L'analyse factorielle pour la modélisation acoustique des systèmes de reconnaissance de la parole

Mohamed Bouallegue

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE
MINISTÈRE DE L'ENSEIGNEMENT
SUPÉRIEUR ET DE LA RECHERCHE

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « Sciences et Agrosciences »
Laboratoire d'Informatique (EA 4128)

L'analyse factorielle pour la modélisation acoustique des systèmes de reconnaissance de la parole

par
Bouallègue Mohamed

Soutenue publiquement le 16 December 2013 devant un jury composé de :

Mme	Régine André-Obrecht	Professeur des Universités, IRIT, Toulouse	Rapporteur
M.	Denis Jouvet	Directeur de Recherche, LORIA - INRIA, Nancy	Rapporteur
M.	—	—	Examinateur
M.	—	—	Examinateur
M.	Georges Linarès	Professeur des Universités, LIA, Avignon	Co-Directeur de thèse
M.	Driss Matrouf	Maître de Conférences (HDR), LIA, Avignon	Directeur de thèse

---

Laboratoire Informatique d'Avignon

Résumé

Dans cette thèse, nous proposons d'utiliser des techniques fondées sur l'analyse factorielle pour la modélisation acoustique pour le traitement automatique de la parole, notamment pour la Reconnaissance Automatique de la parole. Nous nous sommes, dans un premier temps, intéressés à la réduction de l'empreinte mémoire des modèles acoustiques. Notre méthode à base d'analyse factorielle a démontré une capacité de mutualisation des paramètres des modèles acoustiques, tout en maintenant des performances similaires à celles des modèles de base. La modélisation proposée nous conduit à décomposer l'ensemble des paramètres des modèles acoustiques en sous-ensembles de paramètres indépendants, ce qui permet une grande flexibilité pour d'éventuelles adaptations (locuteurs, genre, nouvelles tâches).

Dans les modélisations actuelles, un état d'un Modèle de Markov Caché (MMC) est représenté par un mélange de Gaussiennes (GMM : Gaussian Mixture Model). Nous proposons, comme alternative, une représentation vectorielle des états : les *facteur d'états*. Ces facteur d'états nous permettent de mesurer efficacement la similarité entre les états des MMC au moyen d'une distance euclidienne, par exemple. Grâce à cette représenation vectorielle, nous proposons une méthode simple et efficace pour la construction de modèles acoustiques avec des états partagés. Cette procédure s'avère encore plus efficace dans le cas de langues peu ou très peu dotées en ressouces et en connaissances linguistiques.

Enfin, nos efforts se sont portés sur la robustesse des systèmes de reconnaissance de la parole face aux variabilités acoustiques, et plus particulièrement celles générées par l'environnement. Nous nous sommes intéressés, dans nos différentes expérimentations, à la variabilité locuteur, à la variabilité canal et au bruit additif. Grâce à notre approche s'appuyant sur l'analyse factorielle, nous avons démontré la possibilité de modéliser ces différents types de variabilité acoustique nuisible comme une composante additive dans le domaine cepstral. Nous soustrayons cette composante des vecteurs cepstraux pour annuler son effet pénalisant pour la reconnaissance de la parole.

Mots-clés : Reconnaissance automatique de la parole, analyse factorielle, modélisation acoustique compacte, classification phonétique, variabilité acoustique.

Abstract

In this thesis, we propose to use techniques based on factor analysis to build acoustic models for automatic speech processing, especially Automatic Speech Recognition (ASR). Frstly, we were interested in reducing the footprint memory of acoustic models. Our factor analysis-based method demonstrated that it is possible to pool the parameters of acoustic models and still maintain performance similar to the one obtained with the baseline models. The proposed modeling leads us to deconstruct the ensemble of the acoustic model parameters into independent parameter sub-sets, which allow a great flexibility for particular adaptations (speakers, genre, new tasks etc.).

With current modeling techniques, the state of a Hidden Markov Model (HMM) is represented by a combination of Gaussians (GMM : Gaussian Mixture Model). We propose as an alternative a vector representation of states : the factors of states. These factors of states enable us to accurately measure the similarity between the states of the HMM by means of an euclidean distance for example. Using this vector representation, we propose a simple and effective method for building acoustic models with shared states. This procedure is even more effective when applied to under-resourced languages.

Finally, we concentrated our efforts on the robustness of the speech recognition systems to acoustic variabilities, particularly those generated by the environment. In our various experiments, we examined speaker variability, channel variability and additive noise. Through our factor analysis-based approach, we demonstrated the possibility of modeling these different types of acoustic variability as an additive component in the cepstral domain. By compensation of this component from the cepstral vectors, we are able to cancel out the harmful effect it has on speech recognition.

Keywords : Automatic speech recognition, factor analysis, compact acoustic modeling, phonetic classification, acoustic variability.

Table des matières

Chapitre 1

Introduction

Depuis presque trente ans, la reconnaissance automatique de la parole (RAP) est un domaine qui a captivé le public ainsi que de nombreux chercheurs. Il est prévisible que la parole fera de plus en plus partie des interfaces multimédia entre un utilisateur et un système automatique. Malheureusement, malgré l'incroyable évolution, les résultats obtenus sont encore loin de l'idéal espéré. En effet, dans les cadres applicatifs réels, les systèmes de RAP (SRAP) sont toujours soumis à de nombreuses difficultés qui freinent l'évolution des performances. Les principales contraintes qui limitent le développement de systèmes robustes dans une application réelle sont :

– *La variabilité acoustique présente dans le signal de parole* : dans un contexte applicatif réel, les SRAP sont soumis à diverses sources de bruit engendrant une dégradation significative des performances. Les sources de variabilité nuisible pour la reconnaissance de la parole ne se limitent pas au bruit environnemental, mais incluent toute variabilité (information) qui n'est pas utile à la tâche envisagée. Par exemple, les informations concernant le locuteur sont inutiles pour la reconnaissance de la parole, et peuvent même être gênantes pour cette tâche.

– *La taille importante des SRAP* : les différentes composantes d'un SRAP présentent un grand défi lors de leur intégration dans des terminaux légers (téléphones mobiles, lecteurs MP3, *etc*). En effet, les modèles sont constitués de plus de 10 millions de paramètres, ce qui représente une complexité incompatible avec les puissances de calcul et l'espace mémoire généralement disponibles dans ce genre d'appareils.

– *La limitation de la taille du vocabulaire d'apprentissage* : d'une façon générale, les performances d'un SRAP dépendent de sa capacité à modéliser les connaissances, ainsi que de la quantité et la qualité des données utilisées pour l'apprentissage des modèles. Les systèmes nécessitent d'importants volumes de données annotées pour l'estimation des modèles acoustiques et linguistiques. En effet, malgré l'évolution continue des techniques mises en oeuvre dans les SRAP, l'amélioration des modèles acoustiques reste souvent liée à l'augmentation des quantités de

données d'apprentissage. Aujourd'hui les systèmes les plus aboutis ont des modèles estimés sur plusieurs milliers d'heures annotées. Mais ces ressources sont loin d'être disponibles pour un grand nombre de langues dans le monde, que l'on nomme communément "peu dotées". Ces langues ne possèdent pas encore suffisamment de ressources (en quantité et en qualité) pour la construction d'un SRAP.

De nombreux travaux de recherche ont été réalisés au cours de ces dernières années pour proposer des solutions adaptées à chaque problème. Malgré ces efforts, l'amélioration de la robustesse du SRAP demeure un sujet de recherche d'actualité.

L'objectif de cette thèse est de proposer des solutions afin de répondre au mieux à l'ensemble de ces problèmes. Nous proposons de nouvelles techniques fondées sur des approches d'analyse factorielle. L'utilisation de techniques à base d'analyse factorielle a connu un grand succès dans le domaine de la vérification du locuteur face à la variabilité acoustique (Kenny et al., 2005b). Ce type d'approches a également été utilisé dans d'autres domaines tels que la reconnaissance de la langue, la vérification du genre vidéo, le traitement d'images, etc. En revanche, il est très peu exploité dans le domaine de la reconnaissance de la parole. Le succès de l'analyse factorielle dans ces différents domaines motive son utilisation dans la recherche de solutions adaptées aux problèmes freinant l'évolution du SRAP. Dans cette thèse, nous proposons d'utiliser des techniques à base d'analyse factorielle afin de résoudre les problèmes de robustesse des SRAP cités ci-dessus. Chaque technique proposée est évaluée expérimentalement.

1.1 Contexte général

Cette thèse s'intéresse à trois axes principaux de recherche : la robustesse des SRAP face aux variabilités acoustiques, la modélisation acoustique compacte et enfin la classification d'unités acoustiques et ses utilisations pour le traitement automatique de la parole. Nous exposons par la suite les problématiques posées dans ces trois thèmes ainsi qu'une introduction à la technique de l'analyse factorielle utilisée tout au long de cette thèse.

1.1.1 Variabilité acoustique dans le SRAP

Le travail effectué dans cette thèse se situe dans le cadre de la reconnaissance automatique de la parole utilisant des modèles statistiques. La reconnaissance de la parole s'inscrit dans le domaine plus général de la reconnaissance des formes. L'idée de base est d'apprendre des formes (des statistiques sur ces formes) pour pouvoir les reconnaître par la suite. La forme reconnue est celle, parmi toutes les formes apprises, qui ressemble le plus à la forme inconnue. Si les formes de test ne subissent pas de distorsions majeures par rapport aux formes apprises avant que les mesures de similarité utilisées ne soient appliquées, le système de reconnaissance de formes atteint les

meilleures performances. En revanche, si les formes de test sont modifiées par des événements inconnus *a priori*, les performances du système de reconnaissance des formes chutent. Dans ce cas, nous parlons de non-concordance entre les conditions d'apprentissage et les conditions de test.

Dans le cadre de la reconnaissance de la parole, deux types de variabilité affectant la réalisation d'un signal de parole peuvent être distinguées : le premier type de variabilité correspond à celui naturellement présent dans la parole est liée aux caractéristiques propres du locuteur ou à l'aspect étranger ou régional du locuteur, au style et à la vitesse de production de la parole, à l'âge du locuteur ou encore, à l'état émotionnel du locuteur. Nous appelons ce type de variabilité, *les variabilités intrinsèques* de la parole. Le deuxième type de variabilité n'est pas lié au locuteur, mais plutôt à son environnement (ouvert ou fermé, bruit ambiant, écho, etc). Ces sources de variabilité sont très pénalisantes pour le développement et l'exploitation à grande échelle des SRAP. Pour cette raison, de nombreuses techniques ont été proposées pour augmenter la robustesse des systèmes, en particulier leur résistance aux bruits (Lim, 1978) (Acero, 1990) (Vaseghi et Milner, 1992) (Gales et Young, 1996) (Holmes et Sedgwick, 1986). L'objectif de ces techniques est de compenser les différences entre les conditions d'apprentissage et les conditions d'utilisation du système.

1.1.2 Modélisation acoustique compacte

Dans le cadre des SRAP, 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. L'unité acoustique la plus utilisée est le phonème dépendant du contexte. Un phonème contextuel est modélisé par un modèle de Markov caché (MMC) gauche-droit à trois états. Pour relier ce modèle aux vecteurs de paramètres acoustiques du signal de parole, à chaque état est associé un mélange de densités de probabilité qui suivent chacune une loi gaussienne (GMM, signifiant *Gaussian Mixture Model*).

La multiplication de modèles contextuels permet de rendre la modélisation acoustique plus précise. Cependant, cette amélioration théorique est limitée dans la pratique par des problèmes d'estimation : la quantité de données disponible pour l'estimation de chaque modèle contextuel se réduit au fur et à mesure de l'augmentation de la complexité des modèles.

En outre, l'augmentation de la taille des modèles acoustiques présente un grand défi lors de l'intégration de SRAP dans des appareils légers tels que les téléphones portables et les lecteurs MP3. En effet, les modèles sont composés de dizaines de millions de paramètres représentant une complexité incompatible avec la puissance de calcul et l'espace mémoire généralement disponible dans ce type d'appareils. Différentes architectures ont été proposées dans la littérature pour réduire l'empreinte mémoire des modèles acoustiques (Huang et al., 1989) (Lee et al., 1990) (Demuynck et al., 1996) (Bocchieri et Mak, 2001) (LEVY, 2006).

1.1.3 Classification phonétique

Depuis les premières applications dans le traitement de la parole, la problématique de calcul de similitudes entre phonèmes (ou allophones) a été posée comme sujet de recherche par la communauté scientifique. Cette mesure est utilisée dans plusieurs applications dans tous les domaines du traitement de la parole : reconnaissance du locuteur, reconnaissance de la langue, synthèse de la parole ou encore analyse de variation de la prononciation.

Dans le domaine de la reconnaissance de la parole, le calcul de similitude est utilisé principalement dans la procédure de *partage des états*. Cette procédure, incontournable dans la modélisation acoustique, consiste à associer la même fonction de densité de probabilité (*probability density function* : pdf) aux états des phonèmes contextuels acoustiquement proches (Young, 1992). La réalisation de cette procédure est confrontée à la difficulté de la définition d'une distance de similarité entre les états des MMC. En effet, chaque état est généralement modélisé par un mélange de gaussiennes. Ce qui rend complexe la définition de distance entre les différents états.

Afin de réduire le nombre de calculs de distances entre états, une méthode fondée sur l'utilisation d'informations linguistiques et phonétiques a été proposée. Il s'agit d'une méthode à base de connaissances phonétiques : elle utilise un arbre de décision où à chaque noeud est associé un certain nombre de questions linguistiques qui permettent de parcourir l'arbre de haut en bas. Cette méthode ne peut être utilisée que dans le cadre de langues pour lesquelles de telles connaissances phonétiques et linguistiques sont disponibles, tout comme de grandes quantités de données d'apprentissage. Pour les langues qualifiées de peu dotées, nous ne disposons pas de telles informations ni d'aussi grandes quantités de données d'apprentissage.

Dans cette thèse, nous allons nous intéresser aux problèmes exposés selon les trois axes ci-dessus. Les solutions proposées sont fondées sur des approches provenant de l'analyse factorielle.

1.2 Travaux réalisés

Dans cette thèse, nos contributions concernent la modélisation acoustique compacte, la classification phonétique et enfin la compensation de variabilité nuisible. Les approches que nous proposons sont toutes issues de l'analyse factorielle.

1.2.1 Modélisation acoustique compacte

L'objectif de la modélisation compacte est de réduire le nombre de paramètres des modèles acoustiques tout en préservant les performances du SRAP. Pour ce faire, nous proposons d'utiliser l'analyse factorielle. La méthode consiste à modéliser l'espace acoustique de la parole par un modèle générique appelé modèle du monde (*Universal Background Models* : UBM), puis à dériver les modèles des différents états des MMCs

depuis ce modèle générique, en mutualisant une large partie des paramètres des modèles.

La première étude expérimentale permet de trouver un point de fonctionnement optimal entre la taille et les performances du modèle. La modélisation proposée permet de décomposer l'ensemble des paramètres des modèles acoustiques en sous-ensembles de paramètres indépendants. Cela donne une grande flexibilité pour d'éventuelles adaptations (adaptation au locuteur, à l'environnement acoustique, etc.). La seconde étude expérimentale consiste donc à exploiter cette décomposition dans les différentes adaptations possibles.

1.2.2 Classification phonétique

Dans la deuxième partie, nous proposons une nouvelle vision de la classification des phonèmes : la représentation vectorielle des états du MMC. Nous obtenons cette représentation à l'aide de l'analyse factorielle. Nous appelons ces vecteurs représentatifs d'états du MMC, *facteurs d'états*. En utilisant ces vecteurs, la classification phonétique est formulée comme un problème de classification usuel dans l'espace R^d. Cette représentation vectorielle permet d'exploiter les résultats scientifiques obtenus au cours de plusieurs années de recherche dans le domaine de l'analyse de données. Ces résultats peuvent servir dans l'analyse de variabilité acoustique et de la variation de l'articulation phonétique. Également, les *facteurs d'états* peuvent être utilisés dans d'autres applications comme la phonétique cliniques, la détection de dialecte ou l'identification automatique de la langue.

Nous montrons que cette représentation vectorielle peut être dans la réalisation de la procédure de *partage d'états* du MMC, utilisée dans le cadre de la modélisation acoustique contextuelle. La plupart des techniques proposées dans la littérature pour réaliser cette procédure nécessitent notamment des connaissances linguistiques qui peuvent ne pas être disponibles pour certaines langues. Notre nouvelle méthode de *partage d'états* s'appuie uniquement sur l'information portée par les *facteurs d'états*. Cette méthode nous permet de contourner le problème d'insuffisance ou d'absence d'informations phonétiques ou linguistiques pour les langues peu dotées.

Dans nos expériences, nous évaluons la pertinence de notre méthode sur la langue française. Ensuite, nous l'appliquons sur deux langues catégorisées comme langues peu dotées : la langue vietnamienne et la langue berbère. Nous montrons aussi l'utilité des *facteurs d'états* dans l'analyse graphique de quelques phénomènes acoustiques.

1.2.3 Compensation de variabilité nuisible

Dans la troisième partie, nous nous intéressons à la robustesse du SRAP face à la variabilité locuteur, variabilité canal et le bruit additif. Dans cette partie nous développons une nouvelle approche de compensation de la variabilité nuisible en nous appuyant sur l'analyse factorielle. Les vecteurs cepstraux sont supposés être générés par l'UBM et les

états des HMM doivent être modélisés par des GMM obtenus à partir de l'UBM par une adaptation MAP. Dans nos expériences, nous nous intéressons à la variable aléatoire, appelée super-vecteur, constituée par la concaténation des moyennes des gaussiennes composant le GMM (dépendant du phonème ou de l'état). L'isolation et l'estimation du bruit se fait en utilisant l'analyse factorielle dans l'espace des super-vecteurs. En effet, le super-vecteur d'un état est donné selon trois composantes : la première composante est indépendante de l'état et de la variabilité nuisible en question, la deuxième composante correspond à l'information phonétique (état d'un MMC ou d'un phonème) et la troisième est une composante correspondant à la variabilité nuisible que nous traitons ici. L'hypothèse fondamentale, dans le formalisme développé, est que la variabilité nuisible est située dans un sous-espace de faible dimension par rapport à la dimension du super-vecteur.

Dans nos expériences, nous étudions plusieurs scénarios liés à la variabilité nuisible pour la reconnaissance de la parole. Dans un premier temps, nous nous intéressons à la variabilité locuteur et la variabilité canal. La variabilité locuteur est une des plus perturbantes pour un système de reconnaissance de la parole. Cette variabilité inclut la variabilité *intra-locuteur* due au mode d'élocution et la variabilité *inter-locuteur* due aux différences entre locuteurs. L'effet de cette variabilité sur le signal est très complexe, certaines des parties étant linéaires et d'autres non. Cependant, on peut considèrer que le changement de locuteur se traduit par des changements du conduit vocal et que cette variabilité peut être considérée comme étant additive dans le domaine cepstral. Pour la variabilité canal, nous désignons la variabilité qui inclut tout changement de condition d'enregistrement, notamment le changement de microphone (ou du téléphone), la position du microphone par rapport à la bouche, ou encore l'endroit ou s'effectue l'enregistrement (la géométrie de l'endroit où se passe l'enregistrement : hall, bureau, ville, etc.).

Dans un second temps, nous nous intéresserons à un autre type de source de nuisance pour le SRAP : le bruit additif. Ce type de bruit est caractérisé par sa non-linéarité avec le signal de parole dans le domaine cepstral. Malgré cette caractéristique, l'objectif de ce travail est de savoir s'il est possible de modéliser le bruit additif (ou une partie du bruit additif) comme une composante additive en utilisant la modélisation d'analyse factorielle.

1.3 Organisation du document

Ce document est organisé en deux grandes parties. La première partie (chapitre 2 et 3) présente le contexte d'étude et les outils utilisés dans les systèmes de reconnaissance de la parole, ainsi que l'origine théorique de l'analyse factorielle.

La seconde partie correspond aux contributions originales réalisées durant cette thèse, à savoir, la modélisation acoustique compacte, la classification phonétique et la modélisation de la variabilité acoustique.

En détails, le chapitre 2 présente un état de l'art sur les SRAP. Nous y établissons

leurs principes généraux et nous introduisons les éléments de base nécessaires à leurs fonctionnements, notamment les modélisations acoustiques, linguistiques et les algorithmes de décodage.

Le chapitre 3 présente les bases théoriques de l'analyse factorielle. Il montre la relation théorique entre les différentes méthodes de réduction de dimension. Une démonstration mathématique de l'estimation des paramètres du modèle d'analyse factorielle est présentée.

Le chapitre 4 est centré sur la modélisation acoustique compacte. Nous présentons dans ce chapitre une nouvelle méthode de mutualisation des paramètres des modèles acoustiques. Un éventail de la modélisation acoustique compacte est présenté dans un premier temps. Ensuite, nous exposons notre méthode de réduction du nombre de paramètres des modèles acoustiques en discutant des différences avec des approches similaires. La modélisation acoustique que nous proposons permet de décomposer l'ensemble des paramètres des modèles en sous-ensembles de paramètres indépendants. Cette caractéristique donne une grande flexibilité dans le processus d'adaptation. Dans la partie expérimentale, nous exposons les résultats des différentes stratégies d'adaptation des modèles compacts.

Dans le chapitre 5 nous présentons une nouvelle vue de la classification des phonèmes. Nous montrerons comment les phonèmes sont caractérisés par des vecteurs estimés par l'approche par analyse factorielle. Nous appelons ces vecteurs *facteurs d'états*. Dans nos expériences, nous proposons d'exploiter les *facteurs d'états* dans la réalisation de la procédure de regroupement d'états du MMC.

Dans la première partie de ce chapitre, nous exposons la méthode d'estimation des *facteurs d'états* et nous détaillons ensuite les différentes étapes de la procédure de regroupement d'états en s'appuyant sur ces facteurs.

Dans la seconde partie, nous appliquons notre méthode sur la langue française, la langue vietnamienne puis la langue berbère. Finalement, nous exposons quelques interprétations graphiques fondées sur les *facteurs d'états*.

Le chapitre 6 se focalise sur la robustesse du SRAP. Dans ce travail, nous décrivons une technique de compensation de la variabilité acoustique nuisible s'appuyant sur l'analyse factorielle. Dans un premier temps, nous nous intéressons à la variablité locuteur et la variabilité canal. Nous commençons par étudier la compensation des deux variablités séparément. Puis, nous tentons de procéder à une compensation conjointe de celles-ci. Nous traiterons par la suite de la variabilité liée au bruit additif. Les méthodes proposées seront expérimentées sur des données artificiellement bruitées ainsi que sur des données représentant un cas réel de conditions défavorables pour la reconnaissance de la parole.

Le chapitre 7 conclut et expose quelques perspectives de notre thèse.

Chapitre 2

Système de reconnaissance de la parole : définitions, modèles et algorithmes

Sommaire

Dans ce chapitre, nous décrivons le cadre général dans lequel se sont effectués nos travaux. Nous présentons le fonctionnement complet d'un système de reconnaissance automatique de la parole (SRAP). Cet état de l'art se concentre sur les SRAP Markoviens utilisant des modèles de langage probabilistes à base de n-grammes. Nous survolons les principes des différents modèles acoustiques et linguistiques ainsi que les algorithmes liés à leur apprentissage. Nous finissons en présentant les algorithmes de décodage, ainsi que les paradigmes d'évaluation des SRAP.

2.1 Principe général d'un système de reconnaissance automatique de la parole

Les systèmes de reconnaissance automatique de la parole (SRAP) ont pour objectif de transcrire un message oral (signal de parole) en texte (suite de mots). Les systèmes de reconnaissance automatique de la parole continue actuels sont fondés sur une approche statistique dont (Jelinek, 1976) a proposé une formalisation, issue de la théorie de l'information. À partir des observations acoustiques X, l'objectif d'un système de reconnaissance est de trouver l'hypothèse $\tilde{W}$ dont la probabilité est maximale, étant donnés les observations X, les modèles acoustiques et le modèle de langage :

$$\tilde{W} = \underset{W}{\operatorname{argmax}}\, P(W/X) = \underset{W}{\operatorname{argmax}}\, \frac{P(X/W)P(W)}{P(X)} \tag{2.1}$$

Dans cette équation, nous pouvons identifier plusieurs facteurs :

- P(X | W) est la probabilité d'observer le signal de parole X, étant donnée W, la suite de mots prononcés. Cette probabilité est calculée en utilisant le modèle acoustique.
- P(W) est la probabilité *a priori* que la suite de mots W ait été générée. Cette probabilité est estimée via le modèle de langage.
- P(X) est la probabilité d'observer le signal de parole X. Cette probabilité est identique pour chaque suite de mots. Elle n'est pas utile pour déterminer la meilleure suite de mots, et peut donc être ignorée.

Un SRAP est constitué de plusieurs composants, comme décrit dans la figure 2.1.

- *Module de paramétrisation* : produit les vecteurs de paramètres acoustiques représentatifs de l'information nécessaire à la reconnaissance de la parole.
- *Modèle acoustique* : calcule la probabilité qu'un phonème (ou syllabe, allophone, etc.) ait généré une séquence de vecteurs de paramètres acoustiques.
- *Le modèle de langage* : permet d'introduire la notion de contraintes linguistiques dans les SRAP. Ce modèle est la représentation statistique de l'enchaînement possible des mots dans une langue donnée. Il permet de guider le décodeur vers les suites de mots les plus probables.

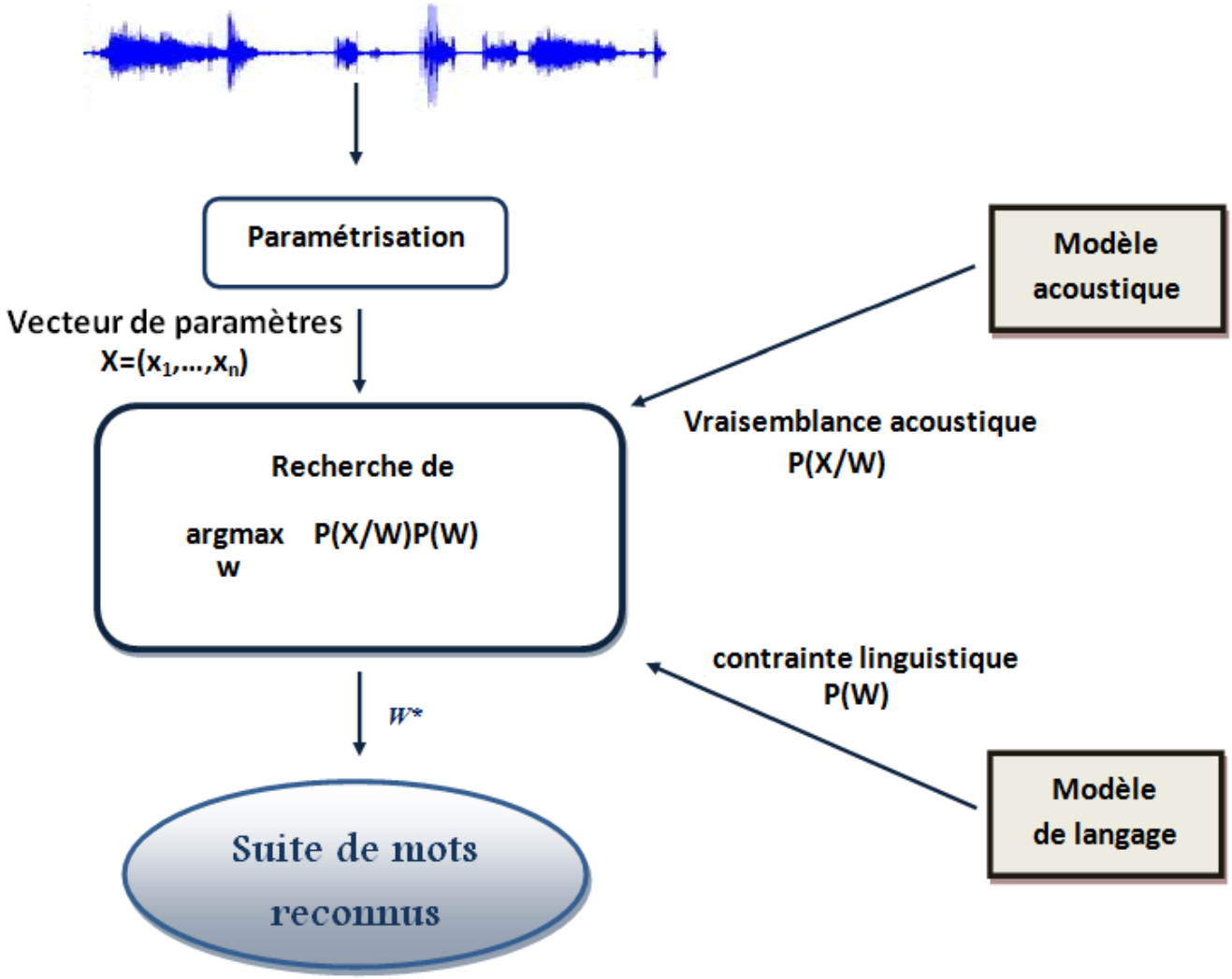

FIGURE 2.1 – *Description d'un système de reconnaissance de la parole.*

2.2 Traitement du signal et paramétrisation acoustique

Le signal de parole est trop redondant et variable pour être utilisé directement dans un système de reconnaissance automatique de la parole. Il doit être traité de manière à extraire au mieux l'information utile à la caractérisation de son contenu linguistique. Généralement, ces paramètres sont estimés à travers des fenêtres glissantes sur le signal. Cette analyse permet d'estimer le signal sur une portion du signal considérée stationnaire : généralement 10 à 30 ms en limitant les effets de bord et les discontinuités du signal via une fenêtre de Hamming.

Les systèmes de reconnaissance de la parole intègrent un module de paramétrisation dont le rôle est de créer des vecteurs de paramètres acoustiques résultant de l'analyse du signal de parole. Les techniques de paramétrisation les plus utilisées dans

les SRAP sont : MFCC (Mel Frequency Cepstral Coefficients) (Davis et Mermelstein, 1980), PLP (Perceptual Linear Prediction) (Hermansky et Jr., 1991), LPCC (Linear Prediction Cepstral Coefficients) (Markel et Gray, Jr., 1976), RASTA-PLP (Relative Spectral PLP) (Hermansky et al., 1992).

