N° d'ordre : 06/2011-D/EL

REPUBLIQUE ALGERIENNE DEMOCRATIQUEET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie
Houari Boumediene U.S.T.H.B.

FACULTÉ D'ELECTRONIQUE ET INFORMATIQUE

THÉSE

Présentée pour l'obtention du grade de DOCTORAT

En : Electronique

Spécialité : Communication Parlée

Par: **Djamel ADDOU**

Thème :

Intégration du Paradigme Multi-Stream MFCC-LSF dans un Système de Reconnaissance de la Parole Distribuée Robuste

Soutenue publiquement le 22/06/2011, devant le jury composé de:

R. TOUHAMI	Professeur	à l'USTHB	Président
M. BOUDRAA	Professeur	à l'USTHB	Directeur de thèse
S.A. SELOUANI	Professeur	à l'U. Moncton	Co-Directeur de thèse
M. GUERTI	Professeur	à l'ENP	Examinateur
A.OULDALI	Professeur	à l'EMP	Examinateur
B. FERGANI	Maitre de Confèrence/A	à l'USTHB	Examinateur

Remerciements

Je tiens en tout premier lieu à remercier Monsieur ***Sid-Ahmed Selouani****, expert en « interfaces de communication en parole embarquée » et directeur de cette thèse. Grâce à lui, j'ai eu la chance d'évoluer, pendant ces dernières années, dans un environnement scientifique stimulant tout en profitant d'une grande liberté dans l'orientation de mon travail. Sur le plan scientifique comme humain, je le remercie pour son écoute attentive, ses conseils avisés et sa grande disponibilité, sans oublier les heures de discussions fructueuses via Skype.*

Je souhaite ensuite remercier chaleureusement Madame ***Malika Boudraa****, également directeur de ce travail de thèse. Ses nombreux conseils et sa très grande expérience dans le domaine du traitement de la parole m'ont beaucoup apporté. Je la remercie vivement pour la confiance qu'elle m'a témoignée et le regard attentif qu'elle n'a cessé de porter sur mon travail.*

Dans le cadre ma thèse, j'ai eu la grande chance de travailler avec Monsieur ***Bachir Boudraa****, Responsable du laboratoire Analyse et Codage du Signal et pionnier de la lecture et révision de tous les travaux scientifiques de notre laboratoire. Nos échanges réguliers ainsi que mon intégration au sein de son équipe m'ont beaucoup apporté et je souhaite le remercier vivement pour cela.*

Je tiens ensuite à exprimer ma gratitude aux différents membres du jury. Je remercie Madame ***Rachida Touhami*** *pour avoir accepté de présider la soutenance, ainsi que Madame* ***Mhania Guerti,*** *Messieurs* ***Abdelaziz Ouldali*** *et* ***Belkacem Fergani*** *pour avoir examiné ce travail.*

Je tiens évidemment à remercier ma famille qui m'a accompagné et soutenu pendant tout mon parcours d'études dont ce travail est, d'une certaine manière, une forme d'aboutissement. Parce que je puise en eux et en leurs attentes une grande partie de ma motivation, je tiens à associer tout particulièrement mon père, ma mère et ma femme à ce travail.Merci infiniment pour vos prières.

Table des matières

Liste des figures

Liste des tableaux

Abréviations

En Anglais :

ANN : Artificiel Neural Network

AR : Autoregressive

ASR : Automatic Speech Recognition

CDMA: Code Division Multiple Access

CE : Conditional Expectation

CI-KLT: Class Independent Karhunen-Loève Transform

DSR : Distributed Speech Recognition

DSR_FE: Distributed Speech Recognition_FrontEnd

DTTS : Distributed Text To Speech

DTW : Dynamic Time Warping

EM : Expectation Maximisation

ETSI : European Télécommunication Standard Institute

EVD : Eigen Value Decomposition

FFT : Fast Fourier Transform

GMM : Gaussian Mixture Model

GSM : Global System for Mobile Communications

HMM : Hidden Markov Model

HTK : Hidden Markov ToolKit

IETF : Internet Engineering Task Foce

KLT : Karhunen-Loève Transform

LDA : Linear Discriminant Analysis

LMR : Linear Multivariate Regression

LPC : Linear Prediction Coefficients

LPCC : Linear Prediction Cepstral Coefficients

LSF : Line Spectral Frequencies

MAP : Maximum A Posteriori

ML : Maximum Likelihood

MLLR : Maximum Likelihood Linear Regression

MFCC : Mel Frequency Cepstral Coefficients

PDA : Personal Digital Assistant

PLP : Perceptual Linear Prediction

ROVER: Recognizer Output Voting Error Reduction

sil : silence model

sp : short pause (silence model)

SVD : Singular Value Decomposition

TTS : Text-To-Speech

UMTS : Universel Mobil Telecommunication System

WER : Word Error Rate

En Français :

EQM : Erreur Quadratique Moyenne

CPM : Combinaison Parallèle de Modèles

MAP : Maximum A Posteriori

MMC : Modèles de Markov Cachés

MMI : MaximalMutual Information

MMIE : Maximum Mutual Information Estimation

RAP : Reconnaissance Automatique de la Parole

RRP : Reconnaissance Robuste de la Parole

RSB : Rapport Signal sur Bruit

SRAP : Systèmes de Reconnaissance Automatique de la Parole

TALN : Traitement Automatique du Langage Naturel

Résumé

Dans ce mémoire, nous présentons une nouvelle approche de l'analyse acoustique basée sur le paradigme multi-stream dans un système de reconnaissance de la parole distribuée (Distributed Speech Recognition). Nous visons à améliorer les performances des systèmes basés sur les modèles de Markov cachés (HMM), en fusionnant les paramètres LSF (*line spectral frequencies*) avec les paramètres conventionnels MFCC *(Mel Frequency Cepstral Coefficient)*, afin de constituer un nouveau vecteur acoustique multi-variable. D'autre part, nous visons à optimiser l'utilisation des flux de paramètres en réduisant la dimension des vecteurs acoustiques tout en améliorant la robustesse du système en utilisant la transformée de Karhunen-Loève (KLT). C'est une technique de décomposition en sous-espaces également utilisée en rehaussement des signaux bruités. L'approche visée par le présent travail constitue une alternative aux codecs DSR-XAFE (XAFE : eXtended Audio Front-End) actuellement disponibles dans les communications mobiles ou GSM. Comparativement aux standards DSR actuels, notre système utilisant l'intégration de nouveaux streams tels que les coefficients LSF, vise une meilleure dimunition du débit ainsi que de meilleures performances aussi bien pour de faibles rapports signal sur bruit (RSB) que pour des environnements changeants, côté client. Pour l'évaluation de notre système LSF-KLT, nous avons effectué nos expériences sur la base de données Aurora de l'ETSI et la plateforme logicielle HTK de l'Université de Cambridge. Nos résultats montrent que pour des signaux de parole hautement bruités, l'intégration des LSF conduit à une amélioration importante.

Mots-clés : Aurora, HMM, KLT, LSF, MFCC, Paradigme multi-Stream, Reconnaissance de la parole distribuée (DSR), RSB.

Abstract

This thesis describes a new noise-robust Distributed Speech Recognition (DSR) front-end. It aims to improve the performance of the systems based on Hidden Markov Models (HMM), using a combination of the conventional Mel-cepstral Coefficients (MFCC) and Line Spectral Frequencies (LSF) in order to form a new multi-variable acoustical vector.

These features are adequately transformed and reduced in a multi-stream scheme using Karhunen-Loève Transform (KLT). The approach presented in this work may constitute an alternative to the DSR-XAFE (eXtended Audio Front-End) available in mobile communications (GSM). Comparatively to the current standard DSR, our approach using the integration of new streams as the LSF coefficients, gives a better compression rate and better performance for low signal to noise ratio (SNR)and for changing environments at the client side.On the other hand, we investigate the performance of our LSF-KLT system in terms of recognition accuracy in adverse conditions as well as in terms of dimensionality reduction. Our results showed that for highly noisy speech, the proposed transformation significantly improve the recognition accuracy when evaluated on Aurora 2 task and using toolkit HTK soft.

Keywords: Aurora, HMM, Distributed Speech Recognition (DSR),KLT, LSF, MFCC, Multi-stream paradigm, SNR.

Introduction générale

Introduction générale

Situation de la problématique

Les systèmes de reconnaissance automatique de la parole (SRAP) sont aujourd'hui bien connus dans le monde de l'informatique et suscitent l'intérêt d'un public de plus en plus large. De ce fait, ces systèmes doivent être opérationnels en situation réelle, dans un contexte de fortes variations de conditions acoustiques. L'environnement, l'utilisateur, les microphones utilisés, etc. sont autant de sources de variabilité qui peuvent faire chuter les performances de ces systèmes. Afin de garder des taux de reconnaissance acceptables, il est alors nécessaire d'imposer des contraintes sur leurs conditions d'utilisation. Le choix du microphone, le débit de parole ou la taille du vocabulaire font partie des paramètres à gérer correctement pour une bonne robustesse des SRAP. Ainsi, un système de reconnaissance de la parole est robuste s'il est capable de garder un bon taux de reconnaissance même si la qualité du signal est dégradée, ou si les caractéristiques acoustiques du signal sont différentes entre la phase d'apprentissage du système et celle d'utilisation.

Cette notion de robustesse est l'un des problèmes majeurs qui limitent l'*utilisabilité* de ces systèmes. Comment rendre les systèmes existants adaptables aux conditions acoustiques et aux caractéristiques changeantes d'élocution en situation réelle, c'est-à-dire hors laboratoire ?

Le problème de dé-bruitage de la parole n'est pas récent. Cependant, il constitue toujours un champ d'étude vaste et encore riche d'idées. L'objectif est de restaurer un signal utile à partir d'observations corrompues par un bruit souvent considéré additif. Cette hypothèse est souvent utilisée, à la fois pour sa simplicité, mais aussi car elle permet de modéliser un grand nombre de situations pratiques. Le signal observé est donc considéré comme la somme du signal de parole et du bruit ambiant. Ce modèle omet tout bruit convolutif, électrique ou de quantification. Il existe des systèmes de reconnaissance de la parole indépendants du locuteur pour des vocabulaires limités et qui offrent des performances intéressantes à condition que l'environnement ne soit pas trop corrompu par le bruit et les interférences.

Toutefois, l'interrogation de bases de données à distance, (téléphone, radio, etc.) nécessite de concevoir des systèmes polyvalents et peu sensibles aux conditions et aux environnements d'opérations ainsi qu'aux bruits ambiants (téléphone cellulaire, cabine téléphonique, voitures, avions, camions, etc.). Or le traitement de la parole en milieu bruité est loin d'être résolu et les systèmes de reconnaissance actuels ne peuvent fonctionner dans de tels environnements sans subir des dégradations très significatives de leurs performances ce qui les rend à toute fin pratique non utilisables par le grand public. Devant les difficultés rencontrées en reconnaissance de la parole et la nécessité de commercialiser rapidement, la majorité des travaux ont porté sur de la parole "propre" libre de tout bruit. Beaucoup d'efforts ont été investis dans la réalisation de systèmes pour lesquels la parole a été enregistrée en "laboratoire". Ceci a engendré un espoir un peu trop grand et trop rapide vis-à-vis de cette technologie et de son utilisation en conditions réelles (hors laboratoire). La méthodologie utilisée a donc biaisé la conception des systèmes de reconnaissance en assumant que le signal de parole est libre d'interférences et de bruits. Cependant nous avons conscience de ces limites et nous nous proposons de travailler à partir de données enregistrées en conditions réelles afin de nous pousser à aborder le problème sous l'angle des conditions réelles et difficiles de fonctionnement.

On peut considérer qu'il existe deux écoles de pensée à cet égard. Une qui prône le "nettoyage" et le "débruitage" préalable à la reconnaissance, ce qui a pour avantage de pouvoir ensuite utiliser l'arsenal d'outils déjà développés en reconnaissance de parole sur des données relativement propres. Et l'autre école de pensée qui assume qu'il n'est pas possible de "débruiter" à priori et que le système doit être capable de traiter directement les interférences et le bruit un peu de la même façon que le fait l'être humain. Cette dernière approche, quoique moins mature, a l'avantage de permettre la conception de systèmes plus versatiles. Nous pensons qu'en fait il y a lieu de comparer et d'évaluer les points forts et les points faibles des deux tendances afin d'élaborer un système de reconnaissance qui pourra éventuellement être mixte.

D’une manière générale, la stratégie prédominante adoptée par les chercheurs pour résoudre ces problèmes consiste à enregistrer des corpus d'apprentissage dans des conditions qui soient les plus proches de celles dans lesquelles les différents systèmes sont utilisés. Cela nécessite de définir des normes et des critères de comparaison des systèmes de reconnaissance.

L'objectif étant d'estimer et de garantir leur bon fonctionnement lors de leur utilisation. Malheureusement, les sources de variabilité qui peuvent affecter les performances d'un système de reconnaissance sont nombreuses et non prévisibles (environnements changeants, bruit ambiant, etc.), et ces systèmes utilisés alors comme des "boîtes noires" statiques [**Gong 1995**], s'adaptent mal aux variations acoustiques.

Toutefois, des travaux mettant l'accent sur des méthodologies nouvelles, visant une meilleure utilisation des données existantes, reçoivent une plus grande attention de la part de la communauté scientifique [**Haton 1997**]. L'adaptation aux conditions acoustiques a ainsi fait l'objet de nombreux travaux qui ont débouché sur des techniques de filtrage et d'adaptation au bruit ambiant, qui améliore sensiblement les performances des systèmes de reconnaissance [**Junqua et Haton 1996**]. Quant aux problèmes d'adaptation aux locuteurs, ils sont de plus en plus considérés comme cruciaux [**Bradford 1995**] pour les nouvelles technologies vocales.

Contexte technologique et objectif poursuivi

De nombreux travaux de recherche visent la mise en œuvre de techniques d'analyse acoustique robustes. Certaines approches proposent d'intégrer au sein du même vecteur acoustique, des paramètres diversifiés basés sur des analyses acoustiques différentes. Il s'agit de ce qu'on appelle l'analyse acoustique multi-variable. Cette dernière vise à couvrir un maximum de situations environnementales possibles et d'assurer ainsi une meilleure robustesse face aux bruits.

C'est dans ce cadre que se situe cette thèse. En effet, nous visons le développement d'une nouvelle technique d'analyse acoustique multi-variable pour la reconnaissance automatique de la parole distribuée. La technique consiste à extraire du signal vocal, des paramètres pertinents pouvant augmenter la robustesse des signaux dans les milieux fortement bruités en mode distribué pour des télécommunications mobiles.

Notre étude s'intègre dans le cadre du développement des réseaux sans fil supportant les standards de troisième génération ainsi que ceux du système universel de télécommunications mobiles(UMTS). Elle vise une approche pour la reconnaissance de la parole distribuée (ou DSR : Distributed Speech Recognition).

La modélisation acoustique par les méthodes les plus performantes de l'état de l'art reste insuffisante; cette faiblesse est un facteur limitant des SRAP. Nous cherchons à améliorer la qualité de la modélisation acoustique, en intégrant certains traitements adaptés à un processus de décodage de la parole. Nous étudierons en particulier les problèmes de l'adaptation du système de reconnaissance aux milieux bruités par une meilleure modélisation des événements acoustiques.

Le concept de la DSR a été initié pour la première fois dans le projet Aurora au sein de l'ETSI (European Télécommunication Standard Institute) et a donné lieu à la normalisation du module de prétraitement de ces systèmes. Dans un contexte de normalisation mondiale, un consortium formé autour du projet de partenariat de 3GPP a recommandé comme codeur décodeur (codec) pour les services de commandes vocales, l'utilisation du système évolué de codage étendu (XAFE : eXtended Audio Front-End). L'approche visée par le présent travail de recherche constitue une alternative aux codecs DSR-XAFE actuellement disponibles dans les communications mobiles ou GSM[1], en proposant d'incorporer une technique d'analyse acoustique multi-variable plus robuste et à plus bas débit, dont l'optimisation est effectuée par une approche sous-espace multi-stream à pondérations variables, afin d'assurer une robustesse du système de reconnaissance dans des environnements acoustiques changeants. Le système développé sera compatible avec le projet de partenariat de 3GPP qu'il soit relatif aux réseaux élaborés sur GSM (Européen), ou sur CDMA[2] (Nord-Américain). Comparativement aux standards de DSR actuels, notre système utilisant le paradigme multi-variable optimisé vise une meilleure réduction de débit et de meilleures performances aussi bien pour des faibles RSB que pour des environnements changeants côté client. L'idée consiste à agir au niveau de l'analyse acoustique du signal vocal afin de trouver le type de paramètres le mieux adapté aux milieux bruités et pslus particulièrement ceux induits par le milieu de transmission sans fils. L'utilisation d'une analyse multi-variable pose le problème d'assignation des poids de chaque source d'information. En effet, il est très important de bien choisir ces poids en fonction des environnements acoustiques changeants auxquels sont confrontés les systèmes de DSR.

Ce mémoire présente donc *un nouveau front-end* basé sur un multi-stream dans un système de reconnaissance de la parole distribuée. Il vise à améliorer les performances des systèmes

[1]Global System for Mobile Communications

[2]Code Division Multiple Access

basés sur les modèles de Markov cachés (MMC) en combinant les paramètres conventionnels MFCC (Mel Frequency Cepstral Coefficient) et les LSF (Line Spectral Frequencies), afin de constituer un nouveau vecteur acoustique multi variable.

Notre objectif au cours de ce travail est d'augmenter la robustesse de la reconnaissance de la parole en mode distribué. Pour ce faire il s'agira :

- d'exploiter le potentiel des LSF à être utilisé comme paramètres complémentaires aux coefficients MFCC classiques;
- d'optimiser l'assignation des poids à chaque variable et voir à quel point cela peut influer sur le rendement du système;
- de tester le nouveau système d'analyse acoustique (front-end) avec différents débits de transmission et de voir dans quelle mesure notre système peut réduire les débits;
- de discuter les performances du système proposé par rapport au basique DSR-FE.

Pour mener à bien ce travail, nous utilisons comme données expérimentales la base de données Aurora fournie par l'ETSI. Il s'agit d'une base de données de chiffres (digits) prononcés en anglais sur réseau GSM. Ces enregistrements sont pris pour différents types de bruits et différents rapports RSB.

Plan du document

Cette thèse est constituée de quatre chapitres. Ceux-ci décrivent certaines méthodes de traitement des signaux pour la reconnaissance de la parole robuste dans un environnement réel.

Le chapitre 1 aborde la nécessité de la réduction de bruit. Il présente une analyse bibliographique portant sur les travaux en reconnaissance de la parole et sur les caractéristiques des méthodes à développer pour obtenir une reconnaissance robuste.

Le chapitre 2 présente, en premier lieu le front-end fourni par l'Institut européen des normes de télécommunications (ETSI). Les sections suivantes sont dédiées à l'approche acoustique multi-variable où nous mettrons l'accent sur les principaux travaux de recherches dans ce cadre.

Le chapitre 3 décrit le traitement statistique et dimensionnel des paramètres acoustiques dans un cadre d'intégration des paramètres multiples dans un système DSR.

Le chapitre 4 examine le potentiel de l'approche acoustique multi-variable, présenté dans les chapitres 2 et 3, à l'aide de plusieurs expériences. Dans ce chapitre, nous effectuons également l'analyse des résultats obtenus.

En conclusion nous ferons un bilan sur l'apport de notre système par rapport au modèle standard de DSR et nous esquisserons les perspectives des recherches futures dans cet axe.

Chapitre 1

Reconnaissance robuste de la parole : Etat de l'art

Chapitre 1

Reconnaissance Robuste de la Parole: État de l'Art

Sommaire

Ce chapitre aborde la nécessité de la réduction de bruit dans la reconnaissance de la parole. Il présente une analyse bibliographique sur les travaux en reconnaissance de la parole en milieu bruité et sur les caractéristiques des méthodes pour obtenir une reconnaissance robuste. Les systèmes actuels de reconnaissance automatique de la parole (SRAP), bien que très performants pour un environnement acoustique déterminé, voient leurs performances se dégrader d'une manière drastique dès que les conditions de test ou d'utilisation changent par rapport aux conditions d'apprentissage. Même si différentes approches sont proposées dans la littérature [**Calliope, 1989** ; **Haton et al. 1991**], le problème reste à ce jour posé avec acuité.

1.1 Reconnaissance automatique de la parole (RAP)

L'objectif principal de la reconnaissance vocale est de permettre à l'utilisateur un dialogue avec la machine qui soit le plus proche possible de celui qu'il tient avec une personne. Les critères de qualité d'un SRAP ont énormément évolué au fil du temps et peuvent se résumer comme suit:

- permettre une interaction naturelle;
- avoir une tolérance suffisante aux bruits;
- accepter les dialogues multi-locuteurs.

En conséquence du grand nombre d'applications, il est naturel qu'il existe plusieurs types de systèmes de reconnaissance de la parole. Chaque système est adapté pour avoir la meilleure performance (taux de reconnaissance) dans le cadre de sa propre tâche. Selon le domaine d'application, les systèmes de reconnaissance de la parole peuvent être classifiés selon les critères suivants:

- Le système dépendant du locuteur (optimisé pour un locuteur bien particulier) ou indépendant du locuteur (pouvant reconnaître n'importe quel utilisateur) ;
- Le système de reconnaissance de la parole continue, des mots isolés (chaque mot est séparé l'un de l'autre par une pause importante) ou des mots clés (dans ce dernier cas la tache consiste à reconnaitre des mots appartenant un petit vocabulaire bien défini et de rejeter tous les autres mots);
- La taille du vocabulaire: petite (100 mots), moyenne (5000 mots) ou grande (20000 mots et plus).

Il y a aussi plusieurs approches de reconnaissance basées sur des méthodes différentes. Onchoisit le type d'approche en fonction de la tâche de reconnaissance.Les premiers succès en reconnaissance de la parole ont été obtenus à la fin des années 70 à l'aide d'une méthode de reconnaissance par comparaison à des exemples. L'algorithme DTW (Dynamic Time Warping) [**Sakoe et Chiba, 1978**] est utilisé pour ce faire. L'application a plutôt ciblé la reconnaissance des mots isolés avec un petit vocabulaire. Le locuteur qui doit se faire comprendre par la machine prononce un ou plusieurs exemples de chacun des mots susceptibles d'être reconnus et le système les analyse sous la forme d'une suite de vecteurs

acoustiques qui seront ensuite sauvegardés. Ces vecteurs correspondent à la transformation d'un signal de parole en une séquence de symboles représentatifs du contenu de celui-ci. Cette séquence est fréquemment représentée par une suite de coefficients issus de modèles mathématiques. L'étape de reconnaissance vocale, à proprement parler, consiste à:

- convertir le signal vocal inconnu à identifier en une suite de vecteurs acoustiques;
- comparer (par superposition graphique, cas de la DTW) la suite des vecteurs acoustiques obtenus à chacune des suites enregistrées lors de la phase d'apprentissage.

Le taux d'erreur est généralement faible si les conditions suivantes sont rencontrées:

- un enregistrement dans un environnement calme;
- une étape d'apprentissage impliquant chaque locuteur;
- une prononciation bien articulée;
- un vocabulaire relativement restreint: moins de 100 mots.

Il est à noter que lorsque le système de reconnaissance visé prévoit une utilisation par plusieurs personnes ou un plus grand vocabulaire, il est nécessaire d'adopter une granularité de vocabulaire (unités de base) plus fine que le mot en utilisant 'une unité de base de plus petite taille telle que le phonème. La reconnaissance phonétique ne se contente plus seulement d'exemples de prononciation des phonèmes de la langue à modéliser. Elle vise plutôt à déduire un modèle applicable pour n'importe quelle voix et ainsi susceptible de supporter un système multi-locuteur. Dans un système à reconnaissance phonétique, on identifie 4 étapes:

- **le traitement du signal:** l'étape qui produit la suite de paramètres issus de modèles mathématiques formant le vecteur acoustique;
- **la modélisation acoustique:** l'étape qui produit une série d'hypothèses phonétiques pour chaque segment de parole et leur associe une vraisemblance;
- **la modélisation lexicale:** l'étape qui force le reconnaisseur vocal à ne prendre en compte que les mots existants dans la langue considérée;
- **la modélisation syntaxique:** l'étape qui force le reconnaisseur vocal à intégrer les contraintes syntaxiques, grammaticales ou même sémantiques de la langue considérée.

L'approche par reconnaissance phonétique offre de bons résultats pour une parole continue, que l'étendue du vocabulaire soit large ou moyenne, et ceci indépendamment du locuteur, à la condition que le signal de parole analysé soit exempt de bruit.

Au début des années 80, l'utilisation des modèles de Markov cachés (HMM : Hidden Markov Model en anglais) a permis de grands progrès [**Levinson et Rabiner, 1983**]. En principe, cette approche n'est qu'une extension statistique de la méthode déterministe DTW. L'utilisation des modèles HMM a permis aussi de passer aux méthodes de reconnaissance par modélisation d'unités de parole, permettant de modéliser des unités de parole de plus petite taille (typiquement les phonèmes), ce qui est fondamental pour construire des systèmes de reconnaissance de la parole "grand vocabulaire".

Dans ce qui suit, cette section d'état de l'art se concentre sur les SRAP Markoviens utilisant des modèles de langages probabilistes. Nous survolons les principes des différents modèles acoustiques ainsi que les principales méthodes d'apprentissage. Nous terminerons en présentant les paradigmes d'évaluation de ces systèmes.

1.1.1 Fonctionnement général d'un SRAP Markovien

Les systèmes de reconnaissance automatique de la parole continue actuels se basent sur une approche statistique dont Jelinek [**Jelinek, 1976**] a proposé une formalisation, issue de la théorie de l'information. À partir des observations acoustiques X, l'objectif d'un moteur de reconnaissance est de trouver la séquence de mots $\widetilde{W}$ la plus probable parmi l'ensemble des séquences possibles. Cette séquence doit maximiser l'équation suivante :

$$\widetilde{W} = arg\,max_w\, P(W/X) \tag{1.1}$$

En appliquant la théorie de Bayes, l'équation devient :

$$\widetilde{W} = argmax_w \frac{P(X/W)P(W)}{P(X)} \tag{1.2}$$

$P(X)$ ne dépend pas d'une valeur particulière de W, ce qui vient du calcul de l'argmax :

$$\widetilde{W} = arg\,max_w\, P(X|W)P(W) \tag{1.3}$$

Le terme *P(W)* est estimé via le modèle de langage et $P(X|W)$ correspond à la probabilité donnée par les modèles acoustiques. Ce type d'approche permet d'intégrer, dans le même processus de décision, les informations acoustiques et linguistiques (figure1.1).

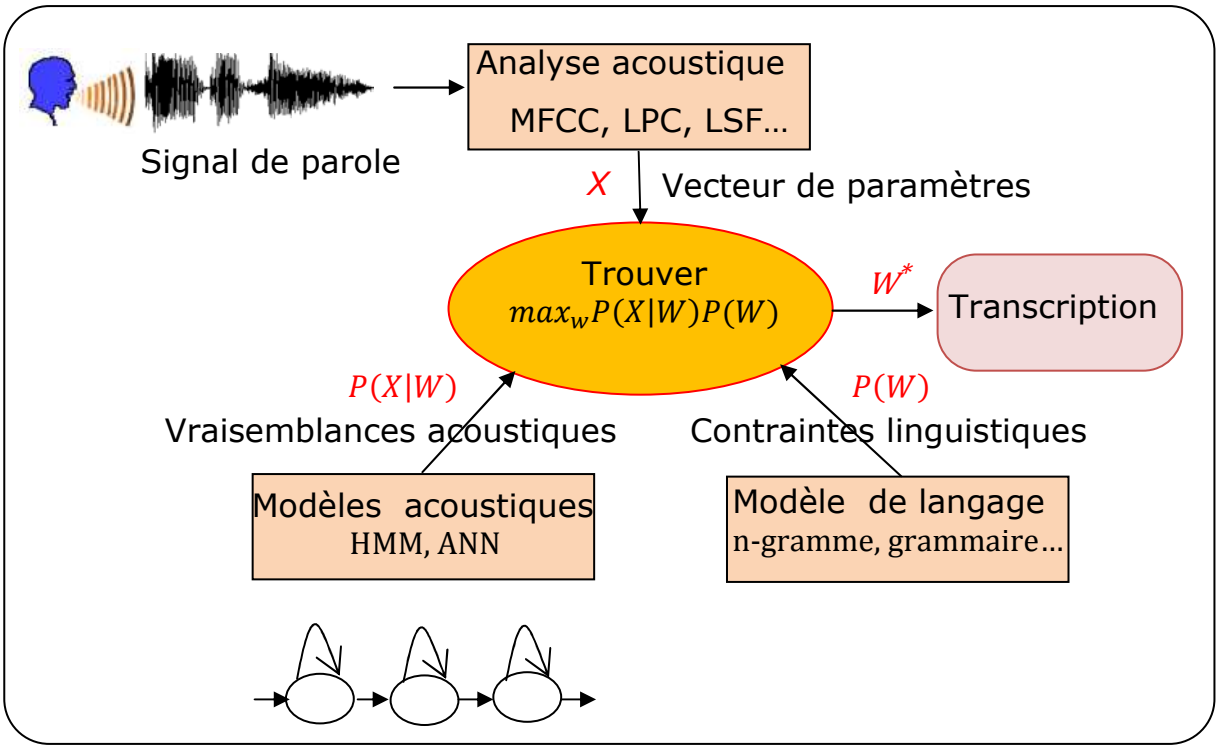


Figure 1.1–Principe générale des SRAP

1.1.2 Modèles et paramètres acoustiques

Le signal de la parole ne peut être exploité directement. En effet, le signal contient de nombreux autres éléments que le message linguistique: des informations liées au locuteur, aux conditions d'enregistrement, etc. Toutes ces informations ne sont pas nécessaires lors du décodage de la parole et rajoutent même du bruit. De plus, la variabilité et la redondance du signal de la parole le rendent difficilement exploitable tel quel. Il est donc nécessaire d'en extraire uniquement les paramètres qui seront dépendants du message linguistique. Généralement, ces paramètres sont estimés via des fenêtres glissantes sur le signal. Cette

analyse par fenêtre permet d'estimer le signal sur une portion du signal jugée stationnaire généralement 10 à 30 ms en limitant les effets de bord et les discontinuités du signal via une fenêtre de Hamming. La majorité des paramètres représentent le spectre fréquentiel et son évolution sur une fenêtre de taille donnée. Les techniques de paramétrisation les plus utilisées sont: LPCC (Linear Prediction Cepstral Coefficients: domaine temporel) [**Markel, 1976**], MFCC (Mel Frequency Cepstral Coefficients: domaine cepstral) [**Davis et Mermelstein, 1980**], PLP (Perceptual Linear Prediction: domaine spectral) [**Hermansky et Cox, 1991**].

