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22:1–15, 1997.

[119] Richard P. Lippmann and Beth A. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages KN37–KN40, Munich, April 1997.

[120] P. Lockwood and J. Boudy. Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11:215–228, 1992.

[121] Beth Logan. *Adaptive Model-Based Speech Enhancement.* PhD thesis, Girton College, University of Cambridge and Cambridge University Engineering Department, 1998.

[122] Beth Logan and Pedro Moreno. Factorial hidden Markov models for speech recognition: Preliminary experiments. Technical Report CRL 97/7, Digital - Cambridge Research Laboratory, 1997.

[123] Beth Logan and Pedro Moreno. Factorial HMMs for acoustic modeling. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 813–816, 1998.

[124] J. Luettin. *Visual Speech and Speaker Recognition.* PhD thesis, University of Sheffield, 1997.

[125] J. Luettin and N.A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, February 1997.

[126] Juergen Luettin. Towards speaker independent continuous speechreading. In *Proc. of EUROSPEECH'97.*, Rhodos, Greece, 1997.

[127] Juergen Luettin and Stéphane Dupont. Continuous audio-visual speech recognition. In *Proc. of Fifth European Conference on Computer Vision*, pages 657–673, Freiburg, Germany, June 1998.

[128] Djamila Mahmoudi and Andrzej Drygajlo. Combined wiener and coherence filtering in wavelet domain for microphone array speech enhancement. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process. 1998*, pages 385–388, 1998.

[129] Brian Mak. Combining ANNs to improve phone recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, Munich, Germany, April 1997.

[130] Jean-Pierre Martens and L. Van Immerseel. An auditory model based on the analysis of envelope patterns. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, 1990.

[131] Jean-Pierre Martens and L. Van Immerseel. Pitch and voiced/unvoiced determination with an auditory model. *The Journal of the Acoustical Society of America*, 91(6):3511–3526, June 1992.

[132] Rainer Martin. An efficient algorithm to estimate the instantaneous SNR of speech signals. In *Eurospeech'93*, pages 1093–1096, 1993.

[133] D. W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[134] D. W. Massaro and M. M. Cohen. Perceiving asynchronous bimodal speech in consonant vowel and vowel syllables. *Speech Comunication*, 13:127–134, 1993.

[135] D. Matrouf. *Adaptation des modèles de Markov cachés pour la reconnaissance de la parole bruitée.* PhD thesis, Université Paris-Sud, 1997.

[136] R. J. McAulay and M.L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE ASSP*, 28:137–145, 1809.

[137] B.L. McKinley and G.H. Whipple. Model based speech pause detection. In *Proc. ICASSP'97*, pages 1179–1182, Munich, 1997.

[138] Philip McMahon, Paul McCourt, and Saeed Vaseghi. Discriminative weighting of multi-resolution sub-band cepstral features for speech recognition. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[139] Jerry M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications and Control.* Prentice Hall Signal Processing Series, ISBN 0-13-120981-7, Englewood Cliffs, NJ 07632, 1995.

[140] Denis Mercier, editor. *Le Livre des Techniques du Son, Tomes 1 & 2.* Editions Fréquences, Diffusion Eyrolles, Paris, France, 1990.

[141] George A. Miller and Patricia E. Nicely. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 22(2):338–352, March 1955.

[142] Naghmeh Nikki Mirghafori. Multi-band speech recognition: A summary of recent work at ICSI. Technical Report TR-97-051, Intl. Comp. Science Institute, Berkeley, CA, 1997.

[143] Naghmeh Nikki Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition.* PhD thesis, International Computer Science Institute, Berkeley, California, January 1999.

[144] Nikki Mirghafori and Nelson Morgan. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[145] Nikki Mirghafori and Nelson Morgan. Transmissions and transitions: A study of two common assumptions in multi-band asr. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 713–716, Seattle, WA, 1998.

[146] Perry Moerland. Mixtures of experts estimate a posteriori probabilities. Technical report, IDIAP, Martigny, Switzerland, 1997.

[147] N. Morgan, C. Wooters, and H. Hermansky. Experiments with temporal resolution for continuous speech recognition with multi-layer perceptrons. In *Proc. of IEEE Workshop on Neural Networks for Signal Processing*, pages 405–410, 1991.

[148] T. Morii and H. Hoshimi. Noise robustness in speaker independent speech recognition. In *Proc. of the Intl. Conf. on Spoken Language Processing.*, pages 1145–1148, November 1990.

[149] Andrew Morris. Personal communication. 1998.

[150] Andrew Morris, Astrid Hagen, and Hervé Bourlard. The full combination sub-bands approach to noise robust HMM/ANN based ASR. In *Proc. of EUROSPEECH*, pages 599–602, 1999.

[151] Joao Neto, Luis Almeida, Mike Hochberg, Ciro Martins, Luis Numes, Steve Renals, and Tony Robinson. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. In *Proc. of EUROSPEECH'95.*, Madrid, Spain, 1995.