Dans le cadre de cette thèse, nous avons utilisé l'analyse PLP afin de paramétriser les signaux utilisés dans nos expérimentations. Cette technique est décrite dans la section suivante.

2.2.1 Analyse par prédiction linéaire perceptuelle

L'analyse par Prédiction Linéaire Perceptuelle (PLP) repose sur un modèle de perception de la parole. Elle est fondée sur le même principe que l'analyse prédictive et intègre trois caractéristiques de la perception (Hermansky, 1990) :

– *Intégration des bandes critiques* : la prédiction linéaire produit la même approximation de l'enveloppe spectrale pour toute la zone de fréquences utiles, ce qui est en contradiction avec le fonctionnement de l'appareil perceptif humain. En effet, l'oreille humaine a la faculté d'intégrer certaines zones de fréquences en bande appelées *bandes critiques*. Les bandes critiques sont réparties selon l'échelle de Bark. Le passage de Bark en Hertz est obtenu en appliquant la transformation suivante :

$$f = 600 sinh(z/6) \tag{2.2}$$

avec *sinh* représentant le sinus hyperbolique[1], f la fréquence en Hertz et z la fréquence en Bark. La nouvelle densité spectrale est échantillonnée selon cette nouvelle échelle, ce qui a pour effet d'augmenter la résolution dans les basses fréquences.

– *Préaccentuation du signal selon une courbe d'isosonie* : des expériences psychoacoustiques ont montré que l'oreille possède des caractéristique non linéaires (Fletcher et Munson, 1933). Pour simuler ce phénomène dans le cadre de l'analyse PLP, la densité spectrale résultante de l'étape précédente est multipliée par une fonction de pondération.
– *Compression en racine cubique* : l'intégration des bandes critiques et de la préaccentutation ne suffisent pas à faire correspondre l'intensité mesurée et l'intensité subjective (la sonie). La loi de Stevens donne la relation entre ces deux mesures :

$$Sonie = (intensite)^{\frac{1}{3}} \tag{2.3}$$

1. http ://homeomath.imingo.net/sinh.htm

2.3 Modélisation acoustique : Modèles de Markov Cachés

Le signal acoustique de 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. Cependant, d'autres unités acoustiques sont envisageables, parmi lesquelles nous pouvons citer le phonème en contexte, la syllabe ou encore la dissyllabe. Plus la taille de l'unité augmente, plus son caractère discriminant s'affirme, mais plus il faut d'unités élémentaires pour couvrir la langue. Dans le cadre d'un système de reconnaissance de la parole continue à grand vocabulaire, l'unité acoustique la plus utilisée est le phonème dépendant du contexte. Lorsque le contexte phonétique comprend le phonème précédent et le phonème suivant, nous parlons de *triphone*.

Une fois l'unité acoustique choisie, le problème de sa modélisation se pose : il s'agit de trouver un modèle représentant une séquence de vecteurs acoustiques dont les caractéristiques spectrales et temporelles peuvent varier considérablement d'un locuteur à l'autre ou entre deux élocutions d'un même locuteur. La solution quasi-exclusivement utilisée est celle des modèles de Markov cachés (MMC) gauche-droite à trois états. Pour relier ce modèle aux vecteurs de paramètres acoustiques du signal de parole, à chaque état est associé une distribution de probabilités modélisant la génération des vecteurs acoustiques par cet état.

2.3.1 Structure d'un MMC

Un MMC est un automate probabiliste contrôlé par deux processus stochastiques. Le premier processus, interne au MMC et donc caché à l'observateur, débute sur l'état initial puis se déplace d'état en état en respectant la topologie du MMC. Le second processus stochastique génère les unités linguistiques correspondant à chaque état parcouru par le premier processus. Un MMC est caractérisé par l'ensemble de paramètres suivant :

- L qui est le nombre d'états du modèle.

- Une matrice A qui permet la définition de la topologie du MMC en indiquant les probabilités de transition d'un état q_i vers un état q_j. Les modèles utilisés en reconnaissance de la parole sont d'ordre 1, c'est-à-dire que la probabilité de passer dans l'état suivant dépend uniquement de l'état courant. La taille de cette matrice est $L \times L$.

- Une matrice B qui contient les densités de probabilité des observations associées à chaque état j du modèle, avec $b_j\ (x_n) = p(x_n|q_j)$ représentant la probabilité d'émettre l'observation x_n étant dans l'état q_j .

- Une matrice P qui donne la distribution de départ des états, c'est-à-dire pour chaque état la probabilité d'être atteint à partir de l'état initial q_0. Cet état est

particulier puisqu'il ne peut émettre d'observation.

Les paramètres d'un MMC sont estimés à partir d'un corpus d'entraînement généralement transcrit manuellement. La transcription manuelle permet d'identifier les segments de parole correspondant à chaque unité linguistique (généralement phonème). La reconnaissance revient à choisir le MMC ayant la plus grande probabilité d'avoir émis le signal en entrée.

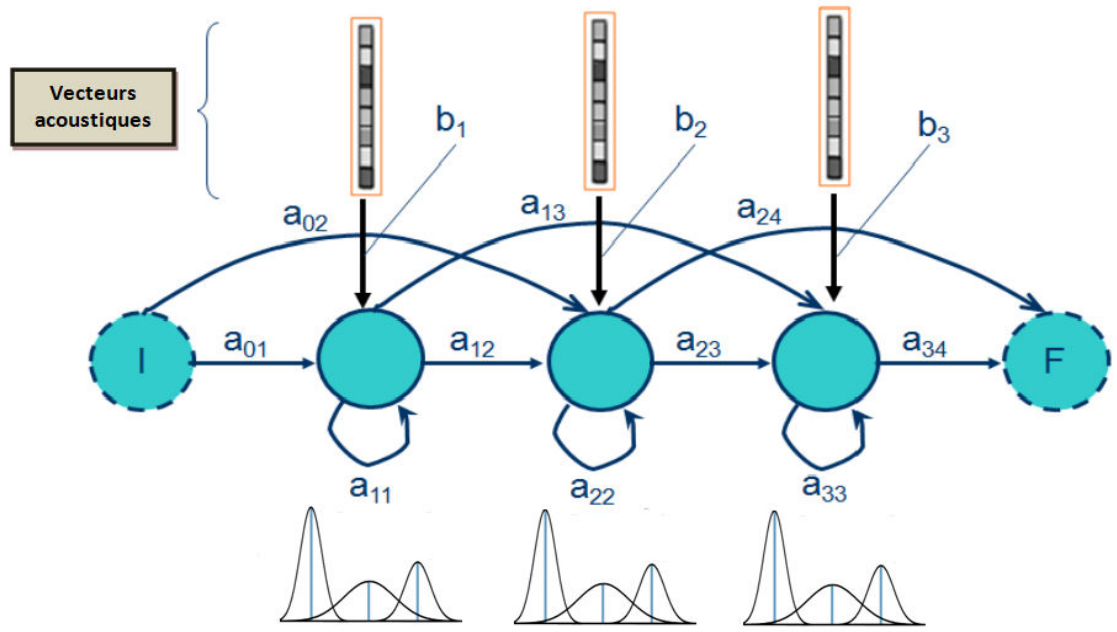

FIGURE 2.2 – *MMC à 5 états dont 3 émetteurs.*

2.3.2 Les mixtures de gaussiennes

Pour relier un MMC aux vecteurs de paramètres acoustiques du signal de parole, à chaque état du MMC est associée une fonction de densité de probabilité. Afin d'approximer cette densité de probabilité, un mélange de gaussiennes (une somme pondérée de gaussiennes) est utilisée. Nous utiliserons le terme communément relayé : Gaussian Mixture Model (GMM). La figure 2.3 montre un exemple de gaussienne bivariée (multivariée de dimension 2). Elle est définie par :

$$N(x|\mu,\Sigma) = \frac{1}{2\pi \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2.4}$$

avec μ la moyenne et Σ est la matrice de covariance. Ces paramètres sont optimisés selon le critère de maximum de vraisemblance pour approcher le plus possible la distribution recherchée. Cette procédure se fait le plus souvent itérativement via l'algorithme d'espérance-maximisation (EM), détaillé dans la section 2.4.1.

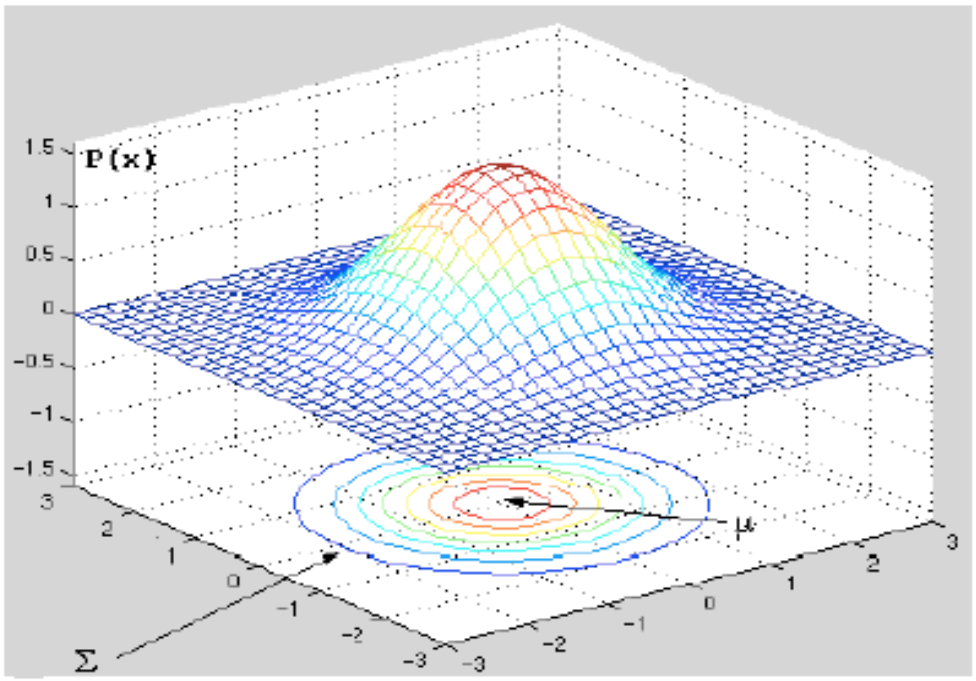

FIGURE 2.3 – *Exemple de densité de probabilité d'une gaussienne bivariée.*

2.4 Apprentissage et adaptation acoustique

Cette section décrit succinctement les principales méthodes d'apprentissage et d'adaptation des modèles acoustiques.

2.4.1 Apprentissage d'un MMC

L'apprentissage du MMC consiste à déterminer les paramètres $\Theta = \{L, A, \{p_i\}, \{b_i\}\}$ optimaux selon le critère de maximum de vraisemblance (Maximum Likelihood - ML).

2.4.1.1 Maximum de vraisemblance (ML)

Soit m un MMC, et Θ l'ensemble des paramètres associés au modèle m. Supposons que m modélise le message linguistique w. Soit $y = (y_1, ..., y_T)$ une observation correspondant à l'élocution du message w. L'approche ML consiste à déterminer les paramètres $\widehat{\Theta}$ maximisant la probabilité que l'observation y soit générée par le modèle m :

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, P(Y = y/m) = \underset{\Theta}{\operatorname{argmax}}\, P(Y = y/\Theta) \tag{2.5}$$

Le modèle MMC est un cas typique des modèles générant des données incomplètes. En effet, étant donnée une observation $y = (y_1, ..., y_T)$ générée par un tel modèle, il est impossible de retrouver avec certitude l'état (ou la gaussienne) qui a généré la trame y_t. Pour cette raison, le problème de la maximisation associé à l'équation 2.5 devient très complexe et il est pratiquement impossible de le résoudre directement. Ce constat a conduit à utiliser l'algorithme EM.

2.4.1.2 Algorithme EM

EM est un algorithme de maximisation proposé 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. Cette méthode nécessite un modèle initial Θ^0, à partir duquel sont alternées plusieurs itérations de deux étapes de l'algorithme :

- **Étape E (Expectation)** : les données d'apprentissage sont complétées en fonction du modèle courant Θ^i.
- **Étape M (Maximisation)** : l'ensemble des paramètres du modèle Θ^{i+1} qui maximise la vraisemblance des nouvelles données complètes connaissant le modèle courant Θ^i est calculé.

Dans la pratique, un alignement forcé est tout d'abord réalisé par rapport aux états au moyen de l'algorithme de Viterbi (Forney, 1973). À partir de cette segmentation des données d'apprentissage par rapport aux états du MMC, nous appliquons l'algorithme EM au niveau de chaque état pour déterminer les paramètres du mélange de gaussiennes correspondant. Ce processus est répété jusqu'à la convergence.

Pour initialiser chaque GMM, deux types d'approches peuvent être suivies. La première consiste à estimer une première gaussienne (moyenne et variance). Ensuite, deux gaussiennes sont créées à partir de la première en faisant varier la moyenne de $\pm\varepsilon$. Puis les paramètres de chaque gaussienne sont réestimés avec les données ayant le plus de vraisemblance. Cette étape est réitérée plusieurs fois pour atteindre le nombre de gaussiennes désiré. La seconde solution consiste à utiliser l'algorithme des *k*-means (ou k-moyennes). Cet algorithme a pour but de partager l'espace acoustique en k parties en essayant de trouver les centres naturels de ces parties. L'objectif est double : minimiser la variance intra-classe tout en maximisant la variance inter-classe.

2.4.2 Adaptation des modèles acoustiques

Les systèmes de RAP doivent faire face à de nombreuses contraintes liées au signal audio, comme par exemple devoir traiter un grand nombre de locuteurs différents, dans des conditions variables d'enregistrement. Le volume de données d'apprentissage ne

peut couvrir toutes les variabilités liées à ces conditions. Pour remédier à ces difficultés, des techniques d'adaptation ont été mises en place. L'idée est d'adapter les modèles déjà appris pour en créer des nouveaux beaucoup plus proches des conditions de test, en utilisant les modèles initiaux et un nombre restreint de nouvelles données. De nombreuses techniques existent, nous verrons succinctement les techniques les plus utilisées pour bien comprendre les enjeux de l'adaptation.

Généralement, l'apprentissage des modèles acoustiques est effectué avec un corpus d'entraînement très varié. L'objectif de ce choix est de reconnaître correctement la parole dans des conditions acoustiques variées. Mais ce modèle générique obtenu ne correspond pas souvent à la tâche pour laquelle le modèle est construit. Afin de pallier ce genre de problèmes, les paramètres du modèle acoustique générique existant peuvent être modifiés de manière à mieux reconnaître les nouvelles données. Cette opération de modification est appelée *adaptation*. Elle permet de résoudre le problème de la dissimilarité entre les données d'apprentissage et les données de test. L'adaptation prend en compte ces différences en modifiant les paramètres du modèle acoustique générique afin qu'ils correspondent à la tâche cible. Par exemple, un modèle adapté à la voix d'un homme (ou d'une femme) fournira de meilleures performances pour reconnaître un message vocal prononcé par un homme (ou une femme).

2.4.3 Avantage et limitation des MMC/GMM

Le cadre des MMC présente plusieurs avantages. D'abord il offre un formalisme mathématique bien établi et un apprentissage automatique des paramètres. Ainsi, il permet de prendre en compte les variations temporelles au sein du signal de parole. Sur la parole avec débit rapide, le MMC permet de sauter un état. Au contraire, sur la parole lente, il est possible de boucler sur un état, ce qui peut conduire à des répétitions d'états (transition a_{11} dans la figure 2.2). Les MMC s'appuient sur des fonctions multigaussiennes permettant de prendre en compte les variabilités locuteurs (sexe, variantes de prononciation, état émotionnel, etc.) autour de la moyenne calculée sur les données d'apprentissage.
Le modèle MMC repose sur un ensemble d'hypothèses simplificatrices. Certaines d'entres elles sont des sources de limitation du modèle que les chercheurs ont tenté de pallier. Elles concernent en particulier :

- *L'hypothèse d'indépendance temporelle des observations* : les données à l'entrée d'un MMC sont supposées être statistiquement indépendantes. Cette hypothèse est irréaliste. En effet, les vecteurs de paramètres acoustiques sont calculés sur des portions de signal d'une durée très petite (en général 30 ms.). Cependant, cette hypothèse simplificatrice est une des origines du succès des MMC pour modéliser la parole[2]. Afin de réduire les effets de cette approximation sans changer les algorithmes d'apprentissage et de décodage, la dépendance entre les trames successives est modélisée en introduisant les paramètres dynamiques (les dérivées premières et secondes des paramètres (Furui, 1986)). La modélisation

2. Nous disposons d'algorithmes de décodage et d'apprentissage très efficaces pour ce type de modèles : l'algorithme de Viterbi pour le décodage et l'algorithme EM ou de Baum pour l'apprentissage.

explicite de la corrélation entre les trames successives de parole a également été étudiée : MMC d'ordre 2 (Mari et al., 1997), MMC conditionnellement gaussiens (Wellekens, 1987) et MMC segmentaux (Russell, 1993). Cependant, ils n'ont jamais montré de gains en performance significatifs et systématiques au point de remplacer les modèles classiques.

– *La modélisation de la durée* : celle-ci est implicite dans un MMC. Un segment plus court a une probabilité plus grande d'être généré par un MMC qu'un segment plus long. De nombreuses tentatives de modélisation explicite de durées ont été proposées dans (Gong et Haton, 1994) (Suaudeau et Andé-Obrecht, 1994).

2.5 Modèle de langage

Le modèle de langage représente un point clef du système de reconnaissance automatique de la parole. Ce modèle permet d'introduire la notion de contraintes linguistiques dans le SRAP. Nous distinguons deux types de modèles de langage. Le premier est le modèle à base de grammaires formelles réalisé par des experts en linguistique. Le second est le modèle statistique utilisant des corpus pour estimer les probabilités d'une suite de mots d'une manière automatique. Le modèle statistique est privilégié dans les systèmes de reconnaissance automatique de la parole car la génération manuelle d'un ensemble de règles décrivant une langue est un processus long, difficile et coûteux. De plus, il s'intègre bien dans le processus de décodage et possède le formalisme statistique du problème de la reconnaissance automatique de la parole.
Le modèle de langage modélise les contraintes liées à une langue, afin d'estimer la probabilité d'une suite de mots :

$$P(W_1^k) = P(w_1) \prod_{i=2}^{k} P(w_i/h_i) \tag{2.6}$$

Où $P(W_1^k)$ est la probabilité de la suite de mots de h_i correspondant à l'historique du mot w_i.
Les modèles de langage stochastiques les plus utilisés sont les modèles n-grammes permettant d'estimer la probabilité *a priori* d'une suite de mots W_k. Dans ce modèle l'historique d'un mot est représenté par les $(n-1)$ mots qui le précèdent :

$$P(W_1^k) = P(w_1) \prod_{i=2}^{n-1} P(w_i/w_1, ..., w_{i-1}) \tag{2.7}$$

Où $P(w_i/w_1, ..., w_{i-1})$ est la probabilité d'avoir le mot w_i sachant les observations $w_1, ..., w_{i-1}$.

Les modèles les plus couramment utilisés dans les SRAP sont les modèles d'ordre 3 ou 4. Dans le cas d'un modèle tri-gramme, l'équation précédente peut être réécrite de la manière suivante :

$$P(W_1^k) = P(w_1)P(w_2/w_1)\prod_{i=3}^{k} P(w_i/w_{i-2}, w_{i-1}) \tag{2.8}$$

Les modèles de langage n-grammes obtiennent de bons résultats grâce à leur souplesse et leur simplicité. Mais la taille et la diversité du corpus d'apprentissage représentent toujours une source de problèmes. Les séquences de mots qui n'apparaissent pas dans le corpus d'apprentissage peuvent avoir des probabilités nulles, ce qui pose un problème pour le calcul de la probabilité de l'équation 2.8. Le lissage des probabilités, par interpolation ou par repli, est la solution la plus usuelle pour ne pas attribuer des probabilités nulles aux séquences de mots non observées. Un survol des méthodes de lissage est présenté dans (Chen et Goodman, 1996; Kneser et Ney, 1995).

La meilleure façon d'évaluer la qualité d'un modèle de langage est de tester ce dernier dans un système de reconnaissance de la parole. Néanmoins, pour des raisons de difficulté de mise en oeuvre, la mesure la plus couramment utilisée est la perplexité. La perplexité est estimée sur un corpus de développement (non inclus lors de la phase d'apprentissage du modèle de langage) pour définir si les modèles choisis modélisent correctement le corpus. Plus la valeur de perplexité est petite, plus le modèle de langage possède des capacités de prédiction. Pour les modèles n-grammes, la perplexité est définie comme suit :

$$PP = 2^{-1/n}\sum_{t=1}^{n} \log P(w_t\mathbf{h}) \tag{2.9}$$

Où $P(w_t\mathbf{h})$ est la probabilité associée au n-gramme $(w_t\mathbf{h})$.

2.6 Algorithme de décodage

Les progrès intervenus au cours des vingt dernières années ont permis d'améliorer les capacités des systèmes de reconnaissance de la parole. Elles sont passées de quelques mots reconnus en mode isolé pour un seul locuteur à des systèmes multi-locuteurs avec plusieurs milliers de mots, en parole continue voire spontanée. Ces différentes conditions engendrent de nouveaux problèmes. Il est ainsi nécessaire de limiter l'espace de recherche, qui croît de manière exponentielle, pour obtenir le bon résultat avec un temps de traitement acceptable. Les algorithmes de reconnaissance incluent donc très souvent des stratégies permettant de choisir un nombre limité d'hypothèses à chaque instant, et ainsi de n'explorer que l'espace suffisant pour trouver la meilleure solution. Les travaux de (Woszczyna, 1998) montrent que de nombreuses stratégies peuvent être intégrées à différents niveaux des SRAP, sans pour autant dégrader les performances.

Lors d'un décodage, le SRAP génère, à partir des probabilités obtenues par le modèle acoustique et le modèle de langage, un ensemble d'hypothèses de mots. Généralement, cet ensemble est codé sous la forme d'un graphe ou treillis de mots. Le décodeur

explore ce graphe pour trouver l'hypothèse qui maximise conjointement les probabilités acoustiques et linguistiques. Comme l'exploration complète du graphe est impossible, l'espace de recherche est limité par des heuristiques. Les algorithmes de reconnaissance incluent donc très souvent des stratégies permettant de choisir un nombre limité d'hypothèses à chaque instant, et ainsi de n'explorer que l'espace suffisant pour trouver la meilleure solution. Les algorithmes les plus fréquemment utilisés sont l'algorithme de Viterbi (Beam Search) présenté dans (Viterbi, 1967) et l'algorithme A^* présenté entre autre dans (Nocera et al., 2002).

2.6.1 Recherche synchrone basée sur un arbre réentrant

Cet l'algorithme est le plus répandu dans les SRAP, dont la mise en oeuvre la plus courante est un Viterbi en faisceaux (beam search) (Ney et al., 1992). Cet algorithme est synchrone car toutes les hypothèses explorées sont évaluées en parallèle sur la même portion de signal. Cette caractéristique permet d'appliquer facilement des heuristiques de coupure de l'espace de recherche. En effet, les heuristiques de cet algorithme écartent les hypothèses partielles dont la vraisemblance est trop faible. Elles réduisent ainsi la complexité de la recherche tout en augmentant sa rapidité.

2.6.2 Recherche synchrone à pile

Dans le cadre de ce travail, nous avons utilisé le SRAP du Laboratoire Informatique d'Avignon (LIA), SPEERAL, qui utilise un algorithme A^*. Il opère sur un treillis de phonèmes et non sur un treillis de mots comme les méthodes classiques. Les réalisations de ce type d'algorithme de recherche se fondent sur des piles qui ordonnent les hypothèses à explorer. L'algorithme A^* utilise une fonction heuristique pour guider la recherche qui dépend à la fois du chemin déjà parcouru et de l'estimation du coût du chemin qui reste à parcourir. L'ordre d'exploration des noeuds est déterminé par la fonction $f(n)$ qui représente une estimation du coût du meilleur chemin passant par le noeud n :

$$f(n) = g(n) + h(n) \tag{2.10}$$

Dans cette équation, $g(n)$ représente le score du chemin pour arriver à l'état courant n et $h(n)$ une estimation du score pour atteindre le noeud final. Cette deuxième fonction est souvent appelée sonde, puisqu'elle permet de guider l'algorithme pour qu'il choisisse le chemin le plus prometteur. La fonction $h(n)$ combine un score purement acoustique et un score d'anticipation linguistique. Cette fonction est critique pour la qualité de l'algorithme de recherche. Le score acoustique est calculé par un décodage acoustico-phonétique, qui est effectué par un algorithme de Viterbi arrière sur le treillis de phonèmes.

L'avantage principal de l'algorithme A^* est de pouvoir fournir en une seule passe les n meilleurs chemins au sein du graphe, il suffit pour cela de garder une pile des n

chemins les plus prometteurs et de les étendre à chaque fois, et ainsi de limiter l'espace de recherche tout en fournissant plusieurs résultats (Soong et Huang, 1991).

2.6.3 Décodage multi-passes

Les SRAP utilisent souvent des stratégies multi-passes. Généralement la première passe génère une transcription qui sera ré-utilisée pour adapter les modèles acoustiques en fonction des locuteurs ou de la qualité d'enregistrement. Les premières passes permettent également de générer des graphes de mots qui peuvent être ré-explorés *a posteriori* avec des modèles de langage plus importants. Ces stratégies en plusieurs passes permettent ainsi d'introduire à chaque itération une information supplémentaire : généralement les informations rajoutées n'auraient pu l'être à l'itération précédente, la quantité d'hypothèses en concurrence étant trop importante.

2.7 Évaluation d'un système de reconnaissance automatique de la parole

En reconnaissance automatique de la parole, la mesure d'évaluation la plus répandue est le taux d'erreur mot (TEM). Le TEM consiste à comparer la phrase reconnue et la phrase de référence (celle qui a effectivement été prononcée). Il est estimé par un algorithme d'alignement dynamique qui permet d'aligner l'hypothèse issue du SRAP avec le texte de référence. Il existe 3 types d'erreurs :

- *Insertion* : ajout d'une hypothèse de mot non présent dans la référence,
- *Substitution* : remplacement d'un mot par un autre,
- *Suppression* : omission d'un mot de la référence.

Le TEM est défini comme suit :

$$TEM = \frac{I + D + S}{W} * 100 \tag{2.11}$$

avec I le nombre d'insertions, D le nombre de suppressions, S le nombre de substitutions et W le nombre de mots 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, d'indexation (Senay, 2011), etc.

Chapitre 3

L'analyse factorielle pour la modélisation acoustique

Sommaire

3.1 Introduction

Dans l'analyse statistique, on utilise le terme générique d'analyse factorielle pour parler de deux types d'analyse légèrement différentes ayant de nombreux liens de parenté : l'analyse en composantes principales et l'analyse factorielle proprement dite.

L'Analyse en Composantes Principales (ACP) est une des techniques les plus utilisées en analyse de données multidimensionnelles. L'ACP réduit un vecteur par projection orthogonale sur le sous-espace, de dimension fixée *a priori*, qui maximise la variance des projetés. Dans le cas où la variabilité traitée est la variabilité totale, sa solution exacte est le sous-espace engendré par les premiers vecteurs propres de la matrice de covariance, dans l'ordre décroissant des valeurs propres. Cette méthode garantit une erreur minimale (erreur de reconstruction) entre vecteurs initiaux et vecteurs projetés au sens euclidien du terme (principe de "moindre inertie"). L'un des inconvénients majeurs de l'ACP est l'absence d'un modèle génératif des données et d'une densité de probabilité associée. En fait, l'ACP suppose implicitement que la distribution des données est un hyper-ellipsoide caractérisé par sa moyenne et sa matrice de covariance globale (Jolliffe, 1986).

Pour remédier à cette limitation, l'analyse en composantes principales probabiliste (ACPP) a été proposée par Tipping et Bishop (Tipping et Bishop, 1999). Dans ce modèle les axes de projection sont déterminés en utilisant le critère du maximum de vraisemblance sur les paramètres d'un modèle à variables latentes. L'ACP probabiliste offre plusieurs avantages par rapport à une ACP standard. Elle permet, par exemple, de calculer la vraisemblance des observations (de test ou d'apprentissage). Le calcul de la vraisemblance permet de donner une idée sur la validité du modèle proposé, selon les applications. Un autre avantage, qui n'est pas des moindres, est la possiblité d'utiliser un certain nombre d'approches d'estimation, comme l'approche bayésienne. L'approche bayésienne permet d'intégrer des connaissances *a priori* sur les distributions des paramètres des modèles.

A la base, l'ACPP est prévue pour des données suivant une distribution monomodale. Pour traiter les données suivant une distribution muti-modale, deux extensions de l'ACPP sont proposées. la première à étais proposée par Tipping et Bishop (Tipping et Bishop, 1999), ils ont considéré un mélange d'ACPP locales, chacune d'elles modélisant efficacement une partie de l'espace de sorte que la génération de chaque observation soit partagée par tous les modèles locaux. Une deuxième extension à était proposée par Kenny (Kenny et al., 2005b) dans le cadre du traitement de la parole. Dans son travail, Kenny propose de travailler dans l'espace formé par la concaténation des moyennes des gaussiennes issues de mixture de gaussiennes qui modélisent les différentes partie de l'espace acoustique. L'avantage de cette modélisation est qu'elle prend en compte la redondance qui peut exister entre les différentes parties de l'espace des observations (les gaussiennes).

Dans ce chapitre nous présentons les origines théoriques de l'analyse factorielle ainsi que la manière dont on estime les paramètres du modèle sont estimés. D'abord, nous exposons la théorie de l'ACP dans la section 3.2. Dans la section 3.3 nous mon-

trons comment l'ACP peut être formulée comme une solution du maximum de vraisemblance d'un modèle à variables latentes particulier. Ensuite, nous exposons le modèle d'analyse factorielle et son application à un modèle de mélange de gaussiennes dans la section 3.4. La méthode de l'estimation des paramètres du modèle d'analyse factorielle est présentée dans la section 3.5.2. Enfin, dans la section 3.6 nous exposons l'utilisation de l'analyse factorielle pour la modélisation de la session dans un système de vérification du locuteur.