1.1.2.1 Modèles de Markov Cachés

Le signal acoustique de la parole est modélisable par un ensemble réduit d'unités acoustiques, qui peuvent être considérées comme des sons élémentaires de la langue. Classiquement, l'unité choisie est le phonème: un mot étant formé par leur concaténation. Des unités plus précises peuvent être employées comme les syllabes, les dissyllabes, les phonèmes en contexte, permettant ainsi de rendre la modélisation plus discriminante, mais cette amélioration théorique est limitée dans la pratique par lacomplexité induite et les problèmes d'estimation. Un compromis souvent employé est l'utilisation de phonèmes contextuels avec partage d'états. Le signal de la parole peut être assimilé à une succession d'unités. Dans le cadre des SRAP Markoviens, les unités acoustiques sont modélisées par des Modèles de Markov Cachés, typiquement des MMC gauche-droite à trois états. A chaque état du modèle de Markov est associée une distribution de probabilité modélisant la génération des vecteurs acoustiques via cet état (figure 1.2). Un MMC est caractérisé par plusieurs paramètres :

- Son nombre d'états N.
- L'ensemble des états du modèle $e = (e_i)_{(1 \le i \le N)}$
- Une matrice de transition entre les états : $A = (a_{ij})_{(1 \le i, j \le N)}$ de taille NxN
- La probabilité d'occupation d'un état à l'instant initial : $(\pi_i)_{(1 \le i \le N)}$: $\pi_i = P(e_1 = e_i)$

La densité de probabilité d'observation b_i associée à l'état e_i : bi qui est généralement modélisée par un modèle à mélange de Gaussiennes

Un MMC est donc représenté par un ensemble de paramètres : $\Phi_{MMC} = (N, A, \{\pi_i\}, \{b_i\})$. Les paramètres du MMC sont estimés empiriquement sur de grands corpus de parole annotés.

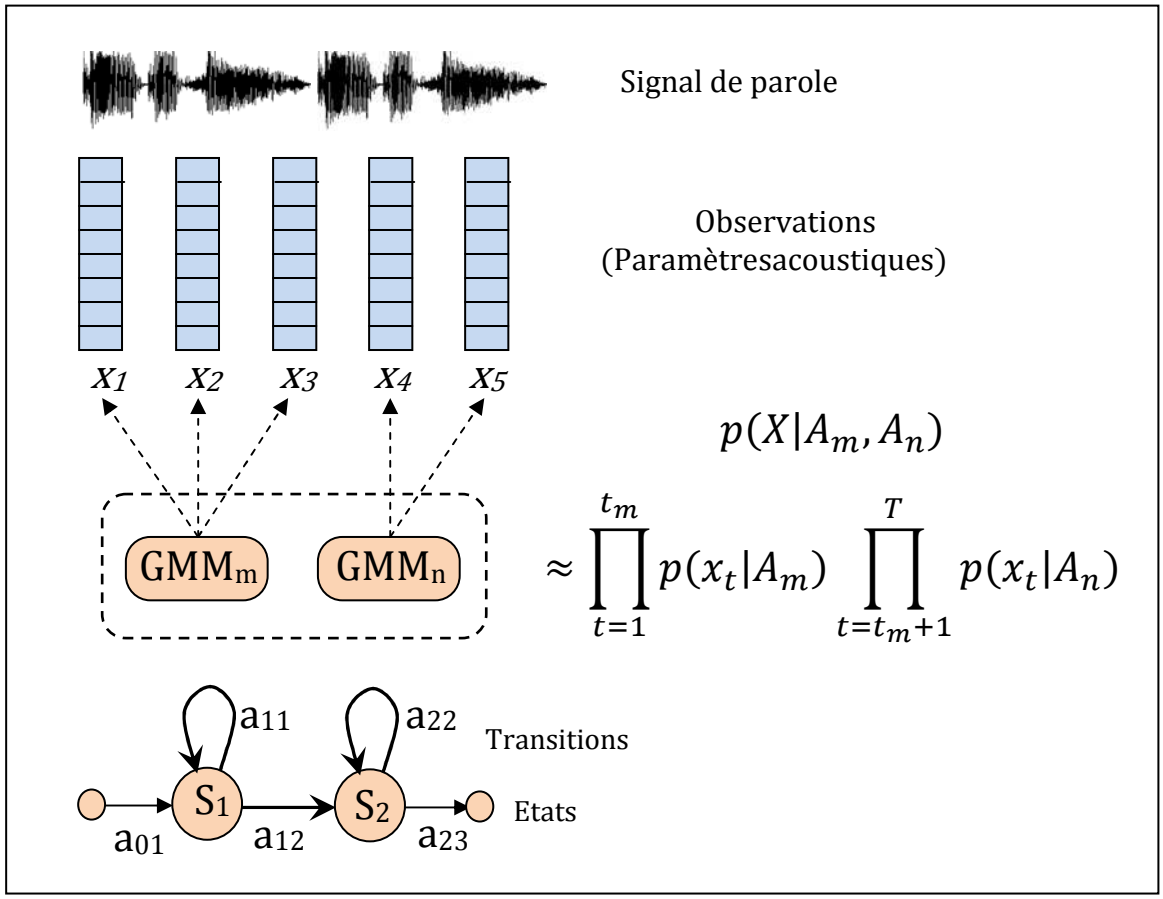


Figure 1.2–Paramétrisation et modèles acoustiques

Nous présentons succintement dans les paragraphes suivants les techniques d'apprentissage et d'adaptation des modèles acoustiques.

1.1.2.2 Apprentissage et adaptation des modèles acoustiques

L'apprentissage des modèles acoustiques d'un SRAP permet de modéliser le messagede la parole avec une quantité de données *a priori* annotées. Les techniques que nous présentons sont celles utilisées les plus couramment pour estimer correctement les paramètres des MMC. Cette partie décrira succintement les principales méthodes d'apprentissage et d'adaptation pour les modèles et paramètres acoustiques.

Apprentissage par maximum de vraisemblance (ML)

L'estimation du maximum de vraisemblance (Maximum Likelihood: ML) est une méthode statistique utilisée pour déterminer les paramètres de la distribution de probabilité d'un échantillon *X* donné. Soit θ l'ensemble des paramètres associés au modèle *m* qui modélise le message *w*. Soit *x* une observation correspondant à l'acoustique du message *w*. Généralement, l'apprentissage consiste à déterminer les paramètres $\tilde{\theta}$ maximisant la probabilité que l'observation *x* soit générée par le modèle *m* :

$$\tilde{\theta} = argmax_{\theta_m} P\big(X = (x|m)\big) = argmax_{\theta} P\big(X = (x|\theta)\big) \qquad (1.4)$$

Le paramètre θ est une variable inconnue à déterminer maximisant la vraisemblance avec l'échantillon *X*.

L'algorithme EM (Expectation et Maximisation)

EM est une méthode de maximisation proposée par [**Dempster et al. 1977**], permettant de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables. L'algorithme EM alterne des étapes d'évaluation de l'espérance (*Expectation*), où la vraisemblance est maximisée en optimisant une fonction qui est l'espérance de la log-vraisemblance sous une distribution conditionnelle sachant les observations $E\left(log\big(P(X,Y;\theta|X,\theta)\big)\right)$, et une étape de maximisation (Maximisation) estimant le maximum de vraisemblance des paramètres, en maximisant la vraisemblance trouvée à l'étape *Expectation*. Les paramètres trouvés grâce à la Maximisation sont réutilisés comme point de départ d'une nouvelle phase d'évaluation de l'espérance: la méthode est réitérée jusqu'à convergence. Cet algorithme se retrouve par exemple dans l'apprentissage des MMC avec l'algorithme de Baum-Welch [**Baum et al. 1970**].

Dans l'équation 1.4, $P\big(X = (x|\theta)\big)$ ne peut être maximisée directement à cause de l'incomplétude des données d'apprentissage. Ce problème est résolu par l'approche EM qui permet en partant de conditions initiales des paramètres θ^0, d'attribuer des valeurs z^l aux données manquantes (*Expectation)* puis de trouver une nouvelle valeur θ^{l+1} des paramètres qui maximisera la vraisemblance des données complètes $P(y^l \mid \theta)$ avec $y^l = (x, z^l)$.

Estimation par information mutuelle maximale (MMIE)

Tandis qu'une estimation ML cherche à maximiser les vraisemblances, MMI (Maximal Mutual Information) est une approche discriminante: le principe général du MMI est de trouver au sein de différentes classes G = {g1, g_2, ..., g_k}, dans un espace vectoriel défini par $F_m = V_1xV_2x...xV_m$, quels sont les paramètres <*X*, g > avec X Є F_m et g Є G qui discriminent le plus ces classes. Il faut donc déterminer quelles sont les composantes de X qui permettent de singulariser chaque classe. La métrique la plus appropriée pour déterminer si une composante peut s'associer à une classe est l'information mutuelle entre les valeurs de la composante et les valeurs contenues dans la classe. Afin de calculer cette quantité, on définit deux variables aléatoires : X_i la composante *i* de X pour un point de données et C la classe d'un point de données. L'information mutuelle entre X_i et C pour tous les points de données est :

$$Im(X_i, C) = H(C) - H(C|X_i) \tag{1.5}$$

Étant donné que P(C) est identique pour toute valeur de i, et que MMI est utilisé pour ordonner, il est suffisant de calculer$P(C|X_i)$:

$$H(C|X_i) = -\sum_{c\epsilon G}\sum_{x_i\epsilon V_i} P(c, x_i) log P(c|x_i) \tag{1.6}$$

Où P(c, x_i) est la probabilité conjointe de voir une donnée de la classe c avec la composante x_i et $P(c|x_i)$est la probabilité d'être dans la classe c de la composante x_i. Dans cette équation, les composantes les plus discriminantes obtiendront le score le plus élevé.

Cette méthode a été introduite par [**Bahl et al. 1986**] afin d'adapter les paramètres de modèles de Markov pour les SRAP. MMIE a été par la suite développée pour les SRAP par [**Valtchev et al. 1997**]. La fonction objective de MMIE est :

$$f(\lambda) = \sum_{r=1}^{R} log \frac{P(w_r)P_\lambda(X_r|M_w)}{\sum_{\widehat{w}} P(\widehat{w}_r)P_\lambda(X_r|M_{\widehat{w}})} \tag{1.7}$$

Où $\widehat{W}$représente toutes les séquences de mots possibles dans la tâche courante, X = X_1, X_2, .., X_r les observations qui correspondent aux mots w_1, w_2, ..,w_R. Les ré-estimations des moyennes et des variances des MMC sont décrites dans [**Gao et al. 2000**].

Autres approches discriminantes : MPE, MWE, MCE

Avec l'augmentation de la puissance de calcul, ainsi que l'amélioration des méthodes d'apprentissage discriminantes telles que MMIE, plusieurs approches se sont développées. MMIE se concentre sur la maximisation des probabilités *a posteriori* des phrases d'apprentissage. MPE (Minimum Phone Error) [**Povey et Woodland, 2002**] et MWE (Minimum Word Error) [**Heigold et al. 2005 ; Yan et al. 2008**] fonctionnent sur un principe similaire à celui du MMI, mais cherchent à minimiser respectivement le taux d'erreur de phonèmes et le taux d'erreur mots. Avec un autre niveau de granularité, a été introduit le MCE (Minimum Classification Error) qui tend à minimiser le taux d'erreur au niveau des phrases.

Adaptation par maximum a posteriori (MAP)

La méthode d'estimation du Maximum *a posteriori* (MAP) peut être utilisée afin d'estimer un certain nombre de paramètres inconnus, comme par exemple les paramètres d'une densité de probabilité, reliés à un échantillon donné. Cette méthode a été introduite dans le cadre de la reconnaissance automatique de la parole par [**Gauvain et Lee, 1994**]. Dans le cas des modèles acoustiques, la méthode d'adaptation MAP permet de modifier les paramètres acoustiques d'un modèle générique pour rapprocher ce dernier du corpus de test. Ceci permet par exemple d'adapter un modèle acoustique générique à un locuteur spécifique. On considère un paramètre q comme étant une variable aléatoire de distribution a priori $P(\theta)$. Le critère de maximum *a posteriori* cherche à maximiser la probabilité *a posteriori* $P(\theta|X)$. En appliquant la règle de Bayes, tout en considérant l'indépendance des échantillons par rapport à θ, l'adaptation de θ consiste à maximiser la valeur de $P(X|\theta)P(\theta)$, soit :

$$\theta_{map} = argmax_{\theta} P(\theta|X) = argmax_{\theta} P(X|\theta)P(\theta) \qquad (1.8)$$

Par ailleurs, l'adaptation MAP est équivalente à un apprentissage par maximum de vraisemblance si la distribution a priori P(θ) est uniforme. L'adaptation MAP obtient de bons résultats et la quantité d'informations nécessaire à l'apprentissage est raisonnable en comparaison d'une approche par maximum de vraisemblance [**Gauvain et Lee, 1994**].

Adaptation par régression linéaire (MLLR)

L'adaptation de modèles par régression linéaire, MLLR (Maximum Likelihood Linear Regression) [**Gales 1997**] est également une méthode communément utilisée pour modéliser des données *a priori*. Dans le cas de l'adaptation MLLR, l'hypothèse est que les paramètres cibles peuvent être obtenus via une transformation linéaire des paramètres initiaux (figure1.3):

$$\hat{\mu} = A_i \mu_i + b_i \, . \tag{1.9}$$

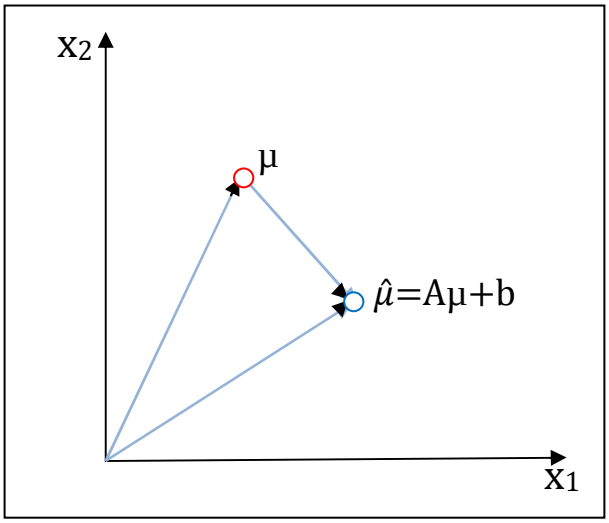


Figure1.3–Apprentissage des paramètres via la régression linéaire MLLR

Avec $\hat{\mu}$ le vecteur cible, μ le vecteur initial, A_i la matrice de transformation et b_i le vecteur d'adaptation. Étant donné une séquence d'observation $X = \{x_1, ..., x_N\}$, l'adaptation MLLR doit trouver le nouvel ensemble de paramètres $\theta = \{A_i, b_i\}_{i=1}^{I}$ qui maximise la vraisemblance $P(X|\theta)$:

$$\hat{\theta} = argmax_{\theta} P(X|\theta). \tag{1.10}$$

Si des données d'apprentissage ne sont pas étiquetées, l'algorithme EM peut être appliqué pour obtenir un ensemble optimal de paramètres. L'une des contraintes de l'adaptation MLLR est la quantité très importante de paramètres à estimer. Une réduction du nombre de paramètres à calculer a été introduite pour résoudre ce problème: la régression par classes.

Cependant l'apprentissage MLLR demande moins de données d'apprentissage que MAP ou ML lorsque le nombre de classes est réduit.

Quelques techniques d'adaptation des paramètres acoustiques

L'apprentissage fMLLR (feature MLLR) présenté par [**Gales, 1997**], contrairement aux méthodes qui modifient les modèles acoustiques (comme la méthode MLLR), s'applique sur les paramètres directement issus de l'observation. Ainsi, les modèles ne sont pas modifiés, et les paramètres sont rapprochés de ceux de l'apprentissage, et ce tout en modifiant leur espace de représentation. Ainsi, des caractéristiques particulières du signal seront dé-bruitées (locuteur, bruit etc.).

Une autre technique présentée par [**Zhan et Waibel, 1997**], VTLN (Vocal Tract Lenght Normalization), s'applique tout comme MLLR sur les paramètres. Cet apprentissage s'appuie sur une normalisation du conduit vocal, qui diffère d'un locuteur à l'autre. Par cette adaptation, les variations de longueur sont éliminées par des filtres modifiant les fréquences.

1.1.3 Évaluation d'un SRAP

Les SRAP sont souvent évalués en termes de taux d'erreur sur les mots, (WER : Word Error Rate). Le WER est basé sur une mesure résultant de la programmation dynamique. L'hypothèse reconnue par le SRAP est alignée avec une hypothèse de référence via un algorithme d'alignement dynamique. Le WER se calcule donc :

$$WER = \frac{S + D + I}{N} . 100, \tag{1.11}$$

Où S correspond aux substitutions, D aux suppressions, I aux insertions et N est le nombre de termes dans la référence.

D'autres métriques ont été introduites, notamment dans le but d'estimer la fidélité sémantique des transcriptions réalisées [**Sarikaya et al. 2005** ; **San-Segundo et al. 2001**], pour des systèmes d'interprétation de dialogue et d'indexation.

Afin d'évaluer la fiabilité de ces mesures statistiques, il convient de calculer un intervalle de confiance relatif au nombre d'échantillons et d'erreurs. Cet intervalle de confiance est calculé en considérant que l'apparition d'une erreur de reconnaissance sur un mot ou sa non

reconnaissance est associée à une variable aléatoire binomiale, dont la distribution dépend des couples (mot reconnu, mot prononcé).

Dans la situation de combinaison de systèmes, il est nécessaire d'estimer la fiabilité des résultats, qui dépend directement de la quantité d'échantillons utilisés. L'intervalle de confiance peut se calculer en supposant qu'une erreur de reconnaissance dépend d'une variable aléatoire binomiale dépendant des couples {*mot reconnu, mot prononcé*}.Dans la littérature, un intervalle de confiance est proposé par [**Saporta, 1990**] :

$$wer_f - u_{\frac{\alpha}{2}}\sqrt{\frac{wer_f(1-wer_f)}{k}} < wer_p < wer_f + u_{\frac{\alpha}{2}}\sqrt{\frac{wer_f(1-wer_f)}{k}}, \qquad (1.12)$$

Où k est le nombre d'échantillons, wer_f est la quantité d'erreurs obtenues sur le corpus de test. α permet de définir l'intervalle de confiance : si a = 95%, alors wer_p sera défini avec une confiance de $\pm$0.05%. $u_{\alpha/2}$ est défini par une table de *Student* : $u_{0.425} = 1.96$.

1.2 Reconnaissance robuste de la parole (RRP)

Les performances des systèmes actuels de reconnaissance automatique de la parole sont satisfaisantes lorsque les systèmes sont évalués sous des conditions contrôlées de laboratoire. Cependant, ces systèmes sont généralement peu robustes, c'est-à-dire que des variations du signal entre les conditions de test et d'apprentissage peuvent provoquer une dégradation significative des taux de reconnaissance, même si ces variations semblent minimes à l'oreille. Les principales sources de variabilité du signal, qui rendent difficile la conception de SRAP robustes, peuvent être classées selon leur provenance, qu'il s'agisse de l'environnement acoustique, de l'équipement d'acquisition du signal, ou encore du locuteur. Le signal est alors perturbé par le bruit ambiant (stationnaire ou non), les distorsions (linéaires ou non) provenant du canal de communication, et les habitudes articulatoires du locuteur. Notons que les séparations entre les différentes classes ne sont pas toujours nettes, l'environnement pouvant par exemple influencer le mode de production de parole. Le tableau 1.1 résume ces différentes sources de variabilité.

TABLEAU 1.1 – Principales causes de variabilité du signal de parole (d'après [**Furui, 1992**])

Environnement	– Bruit corrélé à la parole : réverbération, réflexion – Bruit non corrélé à la parole : bruit additif (stationnaire, non stationnaire)
Locuteur	– Attributs du locuteur : sexe, âge, dialecte. – Mode d'expression : soufflement, bruit des lèvres, stress, effet Lombard, rythme d'élocution, puissance sonore, fréquence fondamentale, locuteur coopératif.
Conditions d'enregistrement	– Microphone – Distance au micro – Filtrage – Matériel de transmission : distorsion, bruit, écho – Matériel d'enregistrement

L'environnement perturbe le signal de parole sous la forme d'un bruit acoustique, que l'on suppose généralement additif. Cette hypothèse est souvent utilisée, à la fois pour sa simplicité, mais aussi car elle permet de couvrir un grand nombre de situations pratiques. Le signal enregistré est donc considéré comme la somme du signal de parole produit par le locuteur et du bruit ambiant.

Les autres types de bruits, tels que les bruits électriques et bruits de quantification sont négligeables dans les applications de RAP. Dautrich et al. [**Dautrich et al. 1983**] sont parmi les premiers à constater la chute des performances d'un système de RAP entraîné dans des conditions calmes et testé dans le bruit: le taux d'erreur de reconnaissance du système, entraîné sur de la parole propre (RSB >40 dB) est multiplié par dix lors d'un test sur de la parole bruitée (RSB = 18 dB). Depuis, la littérature fournit une pléthore d'observations analogues, et à titre d'exemple nous indiquons (figure 1.4), l'évolution des taux de reconnaissance en fonction du niveau de bruit présent lors du test, pour un système de

reconnaissance de parole continue entraîné à partir de parole propre. Une telle évolution s'avère caractéristique du problème de la reconnaissance de la parole dans le bruit.

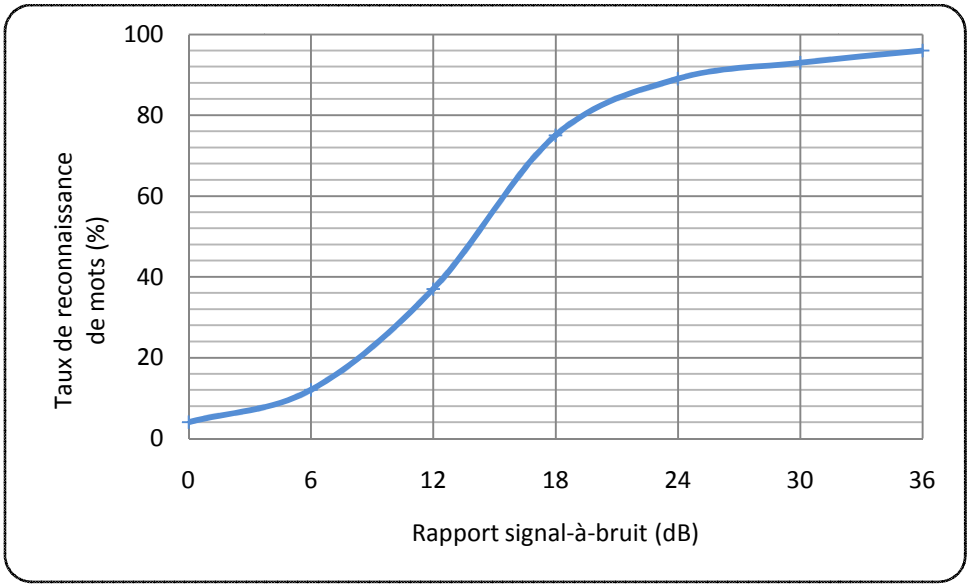


Figure 1.4– Évolution du taux de reconnaissance d'un SRAP entraîné en milieu calme (RSB>40 dB), en fonction du RSB lors du test [**Siohan et al. 1995**].

Les systèmes de communication orale sont généralement utilisés dans des environnements bruités: usines, lieux publics, habitacle de voiture, d'avion, parole téléphonique, etc. Plus les systèmes de RAP seront robustes, plus le nombre de leurs applications potentielles augmentera. L'absence de robustesse au bruit apparaît comme le principal obstacle au développement commercial de telles applications. L'amélioration de la robustesse des systèmes est donc un thème majeur de recherche, faisant appel à des connaissances pluridisciplinaires (traitement du signal, reconnaissance des formes, intelligence artificielle), essentiel pour permettre le développement d'applications en environnements réels.

1.2.1 Représentations robustes pour la RAP

Effectuer l'apprentissage d'un système de reconnaissance de parole dans des conditions bruitées est une opération rarement envisageable en pratique. En effet, il est d'une part très

coûteux d'enregistrer un corpus d'apprentissage dans le bruit. De plus, il est difficile de prévoir lors de l'apprentissage quelles seront les conditions de bruit lors de l'utilisation du système. Dans un tel cas, il devient également délicat d'autoriser une variation du RSB ou du type de bruit lors du test. Le problème à aborder est donc le suivant : étant donné un système de reconnaissance de parole continue entraîné à partir de parole propre, quelles méthodes et techniques peut-on mettre en œuvre pour améliorer la robustesse au bruit du système, c'est-à-dire pour que le système reconnaisse correctement de la parole prononcée en environnement réel, *a priori* inconnu?

Les performances d'un SRAP subissent de lourdes dégradations lorsqu'il est utilisé dans un milieu acoustique (de test) qui diffère de son milieu d'apprentissage. La différence entre ces deux milieux est, la plupart du temps, provoquée par des sources de bruits qui interagissent avec signal de parole de test. Ces sources évoluent au cours du temps et parfois rapidement à l'échelle d'une phrase. De plus, très peu d'informations sont disponibles quant à leur nature, ce qui rend la modélisation de cette interaction très difficile. Le combat contre le bruit se réalise à tous les niveaux, aussi bien dans l'analyse acoustique du signal que dans les méthodes de décodage phonétique, et de très nombreuses méthodes dites "robustes" existent donc dans la littérature. Un état de l'art de ces méthodes peut être trouvé dans [**Junqua, 1996**]. Nous nous contentons de présenter rapidement ici celles qui sont les plus répandues ou qui ont un lien direct avec notre travail.

Une première catégorie de méthodes robustes tente d'éliminer le bruit du signal avant de l'envoyer à un reconnaisseur automatique. Ces méthodes sont dites de "débruitage". En fait, le débruitage n'est pas seulement intéressant pour la reconnaissance de la parole, mais aussi pour l'intelligibilité humaine: ainsi, un certain nombre de recherches se consacrent exclusivement à enlever le bruit du signal de façon à le rendre plus compréhensible et "agréable" à l'oreille. Ceci peut être appliqué par exemple dans un réceptacle téléphonique ou dans les implants auditifs qui aident les malentendants à percevoir la parole. La plus connue des méthodes de débruitage est celle dite de "*soustraction spectrale*" [**Boll, 1979**] qui estime le spectre du bruit dans les zones de silence puis qui soustrait ce spectre à celui du signal bruité. Cette méthode de base a connu de nombreuses améliorations, comme par exemple celle proposée par Ephraïm et al. [**Ephraïm, 1995**]. Le principe de base de cette dernière consiste à utiliser un autre espace de projection que le spectre, ce nouvel espace étant défini par les vecteurs

propres de la matrice de covariance du signal. Néanmoins, si ces méthodes permettent généralement d'améliorer l'intelligibilité du signal, elles ne sont pas systématiquement bénéfiques pour les taux de reconnaissance des SRAP.