[152] comp.speech FAQ, Noisex-92. `http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html`, August 1996.

[153] *Proceedings of Workshop on Robust Methods for Seech Recognition in Adverse Conditions.* NOKIA, COST249, IEEE, Tampere, Finland, 1999.

[154] J. A. Nolazco-Flores and S. J. Young. Continuous speech recognition in noise using spectral subtraction and HMM adaptation. In *Proc. of ICASSP'94*, pages 409–412, Adelaide, Australia, April 1994.

[155] Shigeki Okawa, Enrico Bocchieri, and Alexandros Potamianos. Multi-band speech recognition in noisy environments. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 641–644, Seattle, WA, 1998.

[156] D. B. Paul. A speaker-stress resistant HMM isolated automatic speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 713–716, 1987.

[157] Douglas B. Paul. An efficient a* stack decoder algorithm for continuous speech recognition with a stochastic language model. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, 1992.

[158] E. D. Petajan. Automatic lipreading to enhance speech recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 40–47, 1985.

[159] Joseph W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1214–1247, September 1993.

[160] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database (release 1.00). In *Proc. of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Crans-Montana, Switzerland, 1997.

[161] Gerasimos Potamianos and Hans Peter Graf. Discriminative training of hmm stream exponents for audio-visual speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 3733–3736, Seattle, WA, 1998.

[162] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallet. The DARPA 1000-words ressource management database for continuous speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 651–654, April 1988.

[163] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall Signal Processing Series, 1993.

[164] Mazin Rahim. Multiple models integration for multi-environment speech recognition. EP 0 881 625, December 1998.

[165] Yuval Raviv and Nathan Intrator. Bootstrapping with noise: An effective regularization technique. *Connection Science Special Issue on Combining Estimators*, 1996.

[166] Steve Renals and Mike Hochberg. Decoder technology for connectionist large vocabulary speech recognition. Technical Report CUED/F-INFENG/TR.186, Cambridge University Engineering Department, 1995.

[167] B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[168] Christophe Ris and Stéphane Dupont. Assessing local noise level estimation methods. *to be published in Speech Communication*, 2000.

[169] J. Robert-Ribes. *Modèles d'intégration audiovisuelle de signaux linguistiques: de la perception humaine à la reconnaissance automatique des voyelles.* PhD thesis, Signal Image Parole, Institut National Polytechnique de Grenoble, 1997.

[170] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Macine Intelligence*, 20(1):23–38, January 1998.

[171] Stuart Russell, John Binder, and Daphne Koller. Adaptive probabilistic networks. Technical Report UCB/CSD-94-824, University of California - Computer Science Division, 1994.

[172] H. Sakoe. Two level DP matching - a dynamic time warping based pattern matching algorithm for continuous speech recognition. *IEEE Transactions of the IECE of Japan*, 3, 1979.

[173] Ruhi Sarikaya and John H. L. Hansen. Robust speech activity detection in the presence of noise. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[174] Volker Schless and Fritz Class. SNR-dependent flooring and noise overestimation for joint application of spectral subtraction and model combination. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[175] Robert V. Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, October 1995.

[176] B. Sheikhzadeb, H. Sameti, L. Deng, and R. L. Brennan. Comparative performance of spectral subtraction and HMM-based speech enhancement strategies with application to hearing aid design. In *Proc. of ICASSP'94*, pages 13–16, Adelaide, Australia, April 1994.

[177] P. L. Silsbee and A. C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.

[178] Peter L. Silsbee and Qin Su. Audiovisual sensory integration using hidden Markov models. In *Speechreading by Humans and Machines: Models, Systems and Applications*. Springer-Verlag, 1996.

[179] K. U. Simmer and A. Wasiljeff. Adaptative microphone arrays for noise suppression in the frequency domain. In *Second Cost 229 Workshop on Adapt. Algo. in Com.*, pages 185–194, France, 1992.

[180] L. Singh and S. Sridharan. Speech enhancement using critical band spectral subtraction. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[181] Olivier Siohan, Cristina Chesta, and Chin-Hui Lee. Hidden Markov Model adaptation using maximum a posteriori linear regression. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions. 1999*, Tampere, Finland, 1999.

[182] P Sivakumaran, A. M. Ariyaeeinia, and J. A. Hewitt. Sub-band based speaker verification using dynamic recombination weights. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[183] P. M. Smeele and et al. Intelligibility of audio-visually desynchronized speech: Asymmetrical effect of phoneme position. In *Proc. of the Intl. Conf. on Spoken Language Processing.*, pages 65–68, Alberta, Canada, 1992.

[184] Johan Smolders, Tom Claes, Gert Sablon, and Dirk Van Compernolle. On the importance of the microphone position for speech recognition in the car. In *Proc. of ICASSP'94*, pages 429–432, 1994.

[185] P. Smyth, D. Heckerman, and M. I. Jordan. Probabilistic independence networks for hidden Markov probability models. Technical Report MSR-TR-96-03, Microsoft Research, 1996.