3.2 L'analyse en composantes principales (ACP)

L'analyse en composantes principales (Jolliffe, 1986) est l'une des techniques les plus populaires pour le traitement, la compression et la visualisation de données multidimensionnelles. Etant donné un ensemble de vecteurs y_i, $i = 1, ..., N$, de dimension d, l'ACP consiste à chercher les axes de projection orthogonaux suivant lesquels la variance est maximale. L'approximation optimale, au sens de l'erreur quadratique moyenne [1], d'un vecteur y_i par un vecteur $\widehat{t}_i$ de dimension $r < d$ est donnée par :

$$\widehat{t}_i = \mathbf{W}_r^t(y_i - \mu) \tag{3.1}$$

où μ est la moyenne des y_i et W_r est la matrice de projection composée des r premiers vecteurs propres de la matrice de covariance des données Σ_y, correspondant aux r plus grandes valeurs propres données dans l'ordre descendant $(\lambda_i)_{i=1,...,r}$. La matrice de covariance des données réduites est diagonale d'éléments $(\lambda_i)_{i=1,...,r}$. L'erreur quadratique de l'approximation est donnée par la somme des valeurs propres écartées (les plus petites) :

$$e^2 = \sum_{i=r+1}^{d} \lambda_i \tag{3.2}$$

Le choix de r peut être basé sur l'équation 3.2, ou d'une manière équivalente sur le choix d'un seuil p entre 0 et 1 tel que :

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq p \tag{3.3}$$

Bien que l'ACP soit une technique très utilisée dans le domaine de l'analyse des données, elle présente l'inconvénient de se baser uniquement sur une approche géométrique très étroite. En effet, son rôle consiste simplement à réajuster le nuage des

1. L'erreur quadratique moyenne (EQM) est une mesure de l'erreur moyenne, pondérée par le carré de l'erreur. Elle permet de répondre à la question, « quelle est la magnitude de l'erreur de la prévision », mais n'indique pas la direction des erreurs. Parce qu'il s'agit d'une quantité au carré, l'EQM est plus influencée par les grandes erreurs que par les petites erreurs. Sa portée est de 0 à l'infini, un score de 0 étant un score parfait.

points ou réordonner les axes de l'espace afin de ne donner de l'importance qu'aux axes de grandes variances. Comme annoncé précédemment, l'inconvénient majeur de l'ACP est l'absence d'un modèle génératif de données et d'une densité de probabilité associée. Pour remédier à ce problème , Tipping et al. (Tipping et Bishop, 1999) ont introduit l'ACP probabiliste (ACPP) qui fera l'objet de la prochaine section.

3.3 L'analyse en composantes principales probabiliste (ACPP)

L'Analyse en Composantes Principales Probabiliste (ACPP) a été présentée pour la première fois par Tipping et Bishop en 1999. Ils ont étendu des travaux précédents sur l'analyse en facteurs et ont montré comment l'ACP pouvait être formulée comme une solution du maximum de vraisemblance d'un modèle à variables latentes particulier. L'ACP probabiliste se dérive d'un modèle de variables cachées, avec les hypothèses d'un bruit isotrope et un *a priori* gaussien (Tipping et Bishop, 1999). Le vecteur observé y de dimension d est généré à partir du vecteur caché t de dimension r suivant l'expression :

$$y = At + \mu + \varepsilon \tag{3.4}$$

où A est une matrice de dimension $d \times r$, μ est la moyenne des données et ε est un bruit gaussien de moyenne nulle et de matrice de covariance $\sigma^2 I$, I étant la matrice identité $d \times d$.

La distribution de y connaissant t suit une densité gaussienne :

$$p(y/t) = (2\pi\sigma^2)^{-\frac{d}{2}} exp\{-\frac{1}{2\sigma^2}\|y - At - \mu\|^2\} \tag{3.5}$$

Un *a priori* gaussien est choisi pour t :

$$p(t) = (2\pi)^{\frac{-r}{2}} exp\{-\frac{1}{2}t^t t\} \tag{3.6}$$

ce qui donne la distribution gaussienne du vecteur y :

$$p(y) = (2\pi)^{\frac{-d}{2}} |C|^{\frac{-1}{2}} exp\{\frac{-1}{2}(y-\mu)^t C^{-1}(y-\mu)\} \tag{3.7}$$

avec $C = \sigma^2 I + AA^t$ une matrice $d \times d$. La distribution *a posteriori* est donnée par :

$$p(t/y) = (\frac{1}{2\pi\sigma^2})^{\frac{r}{2}} |M|^{\frac{1}{2}} exp[\frac{-1}{2}\{t - M^{-1}A^t(y-\mu)\}^t(\sigma^{-2}M)\{t - M^{-1}A^t(y-\mu)\}] \tag{3.8}$$

avec M une matrice de dimension $r \times r$.

$$M = \sigma^2 I + A^t A \tag{3.9}$$

La log-vraisemblance des données par rapport au modèle est :

$$\mathcal{L} = \sum_{i=1}^{N} \ln\{p(y_i)\} \tag{3.10}$$

$$= -\frac{N}{2}\{d\ln(2\pi) + \ln|C| + tr(C^{-1}\Sigma_y)\} \tag{3.11}$$

où $\Sigma_y = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mu)(y_n - \mu)^t$ est la matrice de covariance des données observées.

La maximisation de $\mathcal{L}$ donne les estimations suivantes des paramètres (Tipping et Bishop, 1999) :

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3.12}$$

$$\sigma^2_{ML} = \frac{1}{d-r}\sum_{j=r+1}^{d} \lambda_j \tag{3.13}$$

$$A_{ML} = U_r(\Lambda_r - \sigma^2_{ML} I)^{\frac{1}{2}}\Re \tag{3.14}$$

où λ_i représente les valeurs propres de Σ_y ordonnées en valeurs décroissantes ($\lambda_1 \geq \lambda_1 \geq ... \geq \lambda_d$). Λ_r est la matrice diagonale des r premières valeurs propres, U_r est la matrice des r premiers vecteurs propres correspondants et $\Re$ est une matrice de rotation orthogonale $r \times r$ quelconque.

Avec l'ACPP, le vecteur d'observation y_i n'est plus caractérisé par un vecteur caché unique sur l'espace réduit, comme c'est le cas de l'ACP classique (équation 5.1), mais par une distribution gaussienne *a posteriori* (équation 3.8). Un estimateur du vecteur caché associé à y_i est donc la moyenne de la loi aléatoire t suivant la densité *a posteriori* :

$$\widehat{t_i} = M^{-1}_{ML} A^t_{ML}(y_i - \mu) \tag{3.15}$$

avec $M_{ML} = \sigma^2 I + A^t_{ML} A_{ML}$(équation 3.9)

Le modèle définie ci dessus permet de résoudre le problème d'absence d'un modèle génératif des données et d'une densité de probabilité associée. Mais ce modèle

reste global et suppose implicitement que la distribution des données soit un hyper-ellipsoïde caractérisé par sa moyenne et sa matrice de covariance globale (Jimenez et Landgrebe, 1995). Ainsi, l'ACPP peut provoquer la perte définitive d'informations caractérisant d'éventuelles structures locales des données. Pour remédier à ce problème Tipping et Bishop ont considéré un mélange d'ACPP locales, chacune d'elles modélisant efficacement une partie de l'espace (une classe) de sorte que la génération de chaque observation soit partagée par tous les modèles locaux. Nous aboutissons ainsi à un modèle global non linéaire. Cette méthode présente l'avantage d'estimer la partition de l'espace en même temps que les ACPP locales les modélisant. Cependant, ce modèle ne permet pas de mettre en évidence la corrélation entre les différents composantes du mélange de gaussiennes qui modélise les différentes parties de l'espace. Kenny a proposé de travailler dans l'espace formé par la concaténation des moyennes des gaussiennes issues de mixture de gaussienne qui modélise les différentes parties de l'espaces (Kenny et al., 2005b). L'avantage de cette modélisation est la prise en compte de l'information de voisinage entre différente classe de l'espace. Par exemple, dans le cas du traitement d'image, ce modèle tient compte de l'information de voisinage entre pixels qui est très importante pour la régularisation des résultats de traitement (Flitti, 2005).

3.4 Analyse Factorielle

L'analyse factorielle est une généralisation de l'ACP Probabiliste. De la même manière que dans l'ACPP, le vecteur observé y de dimension d est généré à partir du vecteur caché t de dimension r suivant l'expression :

$$y = At + \mu + \varepsilon \tag{3.16}$$

La différence avec l'ACPP se situe dans l'hypothèse faite concernant la distribution du bruit ε. Dans le cas du l'analyse factorielle, le bruit est supposé être gaussien de moyenne nulle et de matrice de covariance Σ diagonale. Les éléments de Σ sont différents d'une composante à l'autre, contrairement à l'ACPP.

3.5 Application à un GMM

Les GMM sont généralement utilisés pour approximer des densités de probabilités. Les GMM sont estimés, généralement, en utilisant le critère du Maximum de vraisemblance, sur des données observées (appelées réalisations). Les données observées englobent différentes caractéristiques, parmi lesquelles se trouve la caractéristique que nous souhaitons modéliser. Par exemple, les trames d'une session de parole contiennent des informations sur le locuteur, la canal, l'état émotionnel du locuteur ou le contenu phonétique.

3.5.1 Définitions, notations et modèle

Lorsque nous nous intéressons aux informations sur le locuteur, nous noterons le super-vecteur associé au GMM correspondant aux données observées m_l. Si par contre nous nous intéressons à l'information canal, le super-vecteur sera noté m_c. D'une manière générale, nous noterons le super-vecteur m_e, avec e étant une caractéristique quelquonque : locuteur, canal, etc.

Les moyennes d'un GMM peuvent être vues comme des variables aléatoires dont les caractéristiques statistiques peuvent être obtenues en utilisant l'analyse factorielle. Une manière de faire est d'appliquer l'équation 3.16 à chacune des moyennes des gaussiennes du mélange :

$$m_e^i = m^i + A^i t^i + \varepsilon \tag{3.17}$$

avec m_e^i le vecteur moyenne de la gaussienne i pour la caractéristique e. L'inconvénient de ce modèle est que les différentes composantes de la mixture de gaussiennes sont traitées indépendamment les unes des autres. Ce qui ne permet pas de mettre en évidence la corrélation pouvant exister entre les différentes régions (gaussiennes) de l'espace des observations. La solution proposée par Kenny (Kenny et al., 2005b) est d'appliquer le modèle d'analyse factorielle à une nouvelle variable aléatoire obtenue par la concaténation des moyennes du GMM. Cette nouvelle variable aléatoire est appelée super-vecteur. Dans ce cadre, tous les GMM sont obtenus à partir d'un GMM global, estimé sur une grande quantité de données et incluant le plus de variabilités possibles. Ce GMM est appellé l'UBM (Universal Backgroud Model). Soit m le super-vecteur associé à l'UBM. le super-vecteur observation m_e est obtenu comme suit :

$$m_e = m + \mathbf{U}\mathbf{x}_e + \varepsilon \tag{3.18}$$

Il est important de souligner le fait que dans ce modèle le vecteur x_e est commun à toutes les gaussiennes. Cette caractéristique permet d'exploiter la redondance pouvant exister entre les différentes gaussiennes. Ce vecteur x_e est une représentation compacte de m_e.

3.5.2 Estimation des paramètres

Dans le modèle de l'équation 3.16, la variable aléatoire en question est le super-vecteur. Il faut donc estimer les super-vecteurs à partir des observations (trames dans le cas de la parole). Ensuite, ces super-vecteurs peuvent être utilisés pour estimer la matrice de projection ($\mathbf{U}$) et les différents vecteurs projetés ($\mathbf{x}_e$). Cependant, il est possible d'estimer ces paramètres directement à partir des observations, sans passer par les super-vecteurs. Cette estimation peut être réalisée en utilisant le critère du maximum de vraisemblance, c'est ce que nous allons décrire dans le reste de cette section.

Dans la suite, nous décrivons la procédé d'estimation de la matrice $\mathbf{U}$ et des vecteurs $\mathbf{x}_e$. Avant tout nous présentons les notations nécessaires au développement de la stratégie d'estimation :

- m_e : super-vecteur associé à la caractéristique e.
- m : super-vecteur du modèle du monde (UBM).
- Σ_g : matrice de covariance de la gaussienne g.
- $\chi(e)$: données d'apprentissage de la caractéristique e.
- $P_{GMM}(\chi(e)|m,\Sigma)$: vraisemblance de $\chi(e)$ étant donné le GMM spécifique au super-vecteur m et de super-matrice de covariance Σ.
- $N_g(e)$: nombre de vecteurs acoustiques d'un enregistrement e appartenant à la gaussienne g.
- G : nombre des gausiennes dans le GMM-UBM
- F : la taille de vecteurs acoustiques.
- Pour chaque gaussienne g nous calculons :

$$S_{X,g}(e) = \sum_t (X_t - \mu_g) \qquad ; \qquad S_{X^tX,g}(e) = \sum_t (X_t - \mu_g)^t (X_t - \mu_g) \tag{3.19}$$

où $\sum_t$ est la somme sur tous les vecteurs acoustiques d'un enregistrement e appartenant à la gaussienne g et μ_g est la moyenne de la gaussienne g du GMM du modèle du monde (UBM).

On prend comme estimation de $\mathbf{x}_e$, sa valeur la plus probable a posteriori (estmation MAP). L'algorithme de son estmation est basé sur la propriété suivante :

Propriété 1 : Dans la phase d'apprentissage, pour chaque enregistrement e, la distribution *a posteriori* de $\mathbf{x}_e$ sachant $\chi(e)$ et les paramètres $\mathbf{U}$ et Σ suit une loi gaussienne de moyenne $l^{-1}(e)^t \mathbf{U}\Sigma^{-1} S_X(e)$ et de matrice de covariance $l^{-1}(e)$.

La matrice $\mathbf{U}$ peut être vue comme une concaténation de matrices $\mathbf{U}_g$. où g désigne l'indice de la gaussienne g dans l'UBM. L'estimation de la $i^{ième}$ ligne de la matrice $\mathbf{U}_g$ repose sur le résultat suivant :

Propriété 2

$$\mathbf{U}_g^i \sum_e N_g(e) E[\mathbf{x}_e^t \mathbf{x}_e)] = \sum_e S_{X,g}^i(e) E[^t\mathbf{x}_e] \tag{3.20}$$

où $S_{X,g}^i(e)$ est la $i^{ième}$ ligne de $S_{X,g}(e)$.

grâce à l'équation 3.20, la matrice $\mathbf{U}$ peut être estimer ligne par ligne. L'algorithme 2 présente la stratégie adoptée pour estimer la matrice $\mathbf{U}$ et les vecteurs x_e. Les démonstrations de ces deux propriétés se trouvent dans l'article (Kenny et al., 2005a)

Algorithm 1: Algorithme d'estimation de la matrice **U** et de vecteur **x**.

pour chaque enregistrement e : $x_{(e)} \leftarrow 0$, $\mathbf{U} \leftarrow random$;
pour $i = 1$ *to* $nb\_iterations$ **faire**
  pour *tous les enregistrements e* **faire**
    Estimation de : $S_{X,g}(e), S_{X^t X,g}(e)$; (eq : 3.19)
    Estimation de : $E[\mathbf{x}_e], E[\mathbf{x}_e.{}^t\mathbf{x}_e]$;
  fin
  Estimation de la matrice **U** ;(eq : 3.20)
fin

3.6 Analyses factorielles pour la vérification de locuteur

Dans cette thèse nous utilisons l'approche de l'analyse factorielle dans la modélisation acoustique pour la reconnaissance automatique de la parole. Le modèle analyse factorielle que nous utilisons est inspiré de la modélisation par analyse factorielle dans le domaine de vérification de locuteur (Kenny et al., 2005b). Dans le Système de Vérification du Locuteur (SVL), cette modélisation a permis d'améliorer la robustesse face à la variabilité session. Dans cette section nous présenterons brièvement le modèle d'analyse factorielle appliqué à la vérification du locuteur.

Dans le cadre d'un SVL, le modèle du monde, (GMM-UBM), représente la génération de vecteurs cepstraux provenant d'une multitude de locuteurs et de sessions. L'estimation des paramètres de ce modèle est réalisée en utilisant l'algorithme EM (Expectation-Maximisaton, voire section 2.4.1.2). Le GMM d'un locuteur donné est obtenu à partir de l'UBM en ré-estimant les moyennes. Les poids et les variances restent inchangés. Cette adaptation des moyennes est réalisée en utilisant l'approche d'analyse factorielle. Pour prendre en compte la variabilité liée au canal d'enregistrement, le modèle de l'équation 3.18 est étendu comme suite :

$$m_{h,l} = m + \mathbf{D}\mathbf{y}_l + \mathbf{U}\mathbf{x}_{h,l} \tag{3.21}$$

Dans ce modèle, le terme $m + Dy_l$, modélise cette fois la part propre au locuteur. **D** est une matrice diagonale de taille $GF \times GF$ et y_l est un vecteur de dimension GF estimé sur les données du locuteur l. **D** satisfait l'équation $\mathbf{I} = \tau \mathbf{D}^t \Sigma^{-1} \mathbf{D}$ où τ est un facteur appelé *relevance factor*. Le facteur $\mathbf{U}\mathbf{x}_{h,l}$ est la composante introduite par l'effet de la session (canal). Les vecteurs colonnes de la matrice **U** ($FG \times r$) représentent une base du sous-espace dans le quel évolue la variabilité session. $\mathbf{x}_{h,l}$ est un vecteur de dimention r contenant les composantes relatives à la session dans ce sous-espace.

La matrice **U** et le vecteur $\mathbf{x}_e$ sont estimés selon la description de la section 3.5.2. Mais, dans ce modèle les statistiques de l'équation 3.19 sont re-calculées comme suit :

$$S_{X,g}(e) = \sum_t (X_t - m_g - \mathbf{D}\mathbf{y})$$

$$S_{X^t X,g}(e) = \sum_t (X_t - \mathbf{m}_g - \mathbf{Dy})^t (X_t - \mathbf{m}_g - \mathbf{Dy})$$

où $\sum_t$ est la somme sur tous les vecteurs acoustiques, du locuteur l, appartenant à la gaussienne g et $\mathbf{m}_g$ est la moyenne de la gaussienne g du GMM-UBM. Dans ce modèle, le vecteur y_l est estimé sur les données du locuteurs l en se basant sur l'équation suivante :

$$\{y_l\}_g = \frac{\tau}{(\tau + N_l(g))} . \mathbf{D}_g . \Sigma^{-1} . [\mathbf{X}_g^l - \sum_{h \in l} . \{\mathbf{m} + \mathbf{U}\mathbf{x}_{(h,l)}\}[g]] \tag{3.22}$$

avec $D_g = \frac{\Sigma_g^{1/2}}{\sqrt{\tau}}$.

3.7 Conclusion

Dans ce chapitre, nous avons exposé les origines théoriques de l'analyse factorielle. L'ACP est avant tout utilisée pour réduire le nombre de dimensions tout en maximisant la variabilité conservée, pour obtenir des facteurs indépendants (non corrélés), ou pour visualiser les données dans un espace à deux ou trois dimensions. L'analyse factorielle est quant à elle utilisée pour identifier une structure latente et pour éventuellement réduire par la suite le nombre de variables mesurées si elles sont redondantes vis-à-vis des facteurs latents.

Nous avons exposé l'ACPP qui représente une extension de l'ACP classique afin de remédier à l'absence d'un modèle génératif des données et d'une densité de probabilité associée. Dans ce modèle les axes de projection sont déterminés en utilisant le critère du maximum de vraisemblance sur les paramètres d'un modèle à variables latentes.

Ensuite, nous avons expliqué les extensions possibles de L'ACPP à des données suivantes une distribution multi-modale. *À* la base, l'ACPP est prévue pour des données suivant une distribution mono-modale. Dans la première extension, Tipping et Bishop (Tipping et Bishop, 1999) ont considéré un mélange d'ACPP locales, chacune d'elles modélisant efficacement une partie de l'espace de sorte que la génération de chaque observation soit partagée par tous les modèles locaux. Dans la deuxième extension (Kenny et al., 2005b) proposent, dans le cadre du traitement de la parole, de travailler dans l'espace formé par la concaténation des moyennes des gaussiennes issues de mixture de gaussiennes qui modélisent les différentes parties de l'espace acoustique.

Chapitre 4

Analyse factorielle pour la modélisation acoustique compacte

Sommaire

4.1 Introduction

Dans les systèmes de reconnaissance automatique de la parole (SRAP), le signal acoustique est modélisable par un ensemble réduit d'unités acoustiques, qui peuvent être considérées comme des sons élémentaires de la langue. L'unité acoustique la plus utilisée est le phonème dépendant du contexte. Lorsque le contexte phonétique comprend le phonème précédent et le phonème suivant, nous parlons de modèles triphones. Dans le cadre des SRAP Markoviens, un phonème contextuel est modélisé par un modèle de Markov caché (MMC) gauche-droite à trois états. Les MMC sont des chaînes n'autorisant que des transitions de la gauche vers la droite afin de tenir compte de l'évolution temporelle du signal de parole. Pour relier ce modèle aux vecteurs de paramètres acoustiques du signal de parole, à chaque état est associé une densité de probabilité. La densité de probabilité pour un état donné est approximée en utilisant un mélange de gaussiennes (GMM : Gaussian Mixture Model), modélisant la génération des vecteurs acoustiques par cet état.

Le triphone permet de rendre la modélisation plus précise, mais cette amélioration théorique est limitée dans la pratique par des problèmes d'estimation. En effet, nous ne pouvons pas modéliser tous les contextes possibles des phonèmes, parce que le nombre de représentations, dans le corpus d'apprentissage, de certains phonèmes contextuels, est insuffisant. Un autre problème lié aux modèles contextuels est leur taille qui augmente rapidement avec le nombre de contextes pris en compte. Cette taille importante présente un grand défi lors de l'intégration de systèmes de reconnaissance de la parole dans des terminaux légers (téléphones mobiles, lecteurs MP3, etc.). En effet, les modèles peuvent avoir plus de 10 millions de paramètres, ce qui représente une complexité incompatible avec les puissances de calcul et l'espace mémoire généralement disponibles dans ces appareils.

Différentes approches ont été proposées dans la littérature pour réduire l'empreinte mémoire des modèles acoustiques, tout en préservant de bonnes performances. L'approche la plus communément utilisée dans les SRAP consiste à associer le même GMM aux états qui sont acoustiquement proches (Young, 1992). Cette mutualisation est le plus souvent appliquée entre modèles contextuels représentant le même phonème, sur des états ayant la même position relative à l'intérieur du MMC. Le regroupement est souvent réalisé par un algorithme de classification fondé sur des arbres de décision.

Une deuxième approche a connu un grand succès dans la modélisation acoustique compacte : le modèle HMM semi-contenu (SCHMM : Semi-Continuous Hidden Markov Model) (Huang et al., 1989). Dans ce modèle, les mixtures de gaussiennes des états des MMC partagent le même dictionnaire de gaussiennes appelé *codebook*. Les états sont différenciés entre eux par un simple vecteur de poids. Ces poids sont généralement ré-estimés, avec des données propres à chaque état. Plusieurs méthodes d'estimation du dictionnaire de gaussiennes d'un SCHMM ont été proposées (Lee et al., 1990) (Demuynck et al., 1996). L'estimation des poids est généralement effectuée par maximisation de la vraisemblance. La mutualisation massive des paramètres, dans le modèle SCHMM, permet de réduire de façon significative l'espace mémoire requis par

le stockage des modèles acoustiques

Dans (Bocchieri et Mak, 2001), les auteurs proposent une nouvelle approche de partage de paramètres pour les SRAP. L'idée principale est de partir d'un MMC existant, puis de séparer les vecteurs acoustiques en différents flux (streams ou subspaces) et ensuite de projeter chaque gaussienne du MMC sur les différents flux. Enfin, les distributions projetées similaires sont regroupées pour réduire le nombre de gaussiennes dans chaque flux. Le modèle obtenu est appelé SCDHMM pour *Subspace Distribution Clustering HMM*. L'utilisation des SDCHMM permet un gain du point de vue du stockage des modèles acoustiques (par rapport à l'utilisation de MMC). (Bocchieri et Mak, 2001) montrent que ce gain augmente en fonction du nombre de flux dans le SCDHMM (réduction de 63 % et 74 % respectivement pour 4 et 20 flux).

Dans ce chapitre, nous présentons une nouvelle méthode de mutualisation des paramètres des modèles acoustiques. La méthode proposée s'appuie sur l'utilisation de l'analyse factorielle. Cette méthode consiste à représenter l'espace acoustique de la parole par un modèle générique appelé modèle du monde (Universal Background Model : UBM), puis à dériver les modèles des différents états des MMC depuis ce modèle générique, en mutualisant une large partie des paramètres du modèle. Nous détaillons cette proposition dans la section 4.2. Avant de présenter nos expériences, nous examinons une nouvelle approche proposée récemment par (Povey et al., 2010) nommée *Subspace Gaussian Mixture Model* (SGMM). Cette méthode a été développée parallèlement à la nôtre et présente de nombreuses similarités.

Cette modélisation possède une grande similarité avec la modélisation d'analyse factorielle que nous proposons. Dans la section 4.3, nous détaillons cette approche (SGMM) en discutant les différences avec notre proposition. Les résultats expérimentaux sont exposés dans la section 4.4.

4.2 Modélisation d'analyse factorielle

Nous présentons, dans ce chapitre, une nouvelle méthode de mutualisation des paramètres acoustiques pour le système de RAP. Cette méthode est inspirée de la modélisation d'analyse factorielle appliquée dans le domaine de la vérification du locuteur (Kenny et al., 2005b). Dans le système de vérification du locuteur (SVL), cette modélisation a permis d'améliorer la robustesse face à la variabilité session [1] qui représente une cause majeure de dégradation des performances. Dans la suite, nous exposons une description de la modélisation d'analyse factorielle dans le cadre de la vérification du locuteur.

1. D'une manière générale, les termes *décalage de session* ou *variabilité session* sont utilisés pour désigner le changement des conditions acoustiques, qui peuvent varier grandement d'une session à l'autre. Le terme *variabilité session* englobe un grand nombre de phénomènes : le canal de transmission, bruit environnant, position du microphone., etc.

4.2.1 Vérification du locuteur

La plupart des systèmes de vérification du locuteur sont fondés sur l'approche *GMM-UBM*. L'UBM est un GMM modélisant la génération de vecteurs d'observation provenant d'une multitude de locuteurs et de conditions acoustiques. L'estimation des paramètres de ce modèle est réalisée en utilisant l'algorithme EM. À partir du modèle UBM, les modèles de locuteurs sont obtenus en utilisant une approche d'adaptation standard (adaptation par maximum *a posteriori* (MAP)). Dans le cadre de la vérification du locuteur, seules les moyennes sont estimées, les poids et les variances restent inchangés. Les modèles de locuteurs résultant contiennent non seulement l'information locuteur, mais aussi l'information sur la session, ce qui n'est pas souhaitable. L'approche d'analyse factorielle propose de séparer les deux composantes *locuteur* et *session*. Dans ce cadre, le modèle pour une session donnée s'écrit comme étant la somme de trois composantes : une composante générale indépendante du locuteur et de la session, une composante dépendante seulement du locuteur, et une composante dépendante seulement de la session. Le modèle du locuteur l dans la session h s'écrit :

$$\mathbf{m}_{(h,l)} = \mathbf{m} + \mathbf{D}\mathbf{y}_l + \mathbf{U}\mathbf{x}_{(h,l)} \tag{4.1}$$

Où $\mathbf{m}$ est le super-vecteur du modèle du monde de dimension $M \times D$, M étant le nombre de gaussiennes dans les GMM et D la dimension de l'espace des paramètres acoustiques. Il est obtenu par la concaténation des moyennes de ses gaussiennes. $\mathbf{m}(h,l)$ est le super-vecteur correspondant au locuteur l dans la session h. $\mathbf{D}$ est une matrice diagonale ($MD \times MD$), $\mathbf{y}_l$ est un vecteur de dimension MD estimé sur les données du locuteur l. Par rapport à une estimation MAP, le modèle FA introduit en plus le terme $\mathbf{U}\mathbf{x}(h,l)$: $\mathbf{U}$ est une matrice de rang faible R ($R \ll MD$), ses vecteurs colonnes forment une base d'un sous-espace dans lequel la variabilité session est la plus forte. $\mathbf{x}(h,l)$ est un vecteur de dimension R contenant les composantes relatives à la session dans ce sous-espace.

4.2.2 Modélisation acoustique compacte

Dans le cadre des SRAP, un phonème contextuel est modélisé par un MMC gauche-droite à trois états. Chaque état est modélisé par un GMM. Habituellement les GMM des états sont estimés indépendamment les uns des autres, ce qui augmente énormément le nombre de paramètres du modèle acoustique. Dans ce travail, nous proposons de dériver tous les GMM des états à partir d'un seul GMM générique en s'appuyant sur la technique d'analyse factorielle.

4.2.2.1 Adaptation des moyennes des états des GMM

Dans le cas de modélisation acoustique pour le SRAP, le modèle du monde représente l'espace acoustique de la parole incluant tous les phonèmes. Les dérivations des

moyennes sont accomplies par une version simplifiée de l'équation 4.1. Le modèle de l'état est décomposé en deux composantes (figure 4.1) : une composante générale commune à tous les états et une composante dépendante de l'état. Le super-vecteur de l'état e est obtenu avec l'équation suivante :

$$\mathbf{m}_{(e)} = \mathbf{m} + \mathbf{U}\mathbf{x}_{(e)} \tag{4.2}$$

Dans ce modèle, $\mathbf{m}$ joue le même rôle que dans l'équation 4.1. $\mathbf{U}\mathbf{x}_{(e)}$ est la composante qui dépend de l'état e, cette composante permettant de distinguer un état d'un autre. La matrice $\mathbf{U}$, de faible rang R, modélise la variabilité inter-états, les R colonnes représentent le sous-espace dans lequel est localisée la variabilité inter-états. Les facteurs d'états $\mathbf{x}_e$ sont des vecteurs de taille R caractérisant les états. La matrice $\mathbf{U}$ et les vecteurs $\mathbf{x}_e$ sont estimés de manière itérative (figure 4.2).

Cette modélisation permet de trouver une solution au problème de l'apprentissage acoustique qui consiste à trouver le bon compromis entre le nombre de paramètres à utiliser pour le modèle et la quantité de données d'apprentissage. En effet, la matrice $\mathbf{U}$ est estimée en utilisant les données de tous les états, alors que les paramètres des facteurs d'états, qui ne nécessitent pas une grande quantité de données, sont estimés sur les données propres à chaque état.