Une autre catégorie de méthodes acceptent au contraire la présence de bruit dans le signal et tentent de créer des modèles ou des algorithmes de décodage qui utilisent essentiellement l'information du signal et non du bruit. C'est typiquement le cas des systèmes Multi-Bandes ou des systèmes hybrides combinant les réseaux de neurones et les HMM [**Bourlard, 1994**]. Les systèmes de cette catégorie n'ont pas pour but de modéliser le bruit, mais essayent plutôt de s'en accommoder.

Une troisième catégorie tente au contraire de modéliser le bruit en même temps que la parole. M. J. Gales [**Gales, 1998**] a ainsi conçu un algorithme dit de "Combinaison Parallèle de Modèles" (CPM), qui modélise la parole non bruitée par un HMM et le bruit par un autre HMM. Ces deux HMM sont ensuite combinés afin de décoder le signal en autorisant les deux modèles à cohabiter. Malheureusement, cette méthode traite essentiellement des bruits stationnaires, i.e. qui ne varient pas " beaucoup" avec le temps, et, comme pour la soustraction spectrale, elle nécessite la connaissance *a priori* du bruit ou au moins d'une partie du signal ne contenant que du bruit pour pouvoir le modéliser.

Enfin, les deux dernières catégories de méthodes sont celles sur lesquelles s'appuie cette thèse. La première définit de nouveaux paramètres qui sont moins influencés par le bruit que les paramètres classiques comme les MFCC. Les plus connus de ces paramètres sont les RASTA-PLP [**Hermansky, 1994**], mais d'autres paramètres utilisant des filtres spécifiques, comme les filtres de Wiener [**Junqua, 1996**] sont également utilisés. Nous pouvons considérer que les méthodes s'appuyant sur la théorie des données manquantes [**Lippmann, 1997a**] se classent aussi dans cette catégorie. La seconde classe de méthodes qui fonde notre approche est basée sur décomposition en sous-espaces du signal de parole. Le but étant de développer un estimateur linéaire non paramétrique du signal de parole propre, obtenu par décomposition du signal observé en deux sous-espaces orthogonaux: le sous-espace signal et le sous-espace bruit. La décomposition est achevée soit par valeurs singulières SVD ou par valeurs propres EVD. Le principe des méthodes de décomposition en sous-espaces du signal, décrit ultérieurement, se fera premièrement en supposant que le bruit est additif, blanc et décorrelé de la parole, et deuxièmement, en raisonnant par rapport à une décomposition en

valeurs propres [**Loizou, 2007**]. La réduction du bruit par cette approche est obtenue par annulation des composantes du sous-espace bruit en premier lieu et en supprimant la contribution du bruit dans le sous-espace signal en second lieu.

1.2.2 Méthodes d'analyses robustes

Nous proposons de classer (Figure 1.5) dans ce type de méthodes les analyses basées sur le modèle d'oreille, ou les analyses tentant d'extraire les caractéristiques les plus pertinentes du signal. L'homme a des capacités très efficaces pour percevoir un signal en faisant abstraction du bruit environnant. C'est pourquoi il est naturel qu'un grand effort soit investi afin d'exploiter les connaissances existantes en perception humaine.

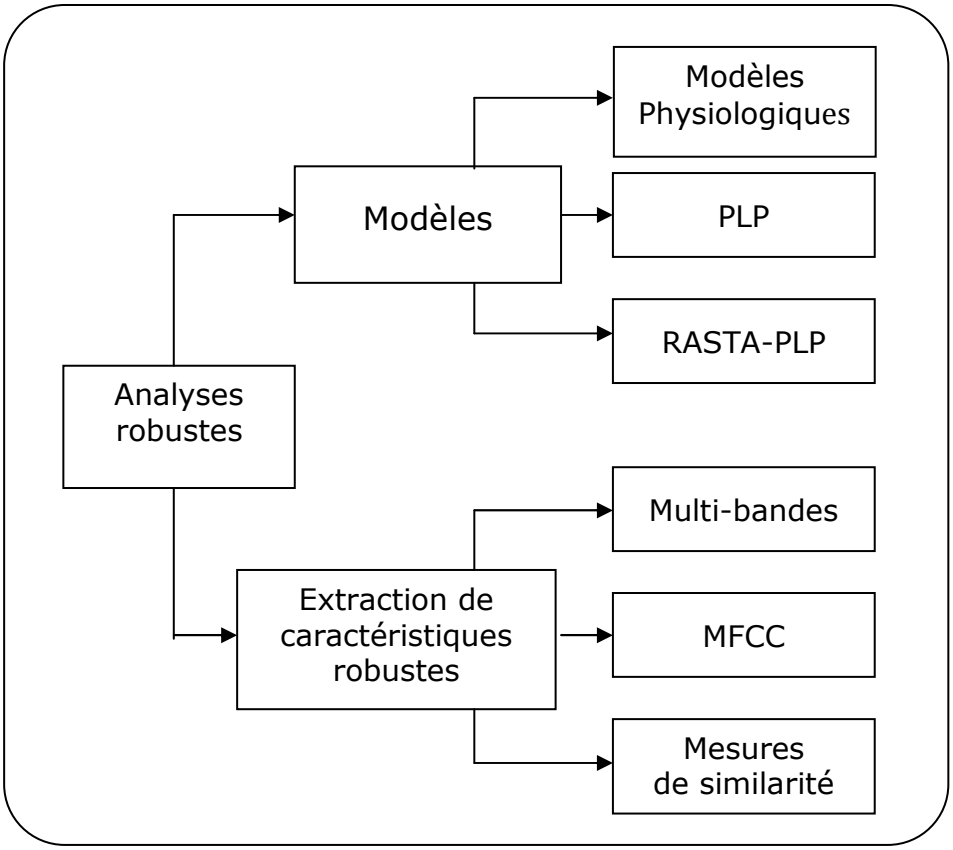


Figure 1.5– Méthodes d'analyses robustes

Contrairement aux SRAP, l'être humain possède des capacités d'analyse phonétique du signal remarquablement robustes aux variations acoustiques dépendant du locuteur, des canaux de transmission, de l'environnement, etc. L'homme est capable de se focaliser sur une source sonore en faisant abstraction de l'environnement. Bien que des facteurs de plus hauts niveaux soient en jeu dans ce processus complexe, il est fort probable que le système auditif ait un rôle important [**Hermansky, 1997**].

Différents modèles d'oreille fonctionnels sont utilisés pour l'analyse acoustique du signal, s'inspirant des capacités du système auditif à percevoir les signaux désirés dont le modèle de Caelen [**Caelen, 1985**] et les modèles de Hermansky [**Hermansky, 1990 et 1997**].

Les coefficients PLP [**Hermansky, 1990**] sont basés sur les concepts psycho-acoustiques connus (intégration des bandes critiques définies par Fletcher, préaccentuation par la courbe d'isotonie et loi de Stevens), et combinent plusieurs techniques d'approximation simulant les caractéristiques de l'audition humaine. Ces méthodes permettent d'améliorer considérablement la robustesse des SRAP sans nécessiter un calcul aussi coûteux que les autres méthodes. La méthode RASTA-PLP [**Hermansky 1997**] est une amélioration de PLP. Celle-ci consiste à ajouter un filtre passe-bande afin de réduire l'influence des bruits convolutifs dû au canal de transmission.

Fletcher a suggéré que le processus de reconnaissance humain soit basé sur des sous-bandes de fréquences qui sont traitées indépendamment les unes des autres [**Allen, 1994**]. Ainsi le décodage d'un message linguistique se fait en parallèle, et de manière indépendante, dans différentes bandes de fréquences, et la décision finale est effectuée en mettant en commun les informations obtenues par chaque analyse. Bourlard et Tibrewala [**Bourlard et Dupont, 1997**] [**Tibrewala et Hermansky, 1997**] ont proposé de décomposer le signal en plusieurs bandes de fréquences et d'associer un système de reconnaissance par sous-bande. Cette méthode présente différents avantages: les sous-bandes sont en quelque sorte spécialisées dans la reconnaissance de certaines classes de sons, et le bruit peut être en partie ignoré en donnant moins d'importance aux sous-bandes dans lesquelles il se trouve [**Dupont et al. 1997**]. Une difficulté apparaît en revanche, lors de l'étape de fusion des sorties pour laquelle il est difficile de trouver les bonnes pondérations.

Enfin, l'analyse MFCC [**Davis et Mermelstein, 1980 ; Junqua, 1987**], ainsi que diverses techniques consistant à tirer partie des caractéristiques du bruit afin d'en limiter l'impact sur le signal de parole [**Martinez et al. 1997**], sont largement étudiées.

Aussi, nous avons fait un état de l'art des méthodes traitant le signal bruité afin que les paramètres découlant de l'analyse acoustique soient le plus proches de ceux reconnus par le système. Ces méthodes font en sorte que le système de reconnaissance n'ait pas à s'adapter aux nouvelles conditions acoustiques, vu les limites de l'analyse cepstrale.

En comparant les méthodes de représentation du signal (MFCC, LPC, PLP etc. …), nous remarquons que les MFCC sont les plus populaires. En effet, depuis plusieurs années, le codage par MFCC est considéré comme la meilleure méthode d'analyse acoustique, quoique les recherches aient montré certaines limites. Parmi ces limites, nous citons:

- Un triplement de la taille des vecteurs de paramètres gérés par les systèmes de reconnaissance vocale comparativement aux systèmes de voix sur IP par exemple. Cette augmentation est coûteuse en ressources de calcul et de mémoire et introduit une redondance de l'information représentée.
- La perte de certaines informations pertinentes lorsque le fonctionnement a lieu dans des environnements bruités.

En même temps, les coefficients MFCC permettent de prendre explicitement en compte la dynamique du signal dans les systèmes de reconnaissance à base de HMM. Ils restent ainsi un choix raisonnable pour les utiliser dans l'analyse acoustique par approche multi-variable.

1.3 Domaines d'applications et problématiques en RAP

Ce mémoire traite essentiellement de la reconnaissance robuste de la parole (RRP). Nous allons donc tout d'abord présenter brièvement les autres domaines de recherche connexes à celui qui nous intéresse (RRP). La RAP est une partie d'un domaine de recherche plus vaste que l'on nomme habituellement le traitement automatique du langage naturel (TALN). De nombreux autres domaines de recherche font également partie du TALN : nous pouvons ainsi citer la reconnaissance automatique du locuteur, l'identification du langage, la synthèse de la parole, le codage et la compression de la parole, mais aussi les systèmes de dialogue,

l'indexation verbale de documents, la modélisation sémantique des textes en vue de leur incorporation en RAP, etc. La RAP emprunte aussi largement à un autre domaine de l'intelligence artificielle qui est la reconnaissance des formes. Ce vaste domaine comprend également la reconnaissance de caractères (manuscrits ou non) et l'analyse des images ou des scènes visuelles. De nombreuses techniques sont partagées par la reconnaissance de la parole et la reconnaissance des images.

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

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

2 **La télématique vocale** : la plupart des grandes compagnies de télécommunications se tournent actuellement vers ce domaine afin de remplacer leurs opérateurs humains par des systèmes automatiques. Leurs rôles restent pour l'instant très simples, par exemple

donner le numéro de téléphone d'un abonné. Ceci est cependant déjà très intéressant pour ces compagnies, car le coût horaire d'un opérateur humain est très élevé comparé à celui d'un logiciel et des économies substantielles peuvent être réalisées de la sorte. De plus, la qualité du service de renseignement téléphonique peut être améliorée car le même logiciel peut facilement gérer plusieurs appels en même temps et l'attente pour les usagers est donc moins grande.

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

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

La RRP est donc très utile dans de nombreux domaines. C'est sans doute pour cette raison qu'un effort financier et humain très important a été consacré à cette recherche au cours de ces dernières années. Aujourd'hui, certains laboratoires s'orientent déjà vers la compréhension du langage naturel du point de vue sémantique, avec tous les problèmes de modélisation que cela pose. Toutefois, nous pensons qu'il ne faut pas pour autant délaisser le niveau acoustico-phonétique en recherche. En effet, il est évident que de nombreux progrès sont encore à réaliser dans ce domaine, notamment au vu des résultats actuellement trop médiocres des systèmes de reconnaissance en milieu bruité. C'est pourquoi la recherche doit encore expérimenter de nouveaux modèles acoustiques, comme nous essayons de le faire dans ce mémoire.

Les études actuelles tentent de trouver des réponses aux problèmes suivants:

- la résistance à l'environnement (bruit, musique, autre locuteur, etc.);
- l'adaptation à la réverbération ou aux caractéristiques du microphone;

- la résistance aux variations de qualité lors de la transmission (réseau IP, réseau téléphonique RTC);
- l'adaptation aux conditions d'élocution: stress, bruit de la respiration, vitesse d'élocution inhabituelle, effet Lombard résultant de problèmes de surdité;
- la modélisation des contraintes syntaxiques et sémantiques de la langue.

L'une des difficultés majeures liées à la RRP provient des nombreuses causes qui peuvent altérer le signal sonore émis. La représentation acoustique d'un phonème dépend fortement du contexte dans lequel il apparaît et doit être prononcé. Quelques exemples illustrent cette problématique:

- l'acoustique peut être différente selon l'environnement sonore ou encore la position et les caractéristiques du microphone utilisé;
- la voix du locuteur peut être différente selon son état physique, émotionnel ou la vitesse de son élocution;
- l'historique sociolinguistique, les dialectes, etc. font fortement varier l'élocution d'une personne.

1.4 Conclusion

Ce chapitre présente une analyse bibliographique sur les travaux en reconnaissance de la parole en milieu bruité et les caractéristiques des méthodes pour obtenir une reconnaissance robuste. Cette notion de robustesse est l'un des problèmes majeurs qui limitent l'utilisabilité de ces systèmes. Et comme le signal de parole se caractérise par une grande redondance, il est à noter que ce n'est pas toutes les informations contenues dans le signal de parole qui sont considérées comme utiles lorsqu'il s'agit d'une application particulière. Ainsi, un traitement efficace de la parole doit considérer l'information utile qui est en lien avec l'application visée.

L'une des finalités de la recherche était d'examiner une solution qui consiste à représenter le signal acoustique en termes de paramètres ou caractéristiques qui sont conçus spécifiquement pour l'application souhaitée. Pour ce faire, le défi est de trouver une méthode fiable qui peut extraire ces paramètres tout en préservant les informations pertinentes. Notre démarche de représentation s'inscrit donc dans une problématique de l'évolution de la communication dans des environnements acoustiques changeants. Dans le chapitre suivant, on décrit l'approche

acoustique multi-variable pour la DSR. L'idée de performance d'un système DSR consiste à agir au niveau de l'analyse acoustique du signal vocal afin de trouver le type de paramètres le mieux adapté aux milieux bruités et plus particulièrement ceux induits par le milieu de transmission sans fil.

Chapitre 2

Intégration des paramètres multiples dans un système DSR

Chapitre 2

Intégration des Paramètres Multiples dans un Système DSR

Sommaire

Il est question d'ouvrir le champ des interfaces hommes-machines utilisant la reconnaissance automatique de la parole à des équipements mobiles de petite taille et aux capacités de calcul et de transmission réduites. Dans ce cadre s'est développée une architecture appelée "Reconnaissance Vocale Distribuée" dans laquelle, pour des raisons de robustesse du SRAP, une partie seulement du processus de reconnaissance vocale est effectuée sur l'équipement mobile; le reste du processus étant réalisé sur un "serveur RAP". L'information résultant de l'analyse effectuée dans le mobile doit donc être transmise au serveur après codage (compression bas-débit, quelques kbit/s) et protégée contre les erreurs de transmission.

Dans ce chapitre, nous présenterons, dans un premier lieu, un aperçu de la norme DSR Front-end Mel-Cepstre. Nous introduirons, par la suite, l'analyse acoustique multi-variable et nous mettrons l'accent sur les principaux travaux de recherche dans ce cadre. Et enfin, nous présenterons notre propre vision ainsi que nos défis au cours de ce projet.

2.1 Traitement distribué de la parole

Du fait de l'utilisation intensive des TIC[1], les technologies vocales sont désormais des applications à part entière. Les logiciels associés font de plus en plus appel aux ressources des réseaux IP. Le but recherché par tout système intégrant une interface vocale est de permettre un accès aisé à un large ensemble de services informatisés sans la nécessité de devoir écrire ou d'avoir un clavier à proximité. Dans le secteur du développement informatique, l'architecture logicielle client/serveur a largement fait ses preuves.

Cette approche de développement, appliquée aux systèmes vocaux, a donné naissance aux outils de dernière génération:

- *Distributed Speech Recognition (DSR)*: système distribué de la reconnaissance vocale;
- *Distributed Text To Speech (DTTS)*: système distribué pour la synthèse vocale.

L'approche mode client/serveur permet de donner aux applications vocales (ASR, TTS) un niveau coût/performance acceptable, rendant ainsi celles-ci plus accessibles. Le bénéfice de cette synergie est double:

- *gain financier*: Partage d'infrastructure matérielle et logicielle côté serveur;
- *adéquation aux performances mobiles*: Décharger vers le serveur les besoins du client en matière de puissance de calcul ou d'autonomie d'énergie.

Le spectre des applications est vaste mais prend particulièrement son sens dans le contexte du mobile. Dans le cas des GSM, PDA ou autres équipements à autonomie d'énergie et puissance de calcul réduites, il est évidemment intéressant de décharger le terminal d'un maximum de tâches liées aux technologies vocales.

Il suffit de prendre l'exemple de l'utilisation comme dictaphone d'un GSM disposant d'un module ASR "en ligne" et permettant ainsi à l'utilisateur de retrouver immédiatement à son arrivée au bureau sous forme textuelle les notes prises au vol lors d'un déplacement.

[1]TIC: Technologies d'Information et de Communication

2.1.1 Reconnaissance de la parole distribuée

Les systèmes DSR peuvent se catégoriser en deux groupes:

- *DSR avec client simple*: Le serveur capture la voix numérique "brute" émise par l'équipement mobile (le client) par le simple enregistrement vocal de la communication. Tout le traitement vocal se réalise côté réseau, le client n'est utilisé que comme microphone (figure 2.1);
- *DSR avec client évolué*: Le traitement vocal est partagé entre le client et le réseau. Le terminal réalise certaines tâches comme l'extraction des caractéristiques (les vecteurs acoustiques) afin de transmettre celles-ci au serveur en lieu et place de l'envoi du canal sonore.

C'est le second modèle qui a actuellement le vent en poupe. En effet, les performances des systèmes de reconnaissance vocale sont fortement dégradées lorsque le traitement vocal ne peut être appliqué sur le signal sonore non modifié. Lorsque la voix est transmise au travers de canaux de communication mobile, la qualité du signal est dégradée par la numérisation et l'équivalent numérique récupéré par le réseau est souvent de faible qualité! Cette dégradation du signal est due à l'utilisation d'un codec (codeur-décodeur) vocal bas débit, ainsi qu'aux erreurs de transmission sur le réseau.

Les systèmes DSR performants résolvent ces problèmes en n'utilisant plus la voix numérique émise par le mobile, mais bien un canal de données fiables au travers duquel on transmet une représentation paramétrée de la parole qui sera exploitable par le système de reconnaissance vocale. Dans cette dernière architecture, le traitement est réellement distribué entre le terminal et le réseau (figure 2.2):

- Le terminal joue le rôle de système d'extraction d'informations acoustiques et transmet celles-ci à l'ASR. Le terminal est alors appelé le *front-end*;
- Le serveur joue le rôle de reconnaisseur en traitant les informations acoustiques reçues du terminal. Le serveur est alors dénommé le *back-end*.

Par le biais de ce processus, le canal de transmission n'affecte plus les performances du système de reconnaissance vocale.

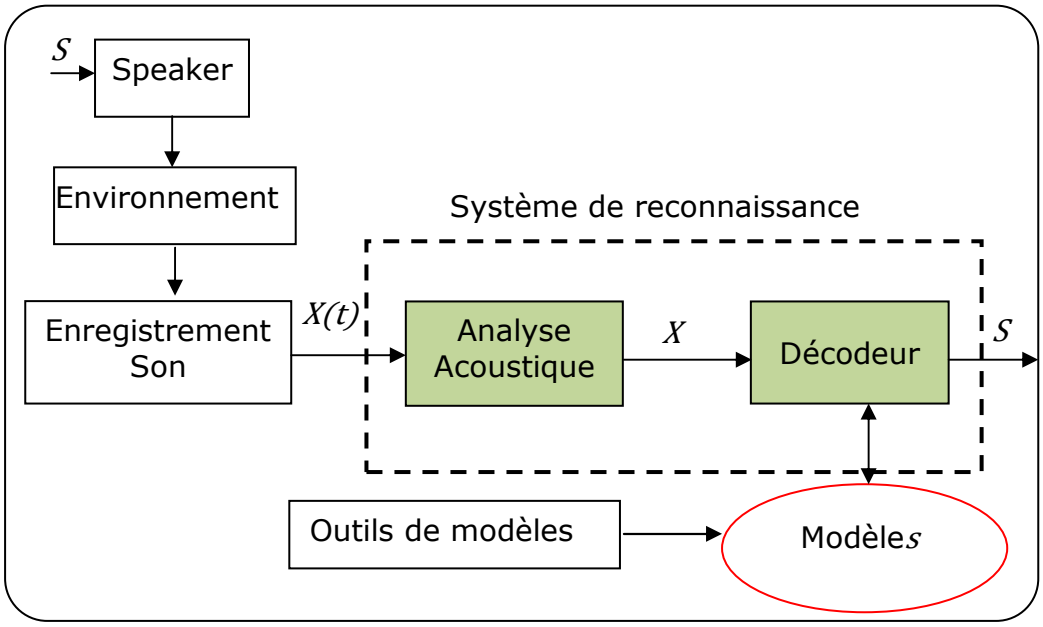


Figure 2.1–Implémentation de la reconnaissance dans le terminal ou réseau

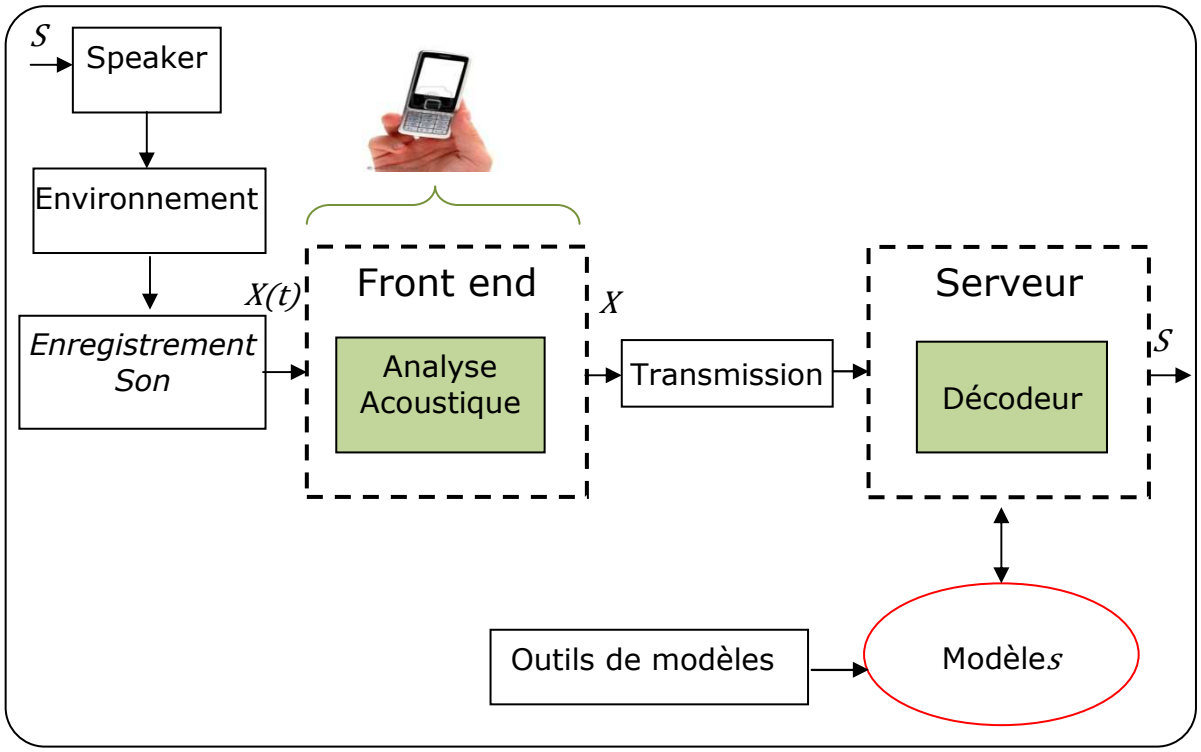


Figure 2.2–Implémentation de la reconnaissance distribuée

Le marché du mobile est large et comprend de nombreux acteurs: fabricants de terminaux, opérateurs de télécoms, fournisseurs de services et éditeurs de logiciels de reconnaissance vocale. Afin de pouvoir mettre en œuvre un système DSR sur un tel marché, une norme commune pour le *front-end,* assurant la compatibilité entre les terminaux et les reconnaisseurs vocaux, est indispensable. Aurora est une norme d'extraction, standardisée par l'ETSI, qui répond à cette demande.

2.1.2 Le standard ETSI Aurora

Dans le projet Aurora, le concept étudié est celui de reconnaissance de la parole distribuée, les travaux étant actuellement menés au sein de l'ETSI. Ces travaux ont déjà donné lieu à la normalisation de l'étage de prétraitement des systèmes de reconnaissance. Une première norme a été publiée en 2001 correspondant à un système de base autour d'un codage MFCC standard à un débit de 4.8 kb/s. On peut ajouter: les travaux menés à l'IETF par le groupe de travail CATS[2] pour normaliser le protocole permettant de contrôler des ressources vocales distribuées. D'autres travaux sont menés au 3GPP pour intégrer le concept DSR dans les réseaux de télécommunication de 3éme génération.

ETSI publie les algorithmes pour la reconnaissance vocale distribuée qui permettent l'accès aux services et aux systèmes de communication sans avoir à taper ou utiliser un clavier. L'algorithme de base, qui se trouve dans les ES 201 108 [**ETSI 2003**] permet l'émission de la parole sur des liens de faible qualité (comme la radio mobile) et convertie en texte pour interagir avec les systèmes automatisés.

Le *front-end* standard Aurora utilise 12 paramètres cepstraux statiques et un paramètre d'énergie calculés toutes les 10 ms. Un étage de quantification est appliqué aux paramètres cepstraux. Le vecteur de paramètres est étendu en ajoutant les dérivées premières et secondes des cepstres. Ce vecteur de dimension 39 est notée MFCC_E_D_A(39)[3].

[2]CATS : Control of ASR and TTS Resources

[3] MFCC_E_D_A(39) :13Paramétres statiques+13Δ+13ΔΔ

2.2 Analyse acoustique par approche multi-variable

L'approche multi-variable appelée en anglais "multi-stream" consiste à traiter l'information de différents angles et à la représenter de plusieurs manières. Ainsi l'utilisation de multiples sources d'information avec la technique multi-stream est susceptible d'accroître les performances et la robustesse du système face aux bruits.

L'analyse acoustique multi-variable repose sur la fusion de plusieurs sources d'informations. Nous distinguons deux manières de combiner les flux au sein d'une trame, soit en affectant à chacun un poids unité, soit en favorisant un des flux en lui assignant des poids qui reflètent son importance au sein de la trame. Cette affectation des poids est faite dans le but d'accroître davantage la robustesse du système multi-variable. Cette technique vient ainsi remédier aux déficiences de la technique classique qui repose sur l'utilisation d'une seule source d'information.

Dans le cadre de l'analyse multi-variable, plusieurs approches ont été proposées. Dans chaque travail un choix particulier des sources d'information à combiner est proposé. Dans ce qui suit, nous présenterons un aperçu sur ces travaux.

Dans [**Schmid et Barnard, 1995**] un algorithme d'extraction des formants est présenté dans le but de les incorporer au niveau de la trame multi-variable. La méthode consiste à trouver dans un premier temps les formants. Ensuite, ces derniers sont sélectionnés par un algorithme de programmation dynamique pour garder uniquement les premières meilleures interprétations des formants. La décision du choix final des bons formants est retardée jusqu'à ce que la recherche phonétique prenne fin, contribuant ainsi à surmonter le manque de robustesse qui caractérise les méthodes traditionnelles d'extraction de formants. Cette méthode a apporté de bons résultats qui pourront être améliorés davantage afin d'aboutir à une extraction des formants plus robustes.