[186] Padhraic Smyth, David Heckerman, and Michael Jordan. Probabilistic independence network for hidden Markov probability models. Technical Report AI 1564 / CBCL 132, Massachusetts Institute of Technology, 1996.

[187] Myung Gyu Song, Hoi In Jung, Kab-Jong Shim, and Hyung Soon Kim. Speech recognition in car noise environments using multiple models according to noise masking levels. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, December 1998.

[188] Helge B.D. Sorensen. A cepstral noise reduction multi-layer neurla network. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 933–936, 1991.

[189] J.M. Steeneken and D.A. Van Leeuwen. Multilingual assessment of speaker independant large vocabulary speech-recognition systems: the SQALE project (speech recognition quality assessment for language engineering). In *Proc. of EUROSPEECH'95.*, Madrid, Spain, 1995.

[190] D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines: Models, Systems and Applications.* NATO ASI Series F, Computer and Systems Sciences, Springer-Verlag, Berlin, 1996.

[191] A. Q. Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London, Series B*, 335:71–78, 1992.

[192] S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. In *Proc. of ICASSP'97*, pages 1255–1258, Munich, 1997.

[193] Sangita Tibrewala and Hynek Hermansky. Multi-band and adaptation approaches to robust speech recognition. In *Proc. of EUROSPEECH*, pages 2619–2622, Rhodes, Greece, 1997.

[194] M.J. Tomlinson, M.J. Russel, and N.M. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 821–824, 1996.

[195] M.J. Tomlinson, M.J. Russell, R.K. Moore, A.P. Buckland, and M.A. Fawley. Modelling asynchrony in speech using elementary single-signal decomposition. In *Proc. of ICASSP'97*, pages 1247–1250, Munich, 1997.

[196] K. Tumer and J. Gosh. Linear and order statistics combiners for pattern classification. In A. Sharkey, editor, *Combining Artificial Neural Nets*, pages 127–162. Springer-Verlag, 1999.

[197] ULg. ULg - acoustics laboratory - the MADRAS project. `http://www.montefiore.ulg.ac.be/services/acous/cedia/madrasen.html`, August 1998.

[198] A.P. Varga and R.K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 845–848, 1990.

[199] A.P. Varga and R.K. Moore. Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition. In *Proc. of EUROSPEECH'91.*, pages 1175–1178, Genova, Italy, 1991.

[200] P. Vary. On the enhancement of noisy speech. In *Proc. of EUSIPCO'83*, pages 327–330, 1983.

[201] Olli Viikki, David Bye, and Kari Laurila. A recursive feature vector normalization approach for robust speech recognition in noise. In *Proc. of ICASSP'98*, pages 733–736, 1998.

[202] K. Wang, S.A. Shamma, and W.J. Byrne. Noise robustness in the auditory representation of speech signals. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, Signal Processing 1993*, pages 335–338, 1993.

[203] C. J. Wellekens, L. Kangasharju, and C. Milesi. The use of Meta-HMM in multistream HMM training for automatic speech recognition. In *Proc. of the Intl. Conf. on Spoken Language Processing.*, Sidney, Australia, 1998.

[204] Gethin Williams and Steve Renals. Confidence measures for hybrid HMM/ANN speech recognition. In *Proc. of EUROSPEECH'97.*, Rhodes, Greece, September 1997.

[205] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[206] Kevin Woods, Philip Kegelmeyer, and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):405–410, April 1997.

[207] S. Wu, M. Shire, S. Greenberg, and N. Morgan. Integrating syllable boundary information into speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 987–990, Munich, April 1997.

[208] Su-Lin Wu, E.D. Kingsbury, Nelson Morgan, and Steven Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 721–724, Seattle, WA, 1998.

[209] Fei Xie and Dirk Van Compernolle. Speech enhancement by nonlinear spectral estimation - a unifying approach. In *Proc. of EUROSPEECH'93*, pages 617–620, 1993.

[210] D. Xu, C. Francourt, and C. Wang. Multi-channel HMM. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 841–844, Atlanta, USA, May 1996.

[211] Lei Xu, Adam Kryzak, and Ching Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):318–435, May 1992.

[212] Nestor Becerra Yoma, Fergus R. McInnes, and Mervyn A. Jack. Weighted viterbi algorithm and state duration modelling for speech recognition in noise. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pages 709–712, Seattle, WA, 1998.

[213] R. Zelinski. A microphone array with adaptative postfiltering for noise reduction in reverberant rooms. In *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process. 1988*, pages 2578–2581, 1988.

[214] Geoffrey Zweig. Dynamic bayesian networks and the concatenation problem in speech recognition. Technical Report UCB/CSD-96-927, University of California - Computer Science Division, 1996.

[215] Geoffrey Zweig and Stuart Russell. Compositional modeling with DPNs. Technical Report UCB/CSD-97-970, University of California - Computer Science Division, 1997.