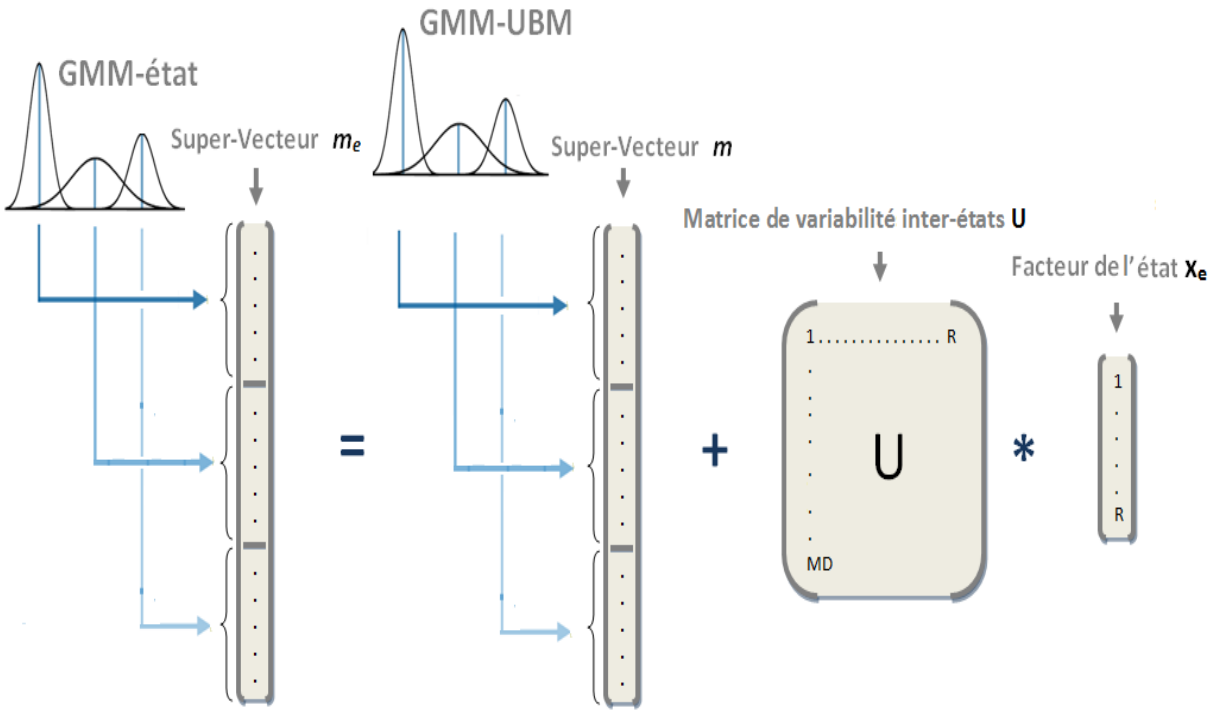

FIGURE 4.1 – *Décomposition du super-vecteur m_e d'un état e en une composante générale commune à tous les états et une composante dépendante de l'état e.*

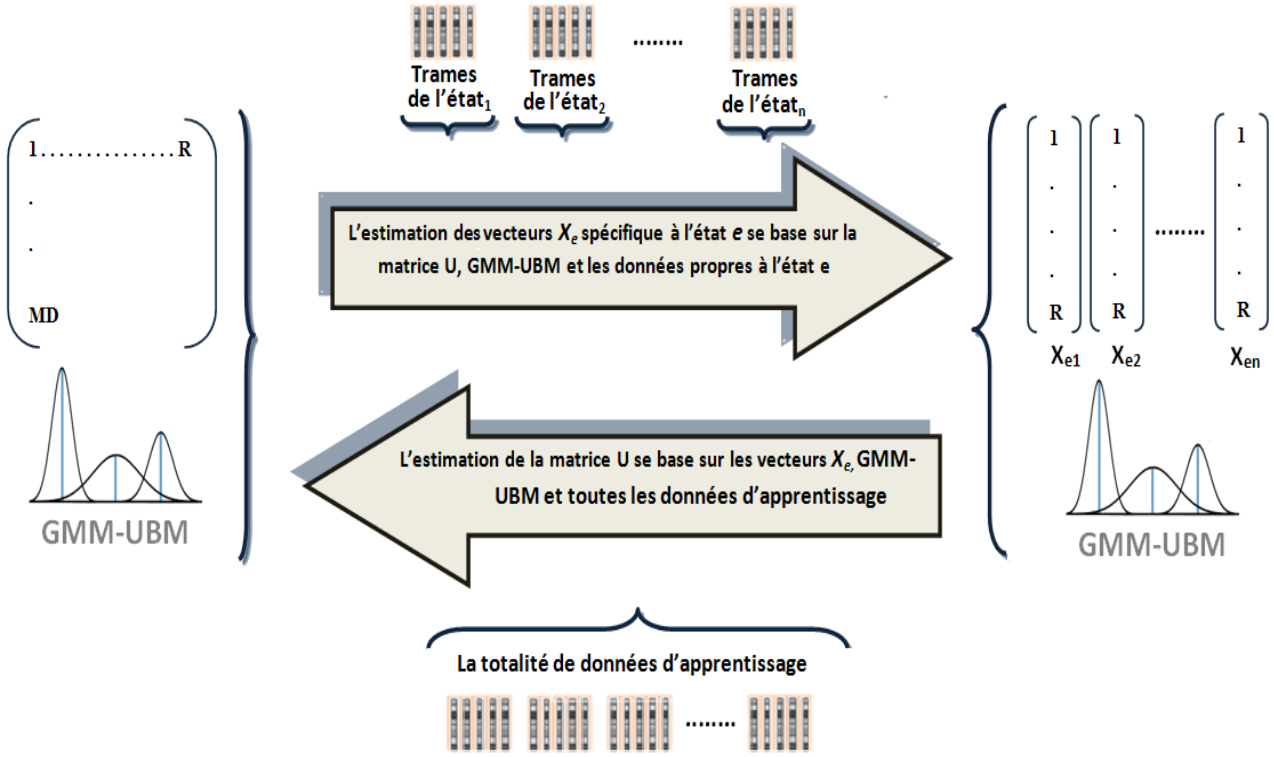

FIGURE 4.2 – Processus itératif de l'estimation de la matrice $\boldsymbol{U}$ et des vecteurs spécifiques à l'état $\boldsymbol{x}_e$.

4.2.2.2 Adaptation des poids des états des GMM

Après l'adaptation des moyennes, nous ré-estimons les poids des gaussiennes par une simple itération de l'algorithme EM (*Expectation-Maximisation*). Soit w_g le poids de la gaussienne g dans le GMM-UBM. Le poids w_g^s de cette gaussienne dans l'état e est calculé comme suit :

$$w_g^e = \frac{\sum_{t \in e} P(g|t)}{N_e} \tag{4.3}$$

où,

$$P(g|t) = \frac{w_g * f(t|g)}{\sum_{g'} w_g * f(t|g')} \tag{4.4}$$

$N_{(e)}$ est le nombre de trames de l'état (e) et $P(g|t)$ est la probabilité que la gaussienne g ait généré la trame t.

Pour une réduction encore plus importante du nombre de paramètres dans un état, nous ne garderons que les **N** gaussiennes possédant les poids les plus importants. **N** est choisi de telle manière que la somme des poids des gaussiennes choisies atteigne un seuil prédéfini. Le modèle compact proposé ici est défini par : le modèle UBM, la

matrice de variabilité inter-états **U**, les vecteurs de facteurs d'états **X** et des vecteurs correspondant aux nouveaux poids **W** :

$$\underbrace{(2*nbp+1)*nbg_{ubm}}_{GMM-UBM} + \underbrace{nbg_{ubm}*nbp*R}_{U} + \underbrace{nbs*R}_{X} + \underbrace{nbs*nbest}_{W} \tag{4.5}$$

où nbg_{ubm} est le nombre de gaussiennes dans l'UBM, nbs est le nombre d'états émetteurs, nbp est la taille des vecteurs cepstraux, et $nbest$ est le nombre des gaussiennes sélectionnées par état. Le nombre N de paramètres du modèle de base est calculé par l'équation suivante :

$$N = nbs * [\ \underbrace{(2*nbp+1)}_{\text{une Gaussienne}} * nbg_s\ ] \tag{4.6}$$

où nbg_s est le nombre de gaussiennes par état. Dans la modélisation proposée, les variances resteront inchangées. Tous les GMM d'états partagent les variances de l'UBM.

Avant de présenter nos expériences, nous examinons une nouvelle approche proposée récemment par (Povey et al., 2010) nommée *Subspace Gaussian Mixture Model* (SGMM). Cette modélisation possède une grande similarité avec la modélisation *analyse factorielle* que nous proposons. Dans la section suivante, nous détaillons cette approche en discutant les différences avec notre proposition.

4.3 Comparaison avec SGMM pour la modélisation acoustique compacte

De manière similaire à la modélisation proposée dans ce travail, dans la modélisation SGMM, les mélanges de gaussiennes des états sont dérivées à partir d'un modèle GMM-UBM. Chaque état est associé à un vecteur $\mathbf{v} \in R^S$ appelé *vector-valued* spécifique à l'état. Un vecteur $\mathbf{v}_j$ associé à un état j est partagé globalement pour estimer les moyennes et les poids des gaussiennes associés à cet état. Dans ce modèle, les moyennes et les poids d'une gaussienne i appartenant à l'état j sont calculés par les équations suivantes :

$$\mu_{ji} = M_i \mathbf{v}_j \tag{4.7}$$

$$w_{ji} = \frac{exp\ \mathbf{w}_i^T \mathbf{v}_j}{\sum_{i'=1}^{I} exp\ \mathbf{w}_{i'}^T \mathbf{v}_j} \tag{4.8}$$

Nous remarquons que la dérivation des moyennes dans le modèle SGMM (équation 4.7) est semblable à la dérivation des moyennes dans notre modèle (équation 4.2). Par contre, l'estimation des poids est plus complexe que notre proposition : c'est une

dérivation fondée sur une combinaison de l'inégalité de type Jensen, des expansions de séries de Taylor de second ordre, et une modification de la résultante de la fonction quadratique auxiliaire qui assure la stabilité tout en conservant le même gradient local. Nous remarquons aussi que le même vecteur v_j est utilisé dans la dérivation des moyennes μ_{ji} et des poids w_{ji}.

Dans (Povey et al., 2010), les auteurs ont montré que ce modèle permet une réduction de 68 % du nombre global des paramètres à estimer par rapport au modèle classique (où les GMM des états sont estimés indépendamment les uns des autres). Mais cette réduction cause une perte de performance du SRAP de 10 % relatif par rapport au modèle classique.

Pour améliorer les performances du modèle SGMM, Povey propose dans (Povey et al., 2010) de multiplier le nombre de vecteurs v_j pour l'état j, ce qui permet d'augmenter le nombre de gaussiennes associées à chaque état. Pour ce faire, les trames de chaque état j sont subdivisées en m classes. Chaque classe C_{jm} se voit associer un vecteur v_{jm}.

Bien que notre proposition et la proposition dans (Povey et al., 2010) aient le même fondement théorique, la différence entre les deux formalismes est considérable. L'approche SGMM est caractérisée, comme nous avons vu auparavant, par sa complexité, alors que notre modélisation est simplifiée par le formalisme qui provient de la vérification du locuteur. Les moyennes sont estimées avec l'analyse factorielle, alors que les poids sont ré-estimés avec une simple itération d'algorithme EM.

4.4 Expérimentations

4.4.1 Cadre expérimental

Les expériences sont réalisées en utilisant le système de RAP grand vocabulaire du LIA, Speech RAL (SPEERAL) (Nocera et al., 2004). Ce système utilise l'algorithme A* pour le décodage et des MMC pour la modélisation acoustique. Le lexique contient 65 000 mots et le modèle de langage est un modèle tri-gramme estimé sur 200 millions de mots du journal *Le Monde* et sur environ 1 million de mots du corpus d'entraînement de la campagne d'évaluation ESTER (Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiodiffusées) (Galliano et al., 2005). Les paramètres acoustiques sont composés de 12 coefficients PLP de l'énergie et de leurs dérivées premières et secondes soit 39 dimensions. Dans les expériences citées dans ce chapitre, seule une passe de décodage est exécutée en trois fois le temps réel. Nous testons notre approche sur 7,5 heures de parole extraites de l'ensemble de test d'ESTER.

Le modèle acoustique de base est un MMC gauche-droite avec 10 002 phonèmes contextuels. Pour modéliser tous ces phonèmes contextuels, nous n'avons besoin que de 36 000 états dans le MMC. Ce nombre est réduit à 3 327 états par l'application de l'approche de partage des états (Young et al., 1994).

4.4.2 Modèle compact générique

L'objectif des premières expériences est de trouver un point de fonctionnement entre le taille et le performance du modèle. Nous pouvons contrôler la taille du modèle à travers le rang de la matrice **U** et le nombre de gaussiennes du GMM-UBM. Nous avons fait varier le rang de **U** et le nombre de gausiennes du GMM-UBM afin d'étudier l'impact de la réduction du nombre de paramètres sur les performances. Dans un premier temps, nous avons choisi de fixer la taille du GMM-UBM à 600 gaussiennes et de faire varier le rang de la matrice **U** (40, 100, 150, 200, 250 et 300). Ce test nous permet de trouver d'une manière empirique la taille optimale du sous-espace où la variabilité inter-états est localisée.

Dans le tableau 4.1, nous exposons les performances en termes de taux d'erreur mot (TEM) du modèle compact obtenu. Dans la première colonne, nous présentons les différentes stations radiophoniques auxquelles appartiennent les fichiers de test. Les sept dernières colonnes montrent les performances du modèle acoustique de base et des six différents modèles compacts. La dernière ligne indique le pourcentage de réduction du nombre de paramètres par rapport à celui du système de base.

	Baseline	40	100	150	200	250	300
RTM	35,50	36,01	35,73	35,72	35,68	35,55	35,55
RFI	25,72	26,57	26,22	25,83	25,71	25,36	25,51
INFO	25,85	28,31	27,74	27,64	27,11	27,13	27,09
CLASSIQUE	21,73	22,63	22,11	22,29	21,85	22,19	22,20
CULTURE	34,12	36,82	36,09	35,69	35,37	35,17	35,18
TEM global	28,87 %	30,45 %	29,98 %	29,73 %	29,51 %	29,45 %	29,46 %
Réduction de la taille du modèle	-	92,1 %	83,8 %	76,9 %	70,9 %	63,0 %	57,1 %

TABLE 4.1 – *Performances du modèle acoustique compact en fonction du rang de la matrice* ***U****.*

Nous remarquons que, pour les fichiers de test appartenant aux radios RTM et RFI, nous conservons les performances de base, alors que nous avons une perte négligeable pour les autres. Cette perte négligeable est accompagnée d'une très grande réduction du nombre des paramètres du modèle acoustique. La figure 4.3 montre l'évolution des performances des modèles compacts, en fonction de leurs tailles par rapport au modèle de base. Nous remarquons que, même avec une réduction de plus 90 % du nombre de paramètres, la dégradation des performances reste très limitée.

Pour trouver le nombre de gaussiennes optimal pour le GMM-UBM utilisé dans notre modélisation acoustique, nous avons fixé le rang de la matrice **U** à 150 et nous avons fait varier le nombre des gaussiennes de 400 à 1 000 (avec un pas de 100). Les résultats obtenus sont présentés dans le tableau 4.2. Le tableau montre que le taux d'erreur mot stagne rapidement à partir de 600 gaussiennes, ce qui signifie que ce nombre est suffisant pour la modélisation de l'espace acoustique.

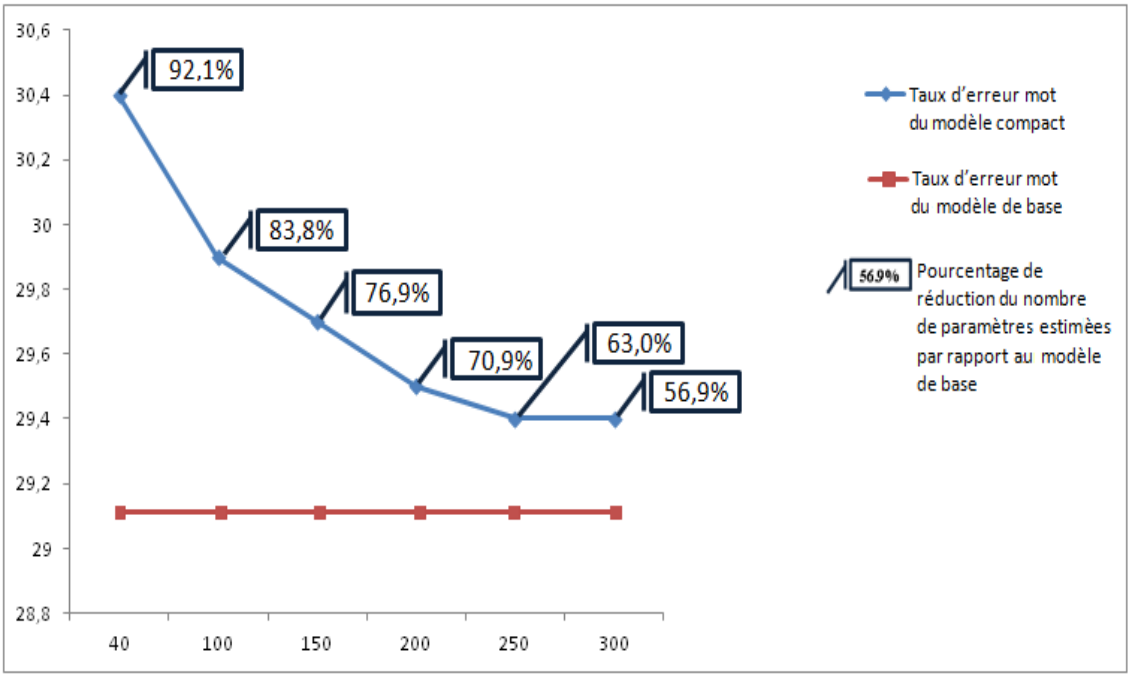

FIGURE 4.3 – *Performance du modèle compact en fonction de sa taille par rapport aux modèles de base.*

4.4.2.1 Modèle à base d'analyse factorielle vs modèle semi-contenu HMM

Dans l'approche SCHMM, les modèles partagent un dictionnaire commun de gaussiennes et les états sont caractérisés par un vecteur de poids généralement estimé par maximisation de la vraisemblance. Cette mutualisation de paramètres permet de réduire de façon très significative l'espace mémoire requis par le stockage des modèles acoustiques.

Afin de comparer les deux approches, nous avons construit deux modèles en utilisant respectivement la modélisation d'analyse factorielle et la modélisation SCHMM. Les deux modèles possèdent le même nombre de paramètres correspondant à une réduction de 89 % par rapport au modèle de base. Les trois dernières colonnes du tableau 4.3 montrent respectivement les performances du modèle de base, du modèle SCHMM et du modèle compact que nous proposons. Nous observons que notre modèle conserve de bonnes performances comparativement au modèle SCHMM (10,88 % de décalage).

4.4.3 Adaptation du modèle compact

La modélisation acoustique proposée ici décompose les paramètres des modèles acoustiques en cinq sous-ensembles de paramètres :

$$\pi = \{\pi_s\}, A = \{a_{s,s'}\}, UBM, \mathbf{U}, X = \{x_s\}, W = \{w_s\}$$

	Baseline	400	500	600	700	800	1000
RTM	35,50	36,26	36,11	35,72	35,72	35,35	35,55
RFI	25,70	26,91	26,91	25,83	25,69	25,43	25,50
INFO	25,80	28,51	28,20	27,64	27,33	27,17	27,10
CLASSIQUE	21,70	23,01	22,20	22,29	21,85	21,78	21,95
CULTURE	34,00	36,80	35,51	35,69	35,21	35,10	35,15
Moyenne	28,8 %	31,07 %	30,31%	29,73 %	29,61 %	29,29 %	29,31 %
Réduction de la taille du modèle	-	80,9 %	77,4 %	76,9 %	70,3 %	66,8 %	59,8 %

TABLE 4.2 – *Évolution des performances du modèle acoustique compact en fonction du nombre du gaussiennes du GMM-UBM.*

	Baseline	SCHMM	Modèle compact
RTM	35,51	44,31	35,81
RFI	25,72	38,36	25,77
INFO	25,97	39,60	28,37
CLASSIQUE	21,70	34,51	22,44
Moyenne	28,06 %	39,97 %	29,09 %
Réduction de la taille du modèle	-	89 %	89 %
Taille absolue	16 821 312	1 850 344	1 850 344

TABLE 4.3 – *Résultats du modèle standard, du système compact SCHMM et du SGMM, en termes de taux d'erreur mot (%).*

Où s indique l'état, et π et A correspondent aux probabilités initiales des états et les probabilités de transitions. UBM est l'ensemble des paramètres du modèle du monde. X est l'ensemble des vecteurs spécifiques aux états s. W est l'ensemble des vecteurs des poids des gaussiennes spécifiques aux état s (voire figure 4.4).

Cette décomposition du modèle acoustique apporte de la flexibilité dans le processus d'adaptation. En effet, nous pouvons adapter ou ré-estimer chaque sous-ensemble de paramètres indépendamment des autres. Nous pouvons adapter chaque sous-ensemble de paramètres au locuteur ou à une nouvelle tâche ayant des caractéristiques acoustiques différentes de celles des données d'apprentissage. Dans la suite, nous présentons les différentes méthodes d'adaptation proposées.

4.4.3.1 Adaptation de l'UBM

Nous rappelons que l'UBM est un mélange de gaussiennes modélisant la génération de vecteurs d'observation provenant d'une multitude du locuteurs et de conditions acoustiques. L'estimation des paramètres de ce modèle est réalisée en utilisant l'algorithme EM. L'estimation ou l'adaptation des paramètres de l'UBM ne nécessite pas de données transcrites. L'adaptation de ce sous-ensemble à une tâche cible permet d'adap-

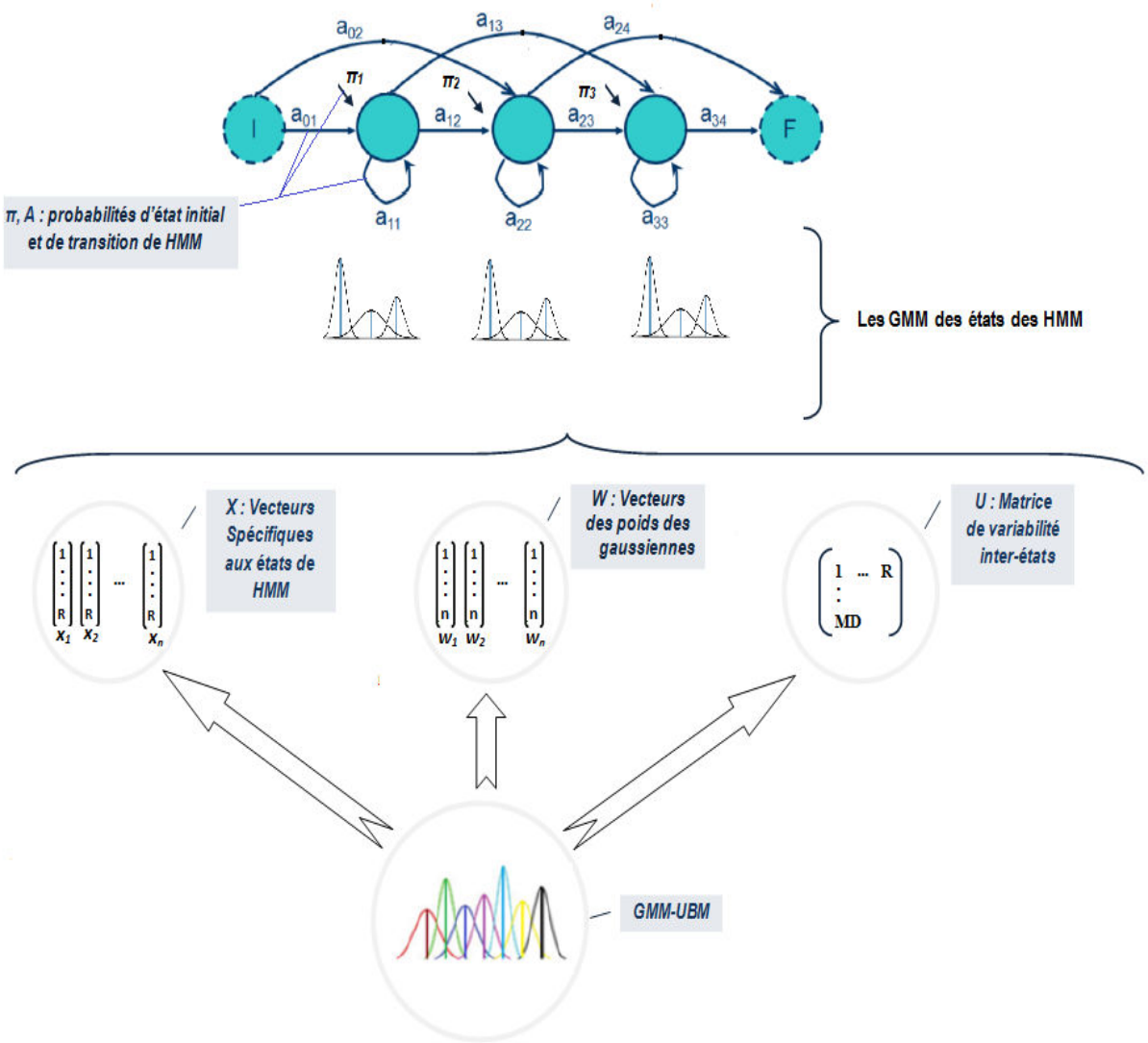

FIGURE 4.4 – *Différents sous-ensembles des paramètres du nouveau modèle acoustique.*

ter le modèle acoustique dans sa globalité à la tâche en question sans avoir besoin de données transcrites.

Dans l'expérience suivante, nous voulons adapter notre modèle acoustique sur des données enregistrées dans des conditions acoustiques bruitées (conférences, colloques, et assemblées générales). Cependant, à cause de l'absence de transcription des fichiers audio, ne nous pouvons pas adapter les MMC. Pour cela, nous adaptons les paramètres de l'UBM sur le corpus de parole non transcrit (50 heures). Ensuite, en utilisant l'UBM obtenu, nous estimons les paramètres de la matrice **U** (rang 150), du vecteur spécifique $\mathbf{x}_s$ et les vecteurs de poids w_s sur le corpus d'apprentissage (ESTER). Le tableau 4.4 montre que le modèle compact adapté nous permet un gain de 5,93 % en absolu en

termes de TEM par rapport au modèle compact générique.

	Baseline	Modèle compact générique	Modèle compact adapté
Test1 (1 heure)	60,35	61,01	53,73
Test2 (1 heure)	49,70	50,11	44,21
Test3 (1 heure)	37,25	38,41	01,51
Total (3 heures)	49,10 %	49,08 %	43,15 %

TABLE 4.4 – *TEM obtenus par la baseline, le modèle compact générique et le modèle compact adapté (GMM-UBM).*

Dans le même ordre d'idée, nous avons essayé d'adapter les paramètres de l'UBM aux différentes radios présentes dans notre corpus d'apprentissage. Les paramètres de l'UBM sont adaptés à chaque fois sur 20 heures sur les radios suivantes : RTM, RFI, et France-INFO. Dans la phase de test, chaque fichier de test est décodé par le modèle correspondant. Le tableau 4.5 montre que le modèle compact adapté nous permet un gain de 1 % absolu en termes de TEM par rapport au modèle compact générique.

	Baseline	Modèle compact générique	Modèle compact adapté
RTM (2 heures)	35,51	35,72	35,81
RFI (1 heure 30 mn)	25,72	25,83	25,07
INFO (2 heures)	25,97	27,64	25,37
Total (6 heures 30 mn)	29,37%	30,08%	29,08%

TABLE 4.5 – *TEM obtenus par le système baseline, le modèle compact générique et le modèle compact adapté.*

4.4.3.2 Adaptation des vecteurs d'états

Le deuxième sous-ensemble que nous pouvons adapter est le sous-ensemble X des vecteurs spécifiques aux états. Dans le modèle adapté, la ré-estimation est faite sur les données d'adaptation choisies. Dans nos expériences, nous avons testé l'adaptation des vecteurs X sur trois types de données : radios, genres (homme / femme) et locuteurs. Dans les deux premiers cas (radios et genres) nous n'avons pas observé d'amélioration des performances des systèmes. L'adaptation du modèle aux locuteurs nous permet de gagner 1,02 % absolu par rapport au modèle compact générique (voir tableau 4.6). L'adaptation est faite sur les données des quatre locuteurs les plus fréquents dans notre corpus d'apprentissage.

4.5 Conclusion

Dans ce chapitre nous avons exposé une nouvelle approche de modélisation acoustique compacte. L'approche proposée est inspirée de la technique *analyse factorielle* utilisée dans le domaine de la vérification des locuteurs. Ce modèle possède une certaine

	Baseline	Modèle compact générique	Modèle compact adapté
locuteur1	28.52	29.11	28.33
locuteur2	30.26	30.30	29.10
locuteur3	22.95	23.20	22.23
locuteur4	27.23	27.30	26.15
Total	27.38%	27.64%	26.62%

TABLE 4.6 – *TEM obtenus par le modèle baseline, le modèle compact générique et le modèle compact adapté (x_s)*

similarité avec la modélisation Subspace GMM développée dans (Povey et al., 2010). Dans nos expériences, nous avons montré que le modèle obtenu permet une réduction de 92 % de la taille des modèles, tout en maintenant des performances similaires à celles du modèle de base.

La modélisation acoustique proposée ici permet de décomposer l'ensemble des paramètres des modèles acoustiques en sous-ensembles de paramètres indépendants. Cela donne une grande flexibilité pour les éventuelles adaptations : locuteurs, nouvelles tâches, etc. Dans nos expériences, nous avons montré deux possibilités d'adaptation. La première est l'adaptation de l'UBM sur des données non transcrites. Dans cette expérience nous avons utilisé des données enregistrées dans des conditions acoustiques bruitées (conférences, colloques, et assemblées générales). Mais, à cause de l'absence de transcription des fichiers audio, les MMC ne peuvent être adaptés. Pour cela, nous adaptons les paramètres de l'UBM sur le corpus de parole non transcrite (50 heures). Cette adaptation a apporté un gain absolu de 5,93 %.