Dans [**Sangita, 1999**], l'auteur présente une nouvelle technique multi-variable pour l'amélioration de la robustesse des signaux face aux bruits. Dans son travail deux modèles sont proposés. Le premier consiste à l'utilisation de différentes bandes de fréquences du spectre du signal de la parole. Ces bandes de fréquences sont traitées indépendamment dans différents flux pour être combinées par la suite. Ce modèle est entraîné sur des données non bruitées et il a apporté une amélioration de 50% sur un corpus de chiffres isolés bruités. Le

second modèle est une extension du premier. En effet, en plus de l'utilisation des bandes de fréquences, une incorporation de l'information temporelle du signal est ajoutée au niveau de chaque bande de fréquence (d'une durée de 200 à 500 msec). Ce modèle a apporté une réduction de 50% sur un corpus de nombres continus testés dans des conditions bruitées et non bruitées.

Habdulla et Kasabov [**Habdulla et Kasabov, 1999**] proposent une trame multi-variable constituée de trois flux de vecteurs indépendants. Le premier est constitué des coefficients cepstraux (12MFCC + énergie), le deuxième contient les 13 dérivées premières des coefficients MFCC et le troisième est composé des 13 dérives secondes des MFCC. Chaque flux est modélisé indépendamment par un HMM. Les trois flux subissent une réduction de taille pour aboutir à une trame de 28 coefficients au lieu de 39 coefficients. Les résultats obtenus par cette technique ont apporté des améliorations bien qu'elle présente des limites vu qu'elle repose sur l'utilisation de la même source d'information séparée en trois sous-flux, ce qui pourrait constituer une redondance d'information.

L'approche multi-variable proposée dans [**Zolany et al. 2000**] combine trois flux, les coefficients MFCC, les coefficients PLP et les caractéristiques de la voix. Deux méthodes sont utilisées pour combiner les différents flux. Une méthode linéaire appelée LDA (Linear Discriminant Analysis) et une autre méthode non linéaire "log-linéaire". Avec la méthode LDA, un seul modèle acoustique est utilisé alors qu'avec la méthode "log-linéaire", trois modèles indépendants sont utilisés, chaque flux est associé à un modèle. Chacune des deux méthodes a apporté des améliorations.

Dans presque tous les travaux cités précédemment, nous remarquons que les coefficients MFCC sont présents. En effet, l'usage complémentaire des coefficients différentiels du premier et du second ordre s'est généralisé depuis une dizaine d'années dans les systèmes de reconnaissance à base de HMM.

Parvenant à ce stade de résultats, les chercheurs sont allés plus loin pour chercher la possibilité d'intégrer d'autres sortes d'information dans la trame multi-variable. Il en est résulté un progrès dans les recherches pour l'optimisation des poids. Plusieurs autres méthodes ont vu le jour dont les principales sont décrites dans ce qui suit.

Dans ce cadre, Selouani et Tolba, ont proposé [**Tolba et al. 2002; Selouani et al. 2003]**, de nouvelles méthodes de codage multi-variable du signal de parole. La méthode consiste à

ajouter aux coefficients MFCC de nouveaux paramètres pour construire une nouvelle trame acoustique plus riche en termes d'information et plus robuste face aux dégradations. La nouvelle trame est constituée de 26 coefficients MFCC au lieu de 39 avec quatre valeurs de formants (F1, F2, F3, F4) et sept indices acoustiques tirés du modèle de l'oreille humaine. Les formants sont les fréquences de résonance du conduit vocal d'un individu. Les indices acoustiques apportent des informations supplémentaires sur la parole comme la gravité/acuité d'un phonème. Pour tester cette nouvelle méthode, une série d'expériences a été menée sur la base de données TIMIT pour différents rapports signal/bruit. Une première expérience, utilisant une trame acoustique constituée des coefficients MFCC+ les formants, a été menée puis une autre utilisant les coefficients MFCC+ les indices acoustiques et une troisième combinant les trois, soient: MFCC+ les formants + les indices acoustiques. Les trois expérimentations ont apporté des améliorations du taux de reconnaissance par rapport à l'utilisation seul des coefficients MFCC. Cette méthode est intéressante, dans la mesure où elle a ouvert un nouveau volet de recherche par l'intégration des formants et les indices acoustiques.

Dans [**Hsieh et al. 2002**] un algorithme de codage de la parole est présenté. Il a pour but l'augmentation de la robustesse des signaux face aux bruits. L'algorithme repose sur l'utilisation de la décomposition en ondelettes ainsi que la modélisation autorégressive d'un signal de parole. L'utilisation des ondelettes dans ce travail a pour objectif de mieux représenter la non-stationnarité du signal. La méthode consiste à générer les paramètres des paquets d'ondelettes ainsi que les paramètres de la modélisation autorégressive. Ces deux derniers paramètres sont alors combinés et transformés à l'aide de la méthode de transformation linéaire LDA pour aboutir à la fin à un seul vecteur de paramètres pour l'analyse acoustique. La méthode est testée en milieu bruité, elle a apporté des améliorations par rapport à l'utilisation des coefficients MFCC seuls. Cette amélioration pourrait être meilleure si elle était complétée par une étude des poids accordés aux différents flux et l'ajout d'une troisième source d'information comme les coefficients MFCC.

Dans [***Tamura et al. 2005b ; Brugger et al. 2006***], une autre variante de trame multi-variable a été proposée, elle combine à la fois le signal audio et le visuel. Cette dernière méthode repose sur l'idée que la parole est un moyen audiovisuel de communication profitant ainsi du mouvement des lèvres. En effet, le message parlé est plus intelligible quand il est accompagné

de la vision du visage du locuteur, et particulièrement quand le milieu de transmission est bruité. Un système de reconnaissance audiovisuelle de la parole résulte de la fusion de deux systèmes mono-modaux audio et vidéo. La trame acoustique est alors constituée de deux flux, le premier est le flux audio classique et le second est le flux visuel. A chaque flux est accordée une pondération qui reflète sa contribution au niveau de la trame. La détermination des poids dans [**Adjoudani et al. 1996**], a fait l'objet d'un problème d'optimisation afin d'avoir la meilleure combinaison de ces deux sources. Une méthode d'optimisation basée sur la probabilité de transmission au niveau du modèle de Markov a été proposée, elle consiste à effectuer un ajustement (égalisation) des probabilités d'émission des différents Modèles de Markov. Cette méthode a été évaluée sur un corpus de chiffres japonais. Une amélioration de 10% a été obtenue par rapport aux résultats obtenus avant optimisation.

Taichi et al. [**Taichi et al. 2006**] ont proposé également une trame multi-variable composée de deux flux combinant les coefficients MFCC classiques et la fréquence fondamentale. Une optimisation des poids accordés aux deux flux de la trame est effectuée. Différents tests sont établis et ils ont montré des améliorations dans le cas d'un bruit blanc.

2.3 Intégration des paramètres multiples

La représentation du signal par des coefficients cepstraux est souvent utilisée en RAP. Bien que les coefficients cepstraux soient utilisés en raison de leurs bonnes propriétés de représentation, notamment la décorrélation des coefficients, ils soufrent de plusieurs limitations. En particulier ils sont sensibles aux conditions d'acquisition du signal et à l'environnement acoustique (problème de robustesse). A cause de cette sensibilité, la performance d'un système de reconnaissance de la parole est dégradée, elle est encore plus dégradée quand les conditions de l'apprentissage et de l'utilisation du système sont différentes. Exploiter les différents aspects de la parole dans la RAP, en intégrant de multiples paramètres, dérivés de sources indépendantes, a fait l'objet de nombreuses études. En combinant plusieurs paramètres, les informations utiles pour la reconnaissance qui risquent d'être perdues au cours du processus d'extraction d'un ensemble donnée de paramètres peuvent être récupérées à partir d'une autre analyse acoustique. Cette méthode ne peut être efficace que si les deux techniques fournissent diverses informations complémentaires sur les caractéristiques de la parole. L'intégration d'informations fortement corrélées, est moins susceptible de donner un avantage

quelconque, étant donné que les informations fournies par les deux techniques sont entièrement redondantes. Il existe des méthodes pour mesurer le degré de corrélation entre les différents paramètres acoustiques [**Ellis et Bilmes, 2000**]. Ces méthodes peuvent être utilisées pour prévoir quelle combinaison serait la plus avantageuse.

Il existe essentiellement quatre approches pour l'intégration des informations acoustiques distinctes :

1 Transformation des vecteurs de paramètres multiples en un seul ensemble de paramètres préalablement à tout processus de reconnaissance (par exemple, en utilisant les méthodes linéaires ou non linéaires).
2 Le paradigme "Multi-Stream" : modèles avec de multiples flots de paramètres acoustiques.
3 Combinaison de paramètres à l'aide des probabilités a posteriori de modèles définies sur chaque ensemble de paramètres telle que la combinaison discriminante de modèles.
4 Combinaison se basant sur le résultat de reconnaissance. Par exemple, le vote des sorties multiples pour réduire l'erreur de reconnaissance, connu aussi sous le nom de ROVER[4].

Dans [**Schulter et al. 2006**], il a été démontré que la transformation d'un ensemble de paramètres multiples en un seul ensemble de paramètres peut parfois conduire à la dégradation des performances. D'autre part, l'idée d'utiliser un système de reconnaissance séparé pour chaque ensemble de paramètres, produisant une nouvelle hypothèse parmi toutes les hypothèses, par le vote pour les occurrences susceptibles d'être prononcées, est nettement plus efficace. Cependant, cette technique n'est pas très efficace en termes de complexité de la procédure menée par le traitement "*Back-end*".

La technique que nous utilisons dans ce mémoire est basée sur le paradigme multi-flots ("multi-stream paradigm") avec les HMM. La combinaison de paramètres dans ce paradigme a l'avantage de la simplicité de calcul aussi bien que la faisabilité et la rapidité d'exécution. Le potentiel de cette technique a été exploité par de nombreux chercheurs

[4] ROVER: Recognizer Output Voting Error Reduction.

2.4 Choix des paramètres acoustiques

Le but de l'analyse acoustique consiste à représenter le signal de parole sous une forme qui est plus adaptée pour la reconnaissance. Le plus souvent, on utilise les représentations suivantes: PLP [**Hermansky, 1990**], MFCC [**Vergin et al. 1999**] ou LPCC[5] [**Rabiner et Juang, 1993**]. Au sein de ce travail, on s'intéressera surtout à la représentation *front-end* MFCC en complémentarité avec celle des LSF.

2.4.1 Les coefficients Mel-Cepstraux

Dans les SRAP, reconnaître l'occurrence prononcée, nécessite la déconvolution du signal d'excitation et de la fonction de transfert du conduit vocal. Une des méthodes largement utilisées dans les SRAP est l'utilisation d'une transformation logarithmique. Cette transformation convertit le produit de deux spectres en une somme de deux signaux et peut considérablement simplifier le processus de déconvolution. Après l'application de la transformation sur les spectres de la parole, il est possible de séparer l'excitation du conduit vocal à l'aide d'un filtre linéaire, car il existe de grandes différences dans les spectres de ces deux signaux. Cela peut être fait après l'application d'une inversion de la transformée de Fourier afin de revenir au domaine temporel de la présentation des signaux. Ce processus est appelé analyse cepstrale et le signal de sortie est appelé le cepstrum. Réparti sur l'échelle Mel, le cepstrum utilise la loi logarithmique de répartition des bandes de fréquences au lieu de la répartition linéaire. Ceci permet une modélisation proche de celle du système auditif humain. En effet, cette modélisation perceptive du spectre à court terme de parole entraine des performances de RAP supérieures des MFCC comparativement à d'autres paramètres. Un schéma explicatif, pour l'extraction des MFCC, est proposé (figure 2.3).

Pour compenser l'absence d'information énergétique, on ajoute aux 12 coefficients mel-cepstraux le logarithme de l'énergie de chaque trame d'analyse de numéro *t*. On obtient donc le vecteur acoustique à 13 composantes $C_t = (C_1(t), . ., C_{12}(t), E(t))^T$ (12 coefficients mel-cepstraux et le logarithme de l'énergie calculées pour la trame t).

[5] LPCC: Linear Predictive Cepstral Coefficients.

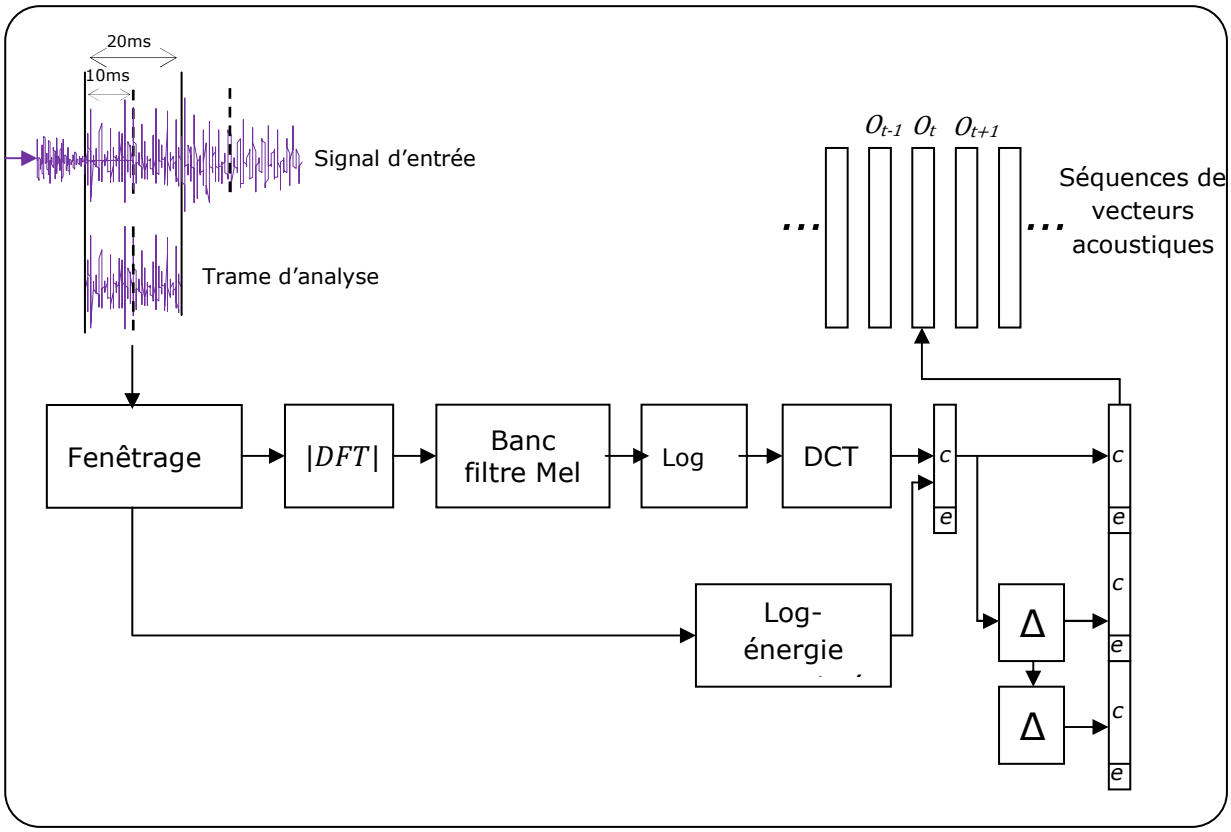


Figure 2.3–Module d'analyse acoustique par la représentation MFCC

Puisque la séquence de ces vecteurs acoustiques est ensuite traitée comme la séquence d'observations d'un HMM, l'information dynamique locale[6] est perdue. Effectivement, dans chaque état, toute l'information est décrite par la distribution correspondante qui ne modélise pas du tout l'ordre des données. Pour garder cette information, on étend ces vecteurs acoustiques à leurs dérivées (temporelles) premières et secondes. Ces paramètres sont souvent appelés coefficients delta et delta-delta et ils peuvent être estimés selon:

$$\Delta C_t = \frac{\sum_{k=-L}^{L} k C_{t+k}}{\sum_{k=-L}^{L} k^2}. \qquad (2.1)$$

La dérivée seconde $\Delta\Delta$ est calculée en itérant deux fois l'expression (2.1). On obtient finalement une suite de vecteurs acoustiques O_t, à 39 composantes chacun.

[6] Il s'agit de la localité dans chaque état d'un HMM.

2.4.2 Les fréquences de raies spectrales

Parmi les nombreuses paramétrisations possibles d'un filtre AR, les coefficients LSF (Line Spectral Frequency) semblent être ceux qui offrent les meilleures propriétés d'interpolation [**Itakura, 1975; Paliwal, 1993**]. Ils ont d'ailleurs été utilisés par Kain dans un but de conversion de voix [**Kain, 2001**].

Les paramètres LSF sont une variante des coefficients LPC reconnue comme ayant de bonnes propriétés d'interpolation. Pour les calculer, il faut d'abord calculer les coefficients LPC. Pour cela, nous procédons comme dans [**McAulay et Quatieri, 1995**]. Le logarithme de la densité spectrale de puissance, représenté par le log des amplitudes A_l est sur-échantillonné en utilisant une interpolation cubique [**Unser et al. 1993**]. Puis, les coefficients du filtre LPC sont estimés par application de l'algorithme de Levinson-Durbin sur les coefficients d'auto-corrélation obtenus par une FFT inverse du carré des amplitudes.

Les coefficients a_k du filtre LPC, $A_p(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ sont donc convertis en coefficients LSFs. Dans cette représentation

$$A_p(z) = \frac{1}{2}\left(P_{P+1}(z) + Q_{p+1}(z)\right), \tag{2.2}$$

avec

$$P_{p+1}(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}), \tag{2.3}$$

$$Q_{p+1}(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}). \tag{2.4}$$

Les coefficients LSF sont extraits à partir des racines complexes des polynômes $P_{p+1}(z)$ et $Q_{p+1}(z)$. Ces fonctions de transfert possèdent deux propriétés très intéressantes [**Saoudi, 1990**]: pour un filtre stable $1/A_p(z)$, toutes les racines de $P_{p+1}(z)$ et $Q_{p+1}(z)$ sont sur le cercle unité et s'alternent deux à deux. D'autre part, ces racines sont conjuguées. En ignorant les racines réelles (1 et -1 selon que p est pair ou impair), le filtre $A_p(z)$ peut être représenté par la séquence $\omega_1, \omega_2, \ldots, \omega_p$ des arguments des racines complexe des filtres $P_{p+1}(z)$ et $Q_{p+1}(z)$ se trouvant sur le demi cercle entre 0 et π. Les paramètres ω_i présentent plusieurs propriétés intéressantes. Tout d'abord, ils sont ordonnés : $0 < \omega_1 < \omega_2 < \ldots < \omega_p < \pi$. Cette relation d'ordre est une condition nécessaire et suffisante pour la stabilité du filtre de synthèse $1/A_p(z)$. En

outre, les coefficients LSF sont robustes car une erreur sur un seul coefficient LSF aura des répercussions sur une région de spectre située au voisinage de la fréquence correspondant à ce coefficient. Les coefficients LSF sont des paramètres fréquentiels. La proximité de deux coefficients fait apparaître un pic dans le spectre d'amplitude assimilable à un formant (figure 2.4). A partir des coefficients LSF il est donc possible d'identifier grossièrement les zones auditivement importantes dans le spectre du signal de façon très aisée [**Sugamura et Itakura 1986**].

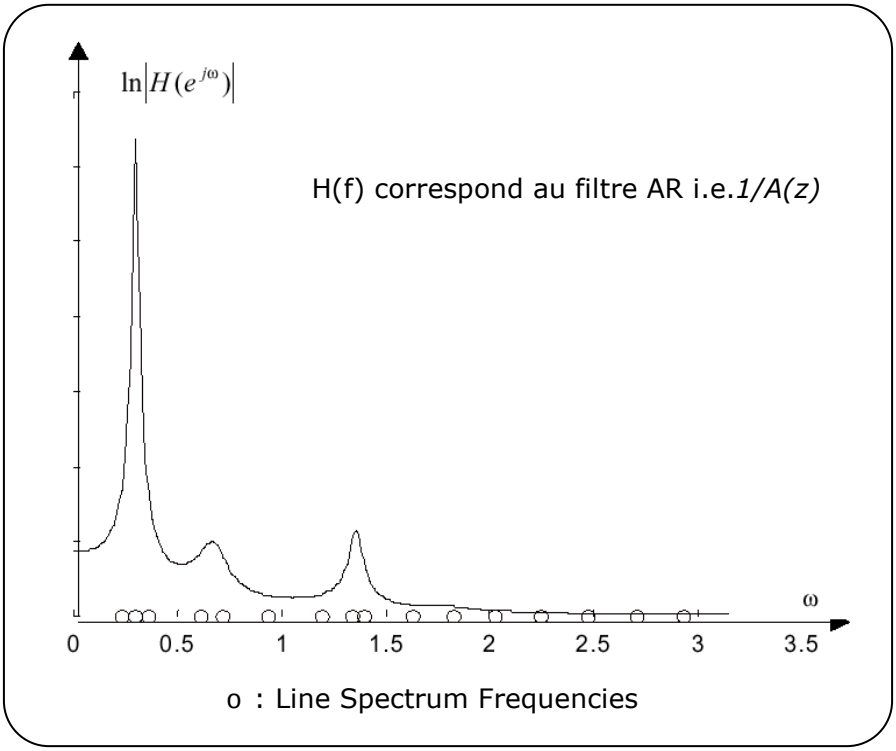


Figure 2.4– Interprétation physique des LSF

Selon la méthode des racines réelles utilisée dans ITU-T Recommandation G.723.1 [**ITU-T 1996**], le signal de parole est divisé en trames. A son tour, chaque trame est subdivisée en quatre sous-trames. L'analyse LPC est donc faite sur une sous-trame de base. Les p coefficients LPC sont transformés en p coefficients LSF correspondants. Cette transformation est effectuée sur la dernière sous-trame. Les LSF des trois autres sous-trames sont calculées en effectuant une interpolation entre la trame courante et précédente. De ce fait, le cercle unité

est divisé en 512 intervalles égaux, de longueur π/256 chacun. Les racines LSF de $P_{p+1}(z)$ et $Q_{p+1}(z)$ sont explorées à travers le cercle unité de 0 à π. L'interpolation linéaire est effectuée sur des intervalles où l'on observe un changement de signe, afin de trouver les zéros des polynômes.

Selon [**Rose et Momayez, 2007**], si le changement de signe apparait entre l'intervalle *l* et *l*-1, l'interpolation de premier ordre est réalisée comme suit:

$$\hat{l} = \left[l - 1 + \frac{|P_{l-1}(z)|}{|P_{l-1}(z)| + |P_l(z)|} \right]. \qquad (2.5)$$

Où $\hat{l}$ est l'index de la solution interpolée et $|P_l(z)|$ est l'amplitude absolue du résultat de l'évaluation de la somme polynomiale à l'intervalle *l* (respectivement *l*-1). Comme les LSF sont intercalés dans la région de 0 à π, seul un zéro est évalué sur *P(z)* à chaque étape. La recherche de la solution suivante est réalisée par évaluation de la différence polynomiale *Q(z)* en commençant par la solution courante. Les LSF sont considérés comme la représentation tendancielle de la connaissance phonétique de la parole et sont prévus d'être relativement robustes dans les cas particuliers de la RAP pour un environnement bruité ou à bande-limitée.

Deux principales raisons expliquent notre choix de considérer les LSF dans les communications mobiles bruitées. La première raison est liée au fait que les régions LSF du spectre demeurent au-dessus du niveau de bruit même pour un RSB très bas, alors que les régions d'énergie faibles tendent à être masquées par l'énergie du bruit. La seconde raison est liée au fait que les LSF sont généralement utilisés dans les schémas conventionnels de codage. Ceci évite l'incorporation de nouveaux paramètres nécessitant des modifications importantes et coûteuses pour les modèles courants et les codecs.

2.5 Conclusion

Motivés par toutes les observations et travaux présentés, nous avons proposé dans ce chapitre une nouvelle trame acoustique multi-variable pour la reconnaissance de la parole en mode distribué, dont l'étude explorant les possibilités offertes par l'approche acoustiques multi-variable et ceux de l'approche dimensionnelle sera traitée au chapitre suivant. Il s'agit d'une fusion de deux sources d'information, à savoir les coefficients MFCC et les LSF.

Nous considérons que les raies spectrales sont bien appropriées à notre application, vu qu'elles représentent le signal de parole en utilisant peu de paramètres (généralement 10) diminuant ainsi la taille de la trame.

Pour les coefficients MFCC, nous choisissons certains coefficients parmi tous ceux qui sont disponibles en tronquant les coefficients différentiels d'ordre deux afin d'éliminer le problème de redondances de l'information. Les différents flux seront modélisés par un seul modèle de Markov. Chaque état du modèle sera composé de l'ensemble de flux choisis et à chaque flux est assigné un poids reflétant sa contribution au niveau de la trame, comme le décriera le chapitre suivant.

Chapitre 3

Traitement amont des paramètres multiples d'un système DSR

Chapitre 3

Traitement amont des paramètres multiples dans un système DSR

Sommaire

Comme dans la plupart des systèmes de reconnaissance, nous utilisons une analyse cepstrale pour extraire les paramètres du signal de parole. Les paramètres cepstraux sont ensuite combinés avec d'autres paramètres: les fréquences de raies spectrales en utilisant une approche statistique multidimensionnelle, pour former le nouveau vecteur acoustique. En fin, on réduit la dimension des vecteurs ainsi obtenus de façon optimale en effectuant la transformée de Karhunen-Loève qui a l'avantage supplémentaire de réduire l'effet du bruit qui pourrait entacher le signal de parole. Ceci résume le présent chapitre, dans un cadre de traitement optimal des paramètres multiples dans un système DSR.

3.1 Analyse statistique markovienne multi-variable

L'hypothèse que le signal de parole est stationnaire pour une courte durée de temps a permis le développement d'algorithmes efficaces pour optimiser le processus de reconnaissance. Ces algorithmes caractérisent les propriétés du signal de parole sous la forme de modèles. Ces modèles sont divisés en deux catégories: les modèles déterministes et les modèles statistiques. Dans un modèle déterministe, ce sont les propriétés simples et connues du signal qui sont exploitées tandis que dans le modèle statistique le signal est considéré comme un processus aléatoire paramétrique. Les modèles statistiques sont plus appropriés pour la modélisation du signal de parole, en raison de la nature aléatoire de ce signal [**Rabiner, 1986**]. Les processus gaussiens, les processus de Poisson, les modèles de Markov, et les modèles de Markov cachés sont quelques-uns des modèles statistiques qui sont utilisés dans les systèmes actuels de traitement de la parole.

Dans la plupart des SRAP, le cadre statistique est basé sur Modèles de Markov Cachés. Une des raisons essentielles pour développer de tels systèmes basés sur les HMM est que les paramètres du modèle peuvent être extraits automatiquement. D'autre part, la simplicité et la faisabilité de l'utilisation des HMM ont fait qu'ils constituent le meilleur modèle statistique pour la RAP.

Dans un SRAP basé sur les HMM, on suppose que la séquence de vecteurs observée correspondant à chaque mot (unité) est générée par un modèle de Markov avec des paramètres inconnus. Dans ce modèle, chaque unité de la parole est représentée par un état. Chaque état est associé à une fonction de distribution de probabilité. La probabilité d'observer le vecteur O_t à l'état j est b_j qui est la fonction de distribution de l'état j.

Afin de simplifier le processus de détermination de la fonction de distribution de probabilité de l'état, certaines distributions définies avec très peu de paramètres sont utilisées. La densité de probabilité la plus populaire utilisée dans la RAP est la densité du mélange gaussien définie comme suit :

$$\mathcal{N}(O_t; \mu_{jm}; \Phi_{jm}) = \frac{1}{\sqrt{(2\pi)^n |\Phi_{jm}|}} exp^{-\frac{1}{2}(O_t-\mu_{jm})' \Phi_{jm}^{-1}(O_t-\mu_{jm})} \quad (3.1)$$

Où $\mathcal{N}$ désigne une gaussienne multi-variable. Avec μ le vecteur de la moyenne et Φ la matrice de covariance. Dans la pratique, un mélange de densités gaussiennes est utilisé pour générer une distribution qui est la plus proche de la distribution réelle des données. Pour l'état *j* associé à un modèle de mélange gaussien, la probabilité d'observer O_t est calculée à partir de:

$$b_j(O_t) = \sum_{m=1}^{M} C_{jm}\, \mathcal{N}\left(O_t; \mu_{jm}; \Phi_{jm}\right). \tag{3.2}$$

Où M est le nombre de composantes du mélange, C_{jm} est le poids de chaque composante *m* de l'état *j*.

3.1.1 Le paradigme multi-stream

Dans l'approche multi-variable acoustique, les différents paramètres obtenus à partir de différentes sources sont concaténés pour former le "multi-stream" utilisé pour l'apprentissage des HMM. Considérons *S* sources d'informations qui fournissent des vecteurs synchrones O_{st} d'observation, où $s = 1,\ldots, S$; *s* indique la source d'informations et *t* l'indice de temps. La dimension des vecteurs d'observation peut varier d'une source à une autre. Chaque séquence de vecteur d'observation fournit des informations au sujet d'une séquence des états cachés. Dans un système de multi-stream, au lieu de produire *S* séquences d'états pour *S* séquences d'observation, seulement une séquence d'état est générée. En réalité, ceci est fait, en introduisant une nouvelle fonction de distribution des états. La fonction de distribution de l'état *j* est définie comme:

$$b_j(O_t) = \prod_{s=1}^{S} \left[b_{js}(O_{st})\right]^{\gamma_{js}}. \tag{3.3}$$

L'exposant γ spécifie la contribution de chaque flux à la distribution globale en mesurant sa distribution correspondante. On suppose que la valeur des γ_{js} satisfait les contraintes suivantes:

$$0 \le \gamma_{js} \le 1 \quad et \quad \sum_{s=1}^{S} \gamma_{js} = 1. \tag{3.4}$$

Dans les HMM, les modèles à mélange de gaussiennes sont utilisés pour représenter la distribution d'émission des états. La probabilité du vecteur O_t à chaque instant t dans l'état j peut être reformulée à partir de l'équation (3.3) comme suit:

$$b_j(O_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M} C_{jsm} \mathcal{N}\left(O_{st}; \mu_{jsm}; \Phi_{jsm}\right) \right]^{\gamma_{js}} . \qquad (3.5)$$

Où M est le nombre de composantes du mélange, C_{jsm} est le $m^{éme}$ poids de la gaussienne de l'état j pour la source S. $\mathcal{N}$ indique le multi-variable gaussien (équation 3.1) avec μ_{jsm} le vecteur moyen et Φ_{jsm} la matrice de covariance.

Le choix des exposants joue un rôle important. La performance du système est sensiblement affectée par les valeurs de γ. La plupart des techniques d'apprentissage pour l'obtention d'exposants optimaux ont été développées dans le domaine logarithmique [**Rose et Momayez, 2007**].

La figure 3.1 illustre comment un modèle de deux flux est formé et mis en fonctionnement. Il peut y avoir des distributions de vecteurs multiples d'observation, qui sont fusionnées pour former une distribution à sortie unique pour l'état j.

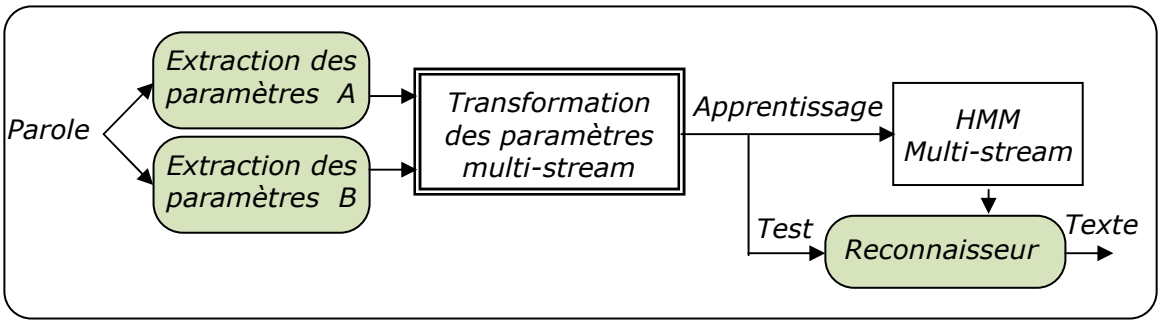


Figure 3.1–Etapes d'apprentissage et test pour un model HMM à deux "streams"

3.1.2 Techniques d'optimisation des poids

L'estimation des valeurs de l'exposant de pondération est une tâche difficile. En raison de la nature aléatoire du signal, il n'y a pas de formalisme fondamental que nous pouvons exploiter et nous nous attendons à ce que les γ_{js} soient une fonction du RSB.

L'approche principale est celle basée sur le maximum de vraisemblance. Bien que la recherche des paramètres HMM soit habituellement effectuée en utilisant le critère du maximum de vraisemblance, il n'est pas possible d'utiliser ce critère pour déterminer les poids des flux. Selon ce critère les poids $\gamma_{js} \in [0,1]$ doivent être choisis de telle sorte que la sommation du logarithme de vraisemblance d'une séquence d'états correspondant à l'observation de vecteur O_t soit maximisée. Dans un exemple simple de deux flots de paramètres, le critère du maximum de vraisemblance implique la maximisation des $\gamma_1 logb^1(O_t^1) + \gamma_2 logb^2(O_t^2)$ sous les contraintes données en équation (3.4). Les termes $b^1(O^1)$ et $b^2(O^2)$ correspondent à la répartition des fonctions du premier et du second flux de paramètres respectivement. Cette maximisation se fait comme suit:

$$\gamma = \begin{cases} 1 & si \quad logb^1(O_t^1) > logb^2(O_t^2) \\ 0 & si \quad logb^1(O_t^1) < logb^2(O_t^2) \end{cases} \tag{3.6}$$

Clairement les poids provenant d'équation (3.6) ne peuvent pas être utilisés pour les modèles multi-flots et ceci indique que le critère du maximum de vraisemblance n'est pas adapté à la formation de poids.

3.1.3 Méthode heuristique du choix du poids

Dans cette approche la détermination des poids optimaux est effectuée d'une manière heuristique ou appelée également par essai-erreur. C'est une méthode très ancienne utilisée pour résoudre divers problèmes. Dans cette méthode les solutions sont obtenues après examen de diverses phases d'essai à travers une recherche irrégulière. Elle convient à un petit ensemble fini de valeurs qui sont examinées variante après variante jusqu'à ce qu'elles soient traitées en totalité ou que la solution soit trouvée.

Dans notre cas, nous avons effectué un ensemble d'expériences d'entraînement et de reconnaissance pour ne conserver à la fin que les poids qui donnent le meilleur taux de

reconnaissance. On présente dans le chapitre suivant, la meilleure combinaison des poids associés aux coefficients MFCC, à leurs dérivées premières et aux LSF.

3.2 Analyse multidimensionnelle

En fin de traitement amont des vecteurs acoustiques considérés, nous visons à optimiser l'utilisation des flux de paramètres en réduisant la dimension de ces vecteurs tout en améliorant la robustesse du système. Une façon efficace d'effectuer cette réduction est d'utiliser la transformée de Karhunen-Loève (KLT) [**Joliffe, 2002**].

3.2.1 La transformée de Karhunen-Loève: Principe et méthode

Connue sous le nom de ACP (Analyse en Composantes Principale), l'idée centrale de la KLT est de réduire la dimension d'un ensemble de données qui se compose d'un grand nombre de variables en corrélation, tout en maintenant autant que possible la variation actuelle dans cet ensemble de données. C'est une technique d'analyse statistique multi-variable qui a été utilisée dans le rehaussement de signaux dégradés par le bruit [**Yi Hu et Loizou, 2002**]. Son principe est basé sur le calcul des vecteurs propres de la matrice de covariance, puis leur agencement selon l'ordre décroissant des valeurs propres correspondantes.

Soit x un échantillon de signal de parole représenté par n coefficients issus d'une analyse acoustique.

$$x = (c_1\ c_2\ \cdots\ c_n) \tag{3.7}$$

m échantillons forment alors une matrice X de dimension [mxn] :

$$X = \begin{bmatrix} c_{11}\ c_{12}\ \cdots\ c_{1n} \\ c_{21}\ c_{22}\cdots\ c_{2n} \\ \cdots \\ c_{m1}\ c_{m2}\ \cdots\ c_{mn} \end{bmatrix} \tag{3.8}$$

Dans un espace de dimension n muni d'un repère, on place les points $x_i = (c_{i1}\ c_{i2}\ \cdots\ c_{in})$. On obtient ainsi une représentation graphique des données: le nuage de points. L'idée de la KLT est de déterminer p vecteurs ($p < n$) tels que la somme des distances euclidiennes des points du nuage au sous-espace engendré par ces p vecteurs soit minimale. On montre alors

que la projection du nuage de points dans ce sous-espace est la meilleure approximation du nuage en dimension *p*. Par exemple, dans le plan (figure 3.2), la KLT détermine la droite de direction *u* qui approche au mieux le nuage de points.

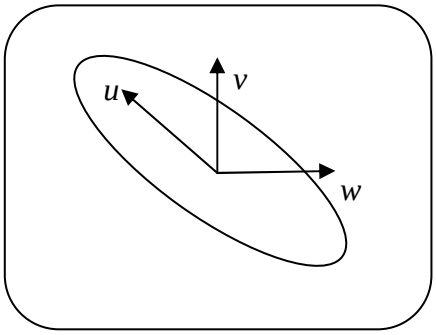


Figure 3.2–Direction (u) de projection d'un nuage de points.

Pour déterminer le sous-espace, la KLT opère sur des données centrées sur zéro. De plus, les données n'étant pas forcément de même nature, on normalise les vecteurs afin de s'affranchir des unités de mesure.

Ainsi, si l'on nomme $\overline{C_J}$ et $\sigma(c_j)$), respectivement la moyenne et l'écart-type des données du caractère c_j, on détermine la matrice *Xc* des données centrées réduites par la formule:

$$X_{C_{ij}} = \frac{X_{ij} - \overline{C_J}}{\sigma(c_j)} \tag{3.9}$$

La KLT calcule ensuite la matrice de corrélation *C* suivante :

$$C = {}^{t}X_C . X_C \tag{3.10}$$

Où ${}^{t}X_C$ représente lareprésente la transformée de *Xc*. *C* est une matrice diagonalisable. On détermine donc les *p* valeurs propres que l'on range par ordre décroissant. Les *p* vecteurs correspondent aux *p* vecteurs propres associés aux *p* premières valeurs propres rangées par ordre décroissant. Chacun de ces vecteurs a une capacité de discrimination relatif à sa valeur propre.

On établit enfin la matrice *P* de projection, appelée KLT, selon les *p* plus grands vecteurs propres (c.-à-d., les *p* directions de plus grandes variances). L'espace généré par les vecteurs correspondant aux valeurs propres d'ordres inférieurs est supposé être très faiblement influencé par le bruit. La projection des vecteurs bruités dans ce sous-espace est le principe des méthodes de décomposition en sous-espaces qui permettent de réaliser le rehaussement. Chaque vecteur X_C de paramètres est alors prétraité pour produire le vecteur optimal X_P selon l'expression:

$$X_P = PX_C. \tag{3.11}$$

La KLT décorréle les paramètres et fournit la plus petite erreur de reconstruction possible parmi toutes les transformations linéaires, c.-à-d. la plus petite erreur quadratique moyenne (EQM) possible entre les vecteurs de données dans l'espace *n* original et les vecteurs de données dans le sous-espace *p* projeté ($p<n$).

3.2.2 Application aux données acoustiques

Nous désirons appliquer une KLT à des données issues d'une analyse acoustique afin de les présenter à un système de reconnaissance basé sur un modèle HMM. Aussi, le but de notre travail est de trouver une sorte de correspondance entre les données de test bruitées et les données d'apprentissage non bruitées. Le signal non-bruité analysé *Anb* est projeté dans un espace par la projection *Pnb* et le vecteur résultant *Dnb* est mis en entrée du réseau HMM multi-stream pour l'apprentissage. Le signal bruité analysé *Ab* est projeté dans un espace par la projection *Pb* et le vecteur résultant *Db* est mis en entrée du réseau pour la reconnaissance (figure 3.3)

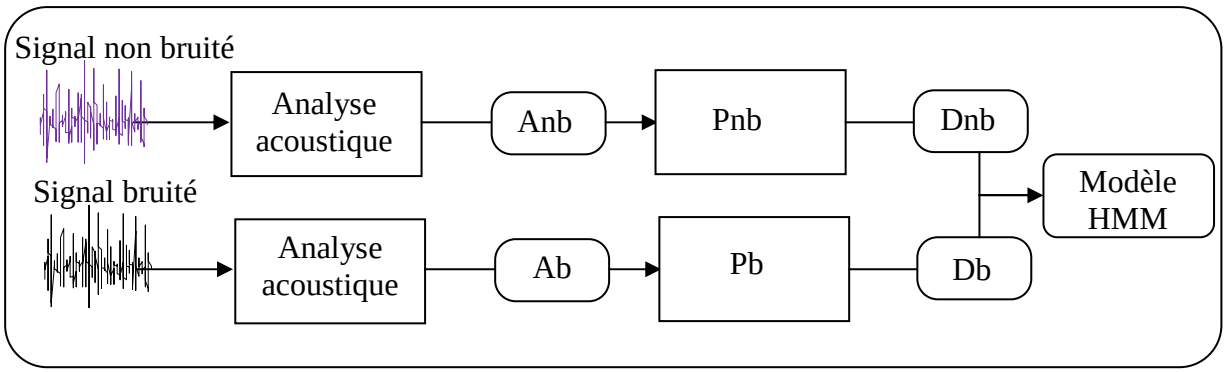


Figure 3.3–Modèle de projection des données bruitées et non bruitées.

Pour que cette reconnaissance se fasse au mieux, il faut réduire la distance entre le vecteur d'apprentissage et le vecteur de test. Pour cela, nous devons trouver la matrice de projection *Pb* de sorte que l'on obtienne des données *Db* équivalentes aux données *Dnb*. Dans notre cas, la base de projection *Pnb* des données *Anb* est une base de vecteurs propres issus d'une KLT.

Cas de la KLT à classes indépendantes (CI-KLT :

Pour plus de détails, concernant notre application, supposons que nous ayons L réalisations d'une unité acoustique, $x=[x_1, x_2,\ldots, x_L]$ dans notre corpus d'apprentissage. Supposons maintenant que, $x_i(m)$ représente la $m^{éme}$ trame de la $i^{éme}$ réalisation (ou observation) qui est pré-accentuée et fenêtrée par la fenêtre de Hamming, où $1 \leq m \leq M_i$ et M_i représente le nombre total de trames du signal x_i. Si le nombre des composantes retenues par la KLT est égal à N, nous aurons une matrice de dimension $N \times M_i$ pour chaque réalisation noté S_i. En concaténant les matrices des différentes observations, nous pouvons représenter l'ensemble du corpus d'apprentissage par la matrice S de dimension $N \times M$, où :

$$M = \sum_{i=1}^{L} M_i. \tag{3.12}$$

La KLT produit une matrice de transformation utilisant la décomposition des valeurs propres de la matrice de covariance C comme suit :

$$\mathrm{C}v_p = \lambda_p v_p, \qquad 1 \leq p \leq N \tag{3.13}$$

Où v_p est le vecteur propre et λ_p sa valeur propre correspondante. Alors la matrice de transformation P est constituée par les vecteurs propres, par ordre de plus grande valeur propre correspondante:

$$P = [v_1, v_2, \ldots, v_N]^T, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N. \tag{3.14}$$

Nous employons la matrice P pour extraire les paramètres. Dans nos expériences, nous avons opté pour une KLT à classes indépendantes (CI-KLT) dans laquelle la matrice de transformation est globale et est déterminée pour toutes les classes, contrairement au cas d'une KLT à classes dépendantes (CD-KLT) [**Sharma et al. 2006**] où nous utilisons une matrice de transformation pour chaque modèle acoustique. Enfin, dans l'étape de

reconnaissance, pour une occurrence inconnue de parole, nous extrayons sa matrice de paramètres pour laquelle nous effectuons la transformation-projection *P* afin d'évaluer le taux de reconnaissance.

3.3 Conclusion

Parmi les nombreuses méthodes abordées dans les sections précédentes, très peu sont utilisées dans les systèmes de reconnaissance opérationnels. En fait, la plupart de ces méthodes sont encore au stade de la recherche expérimentale et n'ont pas montré de résultats suffisamment convaincants pour être intégrées.

La méthode la plus couramment utilisée est certainement l'accroissement de la taille de la base de données d'apprentissage. Avec l'augmentation de la puissance de calcul et de la taille de la mémoire, il devient possible de fournir une très grande quantité d'information lors de l'apprentissage afin d'avoir un système performant dans de multiples situations de test.

Cependant, malgré la taille de plus en plus considérable des bases d'apprentissage, celles-ci ne sont qu'un échantillon très restreint de l'ensemble des variabilités possibles du signal de parole. Devant la difficulté de ce problème, différents chercheurs explorent de nouvelles voies de recherche pour une meilleure compréhension des problèmes liés à la parole. Dans ce chapitre de cette thèse, nous avons exploré les possibilités offertes par l'approche acoustiques multi-variable et ceux de l'approche dimensionnelle.

Des tailles différentes des nouveaux vecteurs multi-variables proposés seront testées sur la base de données standard Aurora de l'ETSI. L'ensemble de résultats sera présenté et discuté dans le quatrième chapitre.

Chapitre 4

Evaluation expérimentale

Chapitre 4

Evaluation expérimentale

Sommaire

Ce chapitre examine le potentiel de l'approche acoustique multi-variable, présenté dans les chapitres précédents, par des expériences ainsi que l'analyse des résultats obtenus. Après une présentation de toutes les ressources nécessaires à la mise en place d'un système DSR, on examinera le potentiel de différents systèmes DSR, en mettant en œuvre les tâches de reconnaissance. Nous présenterons, alors, la base de données et les protocoles utilisés pour la validation expérimentale de notre système LSF-KLT.

4.1 Plan d'expériences

Nous rappelons qu'un de nos objectifs principaux est d'évaluer les performances de l'analyse acoustique multi-flots de paramètres en entrée d'une RAP distribuée. Le protocole expérimental et la base de données Aurora fournis par l'ETSI sont pris comme base pour évaluer notre approche d'analyse acoustique en la comparant au standard DSR de l'ETSI.

Les expériences que nous avons menées sont basées sur les performances d'un système de reconnaissance ayant appris 11 classes TI Digits dans un environnement non bruité (clean) et que nous testons dans un environnement bruité.

Dans un premier temps, nous avons évalué l'incorporation, dans le système DSR de base, des paramètres LSF dans une approche les combinant avec les MFCC. Les résultats des différentes stratégies de cette combinaison sont présentés, discutés et analysés.

Concernant les performances de la reconnaissance robuste, nous avons exploré, dans une seconde phase, les poids des flots de paramètres qui déterminent comment chaque flot de paramètres contribue dans le modèle à la probabilité globale. Nous présentons aussi, les différentes améliorations que nous avons apportées à la méthode heuristique.

En fin, partant de la meilleure combinaison trouvée, nous avons testé l'effet de la projection par KLT des données bruitées et non bruitées, notamment l'influence du choix de la dimension de l'espace de projection.

4.2 Implémentation d'un système reconnaissance de parole

Pour réaliser un SRAP, il est nécessaire de disposer les éléments suivants :

- Un corpus de parole,
- Un outil d'analyse acoustique à partir du signal vocal,
- Une plateforme de modélisations HMM, phonétique et de langage.

Les différentes sections de ce chapitre, fournissent des informations au sujet de chacun de ces éléments.

4.2.1 Corpus de parole

La base de données Aurora est dérivée de la base de données de TIDigits contenant les enregistrements de locuteurs adultes Nord-Américains (hommes et femmes) éditant des séquences continues pouvant aller jusqu'à 7 chiffres. Les données originales échantillonnées à 20 kHz ont été sous-échantillonnées à 8 kHz en utilisant un filtre passe-bas ayant une bande passante entre 0 et 4 kHz.

Apprentissage

L'ensemble d'apprentissage contient des données bruitées et non bruitées (mode clean). Les mêmes expressions sont prononcées dans les deux types de données. Ces données incluent 8840 expressions parlées par 55 femmes et 55 hommes dérivées du corpus d'apprentissage TIDigits filtré avec la caractéristique G.712 [**ITU G.712 1996**]. Pour les données bruitées, les 8840 expressions sont réparties en 20 sous-ensembles contenant 422 phrases chacun. Les sous-ensembles représentent quatre différents types de bruit avec des RSB s'étendant de -5 dB à 20 dB en croissant de 5dB. Les quatre bruits sont ceux recueillis, respectivement, dans le hall d'une gare, lors de réceptions(conversations), dans un hall d'exposition et dans une voiture. Ces bruits sont également filtrés avec la caractéristique G.712.

Test

Deux ensembles de test dénommés A et B ont été définis par ETSI. Chaque ensemble contient quatre sous-ensembles composés chacun de 1001 occurrences prononcées par55 hommes et 55 femmes. Ces occurrences sont extraites de TI Digits. Un signal de bruit est artificiellement ajouté à chaque sous-ensemble avec des RSB s'étendant de-5 dB à 20 dB avec un pas croissant de 5 dB. Les données non bruitées sont également prises en compte. Dans le premier ensemble, A, les quatre bruits collectés : dans le hall d'une gare, lors de réceptions (conversations), d'un hall d'exposition, et dans une voiture, sont ajoutés aux 4 sous-ensembles. Au total, l'ensemble formé inclut 28028 occurrences. Le deuxième ensemble de test, B, est semblable à A, excepté les bruits utilisés. Il s'agit de ceux collectés dans un restaurant, dans une rue, dans un aéroport et dans une gare.

4.2.2 Outils d'Analyse Acoustique

Pour extraire les MFCC à partir du signal de parole, nous avons utilisé la technique qui a été développée dans le projet Aurora de l'ETSI. Cette technique permet l'extraction de paramètres Mel-cepstrum Front End pour la DSR [**ETSI, 2003**]. Le processus d'extraction est décrit dans la section dédiée à la MFCC (cf. section 2.4.1). Pour extraire les paramètres LSF, le filtre LPC, l'algorithme UIT [**ITU-T, 1996**] et le soft MATLAB ont été utilisés. Cette technique d'extraction est décrite dans la section 2.4.2.

4.2.3 La boite à outil HTK

La boite à outils des modèles de Markov cachés (HTK) est utilisée principalement pour créer les modèles et effectuer la reconnaissance [**Young, 1993**]. HTK est un utilitaire comprenant des outils pour: l'analyse de la parole, l'apprentissage HMM, la modélisation du langage, le test, et l'analyse des résultats (voir annexe B). Dans ce travail le processus d'apprentissage HMM et l'identification sont effectués avec les outils de HTK.

4.3 Analyse expérimentale

Cette section étudie les performances des différents systèmes DSR via différentes expériences. Toutes les expériences menées dans cette section sont basées sur l'apprentissage des HMM dans le mode clean de la base de données d'apprentissage Aurora. Les performances des systèmes ayant différentes structures sont déterminées par les tâches de reconnaissance sur les tests A et B en milieu bruité. Pour être conforme à la norme standard ETSI [**ETSI, 2003**], le mot entier des TIDigits est utilisé comme modèle de HMM, pour représenter les chiffres. Chaque modèle de mot se compose de 16 états avec 3 mixtures de gaussiennes par état. Deux modèles de silence sont également considérés. L'un d'eux a une durée relativement longue, modélisant les pauses avant et après les élocutions, et constitué de 3 états et 6 mixtures de gaussiennes par état. Le second est un état singulier de HMM lié à l'état moyen du premier modèle de silence, représentant les courtes pauses entre les mots.

Une application du protocole expérimental décrivant la phase apprentissage et test du modèle utilisé est décrite en annexe A.

4.3.1 Description du système de base

Le système de base que nous avons utilisé dans nos expériences d'évaluation est celui fourni par Aurora. Cependant, en raison des différences dans les processeurs des systèmes d'exploitation des ordinateurs utilisés pouvant causer de légères différences dans les résultats, les expériences ont été relancées sur un même ordinateur où est hébergée notre application.

Les coefficients cepstraux sont calculés toutes les 10 ms sur une largeur d'une fenêtre de Hamming de 25 ms. L'énergie de la trame est ajoutée aux coefficients statiques. Par la suite, les premières et secondes dérivées et les variations de leurs énergies correspondantes, sont calculées et ajoutées au vecteur statique.

Le premier coefficient cepstral C_0 et le coefficient de log-énergie fournissent des informations similaires. Aussi, C_0 n'est pas inclus dans le vecteur de caractéristiques. Ainsi, le système de base est défini par un vecteur de dimension 39, comprenant 12 coefficients cepstraux (sans C_0) ainsi que le log-énergie en plus des composants deltas et accélérations correspondants. Ce vecteur noté MFCC_E_D_A(39) est considéré comme le *front-end conventionnel* par le standard DSR de l'ETSI. Les phases d'apprentissage et de reconnaissance des HMM sont réalisées par HTK (voir Annexe A). On divise le vecteur d'observation en flux multiples avec des pondérations égales.

Les résultats obtenus sont présentés dans la table 4.1. Comme prévu les paramètres MFCC sont très efficaces en reconnaissance de la parole non bruitée mais leur performance se dégrade nettement à mesure que le niveau du bruit augmente.

4.3.2 Identification par les LSF

Dans ces expériences, nous étudions l'impact de l'utilisation des paramètres LSF sur la robustesse du système DSR en milieu bruité. Aussi, pour évaluer l'effet du nombre de paramètres LSF dans un système DSR, les expériences sont réalisées sur deux vecteurs acoustiques de dimensions 30 et 36, et comprenant respectivement 10 et 12 coefficients LSF. Ces vecteurs seront notés LSF_D_A(30) et LSF_D_A(36). On remarque, aisément, les trois flux utilisés pour ces vecteurs, à savoir les LSF, les premières et les secondes dérivées.

Tableau 4.1–Taux de reconnaissance (en %) obtenus par le système de base DSR-FE (Test A & B; Corpus AURORA).

Rapport signal/bruit		Clean	20dB	15dB	10dB	5dB	0dB	-5dB
Test A	Hall d'une gare	98.93	97.05	93.49	78.72	52.16	26.01	11.18
	Reception	99.00	90.15	73.76	49.43	26.81	9.28	1.57
	Voiture	98.96	97.41	90.04	67.01	34.09	14.46	9.39
	Hall d'exposition	99.20	96.39	92.04	75.66	44.83	18.05	9.60
Test B	Restaurant	98.93	89.99	76.24	54.77	31.01	10.96	3.47
	Rue	98.99	95.74	88.45	67.11	38.45	17.84	10.46
	Aéroport	98.96	90.64	77.01	53.86	30.33	14.41	8.23
	Station de Train	99.20	94.72	83.65	60.29	27.92	11.57	8.45

Les tables 4.2 et 4.3 montrent que pour, les cas des bruits de réception et voiture (Test A) et ceux de restaurant et aéroport (Test B), lorsque le rapport RSB est inférieur à 10dB, l'utilisation d'un *front-end LSF* avec un vecteur acoustique de dimension 30 conduit à une amélioration significative du taux de reconnaissance. Cette amélioration atteint les 6%, dans le cas de -5dB, comparativement au cas du vecteur de base MFCC_E_D_A de dimension 39. Cependant, on note que dans des conditions où le RSB est élevé (faiblement bruité ou le mode clean), le *front-end MFCC* de base (DSR-FE) présente une meilleure performance.

Tableau 4.2–Taux de reconnaissance (en %) obtenu par le système DSR utilisant un "*front end LSF*" sur le répertoire "Test A" d'Aurora.

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Réception	MFCC_E_D_A(39)	**90.15**	**73.76**	**49.43**	26.81	9.28	1.57
	LSF_D_A(30)	73.94	62.58	49.15	**29.99**	**13.97**	**7.65**
	LSF_D_A(36)	70.59	60.55	46.52	28.17	12.16	6.95
Voiture	MFCC_E_D_A(39)	**97.41**	**90.04**	**67.01**	**34.09**	**14.46**	**9.39**
	LSF_D_A(30)	79.24	69.43	48.17	23.68	13.00	8.71
	LSF_D_A(36)	79.51	68.51	47.39	22.55	12.35	8.47

Tableau 4.3–Taux de reconnaissance (en %) obtenu par le système DSR utilisant un "*front end LSF*" sur le répertoire "Test B" d'Aurora.

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Restaurant	MFCC_E_D_A(39)	**89.99**	**76.24**	**54.77**	31.01	10.96	3.47
	LSF_D_A(30)	74.33	65.19	53.21	**32.51**	**15.44**	**8.87**
	LSF_D_A(36)	74.18	65.89	53.12	31.44	15.26	8.20
Aéroport	MFCC_E_D_A(39)	**90.64**	**77.01**	**53.86**	**30.33**	14.41	8.23
	LSF_D_A(30)	80.79	63.10	51.31	27.10	**16.61**	**9.72**
	LSF_D_A(36)	79.85	61.89	50.67	28.45	16.19	9.05

Ces résultats montrent l'intérêt d'utiliser deux front-end concomitants: Un *front-end LSF* pour des applications fortement bruitées, et l'actuel DSR-FE sous des conditions relativement moins bruitées. L'estimation du RSB est donc nécessaire, afin d'aiguiller d'un front à un autre.

4.3.3 Intégration de paramètres complémentaires

Nous avons évalué l'amélioration apportée par l'incorporation, dans le système DSR de base, des LSF dans une approche les combinant avec les MFCC. Les résultats des différentes stratégies de cette combinaison sont présentés, discutés et analysés.