La deuxième proposition d'adaptation concerne les vecteurs d'états x_s. Dans cette expérience, nous avons testé l'adaptation des vecteurs X sur trois types de données : radios, genres (homme / femme) et locuteurs. Dans les deux premiers cas (radios et genres) nous n'avons pas observé d'amélioration. L'adaptation du modèle aux locuteurs a permis un gain de 1.02 % en absolu par rapport au modèle compact générique.

Chapitre 5

Analyse factorielle pour une représentation vectorielle des états des Modèles de Markov Cachés

Sommaire

5.1 Introduction

Dans les applications liées au traitement automatique de la parole (reconnaissance automatique de la parole, synthèse de la parole, système d'indexation audio, phonétique clinique, etc.) le signal acoustique est modélisé par un ensemble d'unités acoustiques. Dans la littérature, plusieurs unités acoustiques ont été proposées pour la segmentation et la modélisation de la parole : les phones [1], les allophones [2], les diphones [3] et les syllabes [4]. L'unité acoustique la plus utilisée est le phonème qui représente le découpage des mots en sous-unités acoustiques. Le phonème est considéré comme étant la plus petite unité distinctive, fonctionnelle et pertinente.

Depuis les premières applications dans le domaine du traitement de la parole, le probléme du calcul de la similitude entre phonèmes (ou allophones) a été posée comme sujet de recherche par la communauté scientifique. En reconnaissance de la parole, cette mesure est utilisée principalement dans la procédure de *partages d'états*. Cette approche, proposée par Young (Young, 1992), consiste à associer la même fonction de densité de probabilité (*probability density function* : pdf) aux états des phonèmes contextuels acoustiquement proches. Le partages d'états est une procédure incontournable dans le processus de modélisation acoustique afin d'avoir un équilibre entre la quantité de données d'apprentissage (pour les phonèmes contextuels) et le nombre de paramètres à estimer dans les modèles acoustiques. Elle s'appuie sur la définition d'une distance entre les états des MMC (mesure de similitude). En supposant que chaque phonème est modélisé par une seule gaussienne, les auteurs dans (Young et Woodland, 1993) ont utilisé la similitude entre phonèmes au moyen de la divergence de Kullback-Leibler [5]. Pour calculer la distance entre deux mélanges de gaussiennes, une expression alternative de la similitude basée sur la distance de Bhattacharyya [6] a été proposée dans (Mak et Barnard, 1996). La distance de Bhattacharyya permet de mesurer la distance théorique entre deux distributions gaussiennes.

Dans le cadre de l'identification automatique de la langue, les auteurs dans (de Mareüil et al., 2000) décrivent un algorithme de classification automatique permettant d'aboutir à un ensemble d'unités acoustiques communes à différentes langues. Chaque phonème de chaque langue est représenté par un MMC. La classification utilise une mesure de similitude entre phonèmes, calculée en utilisant les vecteurs acoustiques (coefficients cepstraux) et les MMC correspondants. Cette classification a été utilisée dans le cadre de l'identification automatique de onze langues. La classification des phonèmes de plusieurs langues est utilisée dans d'autres domaines comme en synthèse vocale multilingue et en phonétique descriptive, didactique ou corrective. La classification de phonèmes peut servir également à l'analyse de variation de la prononciation. Dans (Liu

1. Les phones : unités sous-phonèmes, qui, fusionnées entre elles, permettent d'obtenir des unités plus longues.
2. Les allophones : différentes réalisations sonores possibles d'un phonème.
3. Les diphones : unités acoustiques qui commencent au milieu de la zone stable d'un phonème et se terminent au milieu de la zone stable du phonème suivant.
4. Les syllabes : permettent d'incorporer les phénomènes co-articulatoires.
5. http ://www.aiaccess.net/french/Glossaires/GlosMod/f_gm_kullbakh̀tm
6. http ://www.cse.yorku.ca

et Fung, 2005), les auteurs s'appuient sur la classification phonétique pour quantifier la proximité entre les phonèmes et les allophones d'une même langue afin de définir des classes de prononciation.

Dans ce travail, nous proposons une nouvelle vision de la classification des phonèmes. Au contraire des travaux mentionnés auparavant, les phonèmes sont caractérisés par des vecteurs estimés par l'approche d'analyse factorielle (Kenny et al., 2005b). Un intérêt indéniable de cette représentation vectorielle est la possibilité de traiter les états par des techniques d'analyse de données, telles que l'analyse factorielle des correspondances. Le fait de représenter les états par des vecteurs permet aussi de mesurer efficacement la similarité entre eux, en utilisant des distances adaptées (distance euclidienne ou distance de Mahalanobis). Il est important de noter que cette tâche (mesure de similarité) était très complexe voire très approximative lorsque les états n'avaient pour représentants que les GMM. Nous proposons ici l'utilisation de cette représentation vectorielle dans la modélisation acoustique pour la reconnaissance de la parole. Nous nous appuyons sur les vecteurs pour réaliser la procédure de regroupement d'états des MMC. Nous appelons les vecteurs représentatifs des états des MMC *facteurs d'états*.

Dans la section 5.2 nous exposons la méthode d'estimation des facteurs d'états. Nous détaillons, dans la section 5.3, les différentes étapes de la procédure de regroupement des états. Nous exposons ensuite les résultats obtenus sur la langue française. Dans la section 5.4, nous montrons l'intérêt de l'application de cette méthode dans la modélisation acoustique pour les langues peu dotées en ressources. Finalement, nous exposons dans la section 5.5 quelques interprétations graphiques fondées sur les *facteurs d'états*.

5.2 Analyse factorielle pour une représentation vectorielle d'états de MMC

Dans la modélisation acoustique s'appuyant sur l'analyse factorielle (chapitre 4), tous les GMM associés aux états des MMC sont dérivés à partir d'un seul modèle générique appelé *modèle du monde* (GMM-UBM). C'est un mélange de gaussiennes modélisant tout l'espace acoustique de la parole. Les moyennes et les poids des états sont obtenus à partir de l'UBM, alors que les variances restent inchangées par rapport à l'UBM. Soit $\mathbf{m}$ le super-vecteur du modèle du monde obtenu par la concaténation des moyennes de ses gaussiennes. Le super-vecteur $\mathbf{m}_e$ de l'état e est obtenu par l'équation suivante :

$$\mathbf{m}_e = \mathbf{m} + \mathbf{U}\mathbf{x}_e \tag{5.1}$$

Dans ce modèle, la matrice $\mathbf{U}$, de faible dimension R, représente le sous-espace de la variabilité inter-états. Elle est estimée sur toutes les données de tous les états des MMC.

$\mathbf{x}_e$ est un vecteur de dimension R estimé sur les données propres à l'état e. C'est un vecteur caractéristique de l'état que nous appelons *facteur d'état*. L'algorithme d'estimation de $\mathbf{U}$ et des $\mathbf{x}_e$ est décrit en détail dans le chapitre 3.5.2.

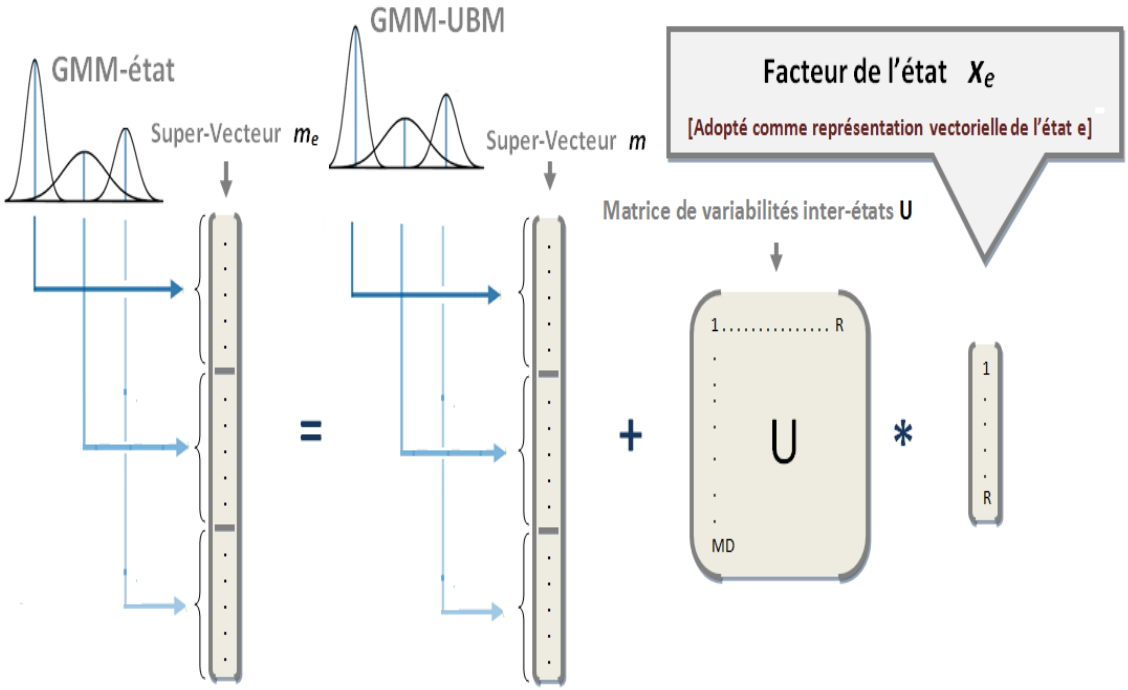

FIGURE 5.1 – *Les facteurs d'états obtenus à partir du modèle d'analyse factorielle.*

Dans le chapitre 4, nous avons montré que les modèles acoustiques compacts obtenus avec le modèle de l'équation 5.1 ont des performances similaires à celles du modèle standard. Ces résultats prouvent que les facteurs d'états $\mathbf{x}_e$, avec leur nombre très limité de paramètres, caractérisent bien les états des MMC. En s'appuyant sur ce résultat, nous proposons d'adopter ces vecteurs comme une représentation vectorielle des états des MMC (figure 5.1). Les facteurs d'états simplifient le calcul des distances entre les états des MMC. En effet, nous remplaçons les mélanges des gaussiennes par un simple vecteur de faible dimension (maximum 100 paramètres). Les distances utilisées généralement dans la classification phonétique sont remplacées par un simple calcul de distance entre vecteurs. Avec la représentation vectorielle, la classification phonétique devient un problème de classification classique dans un espace R^d. Par ailleurs, cette représentation vectorielle permettra d'exploiter les résultats scientifiques obtenus au cours de plusieurs années de recherche dans le domaine de l'analyse de données. Ces résultats peuvent servir dans l'analyse de la variabilité acoustique et de la variation de l'articulation phonétique. Également, les facteurs d'états peuvent être utilisés dans d'autres applications comme la phonétique clinique, la détection de dialectes ou l'identification automatique de la langue.

5.3 Procédure de regroupement d'états basée sur des facteurs d'états

Dans la littérature, le regroupement est souvent réalisé par un algorithme de classification utilisant les arbres de décision (Reichl et Chou, 1998). Plusieurs inconvénients sont inhérents à l'utilisation de cette méthode : elle nécessite notamment des connaissances linguistiques (pour construire le jeu de questions), qui peuvent ne pas être disponibles pour certaines langues. De plus, le temps de calcul est long à cause de l'évaluation des vraisemblances pour chaque question à chaque noeud de la structure de l'arbre. Comme alternative, nous proposons d'utiliser les facteurs d'états dans le regroupement des états des MMC. Avec les facteurs d'états, le calcul de la probabilité est remplacé par un simple calcul de distance entre vecteurs et le regroupement des états peut être formulé comme un problème de classification classique dans R^d. Cette méthode ne nécessite plus de connaissances phonétiques ou linguistiques, ce qui nous permet de contourner le problème d'insuffisance ou d'absence de ce type d'information pour les langues peu dotées. Dans la suite, nous exposons les différentes étapes de réalisation du partage d'états en utilisant les facteurs d'états.

5.3.1 Procédure de partage d'états des MMC basée sur les facteurs d'états

Nous présentons dans cette section les étapes nécessaires pour réaliser le regroupement d'états en utilisant les facteurs d'états.

- *Étape 1* : utilisation d'un système de reconnaissance de la parole existant pour réaliser la segmentation par rapport aux états (alignement forcé). Dans cette étape, nous utilisons l'algorithme de Viterbi. Dans cette segmentation, chaque phonème indépendant du contexte est représenté par trois états.

- *Étape 2* : contextualisation des phonèmes dans la segmentation obtenue dans l'étape 1. En effet, nous associons à chaque phonème, le phonème précédent et le phonème suivant. L'objectif de cette étape est de trouver tous les phonèmes contextuels dans notre corpus d'apprentissage.

- *Étape 3* : estimation des facteurs d'états $\mathbf{x}_s$ pour chaque état avec l'équation 5.1. Une bonne estimation des $\mathbf{x}_s$ nécessite un nombre minimum de trames pour chaque état. Pour cela nous ignorons les états qui sont rarement observés dans le corpus d'apprentissage et nous ne calculons les facteurs d'états que pour les états qui ont suffisamment de trames (dans nos expériences, nous n'avons traité que les états ayant plus de 50 trames).

- *Étape 4* : utilisation des facteurs d'états obtenus précédemment pour classifier les états en utilisant l'algorithme de classification non-supervisée *k*-means[7]. Cette

7. L'algorithme *k*-means permet une classification non-hiérarchique en minimisant la variance intra-classe basée sur la distance Euclidienne ; cette procédure est un moyen simple pour classifier un ensemble

classification permet de regrouper les états dépendants du contexte qui sont acoustiquement proches. Nous appelons les classes obtenues, *classe-états*.

– *Étape 5* : modélisation des états appartenant à la même classe par un seul GMM appelé *GMM-classe-état*. Ces GMM sont dérivés à partir du GMM-UBM en utilisant une adaptation par maximum *a posteriori* (MAP) au moyen des données appartenant à chaque classe-état (obtenu dans l'étape 4).

– *Étape 6* : affectation des états non classifiés dans l'étape 4 (qui ont moins de 50 trames) à la classe la plus proche. Pour ce faire, nous calculons la vraisemblance des trames de chaque état non classifié sur les GMM-classe-états obtenus dans l'étape 5. Ensuite, nous assemblons ces trames avec les trames de la classe la plus vraisemblable.

– *Étape 7* : adaptation de chaque ancienne GMM-classe-état obtenue dans l'étape 5 sur les données de la classe-état correspondante incluant les données des états nouvellement classifiés.

– *Étape 8* : construction des MMC en utilisant les GMM-classe-états obtenues dans l'étape précédente.

– *Étape 9* : amélioration des performances de notre modèle an appliquant la procédure récursive standard de ré-alignement/ré-estimation des paramètres.

La figure 5.2 présente les différentes étapes de cette procédure.

5.3.2 Résultats expérimentaux

Dans ces expériences, nous avons utilisé le système SPEERAL avec un processus de transcription en deux passes (voir section 4.4.1). Les nouveaux modèles acoustiques sont estimés et évalués sur le corpus d'évaluation ESTER (décrit dans la section 4.4.1). Les modèles acoustiques de base sont des MMC gauche-droite avec 13 316 phonèmes contextuelles. Pour modéliser tous ces phonèmes contextuels indépendamment, nous avons besoin de 39 948 états dans les MMC. Ce nombre est réduit à 5 050 états par l'application du regroupement d'états fondé sur des arbres de décision. Cette approche utilise des questions linguistiques pour construire les arbres de décision.

Dans une première expérience, nous avons cherché empiriquement le nombre optimal de coefficients dans les facteurs d'états. Nous rappelons que nous nous appuyons sur ces facteurs pour réaliser la procédure du partage d'états des MMC. Nous avons fait varier le nombre des paramètres des facteurs d'états $\mathbf{x}_e$ de 20 à 140 (avec un pas de 20). Nous nous sommes appuyés systématiquement sur ces vecteurs pour construire

de données dans un certain nombre de classes (noté k) préalablement fixé.

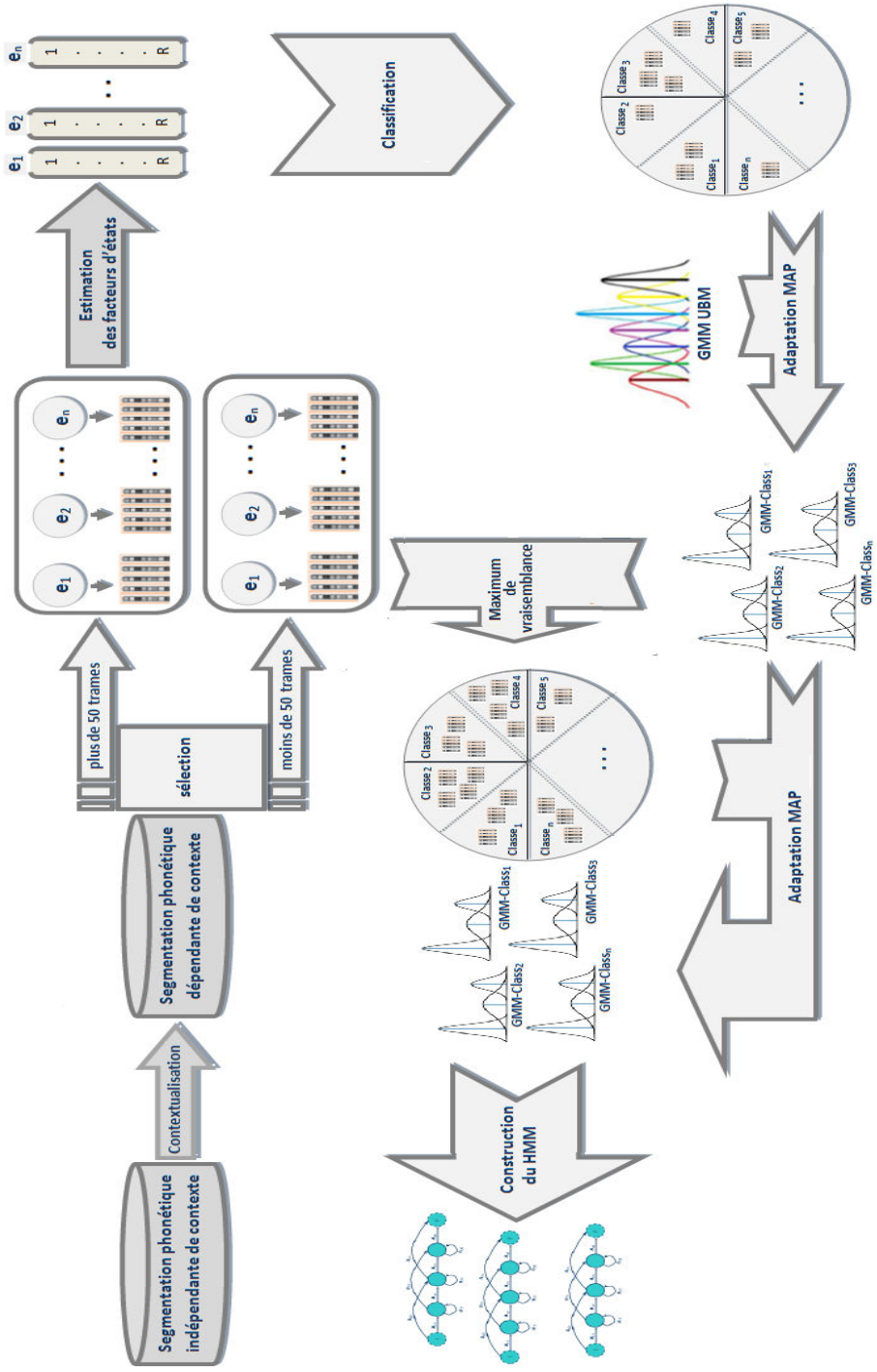

FIGURE 5.2 – *Les différentes étapes de modélisation acoustique s'appuyant sur les facteurs d'états.*

des modèles acoustiques contextuels (de 3 500 états). Le tableau 5.1 montre les performances de ces différents modèles en termes de taux d'erreur mot.

	F.Inter	RFI	TVME	AFRICA	Total
Baseline (5 050)	30,88	16,62	25,74	29,42	27,50
20	31,63	18,11	26,58	30,25	28,57
40	31,56	17,31	25,91	29,43	28,11
60	31,85	17,03	25,58	29,15	27,94
80	31,87	16,89	25,23	29,30	27,74
100	31,77	17,13	25,23	29,05	27,73
120	31,81	17,13	25,19	29,11	27,86
140	31,81	17,25	25,10	29,07	27,83
160	31,71	17,37	25,30	29,10	28,03

TABLE 5.1 – *Influence de la taille de facteurs d'états sur la performance des modèles.*

Nous remarquons que les taux d'erreur mot des différents modèles sont très proches du modèle *Baseline*. Ces résultats montrent que des facteurs d'états avec 40 paramètres nous permettent de bien classifier les états des MMC. Ceci confirme la justesse du choix de ces facteurs comme des vecteurs caractéristiques d'états.

Dans une deuxième expérience, nous avons fixé le nombre de paramètres des facteurs d'états à 60. Nous avons testé différents nombres d'états afin de trouver la meilleure taille du modèle acoustique (en termes de nombre d'états). Les performances des différents modèles sont comparées avec le modèle de base ayant 5 050 états. Le tableau 5.2 expose les performances des différents modèles. La première colonne montre le nombre d'états par modèle. La dernière colonne montre le gain obtenu par rapport au modèle de base. Le meilleur gain est de 0,89 % obtenu avec un modèle de 6 000 états.

	F.Inter	RFI	TVME	AFRICA	Total	Gain absolu
Modèle de base (5 050)	30,88	16,62	25,74	29,42	27,50	-
2 500	31,97	17,56	26,05	29,66	28,43	-0,93
3 500	31,55	17,19	26,12	29,00	27,94	-0,44
4 500	31,15	16,89	25,73	29,57	27,59	-0,09
5 000	31,00	16,29	25,53	29,49	27,33	0,17
5 500	30,70	16,13	25,30	29,28	26,91	0,59
6 000	30,10	16,09	24,90	28,89	26,61	**0,89**
6 250	30,20	16,39	24,70	28,70	26,65	0,85
6 500	30,17	16,38	25,17	28,81	26,74	0,76
6 750	30,17	16,45	25,07	28,81	26,69	0,81
7 000	30,30	16,70	25,17	28,50	26,87	0,63
8 000	33,63	16,75	25,10	28,75	27,09	0,41

TABLE 5.2 – *Résultats en TEM du modèle standard et des nouveaux modèles avec différentes tailles.*

Nous rappelons que les GMM d'états des MMC sont dérivés d'un seul modèle du

monde avec une adaptation MAP. Pour améliorer les performances, nous proposons de modéliser indépendamment chaque état en utilisant l'algorithme EM. Dans notre procédure de modélisation, une fois la classification terminée, nous estimons un GMM pour chaque classe d'états avec l'algorithme EM. Avec cette modélisation, nous augmentons le nombre de paramètres du modèle, ce qui améliore la performance du système de 0,95 % par rapport à l'ancienne modélisation. Cette amélioration nous permet un gain total de 1,84 % par rapport au modèle de base. Les résultats sont exposés dans le tableau 5.3.

	Modèle de base	Modèle MAP	Modèle EM
F.Inter	30,88	30,10	29,38
RFI	16,62	16,09	15,84
TVME	25,74	24,90	24,77
AFRICA	29,42	28,89	27,47
Total	27,50	26,61	25,66
Gain absolu	-	0,89	1,84

TABLE 5.3 – *Performances du modèle MAP et du modèle EM comparées avec le modèle de base.*

5.4 Modèle indépendant du contexte pour les langues peu dotées basé sur les facteurs d'états

5.4.1 Problématique

Parmi les 6 912 langues parlées dans le monde, seul un tout petit nombre d'entre-elles possède les ressources nécessaires pour implémenter des technologies issues du traitement du langage naturel. Les autres langues, catégorisées *peu dotées*, sont les langues minoritaires ou les langues venant de pays en voie de développement. Elles sont moins abordées par la communauté du traitement automatique de la langue naturelle. Une langue peu dotée est définie comme une langue qui ne possède pas encore beaucoup (en quantité et en qualité) de ressources linguistiques pour la construction d'un système de reconnaissance automatique de la parole, particulièrement dans un contexte d'apprentissage statistique où les données doivent être disponibles en grande quantité. Au cours des dernières années, les langues peu dotées ont attiré une attention croissante dans la communauté du Traitement Automatique de la Langue Naturelle (TALN). Plusieurs travaux de recherche, dans différentes disciplines, sont réalisés pour faire sortir ces langues de leur isolement technologique. Parmi ces travaux : l'informatisation des langues peu dotées (Berment, 2004), la construction automatique de corpus textuels pour langues minoritaires (Scannell, 2003) (Ghani et al., 2005) ou encore la modélisation de la langue et du lexique (Tachbelie et al., 2012).

Au niveau de la modélisation acoustique, plusieurs travaux sont proposés dans la littérature pour contourner les difficultés d'absence partielle ou totale de ressources

linguistiques et informatiques. La solution la plus utilisée aujourd'hui est le *bootstraping* (Osterholtz et al., 1992). Cette solution consiste à obtenir un tableau de correspondances phonétiques (*phone mapping*) entre une ou plusieurs langues sources et la langue cible (la langue peu dotée). Ensuite, nous dupliquons les modèles acoustiques des phonèmes de la langue source pour obtenir des modèles acoustiques en langue cible. Nous distinguons deux classes de méthodes pour réaliser cette solution : la première classe comprend les méthodes manuelles à base de connaissances linguistiques et phonétiques (Le et Besacier, 2009) (Schultz et Waibel, 1998). Elles consistent à chercher les couples de phonèmes source/cible les plus proches dans le tableau d'Alphabet Phonétique International [8](API) (voir figure 5.3). Ces méthodes nécessitent des connaissances acoustiques et phonétiques des deux langues source et cible. Dans une deuxième classe, nous trouvons les méthodes automatiques (Anderson et al., 1994; Constantinescu et Chollet, 1997). Ces méthodes utilisent un modèle de la langue source et un corpus vocal étiqueté de la langue cible pour calculer la matrice de confusion entre les phonèmes et trouver le tableau de correspondances phonétiques. Ces méthodes sont utilisées pour construire un modèle acoustique indépendant du contexte. Pour construire des modèles acoustiques contextuels, des travaux comme (Beulen et Ney, 1998; Singh et al., 1999) proposent d'utiliser la procédure classique s'appuyant sur les arbres de décision. Les questions nécessaires à la construction des arbres sont générées automatiquement. D'autres travaux proposent de combiner les modèles acoustiques (au niveau des phonèmes contextuels) de différentes langues sources afin d'obtenir un modèle contextuelle de la langue cible (Beyerlein, 1998; Schultz et Waibel, 2001).

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

FIGURE 5.3 – *API pour les consonnes et les voyelles (IPA, 1999).*

La plupart des travaux de modélisation acoustique contextuelle, cités auparavant,

8. L'alphabet phonétique international (API) est un alphabet utilisé pour la transcription phonétique des sons du langage parlé. Contrairement aux nombreuses autres méthodes de transcription qui se limitent à des familles de langues, l'API est prévu pour couvrir l'ensemble des langues du monde. Développé par des phonéticiens britanniques et français sous les auspices de l'Association phonétique internationale, il a été publié en 1888. Sa dernière révision date de 2005.

sont basés sur l'idée d'association entre une ou plusieurs langues sources et la langue cible. Les différentes applications montrent que l'association à base de connaissances linguistiques donne de meilleurs résultats que les méthodes automatiques basées sur le calcul de distance entre modèles. Mais ces solutions (à base de connaissances linguistiques) se retrouvent toujours face au problème de la couverture phonétique. De plus, elles nécessitent des connaissances linguistiques et phonétiques des deux langues source et cible, qui ne sont pas toujours disponible (en quantité et en qualité) pour la langue cible. Afin de contourner ces difficultés, nous proposons d'appliquer la méthode de modélisation acoustique contextuelle proposée dans la section 5.3. Dans la suite nous montrons les résultats obtenus sur la langue vietnamienne et la langue berbère.

5.4.2 Résultats expérimentaux

5.4.2.1 Vietnamien

La langue vietnamienne appartient au groupe *Viet-Muong*, qui est un membre de la branche *mon-khmer* et fait partie de la branche *austro-asiatique*[9]. Elle est parlée par environ 82 millions de personnes, principalement au Vietnam. La langue vietnamienne a été écrite avec un code *Siniform*, mais un système d'écriture fondé sur le latin a été introduit au 17^{eme} siècle. Ce système d'écriture est le plus utilisé aujourd'hui. Le vietnamien est une langue tonale qui possède six tons avec des caractères accentués pour les tons.

Le vietnamien possède 41 phonèmes, dont 23 consonnes, 13 voyelles simples, 3 diphtongues et 2 semi-voyelles, représentées par 29 lettres dans l'alphabet (N. Thi, 2006) (voir figure 5.4).

Voyelles	a ă â e ê i o ô ơ u ư y
Diphtongues	iê yê ia ya ươ ưa uô ua
Semi-voyelles	o u i y
Consonnes	b c d đ g h k l m n p q r s t v x
Combinaisons de consonnes	ch gh kh ng ngh nh ph th tr gi qu

FIGURE 5.4 – *Les consonnes et voyelles du vietnamien.*

Chaque syllabe vietnamienne se compose de 5 éléments : le ton, le premier son, le son intercalé, le son noyau et le dernier son (R. Beaufort, 2006). Suivant leurs positions

9. http ://www.ethnologue.com

dans les syllabes, les lettres jouent des rôles différents. Selon ce rôle, il y a 5 systèmes phonétiques : le système du premier son, le système du son intercalé, le système du son noyau, le système du dernier son et le système du ton. La structure de la syllabe est présentée à la Figure 5.5.