Concernant, le cas du *front-end* proposé dans le cadre de l'approche multi variable, les 12 coefficients MFCC et leurs dérivées premières, sans la composante d'énergie, constituent le premier et second flux. Les 10 coefficients LSF sont pris comme troisième flux. Ces flux multiples ont des pondérations égales. Ce *nouveau front-end* sera noté MFCC_D_LSF (34), où 34 indique sa dimension. Les LSF ajoutés produisent un ensemble multidimensionnel de paramètres et remplacent les composantes accélérations et énergies du front-end conventionnel.

Par contre, le vecteur utilisant la composante d'énergie sera noté MFCC_E_D_LSF (36), où 36 représentera sa dimension. Pour mener une comparaison significative, nous avons effectué, simultanément, les mêmes expériences sur les deux tests A et B pour différents bruits, avec les deux variantes de la trame acoustique (avec et sans la composante d'énergie), relevant ainsi l'influence de la composante d'énergie. Ainsi, les meilleurs scores du taux de reconnaissance obtenus définiront la nouvelle référence (comparativement à celle de DSR-FE) à laquelle nous évaluons d'autres types de vecteurs acoustiques par la suite.

Les résultats donnés dans les tables 4.4 et 4.5 montrent que l'utilisation de l'approche multi-variable MFCC-LSF conduit à une amélioration significative du taux de reconnaissance par rapport au *front-end* basique. À titre d'exemple, nous remarquons que notre approche est d'autant meilleure lorsque le RSB diminue (inférieur à 10 dB). La substitution des composantes accélérations et énergies par les LSF dans le vecteur de base conduit à une amélioration du taux de reconnaissance et à une dimension de vecteur conséquente. On atteint les meilleurs résultats dans le cas d'une trame de 34 coefficients.

Tableau 4.4–Taux de reconnaissance(en %) du système DSR de base et ceux utilisant l'approche multi-variable MFCC-LSF sur le répertoire "Test A"

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Réception	MFCC_E_D_A(39)	**90.15**	73.76	49.43	26.81	9.28	1.57
	MFCC_E_D_LSF (36)	85.76	68.68	44.17	23.55	11.76	7.41
	MFCC_D_LSF(34)	82.92	**75.48**	**61.00**	**37.94**	**18.32**	**9.52**
Voiture	MFCC_E_D_A(39)	**97.41**	**90.04**	**67.01**	**34.09**	14.46	9.39
	MFCC_E_D_LSF (36)	92.75	80.97	58.60	27.59	13.12	8.38
	MFCC_D_LSF(34)	88.46	77.87	53.18	24.63	**15.57**	**10.23**

Tableau 4.5–Taux de reconnaissance(en %) du système DSR de base et ceux utilisant l'approche multi-variable MFCC-LSF sur le répertoire "Test B"

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Restaurant	MFCC_E_D_A(39)	**89.99**	**76.24**	54.77	31.01	10.96	3.47
	MFCC_E_D_LSF (36)	86.43	71.38	49.86	26.53	11.79	7.65
	MFCC_D_LSF(34)	81.89	75.93	**62.02**	**38.62**	**18.70**	**9.64**
Aéroport	MFCC_E_D_A(39)	**90.64**	**77.01**	**53.86**	30.33	14.41	8.23
	MFCC_E_D_LSF (36)	81.57	63.61	42.38	24.04	12.20	8.38
	MFCC_D_LSF(34)	74.77	66.03	51.00	**33.46**	**17.63**	**9.07**

4.3.4 Optimisation par intégration des paramètres similaires aux LSF

L'idée de combinaison des paramètres similaires aux LSF avec les coefficients MFCC classiques, vise l'enrichissement de la trame acoustique afin d'augmenter sa robustesse face aux bruits. Ainsi, nous avons évalué l'incorporation de ces paramètres, présentés dans cette section, dans les systèmes DSR. Les résultats des différentes trames acoustiques proposées incluant cette nouvelle combinaison sont présentés, comparés et discutés.

4.3.4.1 Extraction des paramètres MLSF

Deux des caractéristiques les plus courantes actuellement utilisées dans la reconnaissance du locuteur sont les (MFCC) et les mel-cepstres provenant des coefficients de prédiction linéaire (MLPCC). MFCC et MLPCC profitent de l'avantage des propriétés de l'oreille humaine, par l'introduction du banc du filtre Mel, la réduction de l'information dans les hautes fréquences. Au contraire, les LSF n'ont pas ce type d'information, ce qui compromet leur performance dans un environnement avec un RSB élevé (tables 4.2 et 4.3).

Pour pallier cet inconvénient, nous avons envisagé le calcul des LSF à partir des spectres Mel d'énergies, afin de vérifier si l'introduction de l'information Mel pourrait améliorer les performances de la DSR_FE. De cette façon, nous proposons de caractériser les nouvelles trames acoustiques avec cette modification de LSF, notés MLSF (Mel Line Spectrum Frequencies), leur estimation et comparaison avec les MFCC sont résumées par la figure 4.2.

Les LSF sont calculés à partir de la transformation des coefficients de prédiction linéaire et ces derniers peuvent être obtenus auprès de l'autocorrélation du signal de la parole. Pour inclure l'information Mel, l'auto corrélation est calculée comme la transformée de Fourier inverse des spectres Mel d'énergies, originaires des coefficients de Mel-autocorrélation.

La figure 4.1 montre les paramètres LSF et MLSF, et aussi leur correspondance en pôles LPC et MLPC (Mel-LPC). En comparant ces deux graphiques, il est possible de confirmer la déformation dans le domaine du spectre imposée par le banc du filtre Mel, et la réduction correspondante des informations à haute fréquences. Les MLSF apportent des informations pertinentes pour pouvoir caractériser les locuteurs mais également les phonèmes. Les tables 4.6 et 4.7 confirment et valident cette performance en taux de reconnaissance de mots.

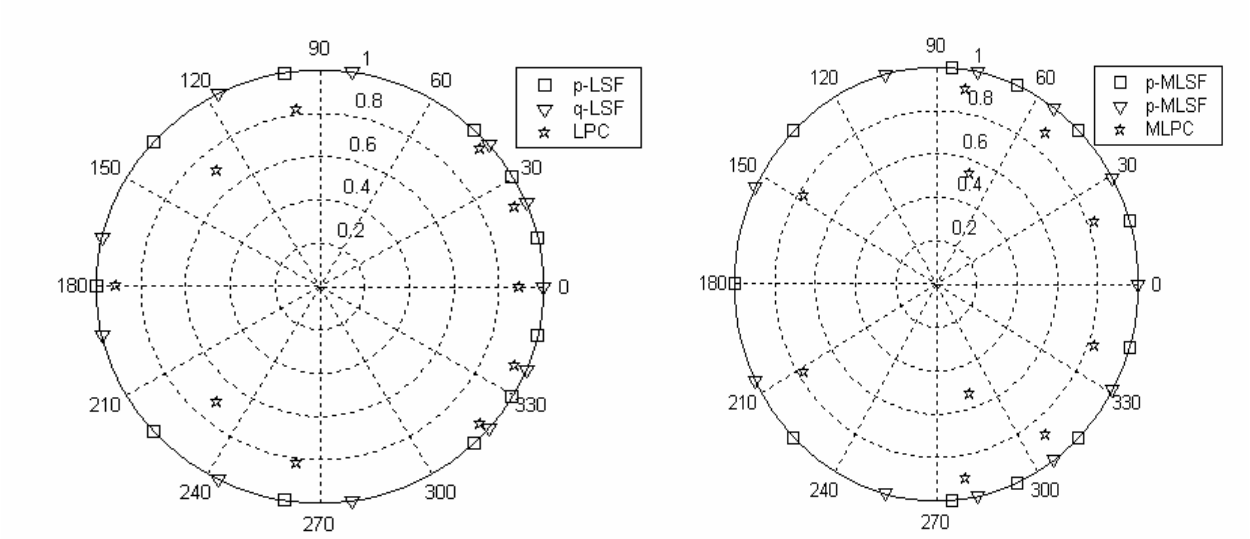


Figure 4.1–Les coefficients LSF et les pôles LPC (à gauche); Les coefficients MLSF et les pôles MLPC (à droite)

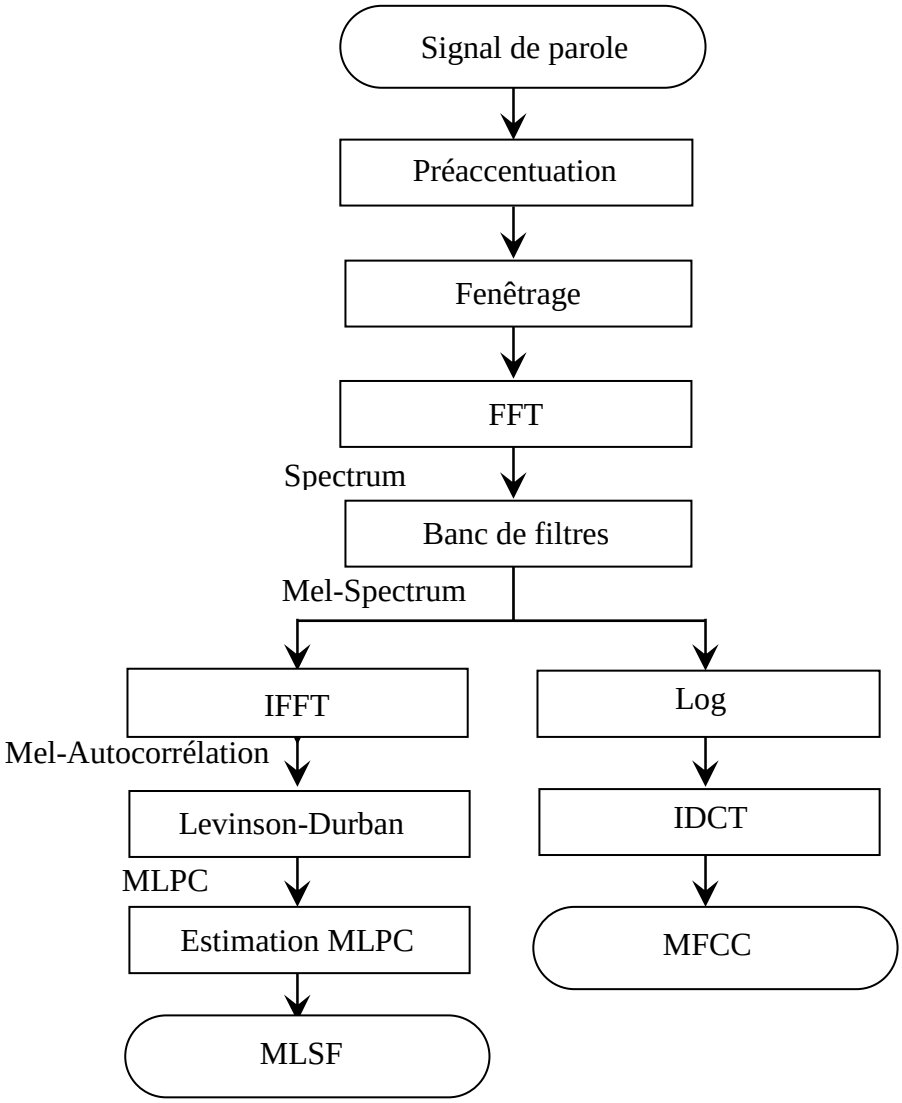


Figure 4.2–Estimation MLSF versus MFCC

Tableau 4.6–Taux de reconnaissance(en %) du système DSR de base et ceux utilisant l'approche multi-variable MFCC-MLSF/LSF^{db} sur "Test A".

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Réception	MFCC_E_D_A(39)	90.15	73.76	49.43	26.81	9.28	1.57
	MFCC_D_LSF (34)	82.92	75.48	61.00	37.94	18.32	9.52
	MFCC_D_MLSF(34)	**94.26**	**88.12**	**72.64**	43.02	19.86	11.00
	MFCC_D_LSF^{db}(34)	86.00	80.74	69.14	**49.67**	**25.30**	**13.09**
Voiture	MFCC_E_D_A(39)	**97.41**	**90.04**	**67.01**	34.09	14.46	9.39
	MFCC_D_LSF (34)	88.46	77.87	53.18	24.63	15.57	10.23
	MFCC_D_MLSF(34)	93.74	83.84	55.17	25.77	16.28	10.56
	MFCC_D_LSF^{db}(34)	86.22	78.65	62.36	**35.19**	**19.56**	**10.59**

Tableau 4.7–Taux de reconnaissance(en %) du système DSR de base et ceux utilisant l'approche multi-variable MFCC-MLSF/LSF^{db} sur "Test B".

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Restaurant	MFCC_E_D_A(39)	89.99	76.24	54.77	31.01	10.96	3.47
	MFCC_D_LSF (34)	81.89	75.93	62.02	38.62	18.70	9.64
	MFCC_D_MLSF(34)	**94.60**	**89.01**	**73.29**	**44.58**	21.68	11.51
	MFCC_D_LSF^{db}(34)	77.99	73.17	61.77	41.42	**24.87**	**12.68**
Aéroport	MFCC_E_D_A(39)	90.64	77.01	53.86	30.33	14.41	8.23
	MFCC_D_LSF (34)	74.77	66.03	51.00	33.46	17.63	9.07
	MFCC_D_MLSF(34)	**94.01**	**88.31**	**73.37**	**46.41**	24.55	12.94
	MFCC_D_LSF^{db}(34)	80.64	73.34	58.87	40.17	**24.60**	**13.42**

4.3.4.2 Extraction de la trame acoustique LSF débruitée

Afin de réduire le bruit, nous proposons de faire appliquer à la trame acoustique proposée, un étage du filtre de Wiener. Les caractéristiques du filtre sont estimées dans le domaine fréquentiel où l'opération du filtrage se fait directement dans le domaine temporel après transformation des caractéristiques du filtre estimé dans le domaine temporel. L'estimation par le filtre est individuellement faite à court segments du signal où deux trames consécutives ont une différence de temps de 10 ms [**Loizou, 2007]**.

La figure 4.3 présente l'estimation des LSF à partir d'une trame débruitée par le filtre de Wiener. Il est, parmi les méthodes de débruitage classiques, les plus utilisées dans la littérature. C'est l'estimateur $W(\nu)$ qui minimise EQM entre le signal d'entrée et celui en sortie (équation 4.1):

$$\mathrm{E}[|e(\nu)|^2] = \mathrm{E}\left[\left|S(\nu) - \hat{S}(\nu)\right|^2\right] \tag{4.1}$$

$$= \mathrm{E}[|S(\nu) - W(\nu)Y(\nu)|^2]. \tag{4.2}$$

L'expression du filtre est donnée par :

$$\mathrm{W}(\nu) = \operatorname{argmin} \mathrm{E}[|S(\nu) - W(\nu)Y(\nu)|^2] \tag{4.3}$$

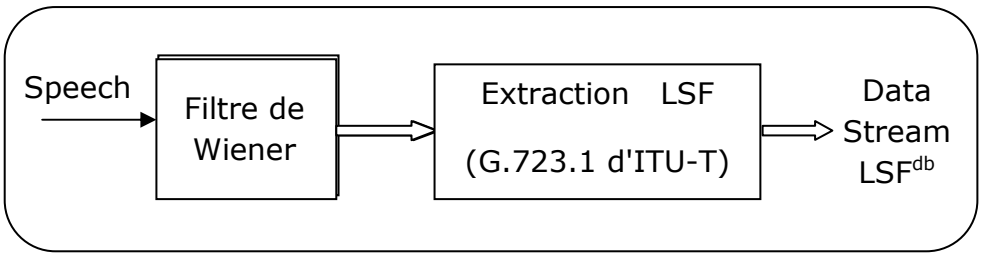


Figure 4.3– Extraction des LSF à partir d'une trame débruitée par un filtrage de Wiener

Aussi, les résultats présentés dans les tables 4.6 et 4.7 montrent que l'extraction des LSF à partir d'une trame débruitée par le filtre de Wiener améliore davantage le taux de

reconnaissance pour différents RSB (en particulier ceux inférieur à 5 dB), comparativement aux DSR-FE et ceux utilisant l'approche proposé MFCC-LSF/MLSF. Le vecteur constitué est noté par MFCC_D_LSF^{db}(34). On remarque bien, les trois flux utilisés pour ces vecteurs, à savoir MFCC, leurs premières dérivées et les LSF débruités comme troisième flux.

4.3.5Optimisation heuristique des poids

Afin d'évaluer l'impact des coefficients de pondération γ_{js} de l'équation (3.3), nous avons effectué d'autres expériences sur le vecteur acoustique multi-flot proposé, en faisant varier ces pondérations tout en satisfaisant à la contrainte définie par l'équation (3.4).

Dans ce système, un ensemble de poids empiriquement optimisés sera lié à tous les états indépendamment de l'unité que chaque état représente. Les poids optimaux retenus pour les trois flux de paramètres, à savoir les MFCC, les delta-MFCC et les LSF, sont : 40%, 40%, et 20% respectivement. Le vecteur constitué sera noté MFCC_D.8_LSF.2 (34), respectivement _MLSF.2 et _LSF^{db}.2, pour le troisième flux des deux autres variantes du vecteur acoustique multi-variable.

Les tables 4.8 et 4.9 indiquent les taux de reconnaissance de mots, obtenus pour le modèle HMM multi-flots de paramètres, avec les poids attachés optimisés. Les résultats montrent que l'incorporation des LSF dans un modèle multi-flots de paramètres dans un système DSR améliore de manière significative les performances de reconnaissance du système dans les environnements bruités. On constate une nette amélioration à partir de 10dB. À ce niveau de RSB, 20% de contribution des LSF, ou paramètres similaires aux LSF, par rapport aux MFCC permet d'améliorer le taux de reconnaissance de manière significative (jusqu'à 25%).

Les résultats obtenus à partir des paramètres similaires aux LSF (MLSF et LSF^{db}), montrent une nette amélioration du taux de reconnaissance lorsque le RSB diminue de moins de 5dB. Elles atteignent les meilleurs résultats dans le cas d'une trame de 34 coefficients, dans presque tous les RSB, par rapport au *front-end* basique et celui de l'approche MFCC-LSF proposé. Ainsi, les poids assignés à chaque flux de la trame proposée influent bien sur le rendement du système.

Tableau 4.8–Taux de reconnaissance (en %) du système DSR de base et ceux utilisant l'approche MFCC-LSF à différentes pondérations sur "Test A"

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Réception	MFCC_E_D_A(39)	90.15	73.76	49.43	26.81	9.28	1.57
	MFCC_D_LSF(34)	82.92	75.48	61.00	37.94	18.32	9.52
	MFCC_D.8_MLSF.2(34)	**93.98**	**88.30**	72.55	44.50	20.86	11.37
	MFCC_D.8_LSF^{db}.2(34)	91.32	86.03	**75.51**	**53.72**	**25.00**	**11.52**
Voiture	MFCC_E_D_A(39)	**97.41**	**90.04**	**67.10**	**34.09**	14.46	9.39
	MFCC_D_LSF(34)	88.46	77.87	53.18	24.63	15.57	10.23
	MFCC_D.8_MLSF.2(34)	94.18	84.64	56.40	27.20	17.15	11.63
	MFCC_D.8_LSF^{db}.2(34)	86.46	78.65	58.75	29.88	**17.95**	**11.96**

Tableau 4.9– Taux de reconnaissance (%) du système DSR de base et ceux utilisant l'approche MFCC-LSF à différentes pondérations sur "Test B"

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Restaurant	MFCC_E_D_A(39)	89.99	76.24	54.77	31.01	10.96	3.47
	MFCC_D_LSF (34)	81.89	75.93	62.02	38.62	18.70	9.64
	MFCC_D.8_MLSF.2(34)	**94.29**	**88.85**	**73.84**	**45.38**	**22.32**	**11.76**
	MFCC_D.8_LSF^{db}.2(34)	81.95	76.24	65.86	41.97	22.29	11.18
Aéroport	MFCC_E_D_A(39)	90.64	77.01	53.86	30.33	14.41	8.23
	MFCC_D_LSF (34)	74.77	66.03	51.00	33.46	17.63	9.07
	MFCC_D.8_MLSF.2(34)	**93.98**	**88.34**	**73.93**	**47.27**	**25.11**	**13.63**
	MFCC_D.8_LSF^{db}.2(34)	81.84	74.44	62.24	42.29	23.89	12.38

4.3.6 Optimisation dimensionnelle

Les vecteurs qui résultent des expériences précédentes peuvent contenir jusqu'à 39 composantes. Dans ce travail, nous visons à optimiser l'utilisation de ces flux de paramètres en réduisant la dimension des vecteurs acoustiques tout en améliorant la robustesse du système. En d'autres termes, l'espace N-dimensionnel de paramètres est transformé en sous-espace P-dimensionnel ($P < N$) en réduisant au maximum la perte d'informations pertinentes. Ainsi en intégrant la KLT dans notre approche, nous réalisons deux objectifs : réduction optimale de paramètres et amélioration de la robustesse, en éliminant les composantes principales bruitées (généralement celles d'ordres supérieurs).

Les résultats des tables 4.10 et 4.11 montrent que la KLT appliquée sur le nouveau vecteur pondéré, réduit et optimise à son tour l'espace original de paramètres. La KLT a ainsi conduit à une meilleure performance avec moins de paramètres. En conditions défavorables, la décomposition par KLT en sous-espaces donne de bons résultats comparativement au front-end conventionnel d'ETSI et ceux de l'approche MFCC-LSF proposés.

Expérimentalement (figure 4.4), l'optimisation a été obtenue pour une dimension $P = 24$, correspond au vecteur noté MFCC_D.8_LSF.2(KLT_24), en utilisant la CI-KLT pour différents types de bruit : réception et voiture pour le test A, restaurant et aéroport pour le test B,et ce pour des RSB variant de -5 dB à 20dB où l'on estime le taux de reconnaissance moyen pour différents cas de nombre de composantes principales retenues. On note, que cet optimum est le même pour les autres variantes du vecteur acoustique. A savoir, MFCC_D.8_LSF^{db}.2(KLT_24) et MFCC_D.8_MLSF.2(KLT_24).

Ainsi en intégrant KLT dans notre approche, nous avons réalisé deux objectifs : Une réduction optimale de paramètres, de la dimension 39 du système de base DSR à la dimension 24 et une amélioration de la robustesse, en éliminant les composantes principales bruitées.

Tableau 4.10–Taux de reconnaissance(en %) du système DSR de base et ceux utilisant l'approche LSF-KLT sur le répertoire "Test A".

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Réception	MFCC_E_D_A(39)	90.15	73.76	49.43	26.81	9.28	1.57
	MFCC_D_LSF (34)	82.92	75.48	61.00	37.94	18.32	9.52
	MFCC_D.8_MLSF.2(KLT22)	93.26	88.33	74.49	48.25	22.46	11.28
	MFCC_D.8_LSF^{db}.2(KLT22)	**93.74**	**89.06**	**75.01**	**50.57**	**23.55**	**11.29**
Voiture	MFCC_E_D_A(39)	**97.41**	**90.04**	**67.10**	**34.09**	14.46	9.39
	MFCC_D_LSF (34)	88.46	77.87	53.18	24.63	15.57	10.23
	MFCC_D.8_MLSF.2(KLT22)	94.18	84.64	56.40	27.20	17.15	11.63
	MFCC_D.8_LSF^{db}.2(KLT22)	86.46	78.65	58.75	29.88	**17.95**	**11.96**

Tableau 4.11– Taux de reconnaissance(en %) du système DSR de base et ceux utilisant l'approche LSF-KLT sur le répertoire "Test B".

Rapport signal/bruit (RSB)		20dB	15dB	10dB	5dB	0dB	-5dB
Restaurant	MFCC_E_D_A(39)	89.99	76.24	54.77	31.01	10.96	3.47
	MFCC_D_LSF(34)	81.89	75.93	62.02	38.62	18.70	9.64
	MFCC_D.8_MLSF.2(KLT22)	92.97	**89.25**	75.87	**50.14**	23.06	11.54
	MFCC_D.8_LSF^{db}.2(KLT22)	**93.00**	88.92	**76.27**	49.62	**22.81**	**11.82**
Aéroport	MFCC_E_D_A(39)	90.64	77.01	53.86	30.33	14.41	8.23
	MFCC_D_LSF (34)	74.77	66.03	51.00	33.46	17.63	9.07
	MFCC_D.8_MLSF.2(KLT22)	**92.34**	**88.19**	75.66	49.45	25.59	12.94
	MFCC_D.8_LSF^{db}.2(KLT22)	92.25	87.56	**76.71**	**51.27**	**26.84**	**13.39**

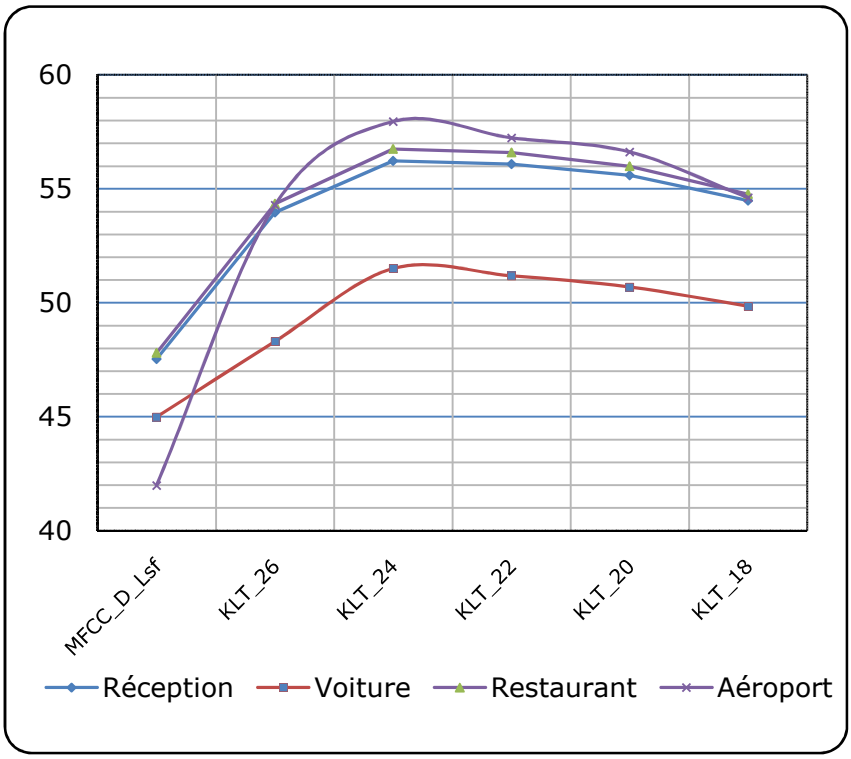
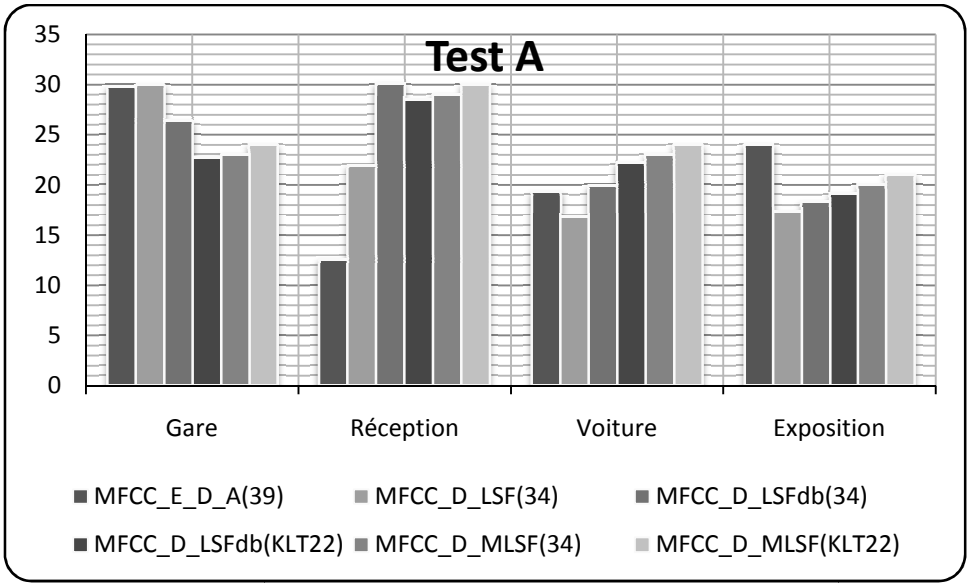


Figure 4.4–Taux de reconnaissance moyens réalisés avec différents types de bruits et pour différents nombres de composantes KLT. Les valeurs de RSB varient de 20 à -5 dB.

4.4 Validation expérimentale du système LSF-KLT

Pour l'évaluation de notre système LSF-KLT, nous avons effectué nos expériences sur la base de données Aurora. Les résultats obtenus montrent que les LSF améliorent la performance de la DSR, comparativement à celle de la DSR front-end de base de l'ETSI utilisant les MFCC seuls. Cette amélioration est d'autant plus importante lorsqu'on effectue un prétraitement par KLT, particulièrement pour des milieux fortement bruités. La figure 4.5 valide le taux de reconnaissance moyen pour trois valeurs de RSB (5, 0 et -5 dB) correspondant aux conditions les plus défavorables, et ce pour les différents bruits des tests A et B.



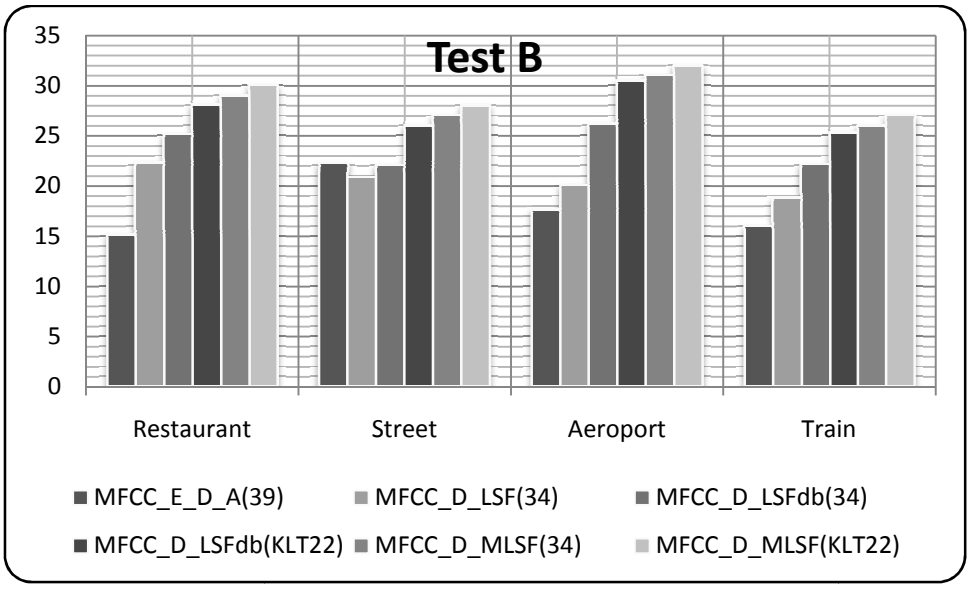


Figure 4.5–Taux de reconnaissance moyens réalisés avec différents types de bruits dans le cas des tests A et B d'Aurora. Les valeurs de RSB varient de -5 à 5 dB.

4.5 Conclusion

Au cours de ce chapitre nous avons présenté la validation de différentes variantes multi-variable proposées par l'établissement des diverses expériences. Des tailles différentes des vecteurs multi-variables sont testées sur la base de données Aurora de l'ETSI. Ces tailles varient en fonction du nombre de paramètres transformés à incorporer. Pour réaliser toutes ces expériences, nous étions confrontés à un problème d'affectation des poids à chaque flux de paramètres de la trame. Nous avons tout d'abord expérimenté une affectation de poids en nous basant sur une heuristique (essai-erreur) qui consiste à effectuer diverses validations croisées pour ne conserver à la fin que les meilleurs poids. Dans une seconde phase, nous avons opté pour une optimisation des paramètres, lorsqu'ils sont pondérés en utilisant la transformée de Kahaurnen-Loeve (KLT). De plus, pour chaque type de vecteurs acoustiques proposés, nous avons étudié l'influence des paramètres MLSF et LSF débruités.

Les résultats obtenus montrent que :

- L'intégration des LSF au niveau de la trame acoustique permet d'obtenir une amélioration du taux de reconnaissance en milieux fortement bruités et cette amélioration est d'autant importante pour des paramètres dérivant des LSF, tel que MLSF et LSF débruités.
- Les poids assignés à chaque flux de la trame influent sur le rendement du système.
- L'application de la KLT a conduit à une réduction optimale de paramètres, passant de la dimension 39 du système de base DSR à la dimension 24.Ceci a réduit davantage le débit et a amélioré la robustesse par élimination des composantes principales bruités.
- Les résultats dont les paramètres sont pondérés, sont influencés par l'incorporation des MLSF et de la trame acoustique LSF débruitée.
- Les résultats obtenus par l'approche LSF-KLT donnent de meilleurs taux de reconnaissance comparativement à ceux du système DSR de base.

Conclusion générale

et

Perspectives

Conclusion générale et Perspectives

Rappel de la problématique

L'une des finalités de ce mémoire était d'examiner une solution qui consiste à représenter le signal acoustique en termes de paramètres (ou caractéristiques) qui sont conçus spécifiquement pour l'application souhaitée. Pour ce faire, le défi était de trouver une méthode fiable pouvant extraire ces paramètres tout en préservant les informations pertinentes. Aussi, la problématique autour de laquelle a été fondée cette recherche concernait l'étude des performances d'un système DSR (Distributed Speech Recognition) en agissant au niveau de l'analyse acoustique du signal vocal. En d'autres termes, il s'agissait de trouver le type de paramètres le mieux adapté aux milieux bruités et plus particulièrement ceux induits par le milieu de transmission sans fil.

Notre démarche de représentation s'inscrit donc dans une problématique de l'évolution de la communication dans des environnements acoustiques changeants. Elle consiste à proposer une approche pour la reconnaissance de la parole distribuée (DSR) dans le contexte du développement des réseaux sans fil.

Contributions principales et résultats obtenus

Deux approches originales sont à la base de la mise en œuvre de notre démarche :

D'abord, nous proposons une analyse statistico-acoustique multi-variable où l'on pose le problème d'assignation des poids de chaque source d'information utilisée (MFCC et LSF ou paramètres similaires aux LSF, en ce qui nous concerne). En effet, il est très important de bien choisir ces poids en fonction des environnements acoustiques changeants auxquels sont confrontés les systèmes DSR. L'objectif dans cette partie est d'améliorer la robustesse des

signaux dans les milieux fortement bruités dans le cadre de cette DSR. Pour ce faire, il s'agissait :

- d'exploiter le potentiel des LSF à être utilisés comme paramètres complémentaires aux coefficients conventionnels MFCC ;
- d'optimiser l'assignation des poids de chaque variable et voir à quel point cela peut influer sur le rendement du système.

Ensuite, dans la mesure où l'intégration des informations acoustiques nécessite des modèles ayant des flots de paramètres (paradigme multi-stream), nous avons proposé une optimisation multidimensionnelle relative des vecteurs acoustiques selon le milieu acoustique ambiant. Pour ce faire, nous avons procédé à une réduction de la dimension de ces vecteurs tout en améliorant la robustesse du système, par l'utilisation de la transformée de Karhunen-Loève(KLT).

Par ailleurs, la boite à outils HTK et la base de données Aurora ont été utilisées pour réaliser et valider aussi bien le système de base que les systèmes proposés. La nouvelle trame multi-variable distribuée a été testée pour différents RSB et les résultats des différentes stratégies de cette combinaison ont été présentés, discutés et analysés. L'ensemble des résultats obtenus répondaient aux problématiques :

1 ***concernant l'impact de l'utilisation des paramètres LSF sur la robustesse du système DSR en milieu bruité :*** L'évaluation de leur incorporation dans le système DSR de base, lors d'une approche les combinant avec les MFCC, a donné des résultats qui ont montré que la combinaison MFCC et LSF dans un seul vecteur de paramètres est recommandée en milieu bruité, par rapport à celle utilisant les MFCC seuls. C'est la raison pour laquelle nous avons proposé une technique de combinaison de multi-flots de paramètres pour combiner ces jeux de paramètres afin de conférer au système de reconnaissance plus de robustesse. Un modèle utilisant notre approche avec trois flots : MFCC, dérivées et LSF a conduit à une amélioration significative, comparativement au système de base utilisant uniquement les MFCC et ce pour les milieux fortement bruités.

2 ***concernant l'évaluation de l'impact des coefficients de pondération sur le vecteur acoustique multi-flot proposé* :** L'utilisation d'une analyse multi-variable pose le problème d'assignation des poids de chaque source d'information à savoir, dans notre cas les MFCC et les LSF ainsi que leurs dérivées. En effet, il est très important de bien

choisir ces poids en fonction des environnements acoustiques changeants auxquels sont confrontés les systèmes de DSR. Pour cela, nous avons prospecté une approche alternative pour l'optimisation des poids, dite approche heuristique.qui est basée sur l'expérience. En utilisant cette approche, une amélioration relative dépassant les 20% a été obtenue.

3 ***Concernant l'optimisation de l'utilisation de ces flux de paramètres, par des transformations adéquates, en réduisant la dimension des vecteurs acoustiques tout en améliorant la robustesse du système*** **:** L'application de la KLT a permis une réduction optimale de paramètres, passant de la dimension 39 (système de base DSR-ETSI) à la dimension 24, ce qui offre un avantage certain en termes de débit et une amélioration de la robustesse, en éliminant les composantes principales bruitées. Les résultats obtenus par l'approche LSF-KLT donnent de meilleurs résultats par rapport à ceux du système DSR de base.

En résumé, cette thèse a permis de tester empiriquement de nouveaux *front-end*, comparativement à celui du système DSR de base de l'ETSI utilisant les MFCC seuls. Nous avons démontré expérimentalement que les LSF peuvent jouer un rôle important en conférant plus de robustesse à un DSR dans un environnement bruité. Il faut souligner que le processus d'extraction des nouveaux paramètres ajoute une certaine complexité au processus d'analyse. Cependant, l'utilisation d'un vecteur de dimension 24 au lieu de 39 diminuera le temps de calcul et la capacité de stockage pour le processus principal effectué sur le serveur (back-end). Ces considérations sont avantageuses dans le contexte de la DSR où robustesse et débit sont des contraintes importantes.

Perspectives générales

L'intention initiale de cette thèse était de contribuer à la conception d'un nouveau *front-end* basé sur un multi-Stream dans un système de reconnaissance de la parole distribuée. Cette contribution visait à améliorer les performances des systèmes basés sur les modèles de Markov cachés (HMM) en combinant les paramètres conventionnels MFCC et les LSF, afin de constituer un nouveau vecteur acoustique multi variable. Au cours de ce mémoire, nous avons d'abord effectué une analyse du codec basique de l'ETSI utilisé en DSR (ETSI-DSR). Nous avons démontré expérimentalement que les LSF augmentent la robustesse d'un tel

système en milieu fortement bruité. Ce travail est en cours d'exploitation pour évaluer l'apport de ces nouveaux paramètres dans un environnement fortement bruité. Cela consiste à mettre au point un *front end* robuste au bruit semblable à celui de l'ETSI *Advanced* ou *Extended front-end.*

Par ailleurs, notre travail ouvre de nombreuses perspectives de recherche telles que :

- le développement d'un module d'analyse acoustique mixte qui consiste à appliquer dans les milieux non bruités et faiblement bruités la technique classique et de basculer, dans la cas des milieux fortement bruités, à notre démarche.
- le développement d'une interface graphique qui sera fournie à l'utilisateur de l'application et permettant d'avoir plus de facilités pour la configuration du système. Cette application intégrera toutes les combinaisons possibles rencontrées en situation réelle.

Bibliographie

RÉFÉRENCES

[Abdulla et Kasabov, 1999] W. H. ABDULLA et N. KASABOV, "Reduced feature-set based parallel CHMM speech recognition systems." ICICS'99, Singapore, 1999.

[Addou et al., 2009] D. ADDOU, S.-A. SELOUANI, M. BOUDRAA, B. BOUDRAA, "Feature combination using multiple spectral cues for robust speech recognition in mobile communications", IEEE-Information Technology of New Generation, Las Vegas, USA, pp.1256-1261, 2009.

[Adjoudani et Benoît, 1996] A. ADJOUDANI, C.BENOÎT, "On the Integration of Auditory and Visual Parameters in an HMM-based ASR." In Speech reading by humans and machines, Hennecke (Springer, Berlin, Germany), pp. 461-471, 1996.

[Allen, 1994] J. B. ALLEN, "How do Humans Process and Recognize Speech?" IEEE Trans. on Speech and Audio Processing, Vol. 2, pp. 567-576, October 1994.

[Aubert et al. 1994] X. AUBERT, C. DUGAST, H. NEY et V. STEINBISS, "Large vocabulary, continuous speech recognition of Wall Street Journal Corpus." In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, volume 2, pages 129–132, Adelaide, Australia, 1994.

[Bahl et al. 1986] L. BAHL, P. BROWN, P. SOUZA, et R. MERCER, "Maximum mutual information estimation of hidden markov model parameters for speech recognition." IEEE International Conference on ICASSP '86. Acoustics, Speech, and Signal Processing, volume 11, pages 49–52, 1986.

[Billa et al. 1997] J. BILLA, K. MA, M. SIU, G. ZAVALIAGOS, "Acoustic modeling work at BBN." In proceedings of the hub5, conversational speech

recognition workshop, NIST, Maryland, 1997.

[Boite et al. 2000] R. BOITE, H. BOURLARD, T. DUTOIT, J. HANCQ et H. LEICH, "Traitement de la parole." Presses polytechniques et universitaires romandes, CH - 1015 Lausanne, 2000.

[Boll, 1979] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction." IEEE Transactions on Acoustics, Speech and Signal Processing, 27:113–120, 1979.

[Bourlard et Morgan, 1994] H. BOURLARD et N. MORGAN, "Connectionist Speech Recognition: A Hybrid Approach." Kluwer Press, 1994.

[Bourlard et al. 1996] H. BOURLARD, S. DUPONT, and C. RIS, "Multi-stream speech recognition." Tech. rep., IDIAP, 1996.

[Bourlard et Dupont, 1997] H. BOURLARD, S. DUPONT, "Subband-based speeh recognition." ICASSP'97, Münich, Allemagne, pp. 1251-1254, 1997.

[Caelen, 1985] J. CAELEN, "Space/Time Data-Information in the A.R.I.A.L. Project Ear Model." Speech Communication 4, pp. 163-179, Elsevier Science Publishers B.V., North-Holland, 1985.

[Calliope, 1989] CALLIOPE (ouvrage collectif). "La parole et son traitement automatique." Collection technique et scientifique des télécommunications, CNET - ENST, Masson, 718 pp, 1989.

[Cohen et al. 1998] P. COHEN, M. JOHNSTON, S. OVIATT, J. CLOW et J. SMITH, "The efficiency of multimodal interaction". In proceedings of ICSLP98, Sidney, Australia, 1998.

[Cole et al. 1995] R. COLE et al. "The Challenge of Spoken Languages Systems: Research Directions for the Nineties". IEEE transactions on speech and audio processing, Vol. 3, N°1, pp. 1-20, 1995.

[Dautrich et al. 1983] B. A. DAUTRICH, L. R. RABINER, et T. B. MARTIN, "On the effects of varying filter bank parameters on isolated word

recognition." IEEE Transactions on Acoustics, Speech and Signal Processing, 31:793–806, 1983.

[Davis et Mermelstein, 1980] S. B. DAVIS and P. MERMELSTEIN, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, pp. 357-366, 1980.

[Dempster et al. 1977] A. P. DEMPSTER, N. M. LAIRD, et D. B. RUBIN, "Maximum Likelihood from incomplete data via the EM Algorithm." Journal of Royal Statistical Society Ser. B, 39:1–39, 1977.

[Dupont et Bourlard, 1997] S. DUPONT et H. BOURLARD, "Using multiple time scales in a multi-stream speech recognition system." EUROSPEECH'97, Rhodes, Greece, 1997.

[Ellis et Bilmes, 2000] D. P. W. ELLIS and J. A. BILMES, "Using mutual information to design feature combinations." In Proceedings ICSLP, vol. 3, pp. 79-82, 2000.

[Ephraïm, 1995] Y. EPHRAIM et H.L. VAN TREES, "A signal subspace approach for speech enhancement." IEEE Trans. Speech and Audio Processing, vol. 3, pp 251–266, 1995.

[ETSI, 2003] ETSI. "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithm." tech. rep., ETSI ES 201 108, 2003.

[Fukunaga, 1990] K. FUKUNAGA, "Introduction to statistical pattern recognition." Second edition, Academic Press, 1990.

[Furui, 1992] S. FURUI, "Recent advances in speech recognition technology at NTT laboratories" Speech Communication, 195–204, 1992.

[Gales, 1997] M.J.F. GALES, "Maximum likelihood linear transformations for hmm-based speech recognition." In CUED, 1997.

[Gales, 1998] M.J.F. GALES, "Predictive model-based compensation schemes for robust speech recognition." Speech Communication, Vol. 25, pp. 49-74, 1998.

[Gao et al. 2000] Y. GAO, B. RAMABHADRAN et M. PICHENY, "New adaptation techniques for large vocabulary continuous speech recognition." In ASR, 2000.

[Garner et Holmes, 1998] P. GARNER and W. HOLMES, "On the robust incorporation of formant features into hidden Markov models for automatic speech recognition". In Proceedings IEEE ICASSP, pp. 14, 1998.

[Gauvain et Lee, 1994] J.L. GAUVAIN et C.H. LEE, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov Chains." IEEE Transactions on Speech and Audio Processing, 2(2):291–298, 1994.

[Gong, 1994] Y. GONG, "Stochastic Trajectory Modeling and Sentence Searching for Continuous Speech Recognition." Rapport de recherche, CRIN – CNRS & INRIA Lorraine, 1994.

[Gong, 1995] Y. GONG, "Speech recognition in noisy environments: a survey". Speech communication, vol. 16, pp. 261–291, 1995.

[Haeb-Umbach et Ney, 1992] R. HAEB-UMBACH and H. NEY. "Linear discriminative analysis for improved large vocabulary continuous speech recognition." In Proceedings ICASSP, vol. 2, pp. 13-16, 1992.

[Haton et al. 1991] J.-P. HATON, J.-M. PIERREL, G. PERENNOU, J. CAELEN et J.-L. GAUVAIN, "Reconnaissance automatique de la parole". Collection AFCET - Dunod informatique, 1991.

[Haton, 1995] J. P. HATON, "Modèles neuronaux et hybrides en reconnaissance de la parole: état des recherches". Fondements et perspectives en traitement automatique de la parole, H. Méloni eds. pp.139-154, 1995.

[Haton, 1997] J. P. HATON, "Proceeding of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels." Pont-à-Mousson, France, 1997.

[Heigold et al. 2005] G. HEIGOLD, W. MACHEREY, R. SCHLUTER et H. NEY, "Minimum exact word error training." In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 186–190, 2005.

[Hermansky, 1990] H. HERMANSKY, "Perceptual linear predictive (plp) analysis of speech." Journal Acoustic Society of America, pp. 1738-1752, 1990.

[Hermansky et Cox, 1991] H. HERMANSKY et Jr. COX, "Perceptual linear predictive (plp) analysis synthesis technique." In IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics Final Program and Paper Summaries, pp. 37–38, 1991.

[Hermansky et Morgan, 1994] H. HERMANSKY and N. Morgan, "Rasta processing of speech." IEEE Trans. Speech Audio Processing, vol. 2, pp. 578-589, 1994.

[Hermansky, 1997] H. HERMANSKY, "Should recognizer have ears? Robust speech recognition for unkown communication channels". Pont-à-Mousson, France, 1997.

[Holmes et al., 1997] J. HOLMES, W. HOLMES, and P. GARNER, "Using formant frequencies in speech recognition." In Proceedings IEEE EUROSPEECH, pp. 2083-2086, 1997.

[Itakura, 1975] F. ITAKURA, "Line spectrum representation of linear predictive coefficients of speech signals," Journal of the Acoustical Society of America, vol. 57, no. 1, 1975.

[ITU-T, 1996] ITU-T Recommendation G.723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," 1996.

[ITU G.712, 1996] ITU Recommendation G.712, "Transmission performance characteristics of pulse code modulation channels", 1996.

[Janin et al. 1999] A. JANIN, D. ELLIS, and N. MORGAN, "Multi-stream speech recognition: ready for prime time." In Proceedings Eurospeech, vol. 2, pp. 591-594, 1999.

[Jelinek, 1976] F. JELINEK, "Continuous speech recognition by statistical methods." 64(4), pp. 532–556, 1976.

[Joliffe, 2002] I. T. JOLLIFFE. "Principal Component Analysis". Second Edition. Springer, 2002.

[Junqua, 1987] J. C. JUNQUA, "Evaluation of ASR Front-ends in Speaker-dependent and Speaker-independent Recognition." Journal of Acoustical Society of America, 81 S1:S93, 1987.

[Junqua, 1989] J.C. JUNQUA, "Contribution à l'amélioration de la robustesse des systèmes de reconnaissance automatique de mots isolés". Thèse de doctorat, Université de NANCY, 1989.

[Junqua, 1996] J.C. JUNQUA, et J. P. HATON, "Robustness in Automatic Speech Recognition: Fundamentals and Applications." The Kluwer International Series in Engineering and Computer Science, 1996.

[Kain, 2001] A. KAIN, "High resolution voice transformation." These, Oregon Health and Science University, 2001.

[Kemp et Waibel, 1999] T. KEMP, A. WAIBEL, "Unsupervised training of a speech recognizer using TV broadcast". In proceedings of ICSLP98, Vol.5, Sidney, pp.2207-2211, 1998.

[Levinson et Rabiner, 1983] S. E. LEVINSON, L. R. RABINER, and M. M. SONDHI. "An introduction to the application of the theory of probabilistic functions of Markov process to automatic speech recognition". The Bell System Technical Journal, 62(4):pp. 1035-1074, 1983.

[Lippmann, 1997] R. P. LIPPMANN et B. A. CARLSON, "Robust speech recognition with time-varying filtering, interruptions, and noise." Ed. Furui S., Juang B.-H. et Chou W., IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 365-372, Santa Barbara, USA, 1997.

[Loizou, 2007] P. LOIZOU, "Speech enhancement: Theory and practice." CRC; 1 édition, 2007.

[Markel et Jr, 1976] J. D. MARKEL et A. H. Gray JR, "Linear prediction of speech." In Communication and Cybernetics. Berlin Heidelberg New York: Springer-Verlag, 1976.

[Martinez et al. 1997] R. MARTINEZ, A. ALVAREZ, V. NIETO, V. RODELLAR et P. GOMEZ, "ASR in highly non-stationary environment using adaptive noise cancelling techniques. Robust Speech Recognition For unknown communication channels." Pont-à-Mousson, France, 1997.

[McAulay et Quatieri 1995] R. MCAULAY et T. QUATIERI, "Speech coding and synthesis." Sinusoidal coding, Elsevier science, pp. 121-173, 1995.

[O'Shaughnessy, 2000] D. O'SHAUGHNESSY, "Speech Communication: Human and Machine". IEEE Press, 2000.

[Paliwal, 1993] K. PALIWAL, "Use of temporal correlation between successive frames in a HMM based speech recognizer." In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol 2, pp. 215–218, Minneapolis, Minnesota, USA, 1993.

[Potamianos et Graf, 1998] G. POTAMIANOS et H. P. GRAF. "Discriminative training of HMM stream exponents for audio-visual speech recognition." In Proceedings IEEE ICASSP, vol. 6, pp. 3733-3736, 1998.

[Povey et Woodland, 2002] D. POVEY et P.C. WOODLAND, "Minimum phone error and i-smoothing for improved discriminative training." In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), Vol. 1, pp. 105–108, 2002.

[Rabiner, 1986] L.R. RABINER et B.H. JUANG, "An introduction to Hidden Markov Models." IEEE Acoustics Speech and Signal Processing Magazine, ASSP-3(1) :4–16, 1986.

[Rabiner, 1989] L. R. RABINER, "A tutorial on hidden Markov model and selected applications in speech recognition." In Proceedings of the IEEE, vol. 77, pp. 257-286, 1989.

[Rabiner et Juang, 1993] L. R. RABINER et B. H. JUANG, "Fundamentals of Speech Recognition." Prentice Hall International Editions, 1993.

[Ravindran et al. 2006] S. RAVINDRAN, D. ANDERSON, and M. SALNEY. "Improving noise robustness of mel frequency cepstral coefficients," in Proc. IEEE ICASSP, vol. 2, 2006.

[Rose et Momayez, 2007] R. ROSE and P. MOMAYEZ, "Integration of multiple feature sets for reducing ambiguity in ASR." in Proceedings IEEE ICASSP, 2007.

[Sarikaya et al. 2005] R. SARIKAYA, Y. GAO, M. PICHENY et H. ERDOGAN, "Semantic confidence measurement for spoken dialog systems."

13(4):pp. 534–545, 2005.

[San-Segundo et al. 2001] R. SAN-SEGUNDO, B. PELLOM, K. HACIOGLU, W. WARD et J.M. PARDO, "Confidence measures for spoken dialogue systems." In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01), Vol. 1, pp. 393–396, 2001.

[Saporta, 1990] G. SAPORTA, "Probabilités analyse des donnés et statistique. Editions Technip, 1990.

[Saoudi, 1990] S. SAOUDI, "Codage de la parole par les paires de raies spectrales." Thèse, Université de Rennes, 1990.

[Schmid et Barnard, 1995] P. SCHMID and E. BARNARD, "Robust, n-best formant tracking." In Proceedings 4th European conference on speech communication and technology, pp. 737-740, 1995.

[Schluter et al. 2006] R. SCHLUTER, A. ZOLNAY, and H. NEY, "Feature combination using linear discriminate analysis and its pitfalls." In Proceedings IEEE interspeech, 2006.

[Sakoe et Chiba, 1978] H. SAKOE and S. CHIBA, "Dynamic programming algorithm optimization for spoken word recognition." IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43-49, 1978.

[Sharma et al. 2006] A. SHARMA, K.K. PALIWAl and G.C. ONWUBOLU, "Class-dependent PCA, MDC and LDA: A combined classifier for pattern classification." Pattern Recognition, vol. 39, no. 7, pp. 1215-1229, 2006.

[Selouani et al. 2002] S.-A. SELOUANI, H. TOLBA and D. O'SHAUGHNESSY, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm." In Proceedings IEEE ICASSP, vol. 1, pp. 797-800,

2002.

[Selouani et al. 2003] H. TOLBA, S.-A. SELOUANI and D. O'SHAUGHNESSY, "Comparative experiments to evaluate the use of auditory-based acoustic distinctive features and formant cues for robust automatic speech recognition in low-SNR car environments." in Proceedings EUROSPEECH, pp. 3085-3088, 2003.

[Siohan et al. 1995] O. SIOHAN, Y. GONG, et J.P. HATON, "Noise adaptation using linear regression for continuous noisy speech recognition." In Proceedings of European Conference on Speech Communication and Technology, Madrid, Spain, 1995.

[Sugamura et Itakura, 1986] N. SUGAMURA et F. ITAKURA, "Speech analysis and synthesis methods developed at ECL in NTT from LPC to LSP." Speech Communication 5(2), pp. 199-215, 1986.

[Tamura et al. 2005a] S. TAMURA, K. IWANO, and S. FURUI, "A stream-weight optimization method for multi-stream hmm based on likelihood value normalization." In Proceedings ICASSP, vol. 1, pp. 469-472, 2005.

[Tamura et al. 2005b] S. TAMURA, K. IWANO, and S. FURUI, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMM." In Proceedings ICASSP, vol. 1, pp. 857-860, 2005.

[Tan Zheng, 2008] TAN ZHENG-HUA, "Automatic speech recognition on mobile devices and over communication networks." Lindberg Borges (Eds.) Springer-Verlag, XX, 404 p. 115, 2008.

[Tibrewala et Hermansky, 1997] S. TIBREWALA et H. HERMANSKY, "Multi-band and Adaptation Approachs to Robust Speech Recognition." Eurospeech'97, 5th European Conference on Speech Communication and Technology, Rhodes, Greece, 1997.

[Unser et al., 1993] M. UNSER, A. ADROUBI et EDEN, "B-spline signal processing." IEEE Transactions on Speech and Audio Processing 41(2), pp. 821-833, 1993.

[Valtchev et al. 1997] V. VALTCHEV, J. ODELL, S. J. WOODLAND, "Mmie training of large vocabulary recognition system." In Speech Communication 22, pp 303-314, 1997.

[Woodland, 1999] P. C. WOODLAND, "Speaker adaptation: techniques and challenges." In proceedings of ASRU, Vol.1, pp.85-89, Keystone 1999.

[Yan et al. 2008] Z.-J. YAN, B. ZHU, Y. HU et R.-H. WANG, «Minimum word classification error training of hmm for automatic speech recognition." In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, pp. 4521–4524, 2008.

[Yi Hu et Loizou, 2006] Y. HU et P. LOIZOU, "Evaluation of objective Measures for speech enhancement". In Proc. Interspeech, pp. 1447–1450, 2006.

[Young, 1993] S. YOUNG, "The HTK hidden markov model toolkit: Design and philosophy." Tech. rep., Cambridge University Engineering Department, Speech Group, Cambridge, 1993.