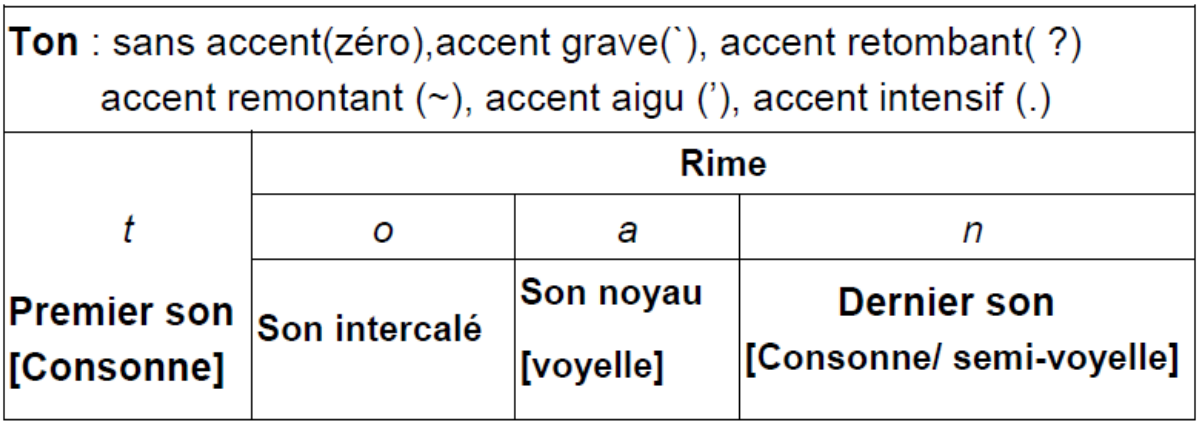

FIGURE 5.5 – *Structure de la syllabe vietnamienne.*

Corpus disponibles

Corpus de parole : dans nos expériences, nous avons utilisé une partie du corpus VNSPEECHCORPUS (Tran, 2003). Ce corpus contient 39 heures de parole enregistrées au centre MICA [10]. En 2005, il contenait 39 locuteurs, 19 femmes et 20 hommes, venant des régions nord, centre et sud du Vietnam. Nous utilisons uniquement les enregistrements de locuteurs d'une langue standard (nord du Vietnam). Au total, environ 9 heures de parole sont enregistrées, ce qui correspond à 18 locuteurs : 10 hommes et 8 femmes. Nous avons utilisé 7 heures pour l'apprentissage des modèles (8 hommes et 6 femmes) et 2 heures pour le corpus de test (2 hommes et 2 femmes).

Corpus de texte : le corpus de texte vietnamien, utilisé pour estimer le modèle de langage, est collecté exclusivement à partir du web et des journaux numériques. Ce corpus est constitué de 2,7 millions de phrases avec 45 millions de syllabes.

Modèle de base : Arbre de décision

Le modèle de phonème indépendant du contexte est obtenu avec le *bootstraping* (Osterholtz et al., 1992). Un tableau de correspondances phonétiques est construit entre les phonèmes vietnamiens et les phonèmes du français en utilisant l'API. Le modèle de phonème contextuel est obtenu par un arbre de décision basé sur des questions générées automatiquement (Singh et al., 1999). Nous avons construit plusieurs modèles acoustiques de différentes tailles afin de trouver le nombre optimal des paramètres du modèle. Nous avons fait varier les nombres d'états par MMC et le nombre de gaussiennes par état. Dans le tableau 5.4, nous exposons les résultats en termes de taux

10. http ://www.mica.edu.vn

d'erreur mot des différents modèles. Chaque colonne contient les nombres d'états dans les MMC. Les lignes contiennent les nombres de gaussiennes par état. Le modèle qui est composé de 600 états avec 64 gaussiennes atteint les meilleures performances (32,70 % de TEM). Ce modèle sera notre modèle de base.

	200 états	400 états	600 états	800 états	1 200 états
32 gaus.	33,11	33,51	34,70	34,90	36,10
64 gaus.	32,81	32,73	**32,70**	33,00	34,90
128 gaus.	33,83	33,80	34,40	35,30	36,40
256 gaus.	34,00	33,97	36,10	35,40	38,40

TABLE 5.4 – *Performances des modèles acoustiques contextuels de différentes tailles où le regroupement des états contextuels est basé sur un arbre de décision.*

Modèle contextuel du vietnamien basé sur les facteurs d'états

Dans ces expériences, nous appliquons la méthode de modélisation contextuelle proposée dans ce chapitre sur le vietnamien. Dans une première expérience, nous avons estimé quatre groupes de facteurs d'états qui ont respectivement 20, 40, 80 et 120 coefficients. Par la suite, nous utilisons chaque groupe de vecteurs pour regrouper les états dans 1000, 1200, 1400, 1600, 1800, 2000 et 2200 classes respectivement. Le nombre de gaussiennes par état est fixé à 128 gaussiennes. Dans le tableau 5.5, nous exposons le TEM obtenu par les différents modèles. Chaque ligne du tableau correspond au nombre d'états des modèles. Dans les colonnes nous exposons les tailles des facteurs d'états que nous avons utilisées dans la phase du regroupement.

	20 coef.	40 coef.	60 coef.	80 coef.	120 coef.
1 000 états	28,80	29,50	36,15	38,50	57,10
1 200 états	28,00	28,80	35,10	38,40	60,70
1 400 états	28,10	28,20	33,80	35,70	52,30
1 600 états	27,20	27,80	33,20	35,10	49,90
1 800 états	26,80	27,20	30,70	34,20	50,50
2 000 états	**26,60**	27,50	30,90	33,80	49,00
2 200 états	26,90	27,50	31,10	33,55	49,24

TABLE 5.5 – *Performances des modèles ayant différents nombres d'états dans les MMC, avec 128 gaussiennes par état.*

Le meilleur modèle donne un gain absolu de 6,10 % par rapport au modèle de base. Les résultats obtenus montrent que les modèles ayant le plus grand nombre d'états obtiennent de meilleurs résultats. En outre, nous observons qu'à partir de 1 800 états le gain devient stationnaire. Ces résultats montrent que, à cause de la quantité limitée de données d'apprentissage disponible pour chaque état, les meilleurs facteurs d'états sont ceux avec la plus faible dimension (20 coefficients). La figure 5.6 montre les performances des modèles contextuels, en fonction du nombre d'états dans les MMC et de la taille des facteurs d'états.

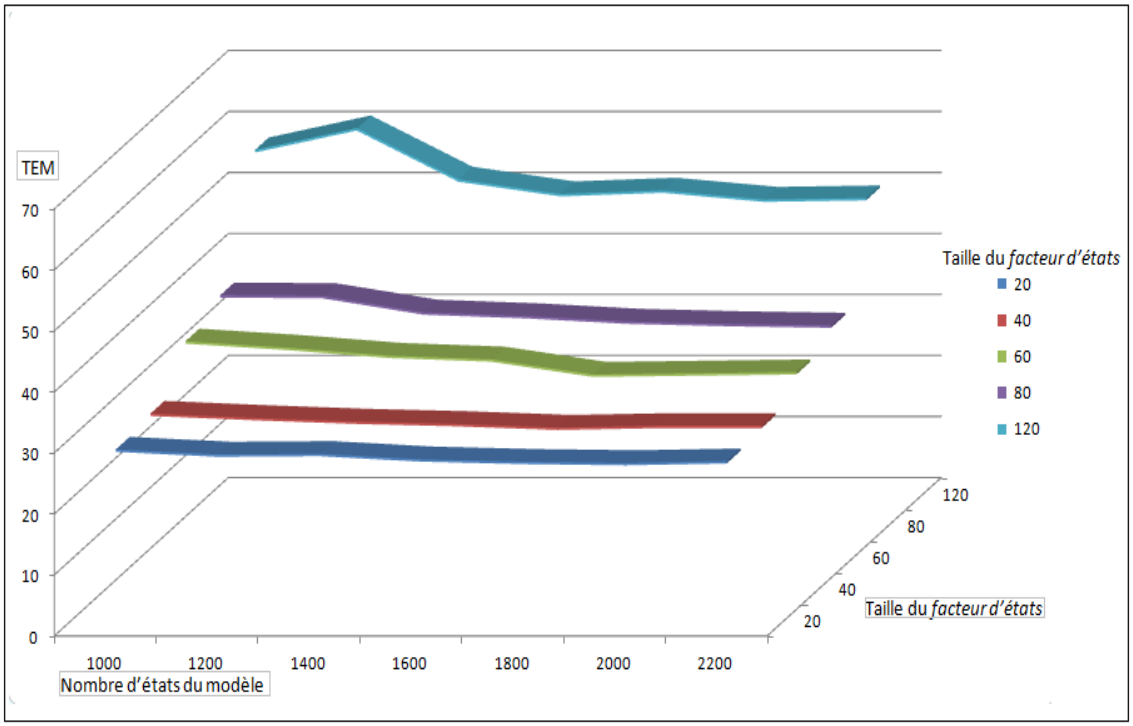

FIGURE 5.6 – *Performances du modèle triphone en fonction de sa taille et de la taille des facteurs d'états utilisés dans la procédure de regroupement.*

Dans une deuxième expérience, nous avons essayé de trouver le nombre optimal de gaussiennes par état. Le nombre de paramètres des facteurs d'états est fixé à 20. Nous fixons le nombre d'états à 1 600, 1 800 et 2 000. Pour chaque modèle, le nombre des gaussiennes varie de 64, 128 et 256 gaussiennes.

	64 gaus.	128 gaus.	256 gaus.
1 600 états	30,60	27,80	26,00
1 800 états	29,80	27,20	25,50
2 000 états	29,30	26,60	**24,90**

TABLE 5.6 – *Performances des modèles en fonction du nombre d'états et du nombre de gaussiennes.*

Le tableau 5.6 montre que le meilleur modèle est celui avec 2 000 états et 256 gaussiennes par état, avec un gain de 7,8 % absolu par rapport au modèle de base. Pour améliorer les performances du système, nous avons testé une approche de regroupement guidé. Nous rappelons que chaque phonème contextuel *ph* est modélisé par trois états. Dans une première étape, nous regroupons ensemble les facteurs d'états qui représentent les différents contextes du même phonème. Nous obtenons alors 38 classes. Dans une deuxième étape, nous utilisons les facteurs d'états pour classifier les états à l'intérieur de chaque classe-phonème *ph*. La figure 5.7 montre un exemple de classification guidée des facteurs d'états. Le nombre de gaussiennes par état est de 256 gaussiennes. Les résultats présentés dans le tableau 5.10 montrent que le regroupement

guidé nous permet un gain de 0,7 % absolu par rapport au regroupement global. Au final, nous obtenons un gain absolu de 8,50 % par rapport au modèle de base.

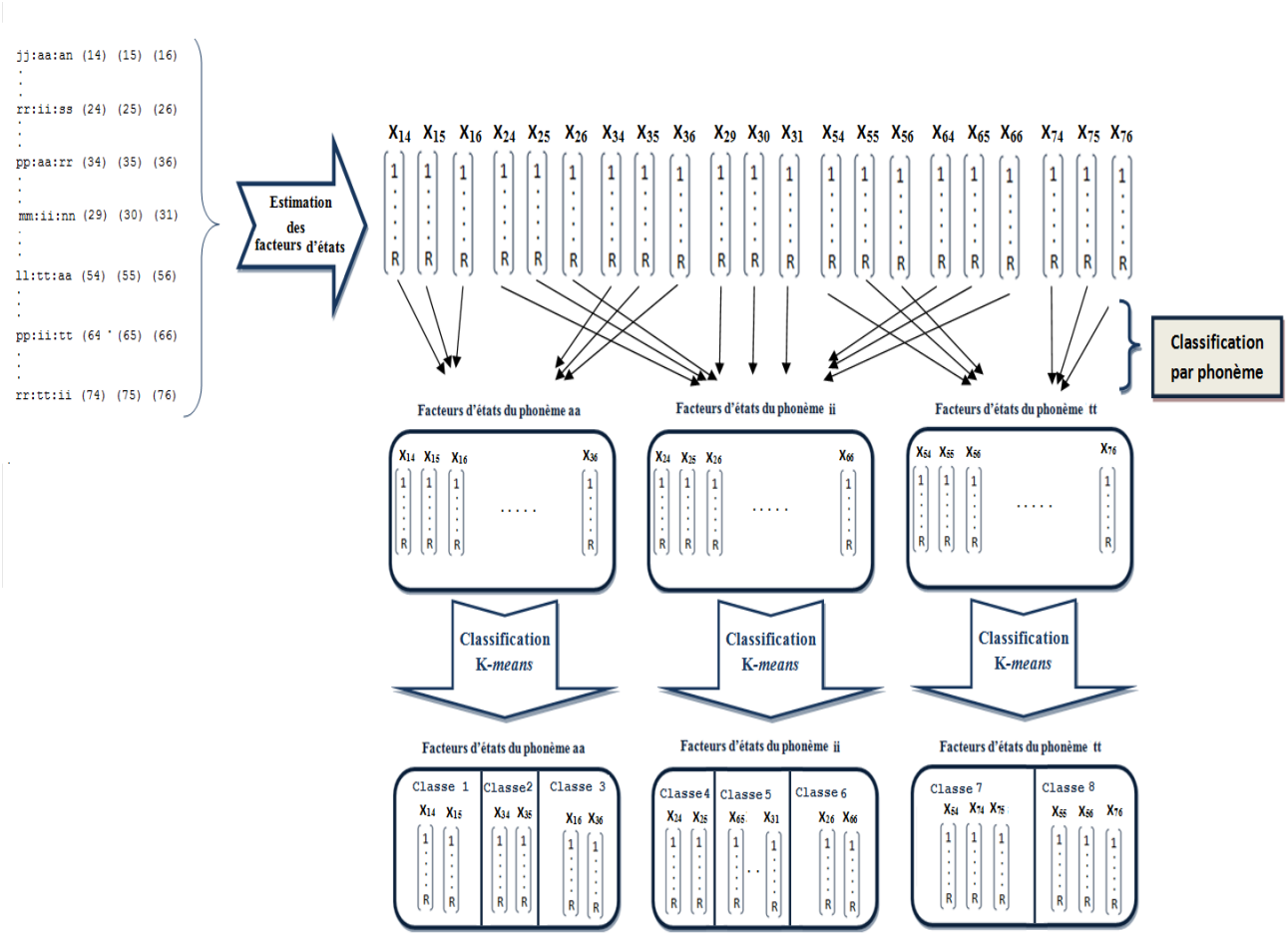

FIGURE 5.7 – *Exemple illustrant la classification guidée des facteurs d'états.*

	1 600 états	1 800 états	2 000 états	2 200 états	2 400 états
TEM	25,00	24,80	24,90	24,30	**24,20**

TABLE 5.7 – *Résultats de l'approche par regroupement guidé.*

5.4.2.2 Berbère

La langue berbère appartient à la famille des langues *chamito-sémitiques* (ou *afro-asiatiques*) (Chaker, 1984). Elle est parlée par environ 30 millions de personnes en Afrique du nord. Bien que le berbère soit une langue essentiellement de tradition

orale, les berbères possèdent, depuis au moins deux millénaires et demi, leur propre système d'écriture appelé *libyco-berbère* (*tifinagh* en berbère). Il s'agit d'un système alphabétique (consonantique) aux usages traditionnellement assez restreints (funéraires, symboliques et ludiques). Depuis le début du 20^{eme} siècle, l'écrit berbère utilise surtout le support de l'alphabet latin (avec diverses adaptations). Le berbère moderne dispose de 43 phonèmes : trois sont des voyelles, deux semi-voyelles et 38 consonnes. Ainsi, cette langue est considérée comme ayant un riche consonantisme et un vocalisme faible. La figure 5.8 montre le tableau du système phonologique berbère proposé par Chaker (Chaker, 1984). Ce tableau résume les points communs des systèmes consonantiques berbères dans leur ensemble, met en évidence les deux caractéristiques principales que partagent tous les systèmes consonantiques berbères, caractéristiques assez spécifiques sur le plan typologique : la gémination et la pharyngalisation.

	LAB.	ALV	POSTALV.	VÉL.	UVUL.	GLOTT.
OCCL.	b	t d		k g		
g.	bb	tt dd		kk gg	qq	
e.		ḍ				
g./e.		ṭṭ				
NAS.	m	n				
g.	mm	nn				
FRIC.	f	s z	ʃ ʒ	ɣ		h
g.	ff	ss zz	ʃʃ ʒʒ			hh
e.		ẓ				
g./e.		ẓẓ				
TRIL.		r				
g.		rr				
LAT.		l				
g.		ll				
APPR.	w			y		
g.[62]	gg^w			gg		

FIGURE 5.8 – *Système phonologique berbère proposé par (Chaker, 1984).*

Corpus disponibles

Corpus de parole : le corpus utilisé est composé de 2 261 fichiers audio correspondant à 19 heures. Il contient 56 locuteurs, 15 hommes et 41 femmes. Ce corpus couvre tous les phonèmes berbères. Nous avons divisé le corpus en deux parties : 14 heures pour le corpus d'apprentissage et 5 heures pour le corpus de test.

Corpus de texte : le corpus de texte berbère utilisé pour estimer le modèle de langage est collecté exclusivement à partir du web et des journaux numériques (le «Tamazight tura» et le «Ayamun»). Ces journaux utilisent le système d'écriture latine. Le corpus est constitué de 67 434 phrases, pour un total de 445 169 mots.

Modèle de base : Arbre de décision

Comme pour la langue vietnamienne, le modèle de phonèmes indépendants du contexte pour le berbère sont obtenu avec le *bootstraping*. Un tableau des correspondances phonétiques est construit entre les phonèmes berbères et les phonèmes du français en utilisant l'API. Le modèle de phonème contextuel est obtenu par un arbre de décision s'appuyant sur des questions générées automatiquement (Singh et al., 1999). Le taux d'erreur mot obtenu par le modèle de base sur les 5 heures de test est de 30,50 %.

Modèle contextuel du berbère basé sur les facteurs d'états

Pour construire un modèle berbère, nous avons suivi la même démarche que pour les expériences menées sur le vietnamien. Nous avons estimé quatre groupes de facteurs d'états ayant respectivement 20, 40, 80 et 100 paramètres. Par la suite, nous utilisons chaque groupe de ces vecteurs pour regrouper les états dans 1 200, 1 400, 1 600, 1 800, 2 000, 2 200, 2 400, 2 600 et 2 700 classes. Dans le tableau 5.8, nous présentons les performances des différents modèles. Le meilleur modèle permet un gain absolu de 5,20 % par rapport au modèle de base.

	20 coef.	40 coef.	60 coef.	80 coef.	100 coef.
1 200 états	27,30	26,90	30,50	38,40	49,20
1 400 états	27,10	26,30	30,80	35,70	48,50
1 600 états	27,20	26,40	31,20	35,10	48,70
1 800 états	26,60	26,20	30,40	34,20	48,30
2 000 états	26,70	26,00	29,80	33,80	47,30
2 200 états	26,30	25,80	29,10	33,55	46,50
2 400 états	26,30	**25,30**	29,70	33,55	46,20
2 600 états	27,10	25,90	31,10	30,50	46,10
2 700 états	26,90	26,10	31,10	30,55	45,70

TABLE 5.8 – *Performances des modèles ayant différents nombres d'états dans les MMC, avec 128 gaussiennes par état.*

Dans la deuxième expérience, nous fixons le nombre de paramètres des facteurs d'états à 40. Le nombre d'états varie de 2 200 à 2 700. Pour chaque modèle, nous augmentons le nombre des gaussiennes de 128 à 512. Le tableau 5.9 montre que le meilleur modèle (celui ayant 2 400 états avec 256 gaussiennes) nous permet un gain de 5,70 % par rapport au modèle de base.

Afin d'améliorer la performance du système, nous avons appliqué l'approche de regroupement de la même manière que sur le vietnamien. En premier lieu, nous avons regroupé les états contextuels selon le phonème. Ensuite, nous avons utilisé les facteurs

	128 gaus.	256 gaus.	512 gaus.
2 200 états	25,80	25,10	24,90
2 400 états	25,30	**24,10**	25,10
2 600 états	25,90	24,10	25,80
2 700 états	26,10	24,90	26,10

TABLE 5.9 – *Performances des modèles en fonction du nombre d'états et du nombre de gaussiennes. Le facteur d'états x_s est de 20 paramètres.*

d'états de taille 20 pour classifier les états de chaque groupe. Le nombre de gaussiennes par état est fixé à 256 gaussiennes. Les résultats présentés dans le tableau 5.10 montrent que ce regroupement guidé nous permet un gain de 1,2 % absolu par rapport au regroupement global et 6,40 % absolu par rapport au modèle de base.

	1 600 états	1 800 états	2 000 états	2 200 états	2 400 états
TEM	25,00	24,80	24,90	24,30	**24,20**

TABLE 5.10 – *Résultats de regroupement guidé pour des modèles ayant différents nombres d'états. Le nombre de gaussiennes par état est de 256 et les facteurs d'états sont de taille 20*

5.5 Analyse graphique basée sur les facteurs d'états

Dans cette section, nous exposons une simple méthode de visualisation graphique d'états des MMC basée sur les facteurs d'états. Pour projeter les facteurs d'états dans une dimension de visualisation de deux dimensions, nous utilisons la procédure mathématique d'Analyse en Composantes Principales (ACP).

L'ACP est une méthode vectorielle linéaire de réduction des dimensions de paramètres non supervisée, choisissant les directions dont la variance intra-groupe est la plus grande. Les données sont alors plus facilement visualisables sur moins de dimensions. L'ACP se calcule à partir de la matrice de covariance des données. Celle-ci est diagonalisée afin d'en extraire les valeurs et vecteurs propres. Les données sont projetées dans l'espace défini par les vecteurs propres. Les valeurs propres, classées dans l'ordre décroissant, correspondent dans l'espace d'arrivée au vecteur propre dont la direction maximise la variance.

Dans une première projection, nous visualisons les résultats de l'algorithme de classification d'états des MMC utilisé dans la section précédente. Nous avons choisi neuf classes d'états. Les composantes principales d'états sont estimées avec la procédure ACP. La figure 5.9 montre la projection dans un espace de deux dimensions. Chaque nuage de points représente les états d'une seule classe (considérés proches acoustiquement). Le centre de l'ellipse est la moyenne de la classe, autour de laquelle les états de la classe sont répartis. Les classes qui se recouvrent (exemple classe 6 et classe 8) sont les classes d'états similaires acoustiquement.

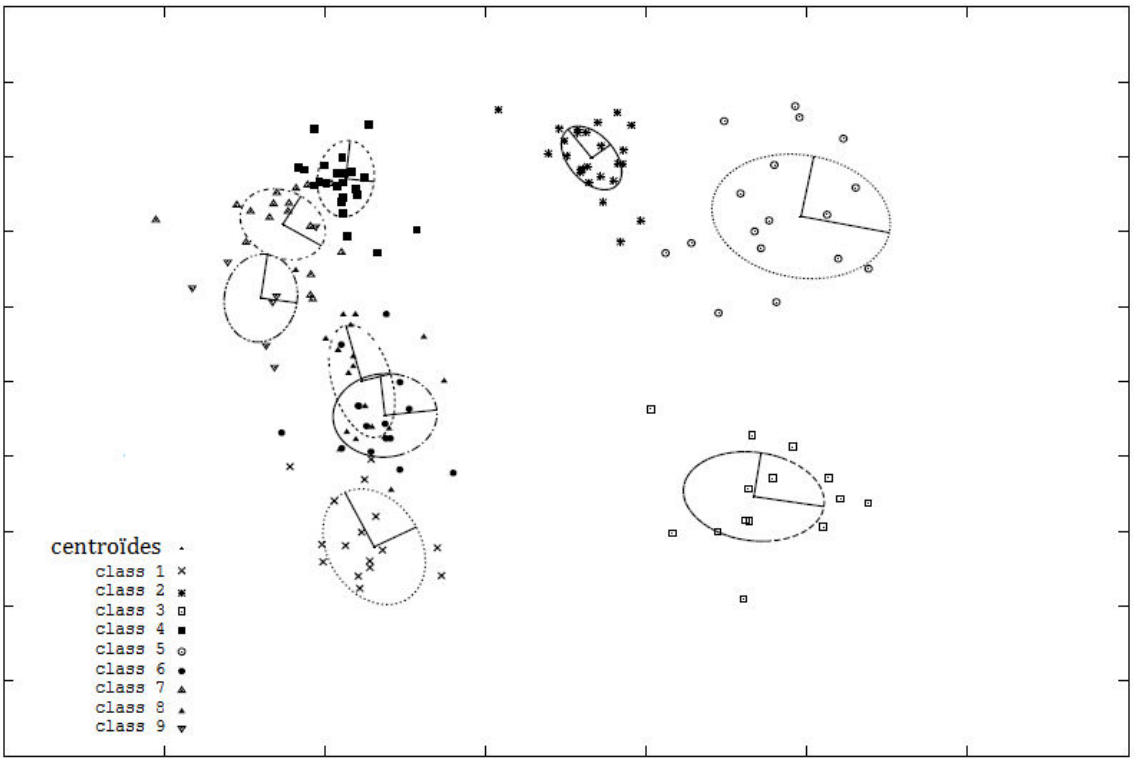

FIGURE 5.9 – *Projection des facteurs d'états appartenant aux neuf classes définies.*

Nous utilisons la même méthode d'estimation des facteurs d'états pour calculer des vecteurs représentatifs des phonèmes appelés *facteurs de phonèmes*. Le modèle de l'équation 5.1 est ré-écrit comme suit :

$$\mathbf{m}_{ph} = \mathbf{m} + \mathbf{U}\mathbf{x}_{ph} \tag{5.2}$$

Avec $\mathbf{m}$ le super-vecteur du modèle du monde, $\mathbf{m}_{ph}$ le super-vecteur du modèle du phoneme ph. La matrice $\mathbf{U}$ représente le sous-espace de la variabilité inter-phonème estimée sur tous les phonèmes et x_{ph} est le facteur de phonèmes estimé sur les données propres au phonème ph.

Pour visualiser l'effet de la variabilité locuteur sur les phonèmes, nous avons estimé les facteurs de phonèmes de plusieurs phonèmes prononcés par plusieurs locuteurs. Notons x_{ph-l} les facteurs de phonèmes ph prononcés par un locuteur l. Nous avons choisi les trois phonèmes *ii*, *aa* et *oo*. Les facteurs de phonèmes estimés sont de dimension 100. Après le calcul de l'ACP, nous projetons les différents vecteurs dans un espace bi-dimensionnel. Dans la figure 5.10, nous observons trois nuages de points pour les trois phonèmes. La dispersion des points des facteurs de phonèmes autour du centre de l'ellipse est due principalement à la variabilité de la prononciation.

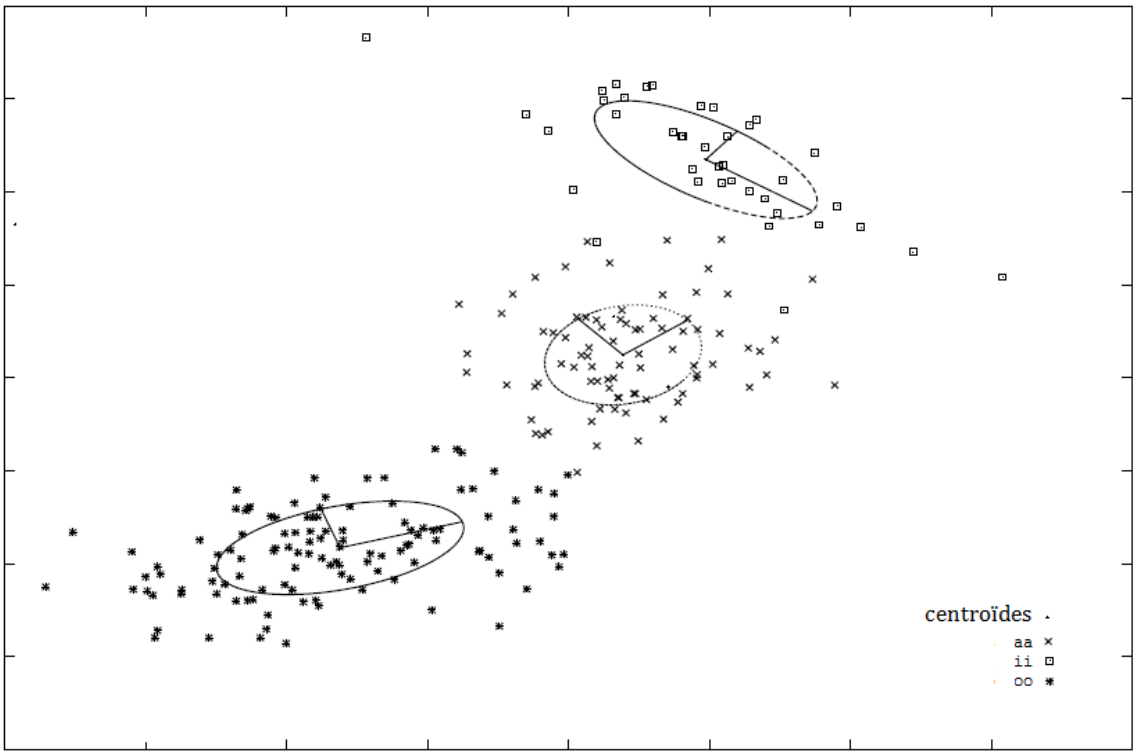

FIGURE 5.10 – *Projection montrant l'effet de la variabilité locuteur sur trois phonèmes différents (aa, ii et oo).*

5.6 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle représentation vectorielle des états des MMC. Cette représentation est obtenue avec le paradigme d'analyse factorielle. Dans la modélisation acoustique contextuelle, nous avons exploité cette représentation vectorielle dans la réalisation de la procédure de regroupement d'états des MMC. La classification d'états dans cette procédure s'appuie seulement sur l'information portée par les vecteurs *facteurs d'états*. L'application de cette méthode à la langue française a montée une amélioration significative des performances. Nous avons étendu l'utilisation des facteurs d'états dans la modélisation acoustique pour les langues peu dotées. La modélisation acoustique pour cette classe de langues souffre du manque de ressources informatiques et linguistiques. Avec les facteurs d'états, nous avons réussi à contourner ces contraintes pour construire des modèles acoustiques dépendants du contexte, qui aménent de bonnes performances (avec un gain relatif de 23 % et 20 % pour la langue vietnamienne et la langue berbère respectivement, par rapport au système de référence obtenu par la technique de l'arbre de décision standard). De plus, nous avons montré que, grâce à la représentation vectorielle, il devient possible d'avoir une visualisation graphique des états (ou des phonèmes). Cette visualisation pourrait être utile pour l'analyse de différents types de variabilités affectant les prononciations des unités phonétiques. Elle pourrait être utilisée, par exemple, dans le domaine de

la phonétique clinique, par des phonéticiens qui désirent étudier certains phénomènes phonologiques particuliers.