[Zhan et Waibel, 1997] P. ZHAN et A. WAIBEL, "Vocal tract length normalization for large vocabulary continuous speech recognition." In CMU-CS, 1997.

Zolnay et al., 2005] A. ZOLNAY, R. SCHLUTER, and H. NEY, "Acoustic feature combination for robust speech recognition." In Proceedings ICASSP, vol. 1, pp. 457-460, 2005.

Annexes

ANNEXE A

Un Système de reconnaissance de la parole sous HTK

Cette annexe a pour objectif de présenter les étapes de la conception d'un système de reconnaissance de la parole à petit vocabulaire en utilisant l'outil HTK.

Pour étudier l'impact du débruitage sur les performances d'un système de reconnaissance de la parole en présence du bruit, nous avons développé, dans le cadre de cette thèse, un système opérationnel, indépendant du locuteur et fondé sur les modèles de Markov cachés. Nous l'avons conçu à partir de la plate forme HTK (Annexe B) de l'Université de Cambridge et sur la base de données de parole Aurora TIdigits. La boîte à outils HTK est efficace, flexible (liberté du choix des options et possibilité d'ajout d'autres modules) et complète dans le sens où elle fournit une documentation très détaillée (le livre HTK [**Young, 2006**] est une encyclopédie dans le domaine).

A ce stade, on conçoit notre système en se basant sur des mots représentés comme unités acoustiques. On commence par définir les ressources nécessaires dont on aura besoin par la suite. On définit, alors, le modèle de langage, appelé aussi lexique ou grammaire (Tableau A.1), qui décrit l'enchaînement des mots dans les phrases. Ensuite, on construit le réseau de mots (wdnet) et le dictionnaire (Tableau A.2) respectivement, grâce aux outils HTK, HParse (Commande A0) et HDMan (Commande A1). Pour la base de données Aurora TIdigits, qui est une base de chiffres en anglais américain, le vocabulaire est assez limité, d'où la facilité de définir le dictionnaire et la grammaire (Tableaux A.1 et A.2).

```
HParse grammaire wdnet                          (A0)
HDMan −m −w wlist −g global.ded −l dlog dict    (A1)
```

où wlist représente la liste des mots (constituant la base de données TIdigits) ordonnés par ordre croissant qui vont être transcrits en mots et sauvegardés dans le dictionnaire dict par la commande HDman. Le fichier dlog contient toutes les statistiques de la phase de construction du dictionnaire, notamment des erreurs s'il y en a.

Tableau A.1—Grammaire de la base Aurora TIdigit (wdnet)

```
$digit = one|two|three|four|five|six|seven|eight|nine|zero|oh ;

   (sil <$digit> sil)
```

Tableau A.2—Dictionnaire de la base Aurora TIdigit (dict)

```
one one
two two
three three
four four
five five
six six
seven seven
eight eight
nine nine
zero zero
oh oh
sp sp
sil sil
```

Une fois qu'on a défini le dictionnaire, la grammaire et la liste des mots, on passe à la description des modèles de Markov cachés. On construit un modèle HMM pour chaque unité acoustique. La topologie HMM choisie est de type gauche-droit à 18 états dont les transitions autorisées sont décrites dans la figure (A.1) et initialisées dans la matrice de transition. La moyenne est initialisée à 0 et la variance à 1. Ces paramètres du modèle HMM seront ré-estimés par la suite lors de la phase d'apprentissage. Le fichier de configuration (Tableau A.3) permet de définir les paramètres indispensables pour la phase de l'analyse acoustique. Le choix s'est porté sur les 12 premiers coefficients MFCC excepté le coefficient C_0 qui est substitué par le logarithme de l'énergie du signal, d'où le terme -E dans le fichier de configuration.

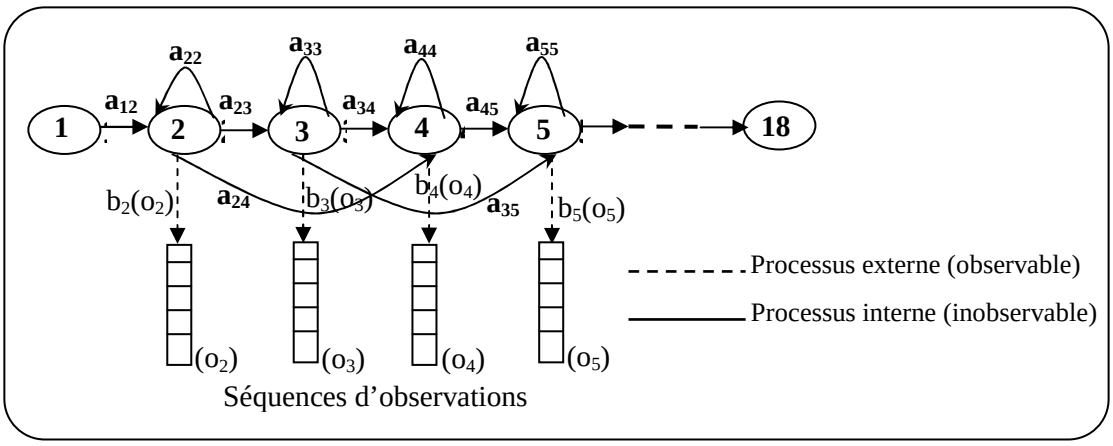


Figure A.1—Exemple de structure à 18 états d'un HMM. Les états de 2 à 17 sont émetteurs alors que l'état initial 1 et l'état final 18 ne génèrent pas d'observations.

Pour chaque coefficient plus l'énergie, on attribue une dérivée première (13 dérivées premières au total) ainsi qu'une dérivée seconde (13 dérivées secondes) pour prendre en compte la dynamique du signal. En somme, on obtient un vecteur acoustique de 39 coefficients correspondant à chaque trame du signal. Ces coefficients sont extraits des fichiers wav et sur des fenêtres de 25ms grâce à l'outil HCopy en se servant du fichier de configuration (Tableau A.3) comme paramètre d'entrée selon la commande A2.

```
HCopy – T 1 –C config –S liste_train.scp      (A2)
```

Tableau A.3—Fichier de configuration pour la phase de l'analyse acoustique

```
SOURCEFORMAT = WAV------> Format des signaux en entrée de la phase
d'analyse acoustique
TARGETKIND = MFCC-E-D-A ------> Type de paramétrisation utilisée
WINDOWSIZE = 250000.0 ------> Durée de la trame (25ms)
TARGETRATE = 100000.0 ------> Périodicité de la trame
PREEMCOEF = 0.97 ------> Coefficient de pré-accentuation
NUMCHANS = 26 ------> Nombre de canaux du banc de filtres Mel
NUMCEPS = 12 ------>Nombre de coefficients cepstraux MFCC
CEPLIFTER = 22 ------>Coefficient de lissage
```

Une étape indispensable, également, concerne la transcription de chaque signal appartenant à la base d'apprentissage. D'habitude, les bases de données de parole sont accompagnées de leur transcription. Cependant, avec la base TIdigits, ce n'est pas le cas. Heureusement, dans notre cas, la transcription n'est pas compliquée, parce que les signaux .wav de cette base portent chacun un nom qui correspond à la phrase prononcée par un certain locuteur. Le résultat de la transcription est sauvegardé dans le fichier words.mlf illustré par le tableau A.4. A partir de ce dernier fichier, on génère une transcription, à travers l'outil HTK, HLEd selon la ligne de commande A3.

```
HLEd -l '*' -d dict -i mkphones0.led words.mlf    (A3)
```

où mkphones0.led est un script permettant de remplacer chaque mot par la prononciation lui correspondant dans le dictionnaire et d'insérer un silence au début et à la fin de chaque expression (mkphones.txt (Tableau A.4).

Tableau A.4—Fichiers de transcription en mots

```
# !MLF !#
"*/9212ZA.lab"
sil
nine
two
one
two
zero
sil
.
"*/54B.lab"
sil
five
four
sil
.
...
```

Le fichier "words.mlf"

```
EX
IS sil sil
DE sp
```

Le fichier "mkphones.txt"

Apprentissage : La phase d'apprentissage permet de constituer la base de données des modèles de référence du système. La qualité de cette modélisation conditionne en grande partie les résultats du système de reconnaissance de la parole. L'apprentissage est réalisé sous

HTK en deux étapes majeures : *l'initialisation et la ré-estimation*. Pour cette raison, deux outils sont souvent sollicités : HCompV et HERest. La phase d'initialisation des modèles HMM par l'outil HCompV (Commande A4) permet de mettre à jour la moyenne et la variance qui valent, avant cette étape, respectivement, 0 et 1 (voir fichier prototype d'initialisation tableau A.5). Cette mise à jour est réalisée sur l'ensemble des données du corpus d'apprentissage permettant d'aboutir, à la fin, à des valeurs globales qui seront clonées pour chaque état des modèles HMM.

```
HCompV −C config −f 0.01 −m −S liste_train −M hmm0 proto (A4)
```

Suite à cette commande, on obtient dans le répertoire hmm0 un nouveau fichier prototype contenant des valeurs globales de la moyenne et de la variance. On copie le contenu de ce fichier autant de fois qu'on a de mots et on stocke le résultat du clonage dans un fichier macro nommé modèles.mmf. Tous les mots seront ainsi initialisés aux mêmes valeurs de moyenne et de variance. Par ailleurs, l'option -f de la commande (A4) permet de générer un fichier vFloor contenant la variance seuil qui est une fraction de la variance globale estimée. L'intérêt de ce seuil est de fixer une limite à la variance lors des étapes d'estimation afin d'´eviter des valeurs aberrantes. A noter également que la mise à jour des variances est effectuée par défaut avec la commande HCompV, tandis que pour ré-estimer la moyenne, l'option -m devient indispensable. Par la suite, le raffinement des modèles HMM consiste à ré-estimer leurs paramètres (moyenne et variance) suivant l'algorithme de Baum Welch[1] grâce à l'outil HERest (Commande A5) et selon trois itérations.

Les modèles ainsi estimés seront sauvegardés dans le répertoire hmm3 (ré-estimation des modèles HMM contenu dans le répertoire hmmi et sauvegarde dans le répertoire hmmi+1 à chaque itération i).

```
HERest −C config −I words.mlf −t 250.0 150.0 1000.0 −S
liste_train −H hmm0/macros − H hmm0/modeles0.mmf − M hmm1
words_list                                                (A5)
```

A ce niveau, on ne considère pas encore le modèle de short pause "sp". Le fichier modeles0 est ainsi une version restreinte de modeles1 dans le sens où on en enlève le phonème "sp".

[1] Détail de cet algorithme dans le livre d'HTK

Tableau A.5—Fichier prototype d'initialisation

```
    <BeginHMM>
     <NumStates> 18 <VecSize> 39 <MFCC_E_D_A> <nullD> <diagC>
     <StreamInfo> 1 39
     <State> 2 <NumMixes> 1
      <Stream> 1
      <Mixture> 1 1.0
        <Mean> 39
    0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
        <Variance> 39
    1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
     <State> 3 <NumMixes> 1
      ...
     <State> 4 <NumMixes> 1
      ...
     <State> 5 <NumMixes> 1

     <State> 6 <NumMixes> 1
      ...
     <State> 7 <NumMixes> 1
      ...
     <State> 8 <NumMixes> 1

     <State> 9 <NumMixes> 1
      ...
     <State> 17 <NumMixes> 1
    <TransP> 18
    0.000e+0  1.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0
0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0
0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  6.000e-1  4.000e-1  0.000e+0
0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  6.000e-1  4.000e-1
0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  6.000e-1
4.000e-1  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0  0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0  0.000e+0  0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0
    0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0  0.000e+0
0.000e+0  0.000e+0  0.000e+0  0.000e+0  6.000e-1  4.000e-1  0.000e+0  0.000e+0  0.000e+0
```

Nombre d'états HMM

Nombre de gaussiennes

Matrice de transition entre états

D'un autre côté le fichier macros est une version de vFloor à laquelle on a ajouté l'entête, ~o <MFCC E D A> <VecSize> 39, définissant le type de paramétrisation et la taille du vecteur MFCC. L'ajout du modèle de silence "sp" aux autres modèles HMM est réalisé différemment. La procédure consiste à l'attacher à l'état central (état 3) du modèle de silence "sil" (figure A.2).

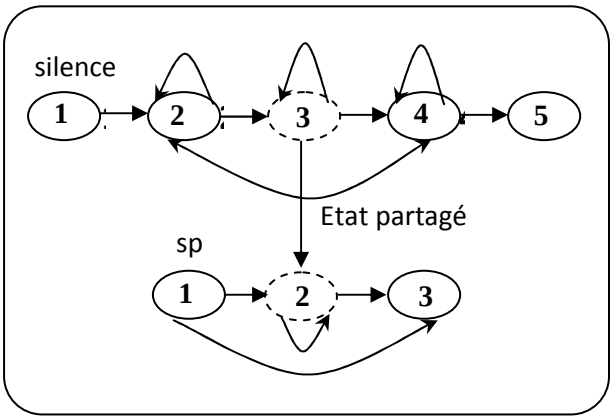


Figure A.2—Fixation du modèle de silence sp

En pratique, on va copier l'état 3 du modèle HMM du phonème "sil" et on va l'attribuer à l'état 2 du modèle de pause "sp". Celui-ci ne possède que 3 états dont le premier et le dernier ne sont pas émetteurs. On initialise la matrice de transition de ce modèle à des valeurs aléatoires qui seront ré-estimées par la suite. A signaler que lors de cette étape et grâce à l'outil HHEd (Commande A6), on ajoute, exclusivement au modèle de silence "sil", une probabilité de transition de l'état 4 à l'état 2 (Tableau A.6).

```
HHEd -H hmm4/macros -H hmm4/modeles0.mmf -M hmm5 sil.hed
word_listsp                                                    (A6)
```

Tableau A.6—Script sil.hed

```
MU 2 (sil.state[2-4].mix)
AT 2 4 0.2 (sil.transP)
AT 4 2 0.2 (sil.transP)
AT 1 3 0.3 (sp.transP)
TI silst (sil.state[3],sp.state[2])
```

La commande (A6) permet d'attacher le modèle de pause "sp" au modèle de silence "sil" selon la figure A.2. Suite à cette commande, on a généré un autre fichier modeles0 dans le répertoire hmm5. Les modèles contenus dans ce fichier seront ré-estimés suite à trois autre itérations de l'algorithme de Baum Welch représenté par l'outil HERest exactement comme lors de l'étape (A5). Les derniers paramètres estimés, à ce stade, sont sauvegardés dans le répertoire hmm8. Ainsi s'achève la phase d'apprentissage des modèles HMM avec une seule gaussienne.

Amélioration des modèles: Les modèles obtenus peuvent être améliorés par utilisation de densités de probabilités d'émission multi-gaussiennes au lieu de se contenter d'une simple loi normale à matrice diagonale. Cela permet d'éviter certaines hypothèses grossières sur la forme de la densité si le nombre de gaussiennes est suffisant. En effet, le choix du nombre optimal de gaussiennes est un problème difficile. En pratique, la seule recommandation donnée est l'augmentation incrémentale et progressive du nombre de gaussiennes jusqu'à atteindre le nombre voulu. Une commande d'HTK HHEd (Commande A7) réalise l'augmentation du nombre de gaussiennes via le script mix2.hed ensuite mix3.hed (Tableau A.7), où on augmente progressivement le nombre de gaussiennes (1, 2, 4, 8, 12, 16). Chaque augmentation de gaussienne est suivie d'un multiple de trois ré-estimations des modèles avec HERest.

```
HHEd -B -H hmm8/macros -H hmm8/modeles0.mmf -M hmm9
mix2_16.hed word_listsp                                    (A7)
```

Tableau A.7—Scripts mix2.hed et mix3.hed

```
MU 2 {one.state [2-17].mix}
MU 2 {two.state[2-17].mix}
MU 2 {three.state[2-17].mix}
MU 2 {four.state[2-17].mix}
MU 2 {five.state[2-17].mix}
MU 2 {six.state[2-17].mix}
MU 2 {seven.state[2-17].mix}
MU 2 {eight.state[2-17].mix}
MU 2 {nine.state[2-17].mix}
MU 2 {oh.state[2-17].mix}
MU 2 {zero.state[2-17].mix}
MU 3 {sil.state[2-4].mix}
MU 3 {sp.state[2].mix}
```

```
MU 3 {one.state[2-17].mix}
MU 3 {two.state[2-17].mix}
MU 3 {three.state[2-17].mix}
MU 3 {four.state[2-17].mix}
MU 3 {five.state[2-17].mix}
MU 3 {six.state[2-17].mix}
MU 3 {seven.state[2-17].mix}
MU 3 {eight.state[2-17].mix}
MU 3 {nine.state[2-17].mix}
MU 3 {oh.state[2-17].mix}
MU 3 {zero.state[2-17].mix}
MU 6 {sil.state[2-4].mix}
MU 6 {sp.state[2].mix}
```

Suite à cette procédure, les modèles sont de plus en plus précis. Le seul inconvénient est la charge des calculs qui augmente à son tour. Les derniers modèles estimés sont sauvegardés dans le répertoire hmm20 (Commande A8).

```
HERest -C Config -I word.mlf -S liste_train -H hmm19/macros -H
hmm19/models0.mmf -M hmm20 word_listsp                    (A8)
```

Reconnaissance: Le processus de décodage consiste à comparer l'image de l'unité à identifier avec celles de la base de référence. Le module de décodage de la parole, HVite, utilise l'algorithme de Viterbi[2] pour trouver la séquence d'états la plus probable correspondant aux paramètres observés et en déduire les unités acoustiques correspondantes. Le décodage est réalisé par l'algorithme de Viterbi sous la contrainte d'un réseau syntaxique et éventuellement d'un modèle de langage.

Nous avons testé les performances de notre système de reconnaissances sur la base TIdigits. A noter que la base de développement nous a permis d'ajuster le paramètre p utilisé par la commande HVite dans (A9) et (A10). Ce paramètre est d'autant plus optimal que le nombre de suppressions S et à peu prés égal au nombre d'insertions I.

```
HVite −H hmm20/macros −H hmm20/modeles –C Config −S liste_test
−l '*' −i resultats_word.mlf −w wdnet –p 0.0 −s 0.0 dict
words_listsp                                              (A9)
```

Enfin, les résultats du décodage sont évalués par alignement dynamique avec les données de référence via l'outil HResults.

```
HResults –I word_test.mlf word_listsp resultats_word.mlf (A10)
```

Nous obtenons, par exemple, le résultat suivant :

```
--------Babble condition: N2_SNR20---MFCC_E_D_A (39) ----------
====================== HTK Results Analysis =======================
  Date: Sun May 23 08:35:27 2010
  Ref : REC\RECOGNIZER\labels\N2.mlf
  Rec : FEAT\test_MFCC0E/result_clean.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=94.81 [H=949, S=52, N=1001]
WORD: %Corr=98.55, Acc=98.25 [H=3260, D=15, S=33, I=10, N=3308]
```

[2] Détail de cet algorithme dans le livre d'HTK.

Ces résultats fournissent les taux de reconnaissance des mots corrects %Corr ainsi que la précision de la reconnaissance de ces mots %Acc. La précision tient compte également des insertions contrairement à %Corr. Lors de nos évaluations, nous tiendrons compte que de %Acc, mais, par abus de langage, nous le noterons taux de reconnaissance.

ANNEXE B La plate-forme de HTK

La plate-forme HTK (Hidden Markov Model Toolkit, ou "boîte à outils de modèles de Markov cachés") a été développée à l'Université de Cambridge par S.J.Young et son équipe [**Young, 1999**]. Elle est constituée d'un ensemble d'outils logiciels qui permettent de construire des systèmes de reconnaissance de la parole continue à base de modèles de Markov cachés. HTK est remarquable par la très grande liberté de choix laissée tout au long de la construction du système de reconnaissance. Les modèles peuvent représenter des mots ou tout type d'unité sub-lexicale, et leur topologie est librement configurable. Les densités de probabilité d'émission, qui sont associées aux états, sont décrites par des multi-gaussiennes. Les modèles sont initialisés avec l'algorithme de Viterbi, puis ré-estimés par l'algorithme optimal de Baum-Welch. Le décodage est réalisé par l'algorithme de Viterbi, sous la contrainte d'un réseau syntaxique défini par l'utilisateur, et le résultat est enfin évalué par alignement dynamique avec la chaîne phonétique ou lexicale.

L'ensemble de ces outils est écrit en langage C, et la documentation détaille leur utilisation et les principes de leur implémentation, ce qui permet d'intégrer de manière efficace les modifications souhaitées dans le système de reconnaissance. Ce système nous servira ensuite à évaluer de manière quantitative l'amélioration apportée par nos traitements spécifiques.

HTK dans sa version 1.4 est structuré en 9 librairies utilisées par 12 outils de base (Tableau B.1). Les outils manipulent des fichiers de différents types: signaux, étiquettes, paramètres, description des modèles et définition de réseaux. Les formats des fichiers de signaux et d'étiquettes des bases de données les plus répandues sont reconnus. Les autres fichiers sont dans un format particulier à HTK, décrit dans le manuel de référence. En particulier, les modèles et les réseaux sont définis dans des fichiers texte, ce qui facilite leur création et leur modification par l'utilisateur. Les options d'utilisation des outils sont transmises en argument sur la ligne de commande. Il est donc facile d'automatiser les processus d'apprentissage et de décodage avec des scripts écrits dans le langage de commande du système d'exploitation (par exemple en C-shell si l'on travaille sous Unix).

Tableau B.1—Librairies et outils de base de HTK [Young 2006]

Librairies		*Outils de base*	
HShell	Interface avec le système d'exploitation	HLEd	Édition des fichiers d'étiquettes
HMath	Procédures mathématiques	HHEd	Édition des modèles
HSigP	Procédures de traitement du signal	HCode	Calcul des paramètres du signal
HDBase	Stockage en mémoire des paramètres	HInit	Initialisation d'un modèle
HSpIO	Gestion des fichiers de données	HRest	Ré-estimation d'un modèle
HLabel	Gestion des fichiers d'étiquettes	HERest	Ré-estimation des modèles enchaînés
HModel	Gestion des modèles	HVite	Décodage en parole continue
HParse	Lecture du réseau syntaxique	HResults	Résultats du décodage
HGraf	Affichage graphique	HList	Affichage des fichiers de données
		HLStats	Calcul de statistiques sur les étiquettes
		HSLab	Affichage du signal et des étiquettes
		HLab2Net	Génération automatique d'un réseau

Les principaux outils de base de HTK s'enchaînent naturellement pour réaliser les différentes étapes d'un système de reconnaissance: calcul des paramètres du signal, apprentissage des modèles, et expériences de reconnaissance (Figure B.1).

Prétraitement des données : Avant l'apprentissage des modèles, il est nécessaire de préparer les données en calculant les paramètres du signal et en étiquetant les phrases d'apprentissage. La représentation du signal est obtenue avec l'outil HCode, qui produit en particulier des coefficients LPCC ou MFCC ainsi que l'énergie. Les coefficients différentiels du premier et du second ordre peuvent être calculés ultérieurement lors de la lecture des fichiers de paramètres, ce qui économise leur stockage en mémoire de masse. Les phrases d'apprentissage doivent être toutes étiquetées en fonction des unités acoustiques modélisées. Les bases de données de parole sont parfois fournies avec un étiquetage phonétique qui ne correspond pas exactement aux unités acoustiques modélisées. L'éditeur HLEd permet alors de modifier les étiquettes, par exemple pour regrouper plusieurs phonèmes différents dans une seule classe ou pour fusionner deux segments adjacents.

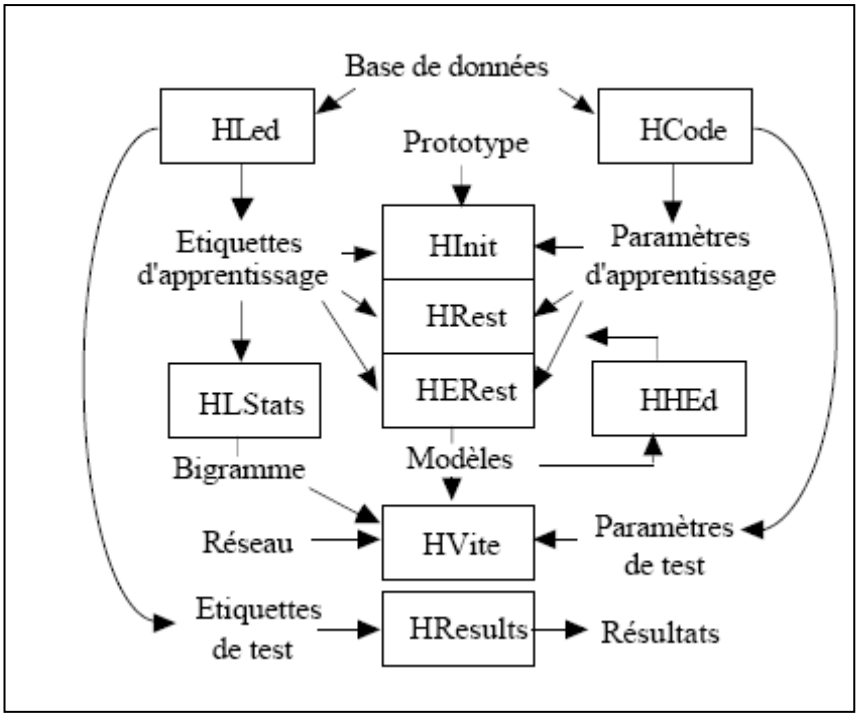


Figure B.1—Structure d'un système de reconnaissance avec HTK [Young, 2006]

Topologie de modèles : Pour chaque unité acoustique, il faut définir un modèle prototype contenant la topologie choisie, à savoir le nombre d'états du modèle, les états entre lesquels les transitions sont possibles, le type de loi de probabilité associée à chaque état. L'état initial et l'état final ont la particularité de ne pas émettre d'observation, mais de servir uniquement à la connexion des modèles en parole continue. Il n'est pas nécessaire que tous les modèles utilisent le même prototype. Les probabilités d'émission sont associées aux états et sont décrites par une combinaison linéaire de gaussiennes multi-variables, caractérisées par leur moyenne et leur matrice de covariance dans l'espace des paramètres. La matrice de covariance est théoriquement symétrique, mais peut être choisie diagonale si l'on suppose l'indépendance entre les composantes des vecteurs de paramètres, et les vecteurs de paramètres peuvent être séparés en flux de données indépendants

Apprentissage : Pour chacune des machines modélisant une unité acoustique, l'outil HInit initialise les probabilités d'émission des états du modèle au moyen de la procédure itérative

des "kmoyennes segmentales" basée sur l'algorithme de Viterbi. Cette phase nécessite l'étiquetage fin des phrases d'apprentissage utilisées, car il faut extraire tous les segments correspondant à l'unité modélisée. Cette fastidieuse segmentation manuelle peut cependant être limitée à une fraction de l'ensemble d'apprentissage, de manière à disposer de quelques représentants pour l'initialisation de chaque modèle acoustique. L'estimation des paramètres d'un modèle est affinée avec HRest, qui applique l'algorithme optimal de Baum-Welch jusqu'à la convergence et ré-estime les probabilités d'émission et de transition. Dans une phase suivante, il est possible d'appliquer plusieurs itérations de l'outil HERest, qui ré-estime simultanément l'ensemble des modèles sur de la parole continue non segmentée.

Les modèles obtenus peuvent être améliorés, en augmentant par exemple le nombre de gaussiennes servant à estimer la probabilité d'émission d'une observation dans un état. Le choix du nombre optimal de gaussiennes est un problème délicat, généralement guidé par des heuristiques. Une commande de l'éditeur de modèles HHEd réalise l'augmentation du nombre de gaussiennes modélisant une densité de probabilité. Les modèles doivent être ensuite ré-estimés par HRest ou HERest. HTK offre aussi des facilités pour travailler avec des modèles contextuels, dépendants des contextes phonétiques gauches ou droits.

Reconnaissance : Le module de décodage de la parole continue, HVite, utilise l'algorithme de Viterbi pour trouver la séquence d'états la plus probable correspondant aux paramètres observés dans un modèle composite, et en déduire les unités acoustiques correspondantes. Le modèle composite autorise la succession des modèles acoustiques en fonction d'une syntaxe choisie par le concepteur du système. Il est possible d'écrire la syntaxe aux niveaux phonétique ou lexical, les mots du lexique pouvant être définis eux-mêmes par la concaténation d'unités sub-lexicales.

Parallèlement à cette syntaxe, HVite tient compte d'un modèle de langage de type bigramme, estimé sur les étiquettes des phrases d'apprentissage par l'outil HLStats. Le résultat du décodage est comparé aux étiquettes de référence par un alignement dynamique réalisé par HResults, afin de compter les étiquettes identifiées, omises, substituées par une autre, et insérées, et de calculer le taux de reconnaissance.