Chapitre 6

Analyse Factorielle pour la compensation de la variabilité nuisible

Sommaire

6.1 Introduction

Aujourd'hui, les systèmes de reconnaissance automatique de la parole (SRAP) permettent d'obtenir de très bonnes performances lorsque les conditions de test sont proches des conditions d'apprentissage du système. Cependant, dans un contexte applicatif réel, les SRAP sont soumis à diverses sources de bruit engendrant une dégradation significative des performances. Les sources de variabilité nuisible dans le signal de la parole ne se limitent pas au bruit environnemental, mais incluent toute variabilité (information) qui n'est pas utile à la tâche envisagée. Par exemple les informations concernant le locuteur sont inutiles pour la reconnaissance de la parole, elle peuvent même être gênantes pour cette tâche. On peut distinguer trois grandes classes de source de variabilité généralement pénalisantes pour la reconnaissance de la parole :

1. L'environnement du locuteur qui produit un bruit ambiant. Ce bruit ambiant est toujours considéré comme étant additif. Il représente l'une des sources de variabilité les plus gênantes pour un système de traitement automatique de la parole. La difficulté liée au traitement du bruit additif dépend de ses caractéristiques, les plus importantes étant la stationnarité et l'intensité. En outre, l'environnement produit un bruit qui peut être corrélé à la parole (réverbération, réflexion).
2. Le locuteur lui-même, selon son état psychologique (stress, état émotionnel) ou physique (fatigue, rhume) modifie les propriétés prosodiques et spectrales du signal de parole. Cette variabilité est appelée variabilité *intra-locuteur*. Ce phénomène s'accentue lorsqu'une même phrase est prononcée par deux locuteurs différents. Dans ce cas nous parlons de variabilité *inter-locuteur*. Les deux variabilités *inter-locuteurs* les plus importantes sont le genre et l'accent (Huang et al., 2001).
3. Les conditions d'enregistrement qui incluent le type de microphone, la distance au microphone, le canal de transmission (distorsion, écho, bruit électronique, etc). Les variabilités liées au type de conditions d'enregistrement ont un effet convolutif dans le domaine temporel, multiplicatif dans le domaine des densités spectrales et additives dans le domaine cepstral. Le bruit qui en résulte est appelé *bruit convolutif*.

Ces sources de variabilité sont très pénalisantes pour le développement et l'exploitation à grande échelle des systèmes de reconnaissance automatique de la parole. Pour cette raison, de nombreuses techniques ont été proposées pour augmenter la robustesse des systèmes, en particulier leur résistance aux bruits. L'objectif de ces techniques est de compenser les différences entre les conditions d'apprentissage et les conditions d'utilisation du système. Deux grandes classes d'approches peuvent être distinguées :

1. La première classe d'approche nécessite généralement une connaissance *a priori* sur la nature du bruit. On suppose que l'apprentissage s'est déroulé dans une ambiance calme et on procède dans ce cas au débruitage des signaux de test. Ce prétraitement peut s'effectuer soit dans l'espace des paramètres spectraux ou cepstraux, soit directement sur le signal observé. Dans cette classe d'approches, nous trouvons la méthode de *soustraction spectrale* qui consiste à soustraire du spectre

de la parole bruitée une estimation du spectre de bruit (Lim, 1978). La méthode de *soustraction du cepstre moyen* consiste à soustraire du signal une estimation du bruit dans le domaine cepstral, correspondant donc au bruit convolutif. Une technique de normalisation cepstrale plus complexe utilise les dictionnaires CDCN (*Codebook Dependent Cepstral Normalisation*). Il s'agit de construire une fonction qui transforme la parole bruitée en parole propre (Acero, 1990). La méthode VTS (*Vector Taylor Series*) va encore plus loin que la CDCN dans la modélisation de la fonction d'environnement en proposant de l'approche par des séries de Taylor dans le domaine log-spectral (Vaseghi et Milner, 1992).

2. La deuxième classe d'approches consiste à ne pas modifier les données de test, mais à adapter les modèles acoustiques, appris sur des signaux propres, pour décoder la parole bruitée. La *combinaison Parallèle des Modèles* (CPM) (Gales et Young, 1996) (Gales, 1992) fait partie de cette classe d'approches. Cette technique suppose que la parole propre est modélisée par des HMM. Les états des HMM modélisant la parole bruitée sont obtenus en combinant les états des HMM modélisant la parole propre et un modèle correspondant au bruit. Plusieurs autres travaux ont été réalisés pour adapter un système de reconnaissance à l'environnement de test bruité (Holmes et Sedgwick, 1986; Roe, 1987; Morii et al., 1990)

Ces deux classes d'approches permettent de compenser les différences entre les conditions d'apprentissage et les conditions d'utilisation du système. Elles peuvent être utilisées conjointement pour appréhender le problème du bruit dans le cadre de la reconnaissance de la parole.

Une autre approche qui a attiré l'attention par son efficacité dans les SRAP est le *signal subspace filtering* (SSF) (Ephraim et Trees, 1995; Dendrinos et al., 1991). Dans cette méthode le bruit est supposé additif, de moyenne nulle, blanc et non corrélé avec le signal de parole. L'idée de base est de subdiviser l'espace d'observation de dimension N en deux sous-espaces. Le premier sous-espace, appelé sous-espace signal, est de dimension $p (p < N)$. Il est le sous-espace (parole + bruit) dans lequel le bruit interfère avec le signal de parole. Le deuxième sous-espace (de dimension $N - p$) est le sous-espace de bruit qui contient uniquement du bruit. La parole propre est obtenue en annulant le sous-espace de bruit et en éliminant la contribution du bruit dans le sous-espace du signal.

Dans ce chapitre, nous développons une nouvelle approche de compensation de la variabilité nuisible à la reconnaissance de la parole. Les vecteurs cepstraux sont supposés être générés par des GMM (de grande taille) dépendants de l'état du MMC (contextuel ou non contextuel). Ici, nous nous intéressons à la variable aléatoire constituée par la concaténation des moyennes des gaussiennes composant le GMM (dépendant du phonème ou de l'état). Cette variable aléatoire sera appelée *super-vecteur*. L'isolation et l'estimation du bruit se fait en utilisant le paradigme de l'Analyse Factorielle dans l'espace des super-vecteurs. Nous supposerons que le bruit est additif dans ce domaine et qu'il est situé dans un sous-espace de très faible dimension (par rapport à la dimension du super-vecteur).

Dans ce chapitre, nous étudierons plusieurs scénarios liés à la variabilité nuisible pour la reconnaissance de la parole. Dans la section 6.2, nous nous intéresserons à deux variabilités : la variabilité locuteur et la variabilité canal. Nous commencerons par étudier la compensation des deux variabilités séparément. Ensuite, nous tenterons de procéder à une compensation conjointe de celles-ci. Dans la section 6.5, nous traiterons de la variabilité liée au bruit additif. Les résultats expérimentaux sont exposés dans la section 6.6.

6.2 Analyse Factorielle pour la modélisation de la variabilité nuisible

Malgré les efforts déployés dans les domaines de la paramétrisation et de la modélisation en vue de la reconnaissance de la parole, les SRAP souffrent d'une manière très importante du changement des conditions acoustiques : changement du matériel d'acquisition, condition d'enregistrement, bruit ambiant, etc. Ces phénomènes introduisent des décalages acoustiques ayant un grand pouvoir de nuisance sur le SRAP. Les solutions proposées agissent à deux niveaux différents du SRAP : au niveau des paramètres acoustiques et au niveau des modèles acoustiques. Dans ce travail nous proposons une méthode de compensation agissant dans l'espace des paramètres acoustiques.

L'approche que nous proposons dans ce chapitre suppose que l'espace acoustique (espace des trames, cesptre) peut être modélisé par un GMM de grande taille (typiquement 512 gaussiennes). Nous appelons ce GMM, l'UBM. Les états des HMM sont modélisés par des GMM qui sont obtenus à partir de l'UBM par une sorte de MAP, par exemple. Pour un GMM donné, nous allons nous intéresser à une variable aléatoire constituée par la concaténation des moyennes des gaussiennes de ce GMM. Nous appelons cette nouvelle variable le super-vecteur. Nous allons développer un super-vecteur d'un état donné selon trois composantes :

1. une composante indépendante de l'état et de la variabilité nuisible en question,
2. une composante correspondant à l'information phonétique (état d'un MMC ou un phonème),
3. une composante correspondant à la variabilité nuisible que nous sommes en train de traiter.

L'hypothèse fondamentale dans le formalisme que nous sommes en train de développer est que la variabilité nuisible est située dans un sous-espace de très faible dimension par rapport à la dimension du super-vecteur. La dimension de l'espace des super-vecteurs est $M \times D$ avec M le nombre des gaussiennes de l'UBM et D la dimension des vecteurs d'observation (vecteur représentant une trame).

Ici, nous traitons la variabilité locuteur et la variabilité canal en utilisant la même modélisation : l'Analyse Factorielle. Cependant, pour des raisons de clarté, nous exposons, dans un premier temps, le modèle pour la variabilité locuteur et dans un second temps, le modèle pour la variabilité canal.

6.2.1 Variabilité locuteur

Comme dit précédemment, la variabilité locuteur est une des plus perturbantes pour un système de reconnaissance de la parole. L'effet de cette variabilité sur le signal est très complexe. Dans cet effet, il y a probablement des parties linéaires et d'autres non-linéaires. Cependant, si l'on considère que le changement du locuteur se traduit par le changement du conduit vocal (modèle source-filtre), alors cette variabilité peut être considérée comme additive dans le domaine cepstral.

Si la variabilité locuteur est considérée comme additive dans le domaine cepstral, alors elle l'est aussi dans l'espace des super-vecteurs. Supposons en plus que cette variabilité locuteur est située dans un sous-espace de faible dimension de l'espace des super-vecteurs. Formellement, désignons par ph l'indice indiquant l'information phonétique et par loc l'indice indiquant l'information locuteur. Soit $m_{ph,loc}$ le super-vecteur associé au phonème ph prononcé par le locuteur loc. Voici la modélisation que nous nous proposons d'utiliser :

$$\mathbf{m}_{(ph,loc)} = \mathbf{m} + \mathbf{D}_{\mathbf{y}_{ph}} + \mathbf{U}\mathbf{x}_{(ph,loc)} \tag{6.1}$$

où $\mathbf{m}$ est le super-vecteur issu de l'UBM. $\mathbf{D}$ est une matrice diagonale $MD \times MD$ et y_{ph} est un vecteur de dimension MD estimé sur les données du phonèmes ph. Le dernier terme $\mathbf{U}\mathbf{x}_{(ph,loc)}$ représente la composante de la variabilité locuteur. La matrice $\mathbf{U}$ est une matrice de rang faible R (avec $R \ll MD$). Ses vecteurs colonnes forment une base d'un sous-espace dans lequel la variabilité locuteur est majoritairement située. La matrice $\mathbf{U}$ est formée par la concaténation des M sous-matrices $\mathbf{U}_g$ de dimensions $D \times R$. Chaque sous-matrice $\mathbf{U}_g$ correspond à la g^{me} gaussienne dans l'UBM (voir figure 6.1). $\mathbf{x}_{(ph,loc)}$ est un vecteur de dimension R contenant les composantes de la variabilité locuteur dans ce sous-espace.

Pour estimer les paramètres de ce modèle, nous devons faire quelques hypothèses : $\mathbf{y}_{ph}$ et $\mathbf{x}_{(ph,loc)}$ sont tous les deux normalement distribués selon $\mathbf{N}(0, \mathbf{I})$. $\mathbf{D}$ est une matrice diagonale de sorte que $\mathbf{D}\mathbf{D}^t$ soit la matrice de covariance *a priori* de la composante liée aux informations linguistiques.

La matrice $\mathbf{U}$ est estimée d'une manière itérative en utilisant l'algorithme EM. Pour chaque étape, $\mathbf{x}_{(ph,loc)}$ est estimé, puis $\mathbf{y}_{ph}$ est estimé pour chaque phonème (utilisant le nouveau $\mathbf{x}_{(ph,loc)}$) et finalement la matrice $\mathbf{U}$ est estimée globalement en s'appuyant sur ces $x_{(ph,loc)}$ et y_{ph}. Les étapes de l'algorithme EM sont décrites plus en profondeur dans (Matrouf et al., 2007).

Pour estimer les paramètres de la matrice $\mathbf{U}$ qui représentent la variabilité locuteur, nous avons besoin d'un corpus permettant de mettre en évidence cette variabilité afin d'estimer les paramètres du modèle correspondant. Pour étudier la variabilité locuteur nous avons besoin d'un corpus dans lequel chaque phonème est prononcé par différents locuteurs. Dans le modèle de l'équation 6.1, le terme $\mathbf{y}_{ph}$ sera estimé en utilisant toutes les données associées au phonème ph, par contre $\mathbf{x}_{(ph,loc)}$ sera estimé sur toutes les données associées au locuteur loc. Pour construire ce corpus, nous regroupons les

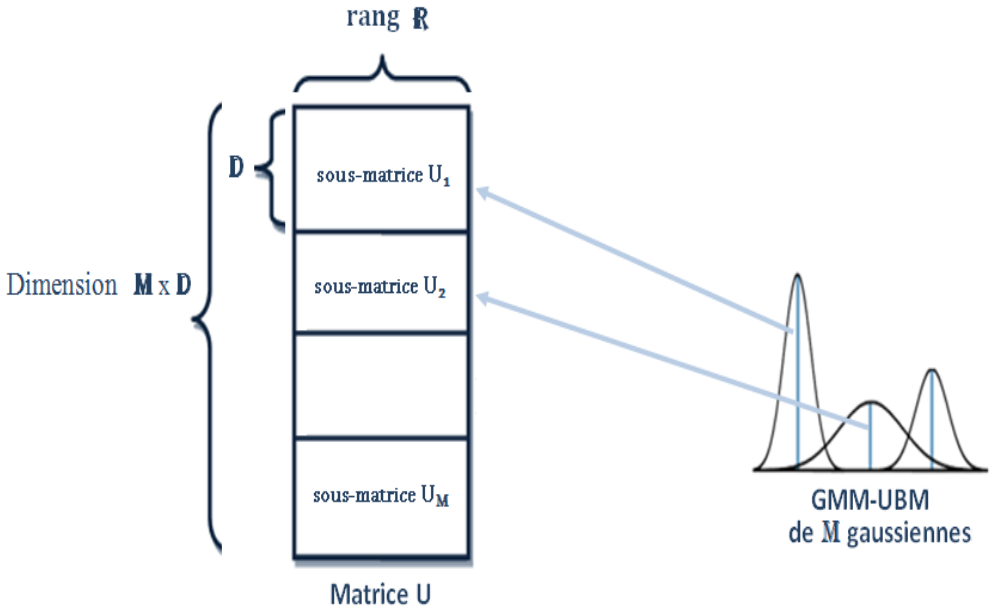

FIGURE 6.1 – *Illustration de la correspondance entre les sous-matrices qui forment la matrice* $\boldsymbol{U}$ *et les gaussiennes de l'UBM.*

trames correspondant à chaque phonème indépendant du contexte ph prononcé par le locuteur loc. Cet ensemble de trames est noté (ph, loc). Dans le corpus d'apprentissage, nous avons au total 36 phonèmes et 150 locuteurs, ce qui nous donne 5 400 couples phonème-locuteur (ph, loc) (voir figure 6.2).

6.2.2 Variabilité canal

Nous désignons par variabilité canal, tout changement dans les conditions d'enregistrement, notamment le changement de microphone ou de téléphone, l'endroit ou s'effectue l'enregistrement, etc. L'effet de cette variabilité est plutôt additif dans le domaine cepstral. De la même manière que dans la section précédente, un phonème (ou un état d'un MMC) est modélisé par un GMM issu d'un UBM. Soit $m_{ph,canal}$ le supervecteur associé au phonème ph prononcé dans les conditions acoustiques du canal $canal$". Le modèle de l'équation 6.1 peut être ré-écrit comme suit :

$$\mathbf{m}_{(ph,canal)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{ph} + \mathbf{U}\mathbf{x}_{(ph,canal)} \tag{6.2}$$

Dans ce modèle, la matrice $\mathbf{U}$ est une matrice de rang faible R (avec $R \ll MD$). Ses vecteurs colonnes forment une base d'un sous-espace dans lequel la variabilité canal est majoritairement située. $\mathbf{x}_{(ph,canal)}$ est un vecteur de dimension R contenant les composantes de la variabilité canal dans ce sous-espace.

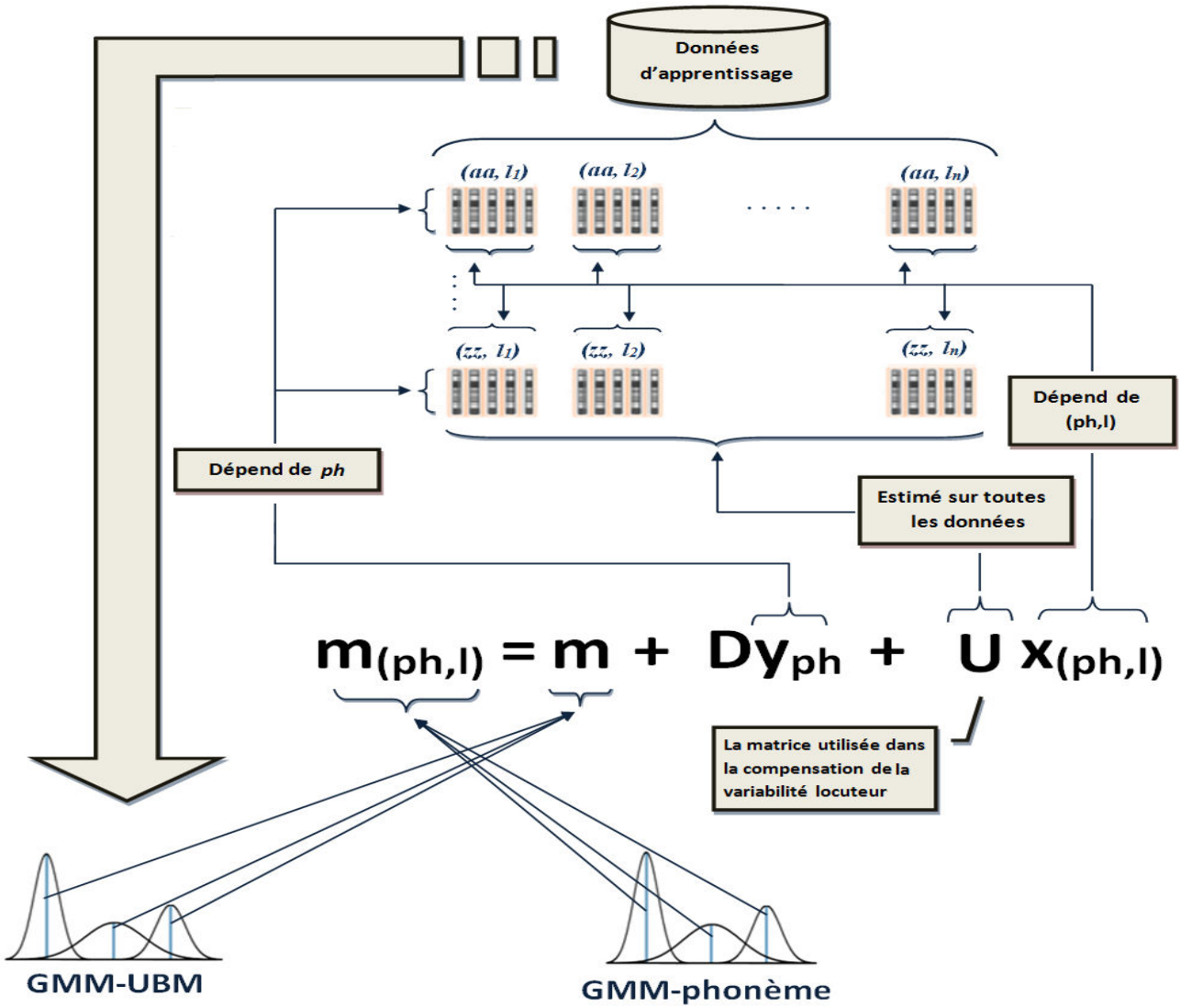

FIGURE 6.2 – *Différentes composantes du modèle de variabilité locuteur.*

Pour étudier la variabilité canal, nous avons besoin d'un corpus dans lequel chaque phonème est prononcé dans différentes conditions acoustiques. Pour construire ce corpus, nous regroupons les trames qui correspondent à chaque phonème *ph* prononcé dans les conditions acoustiques du canal *canal*. Cet ensemble de trames est noté $(ph, canal)$.

Dans le modèle de l'équation 6.2, le terme $\mathbf{y}_{ph}$ sera estimé en utilisant toutes les données associées au phonème *ph*, par contre $\mathbf{x}_{(ph,canal)}$ sera estimé sur toutes les données associées au canal *canal*. Nous avons utilisé le même nom (**U**) pour le sous-espace de la variabilité locuteur et la variabilité canal, et ce malgré le fait qu'elles soient fondamentalement différentes. Ce choix est fait afin de souligner le fait que ces variabilités sont

estimées en utilisant la même procédure, mais sur des données structurées différemment : corpus locuteur et corpus canal.

6.3 Compensation de la variabilité nuisible sur les trames acoustiques

6.3.1 Compensation de la variabilité locuteur

Une fois la matrice **U** obtenue, nous pouvons passer à l'étape de la compensation de variabilité locuteur. Théoriquement, nous devons estimer le facteur de la variabilité locuteur $\mathbf{x}_{(ph,loc)}$ pour chaque phonème présent dans l'ensemble des phrases de test (le modèle obtenu par l'équation 6.1 suppose que l'effet locuteur dépende du phonème). En pratique, en phase de test, ceci n'est pas réalisable en raison du manque de données pour un phonème *ph* dans une phrase : pour un phonème donné, nous ne disposons pas d'assez de trames pour estimer les différentes composantes du modèle. Pour contourner ce problème, nous supposons que l'effet du locuteur est le même pour tous les phonèmes. Ceci implique d'estimer les facteurs de la variabilité locuteur globalement sur l'ensemble des phrases du test. En phase d'utilisation, nous procédons à une seule itération de l'algorithme d'estimation : seul le terme $\mathbf{x}_{(ph,loc)}$ est estimé. En d'autres termes, le modèle de l'équation 6.2 est remplacé par le modèle suivant :

$$\mathbf{m}_{(ph,loc)} = \mathbf{m} + \mathbf{U}\mathbf{x}_{(ph,loc)} \tag{6.3}$$

Ce modèle pour la g^{eme} gaussienne peut être exprimé comme suit :

$$\mathbf{m}^{g}_{(ph,loc)} = \mathbf{m}^{g} + \mathbf{U}\mathbf{x}^{g}_{(ph,loc)} \tag{6.4}$$

Après le calcul de la composante $\mathbf{U}\mathbf{x}_{(ph,loc)}$ sur l'ensemble des phrases d'apprentissage, nous compensons la variabilité *loc* au niveau de chaque trame t. La trame propre $\hat{t}$ est obtenue par l'équation suivante :

$$\hat{t} = t - \sum_{g=1}^{M} \gamma_g(t) \cdot \{\mathbf{U}_{[g]} \cdot \mathbf{x}_{(phrase,loc)}\} \tag{6.5}$$

où M est le nombre de gaussiennes dans le GMM-UBM, et $\gamma_g(t)$ est la probabilité *a posteriori* de la gaussienne g donnée par la trame t. Ces probabilités sont estimées en utilisant le GMM-UBM. $\mathbf{U}_{[g]} \cdot \mathbf{x}_{(phrase,loc)}$ est la composante de la variabilité locuteur *loc* estimée sur l'ensemble des phrases à décoder. $\mathbf{U}_{[g]}$ est la g^{eme} sous-matrice dans la matrice **U**, avec g correspondant à l'indice d'une gaussienne dans le GMM-UBM.

La procédure de compensation de la variabilité locuteur est appliquée sur les données d'apprentissage et sur les données de test. L'estimation des modèles acoustiques se fait ensuite en utilisant les trames compensées.

6.3.2 Compensation de la variabilité canal

De manière similaire à la compensation de la variabilité locuteur, la compensation de la variabilité canal est appliquée sur les données d'apprentissage et sur les données de test. En premier lieu, nous estimons les facteurs de la variabilité canal globalement sur l'ensemble des phrases de test. En phase de test, nous n'estimons pas les composantes y_{ph}, le modèle utilisé s'écrit alors comme suit :

$$\mathbf{m}_{(ph,canal)} = \mathbf{m} + \mathbf{U}\mathbf{x}_{ph,canal} \tag{6.6}$$

Ce modèle s'écrit au niveau de la gaussienne comme suit :

$$\mathbf{m}^{g}_{(ph,canal)} = \mathbf{m}^{g} + \mathbf{U}\mathbf{x}^{g}_{(ph,canal)} \tag{6.7}$$

Après le calcul de la composante $\mathbf{U}\mathbf{x}_{(ph,canal)}$ sur l'ensemble des phrases d'apprentissage, nous compensons la variabilité canal au niveau de chaque trame. La trame propre $\hat{t}$ est obtenue par l'équation suivante :

$$\hat{t} = t - \sum_{g=1}^{M} \gamma_g(t) \cdot \{\mathbf{U}_{[g]} \cdot \mathbf{x}_{(ph,canal)}\} \tag{6.8}$$

où M est le nombre de gaussiennes dans le GMM-UBM, et $\gamma_g(t)$ est la probabilité *a posteriori* de la gaussienne g étant donnée la trame t. Ces probabilités sont estimées en utilisant le GMM-UBM. $\mathbf{U}_{[g]} \cdot \mathbf{x}_{(ph,v)}$ est la composante de la variabilité canal estimée sur l'ensemble des phrases de test. $\mathbf{U}_{[g]}$ est la g^{eme} sous-matrice dans $\mathbf{U}$, avec g correspondant à l'indice d'une gaussienne dans le GMM-UBM.

Dans la modélisation proposée ci-dessus, les variabilités locuteur et canal sont traitées indépendamment l'une de l'autre. Les deux matrices de variabilité sont estimées indépendamment en utilisant deux corpus structurés différemment. En réalité, les deux variabilités affectent le signal de parole simultanément. Pour cette raison, nous proposons d'intégrer les deux variabilités dans le même modèle et d'estimer conjointement leurs paramètres.

6.4 Traitement conjointe de la variabilité locuteur et de la variabilité canal

Dans cette section, nous présentons notre proposition de modélisation et de compensation conjointe de la variabilité locuteur et de la variabilité canal.

6.4.1 Modélisation conjointe

L'idée de modéliser conjointement deux variabilités différentes a été proposée dans (Kenny, 2006) pour le domaine de la vérification de locuteurs. Dans cette approche, appelée *Joint Factor Analysis* (JFA), l'auteur propose de modéliser sur le même corpus deux variabilités : l'effet locuteur et l'effet canal. De manière similaire, nous proposons de modéliser les variabilités locuteur et canal conjointement, mais sur deux corpus différents : le corpus de la *variabilité-locuteur* et le corpus de la *variabilité-canal*. Afin de réaliser cette idée, nous proposons une version étendue de l'équation du modèle 6.1 formulée comme suit :

$$\mathbf{m}_{(ph,l,c)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{ph} + \mathbf{U}\mathbf{x}_{(ph,loc)} + \mathbf{V}\mathbf{z}_{ph,canal} \tag{6.9}$$

Comme dans le modèle précèdent, les vecteurs colonnes de la matrice $\mathbf{U}$ (de rang $R \ll MD$) forment une base d'un sous-espace dans lequel la variabilité locuteur est située. Le vecteur $\mathbf{x}_{(ph,l)}$, de dimension R, contenant les composantes relatives à la variabilité locuteur dans ce sous-espace. Semblablement, les vecteurs colonnes de la matrice $\mathbf{V}$ (de rang $R^{'} \ll MD$) forment une base d'un sous-espace dans lequel la variabilité canal est située. Les vecteurs $\mathbf{x}_{(ph,c)}$, de dimension $R^{'}$, contiennent les composantes relatives à la variabilité canal dans ce sous-espace.

Dans ce modèle, les paramètres des deux matrices $\mathbf{U}$ et $\mathbf{V}$ sont estimés conjointement avec la procédure présentée dans l'algorithme 2. Dans une première étape, nous estimons la matrice $\mathbf{U}$ sur les données du corpus *variabilité-locuteur*. Pour ce faire, nous estimons d'abord $\mathbf{x}_{loc}$ et $\mathbf{z}_{canal}$, ensuite nous estimons $\mathbf{y}_{ph}$ pour chaque phonème en utilisant les vecteurs $\mathbf{x}$ et $\mathbf{z}$. Finalement, nous estimons $\mathbf{U}$ globalement en s'appuyant sur les vecteurs $\mathbf{x}$, $\mathbf{y}$ et $\mathbf{z}$. Dans une seconde étape, nous estimons de manière similaire la matrice $\mathbf{V}$ sur le corpus *variabilité-canal*. Nous estimons d'abord les vecteurs $\mathbf{x}_{loc}$ et $\mathbf{z}_{canal}$, ensuite nous estimons $\mathbf{y}_{ph}$ pour chaque phonème en utilisant les nouveaux $\mathbf{x}$ et $\mathbf{z}$. Finalement, la matrice $\mathbf{V}$ est estimée globalement en utilisant les vecteurs $\mathbf{x}$, $\mathbf{y}$ et $\mathbf{z}$ variables. Nous répétons ces deux étapes dans une procédure itérative pour améliorer l'estimation de $\mathbf{U}$ et $\mathbf{V}$.

Afin de calculer les paramètres des matrices $\mathbf{U}$ et $\mathbf{V}$ et des variables latentes $\mathbf{x}_{ph,loc}$ et $\mathbf{z}_{ph,canal}$, nous avons besoin de calculer les statistiques d'ordre zéro et les statistiques du premier ordre sur le GMM-UBM.

Soient $\mathbf{N}^{ph}$, $\mathbf{N}^{(ph,loc)}$, $\mathbf{X}^{ph}$ et $\mathbf{X}^{(ph,loc)}$ les statistiques d'ordre zéro et du premier ordre dépendant du phonème et du locuteur. Ces statistiques sont calculées sur le corpus *variabilité-locuteur* :

$$\mathbf{N}_g^{ph} = \sum_{t \in ph} \gamma_g(t); \qquad \mathbf{N}_g^{(ph,loc)} = \sum_{t \in (ph,loc)} \gamma_g(t) \tag{6.10}$$

$$\mathbf{X}_g^{ph} = \sum_{t \in ph} \gamma_g(t) \cdot t; \qquad \mathbf{X}_g^{(ph,loc)} = \sum_{t \in (ph,loc)} \gamma_g(t) \cdot t \tag{6.11}$$

où $\gamma_g(t)$ est la probabilité *a posteriori* d'une gaussienne g pour une observation t. $\Sigma_{t\in ph}$ correspond à la somme de toutes les trames du phonème ph et $\Sigma_{t\in(ph,loc)}$ correspond à la somme de toutes les trames du phonème ph prononcé par un locuteur loc.

De manière similaire, nous calculons les statistiques d'ordre zéro et du premier ordre sur le corpus *variabilité-canal*. Soient $M_g^{(ph,c)}$, $M_g^{(ph,c)}$, Z_g^{ph} et $Z_g^{(ph,c)}$ les statistiques d'ordre zéro et de premier ordre dépendant du phonème et des canaux :

$$\mathbf{M}_g^{ph} = \sum_{t\in ph} \gamma_g(t); \qquad \mathbf{M}_g^{(ph,c)} = \sum_{t\in(ph,c)} \gamma_g(t) \tag{6.12}$$

$$\mathbf{Z}_g^{ph} = \sum_{t\in ph} \gamma_g(t)\cdot t; \qquad \mathbf{Z}_g^{(ph,c)} = \sum_{t\in(ph,c)} \gamma_g(t)\cdot t \tag{6.13}$$

où $\gamma_g(t)$ est la probabilité *a posteriori* d'une gaussienne g pour une observation t. $\Sigma_{t\in ph}$ correspond à la somme de toutes les trames du phonème ph et $\Sigma_{t\in(ph,c)}$ correspond à la somme de toutes les trames du phonème ph prononcé dans la condition *canal*.

Ensuite, nous calculons les statistiques $\overline{\mathbf{X}}^{ph}$, $\overline{\mathbf{X}}^{(ph,loc)}$, $\overline{\mathbf{Z}}^{ph}$ et $\overline{\mathbf{Z}}^{(ph,c)}$ en tenant compte des matrices $\mathbf{U}$ et $\mathbf{V}$:

$$\begin{aligned}
\overline{\mathbf{X}}_g^{ph} &= \mathbf{X}_g^{ph} - \sum_{h\in l} \mathbf{N}_g^{(ph,loc)} \cdot \{m + \mathbf{U}\mathbf{x}_{(ph,loc)} + \mathbf{V}\mathbf{z}_{(ph,c)}\}_g \\
\overline{\mathbf{X}}_g^{(ph,loc)} &= \mathbf{N}_g^{(ph,loc)} - \mathbf{N}_g^{(ph,loc)} \cdot \{m + \mathbf{D}\mathbf{y}_{ph} + \mathbf{V}\mathbf{z}_{ph,c}\}_g \\
\overline{\mathbf{Z}}_g^{ph} &= \mathbf{Z}_g^{ph} - \sum_{h\in c} \mathbf{M}_g^{(ph,c)} \cdot \{m + \mathbf{U}\mathbf{x}_{(ph,loc)} + \mathbf{V}\mathbf{z}_{(ph,loc)}\}_g \\
\overline{\mathbf{Z}}_g^{(ph,c)} &= \mathbf{Z}_g^{(ph,c)} - \mathbf{M}_g^{(ph,c)} \cdot \{m + \mathbf{D}\mathbf{y}_{ph} + \mathbf{U}\mathbf{x}_{ph,loc}\}_g
\end{aligned} \tag{6.14}$$

Désignons $L_{(ph,loc)}$ et $P_{(ph,c)}$ deux matrices de dimension $\mathbf{R}\times\mathbf{R}$ et $\mathbf{B}_{(ph,loc)}$, $Q_{(ph,c)}$ deux vecteurs de dimension R, définis par :

$$\begin{aligned}
B_{(ph,loc)} &= \sum_{g\in UBM} \mathbf{U}_g^T \cdot \Sigma_g^{-1} \cdot \overline{\mathbf{X}_g^{ph,loc}} \\
L_{(ph,loc)} &= I + \sum_{g\in UBM} \mathbf{N}_g^{(ph,loc)} \cdot \mathbf{U}_g^T \cdot \Sigma_g^{-1} \cdot \mathbf{U}_g \\
Q_{(ph,c)} &= \sum_{g\in UBM} \mathbf{V}_g^T \cdot \Sigma_g^{-1} \cdot \overline{\mathbf{Z}_g^{(ph,c)}} \\
P_{(ph,c)} &= I + \sum_{g\in UBM} \mathbf{M}_g^{(ph,c)} \cdot \mathbf{V}_g^T \cdot \Sigma_g^{-1} \cdot \mathbf{V}_g
\end{aligned} \tag{6.15}$$

Où Σ_g est la matrice de covariance de la composante g du GMM-UBM. En utilisant $L_{(ph,loc)}$, $B_{(ph,loc)}$, $P_{(ph,c)}$ et $Q_{(ph,c)}$, nous pouvons obtenir $\mathbf{x}_{(ph,loc)}$, $\mathbf{z}_{(ph,c)}$ et $\mathbf{y}_{ph}$ depuis les équations suivantes :

$$\begin{aligned} \mathbf{z}_{(ph,c)} &= P_{(ph,c)}^{-1} \cdot Q_{(ph,c)} \\ \mathbf{x}_{(ph,loc)} &= L_{(ph,loc)}^{-1} \cdot B_{(ph,loc)} \\ \mathbf{y}_{ph} &= \frac{\tau}{\tau + \mathbf{N}_g} \cdot \mathbf{D}_g \Sigma_g^{-1} \cdot \overline{\mathbf{X}_g^{(ph,loc)}} \cdot \overline{\mathbf{Z}_g^{(ph,c)}} \end{aligned} \tag{6.16}$$

où $\mathbf{D}_g = (1/\sqrt{\tau})\Sigma_g^{1/2}$.

Finalement les matrices $\mathbf{U}$ et $\mathbf{V}$ peuvent être estimées ligne par ligne, avec $\mathbf{U}_g^i$ et $\mathbf{V}_g^i$ les i^{th} ligne de $\mathbf{U}_g$ et $\mathbf{V}_g$; donc :

$$\begin{aligned} \mathbf{U}_g^i &= \mathcal{L}(g)^{-1} \cdot \mathcal{R}^i(g) \\ \mathbf{V}_g^i &= \mathcal{P}(g)^{-1} \cdot \mathcal{Q}^i(g) \end{aligned} \tag{6.17}$$

où $\mathcal{L}(g)$, $\mathcal{R}^i(g)$, $\mathcal{P}(g)$ et $\mathcal{Q}^i(g)$ sont obtenus par :

$$\begin{aligned} \mathcal{L}(g) &= \sum_s \sum_{h \in s} \mathbf{N}_g^{(ph,loc)} \cdot (\mathbf{L}_{(ph,loc)}^{-1} + \mathbf{x}_{(ph,loc)} \mathbf{x}_{(ph,loc)}^T) \\ \mathcal{R}^i(g) &= \sum_s \sum_{h \in s} \overline{\mathbf{X}}_g^{(s,ph)}[i] \cdot \mathbf{x}_{(ph,loc)} \\ \mathcal{P}(g) &= \sum_s \sum_{h \in s} \mathbf{M}_g^{(ph,c)} \cdot (P_{(ph,c)}^{-1} + \mathbf{x}_{(ph,c)} \mathbf{x}_{(ph,c)}^T) \\ \mathcal{Q}^i(g) &= \sum_s \sum_{h \in s} \overline{\mathbf{Z}}_g^{(ph,c)}[i] \cdot \mathbf{x}_{(ph,c)} \end{aligned} \tag{6.18}$$

Algorithm 2: Algorithme d'estimation des matrices **U** et **V**

Pour chaque *phonème ph* et (l, c) : $\mathbf{y}_{ph} \leftarrow 0$, $\mathbf{x}_{(ph,loc)} \leftarrow 0$, $\mathbf{z}_{(ph,c)} \leftarrow 0$;
$\mathbf{U} \leftarrow$ *random* (**U** est initialisée aléatoirement);
$\mathbf{V} \leftarrow$ *random* (**V** est initialisée aléatoirement);
Estimation des statistiques : $\mathbf{N}^{ph}$, $\mathbf{N}^{(ph,loc)}$, $\mathbf{X}^{ph}$, $\mathbf{X}^{(ph,loc)}$ sur le corpus *variabilité-locuteur* ;
Estimation des statistiques : $\mathbf{M}^{ph}$, $\mathbf{M}^{(ph,c)}$, $\mathbf{Z}^{ph}$, $\mathbf{Z}^{(ph,c)}$ sur le corpus *variabilité-canal*;
pour $i = 1$ *to* $nb - iterations$ **faire**
  pour *chaque locuteur* (loc) *et phonème ph de corpus* variabilite-locuteur **faire**
    Centrer les statistiques : $\overline{\mathbf{Z}}^{(ph,c)}$;
    Centrer les statistiques : $\overline{\mathbf{X}}^{(ph,loc)}$;
    Estimer $L^{-1}_{(ph,loc)}$ et $B_{(ph,loc)}$;
    Estimer $\mathbf{z}_{(ph,c)}$;
    Estimer $\mathbf{x}_{(ph,loc)}$;
    Centrer les statistiques : $\overline{Z}^{ph}$;
    Centrer les statistiques : $\overline{X}^{ph}$;
    Estimer $\mathbf{y}_{ph}$;
  fin
  Estimer la matrice **U** ;
  pour *chaque canal c et phonème ph de corpus* variabilite-canal **faire**
    Centrer les statistiques : $\overline{\mathbf{Z}}^{(ph,c)}$;
    Centrer les statistiques : $\overline{\mathbf{X}}^{(ph,loc)}$;
    Estimer $P^{-1}_{(ph,c)}$ et $Q_{(ph,c)}$;
    Estimer $\mathbf{z}_{(ph,c)}$;
    Estimer $\mathbf{x}_{(ph,loc)}$;
    Centrer les statistiques : $\overline{\mathbf{Z}}^{c}$;
    Centrer les statistiques : $\overline{\mathbf{X}}^{loc}$;
    Estimer $\mathbf{y}_{ph}$;
  fin
  Estimer la matrice **V** ;
fin

6.4.2 Compensation conjointe

De manière similaire à la compensation indépendante des variabilités (section 6.3), nous estimons les facteurs de la variabilité locuteur $\mathbf{x}_{ph,loc}$ et de la variabilité canal $\mathbf{z}_{ph,c}$

globalement sur l'ensemble des phrases à décoder.

Après le calcul des composante $\mathbf{U}\mathbf{x}_{ph,loc}$ et $\mathbf{V}\mathbf{z}_{ph,c}$ sur l'ensemble des phrases d'apprentissage, nous compensons simultanément les deux variabilités locuteur et canal sur chaque trame. La trame propre $\hat{t}$ est obtenue par l'équation suivante :

$$\hat{t} = t - \sum_{g=1}^{\mathbf{M}} \gamma_g(t) \cdot (\{\mathbf{U}_{[g]} \cdot \mathbf{x}_{ph,loc}\} + \{\mathbf{V}_{[g]} \cdot \mathbf{z}_{ph,c}\}) \tag{6.19}$$

où M est le nombre de gaussiennes dans le GMM-UBM, et $\gamma_g(t)$ est la probabilité *a posteriori* de la gaussienne g donnée par la trame t. Ces probabilités sont estimées en utilisant le GMM-UBM. $\mathbf{U}_{[g]} \cdot \mathbf{x}_{ph,loc}$ et $\mathbf{V}_{[g]} \cdot \mathbf{z}_{ph,c}$ sont les composantes des variabilités locuteur et canal calculées sur l'ensemble des phrases à décoder. $\mathbf{U}_{[g]}$ et $\mathbf{V}_{[g]}$ sont les g^{eme} sous-matrice dans $\mathbf{U}$ et $\mathbf{V}$ respectivement, avec g correspondant à l'indice d'une gaussienne dans le GMM-UBM.

6.5 Analyse Factorielle pour la compensation du bruit additif

Dans cette section, nous avons orienté notre intérêt vers une autre source de dégradation des performances des SRAP : le bruit additif. La difficulté liée au traitement des bruits additifs dépend de leurs caractéristiques (stationnarité, intensité). Le bruit ambiant est additif dans le domaine temporel et spectral. Dans le domaine cepstral, l'effet du bruit ambiant est théoriquement non-linéaire.

Malgré cette caractéristique de non-linéarité entre le bruit ambiant et le signal de la parole dans le domaine cepstral, nous allons tenter de savoir s'il est possible de modéliser le bruit ambiant comme une composante additive localisée dans un sous-espace de faible dimension (dans l'espace de super-vecteurs). Désignons par $\mathbf{m}_{(ph,b)}$ le super-vecteur correspondant au phonème ph dans la présence du bruit additif b. Le modèle d'analyse factorielle s'écrit comme suit :

$$\mathbf{m}_{(ph,b)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{ph} + \mathbf{U}\mathbf{x}_{(ph,b)} \tag{6.20}$$

Dans ce modèle, $\mathbf{m}$, $\mathbf{D}$ et $\mathbf{y}_{ph}$ jouent le même rôle que dans l'équation 6.1. Les vecteurs colonnes de la matrice $\mathbf{U}$ sont les vecteurs générateurs du sous-espace de bruit. $\mathbf{x}_{ph,b}$ est le facteur du bruit, contenant les composantes relatives au bruit dans ce sous-espace.

Pour estimer la matrice $\mathbf{U}$ qui représente le sous-espace du bruit, nous avons besoin d'un grand corpus de parole bruitée avec différents types de bruit. Comme ce type de corpus n'est pas disponible, nous avons décidé de bruiter artificiellement notre corpus d'apprentissage de parole propre en utilisant plusieurs types de bruit. Nous avons choisi quatre types différents de bruit : bruit de foule (brouhaha), bruit de voitures, bruit de supermarchés et bruit de machines (moteur de bateau, ventilateur, machine

d'usines). Nous utilisons différents enregistrements de chaque bruit pour bruiter le corpus d'apprentissage. Cette étape de bruitage est réalisée dans le domaine temporel. Ensuite, nous avons regroupé les trames correspondant au même phonème ph avec la présence de bruit b pour obtenir l'ensemble (ph, b). Sur ce corpus, nous estimons les paramètres des différentes composantes du modèle de l'équation 6.20. La figure 6.3 montre les étapes de l'estimation de la matrice $\mathbf{U}$.

Dans l'étape de compensation, nous estimons la composante $x_{ph,b}$ du bruit additif sur l'ensemble des phrases à décoder. Les trames propres $\hat{t}$ sont obtenues par l'équation suivante :

$$\hat{t} = t - \sum_{g=1}^{M} \gamma_g(t) \cdot \{\mathbf{U}_{[g]} \cdot \mathbf{x}_{phrase,b}\} \tag{6.21}$$

où M est le nombre de gaussiennes dans le GMM-UBM, et $\gamma_g(t)$ est la probabilité *a posteriori* de la gaussienne g donnée par la trame t. Ces probabilités sont estimées en utilisant le GMM-UBM. $\mathbf{U}\mathbf{x}_{ph,b}$ est la composante du bruit additif estimée sur l'ensemble des phrases à décoder. $\mathbf{U}_{[g]}$ est la g^{eme} sous-matrice dans $\mathbf{U}$, avec g correspondant à l'indice d'une gaussienne dans le GMM-UBM.

6.6 Expérimentations

6.6.1 Système et corpus

Dans ces expériences, nous utilisons le système SPEERAL, décrit dans la section 4.4.1. Le processus de transcription se compose de deux passes :

1. La première passe utilise les modèles acoustiques correspondant au genre et à la bande passante détectés par le processus de segmentation et utilisant un modèle de langage tri-gramme.
2. La seconde passe applique une transformation de type *Maximum Likelihood Linear Régression* (MLLR) par locuteur ou par segment, et utilise le même modèle de langage tri-gramme que la première passe.

Les performances du système sont évaluées sur le corpus d'évaluation ESTER. Les données sont composées de 18 fichiers audio pour une durée totale de 10 heures.

6.6.2 Compensation de la variabilité locuteur et de la variabilité canal

Dans ces expériences, Le GMM-UBM utilisé dans le modèle d'analyse factorielle est composé de 600 gaussiennes. Le rang des matrices $\mathbf{U}$ et $\mathbf{V}$ est fixé à 60. Nous rappelons que nous avons proposé deux méthodes pour compenser les variabilités locuteur et canal : compensation indépendante et compensation conjointe. Tout d'abord, nous

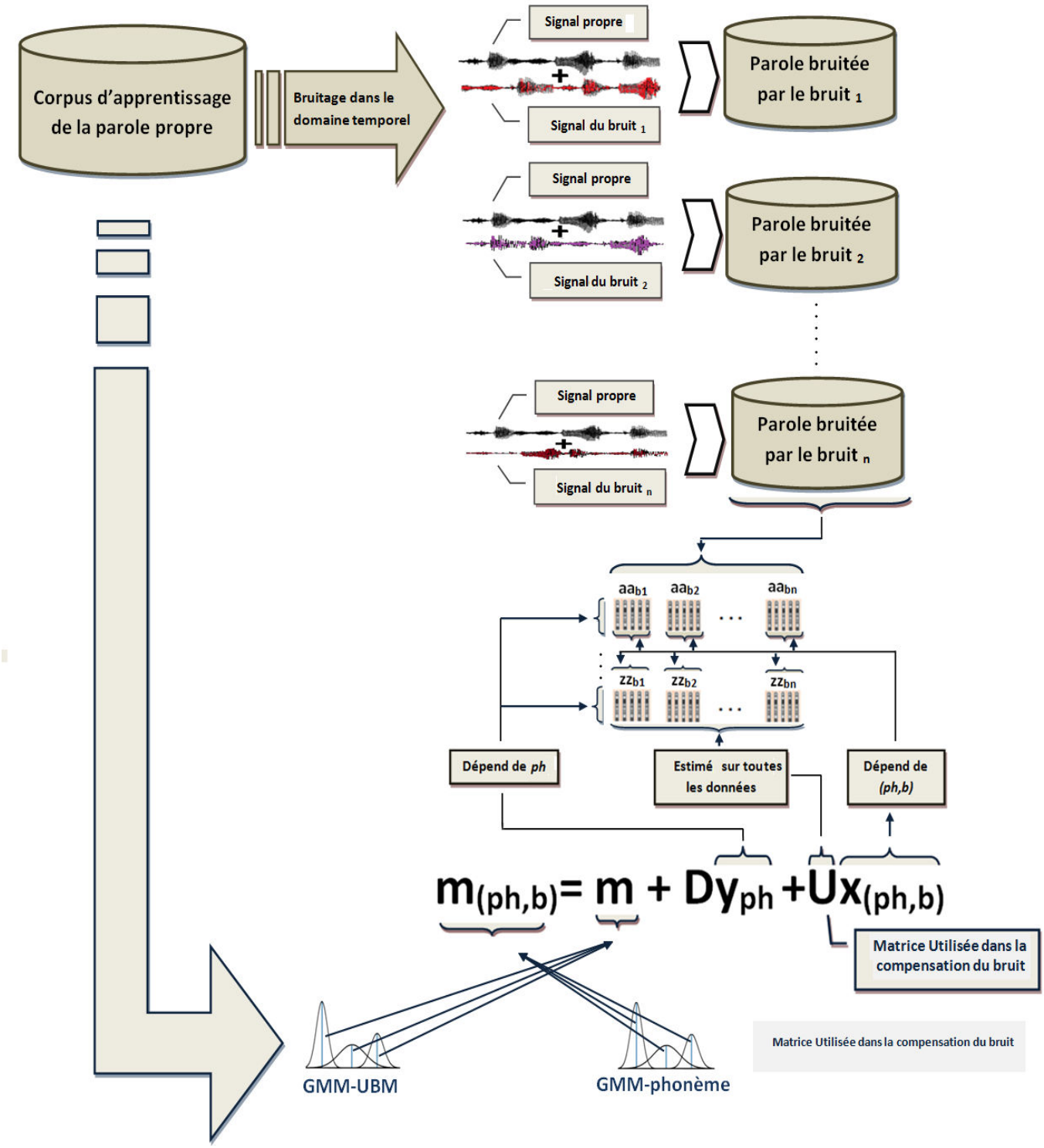

FIGURE 6.3 – *Étapes d'estimation de la matrice* $\boldsymbol{U}$.

appliquons la méthode de compensation simple, ensuite nous testerons l'effet de la compensation conjointe sur les résultats du SRAP.

6.6.2.1 Compensation simple

Dans une première étape, nous allons tester la méthode de compensation simple. Nous utilisons les deux matrices **U** et **V** estimées indépendamment. Les vecteurs de la matrice **U** forment la base d'un sous-espace dans laquelle est située la variabilité locuteur, alors que les vecteurs de la matrice **V** forment la base d'un sous-espace dans laquelle la variabilité canal est située.

En premier lieu, nous calculons le vecteur $\mathbf{x}_{phrase,loc}$ sur l'ensemble des phrases de notre corpus d'apprentissage. Ensuite, nous soustrayons la composante $\mathbf{U}\mathbf{x}_{phrase,loc}$, qui représente la variabilité locuteur, des trames acoustiques, en s'appuyant sur l'équation 6.8. Une fois la compensation terminée, nous utilisons les trames acoustiques propres pour estimer les paramètres de notre modèle acoustique.

Dans la phase de test, nous utilisons la même matrice **U** dans le calcul du vecteur $\mathbf{x}_{phrase,loc}$ sur l'ensemble des phrases de test. En se basant sur la même équation 6.8, nous soustrayons la composante $\mathbf{U}\mathbf{x}_{phrase,loc}$ de variabilité locuteur de chaque trame du corpus de test. Finalement, nous décodons les trames de test normalisées en utilisant le modèle acoustique estimé sur les données également normalisées.

De la même manière, nous utilisons la matrice **V** dans le calcul du vecteur $\mathbf{x}_{phrase,c}$ sur l'ensemble des phrases d'apprentissage. Ensuite, nous soustrayons la composante $\mathbf{V}\mathbf{x}_{phrase,c}$, qui représente la variabilité canal, des trames acoustiques, en se basant sur l'équation 6.8. Une fois la compensation terminée, nous utilisons les trames acoustiques propres pour estimer les paramètres de notre modèle acoustique. Dans la phase de test, nous utilisons la même matrice **V** dans le calcul du vecteur $\mathbf{x}_{phrase,c}$ sur l'ensemble des phrases de test. En nous basant sur la même équation 6.8, nous soustrayons la composante $\mathbf{V}\mathbf{x}_{phrase,loc}$ de variabilité locuteur de chaque trame du corpus de test. Finalement, nous décodons les trames de test normalisées en utilisant le modèle acoustique estimé sur les données également normalisées.

Les résultats des décodages obtenus dans les deux cas (compensation de la variabilité locuteur et compensation de la variabilité canal) sont comparés avec les résultats obtenus avec le système baseline entraîné sur des données non normalisées. Dans le tableau 6.1 nous exposons les résultats, en termes de taux d'erreur mot (TEM), de trois systèmes : baseline, compensation de la variabilité locuteur (*Comp-locuteur*) et compensation de la variabilité canal (*Comp-canal*).

Dans la première passe, nous observons une amélioration de 1,1 % et 1,0 % en absolu, respectivement pour les modèles *Comp-locuteur* et *Comp-canal*.Nous constatons également une amélioration du TEM de 0,8 % et 0,6 % dans la deuxième passe. Les gains dans la deuxième passe apparaissent cependant moins importants que dans la première passe. Ceci peut être expliqué par l'adaptation MLLR appliquée dans la

	Passe 1	Passe 2
Baseline	29,6	27,5
Comp-locuteur	28,5	26,9
Comp-canal	28,6	26,7

TABLE 6.1 – *Performances, en termes de TEM, de trois systèmes : baseline, compensation de la variabilité locuteur (Comp-locuteur) et compensation de la variabilité canal (Comp-canal).*

deuxième passe. En effet, la technique MLLR adapte le modèle acoustique à un locuteur particulier, capturant les relations entre le modèle original et le locuteur courant ou l'environnement acoustique.

6.6.2.2 Compensation conjointe

Dans une seconde expérience, nous allons tester la méthode de compensation conjointe. Les deux matrices **U** et **V** utilisées pour la compensation sont estimées conjointement. D'abord, nous estimons les composantes $\mathbf{U}_{\mathbf{x}_{ph,loc}}$ et $\mathbf{V}_{\mathbf{z}_{ph,c}}$, qui représentent respectivement la variabilité locuteur et canal, sur l'ensemble des phrases de notre corpus d'apprentissage. Ensuite, nous soustrayons ces deux composantes des trames acoustiques en utilisant l'équation 6.19. Nous utilisons les trames normalisées obtenues pour estimer les paramètres du modèle acoustique. Dans la phase de test, nous utilisons les mêmes matrices **U** et **V** pour calculer les composantes $\mathbf{U}_{\mathbf{x}_{phrase,loc}}$ et $\mathbf{V}_{\mathbf{z}_{phrase,loc}}$ sur l'ensemble des phrases de test. En nous appuyant sur la même équation 6.19, nous soustrayons ces composantes des variabilités locuteur et canal de chaque trame du corpus de test. Finalement, nous décodons les trames de test normalisées en utilisant le modèle acoustique estimé sur les données également normalisées.

La performance du système intégrant la compensation conjointe de la variabilité locuteur et de la variabilité canal, est comparée avec celle obtenue avec un système de base entraîné sur des données non normalisées. Dans le tableau 6.2, nous exposons ces résultats, en termes de taux d'erreur mot (TEM), pour ces deux systèmes (baseline et *Comp-locuteur-canal*). Nous constatons que le système *Comp-locuteur-canal* permet d'obtenir, en deuxième passe, un gain absolu de 1,3 % en termes de TEM par rapport au système de base.

	Passe 1	Passe 2
Baseline	29,6	27,5
Comp-locuteur-canal	28,0	26,2

TABLE 6.2 – *Résultats obtenus, en termes de TEM, par le système de base (Baseline) et le système intégrant la compensation conjointe de la variabilité locuteur et de la variabilité canal.*

Afin d'analyser plus précisément l'effet de la compensation conjointe, nous avons évalué son effet sur chaque phrase de test. Pour ce faire, nous avons trié les phrases de test suivant leur TEM obtenus sur le système baseline. Nous ordonnons également les phrases obtenues par le système *Comp-locuteur-canal* de la même façon en 11 intervalles.

Le tableau 6.3 permet de comparer le TEM des phrases en fonction du système utilisé (*Baseline* et *Comp-locuteur-canal*). Les différences entre les deux systèmes sont reportées dans la colonne *Gain TEM*.

	Baseline	Norm-locuteur-canal	Gain TEM
0-5	0,35	2,46	-2,11
5-10	7,19	8,52	-1,33
10-15	12,73	13,44	-0,70
15-20	17,60	18,36	-0,75
20-25	21,52	20,91	0,61
25-30	26,71	25,80	0,91
30-35	32,18	29,79	2,39
35-40	37,01	35,79	1,22
40-45	41,85	39,45	2,39
45-50	46,36	44,00	2,36
50-100	68,28	62,54	5,74

TABLE 6.3 – *Résultats pour chaque rangée selon le TEM.*

Nous pouvons observer que, pour les phrases appartenant à l'intervalle 0-20, la compensation détériore les résultats du SRAP par rapport au système de base, alors que la compensation permet un gain variable pour les phrases appartenant à l'intervalle 30-100. Ce gain est particulièrement important sur les rangées 50 à 100 avec une réduction absolue du TEM de 5,74 %. Nous concluons que la compensation permet le gain le plus important sur les phrases étant les plus bruitées, et ayant un TEM important avec le système de base. Cependant, sur les phrases obtenant un TEM bas, la compensation n'apporte aucune amélioration, et peut même détériorer les résultats.

6.6.3 Compensation du bruit additif

Dans cette expérience, nous allons appliquer la méthode proposée pour compenser le bruit additif. Le bruit additif est caractérisé par sa non-linéarité avec le signal de parole dans le domaine cepstral. Malgré ce fait théorique, nous allons tenter de savoir s'il est possible de modéliser le bruit additif (ou une partie du bruit additif) comme une composante additive $\mathbf{U}\mathbf{x}_{ph,b}$ en utilisant l'équation du modèle de l'analyse factorielle 6.20. La compensation n'est appliquée que sur les données (vecteurs cepstraux) de test en utilisant l'équation 6.21. Le GMM-UBM utilisé dans ces expériences est de 600 gaussiennes. Le rang de la matrice $\mathbf{U}$ est fixé à 60.

Une fois la matrice $\mathbf{U}$ estimée (voir section 6.5), nous passons à l'étape de la compensation. La compensation est testée sur deux types de données, dans un premier temps sur des données enregistrées dans des conditions bruitées, et dans un second temps sur des données bruitées artificiellement.

6.6.3.1 Expérimentation sur des données enregistrées dans des conditions bruitées

La première expérience sur la compensation est réalisée dans le contexte d'un corpus représentant un cas réel de conditions défavorables pour la reconnaissance vocale. Ce corpus est formé de fichiers enregistrés au cours de conférences et d'assemblées générales dans des conditions acoustiques dégradées et très bruitées : des microphones mal placés (la distance et l'orientation du locuteur par rapport au microphone), du bruit ambiant, de l'écho, etc.

Sur l'ensemble des segments de parole à décoder (chaque segment représente une phrase), nous calculons les paramètres des vecteurs $\mathbf{x}_{phrase,b}$ qui contiennent les composantes relatives au bruit additif dans le sous-espace formé par les vecteurs colonnes de la matrice **U**. Cette matrice est estimée sur le corpus ESTER bruité avec différents types de bruit. La compensation est réalisée par la soustraction de la composante $\mathbf{U}\mathbf{x}_{phrase,b}$ au niveau des vecteurs cepstraux en utilisant l'équation 6.21. Une fois la compensation terminée, nous passons à l'étape de décodage. Les résultats de décodage obtenus sont comparés avec la sortie du SRAP sur les mêmes fichiers bruités, mais sans appliquer la méthode de compensation.

Le tableau 6.4 montre les résultats, en termes de TEM, pour les différents enregistrements de test. Dans la deuxième colonne, nous trouvons les TEM obtenus sur les fichiers de test bruités. La troisième colonne expose les TEM de ces mêmes fichiers, mais avec compensation du bruit.

	Sans compensation	Avec compensation
Test 1	59.11	51.22
Test 2	63.11	54.63
Test 3	48.41	44.60
Test 4	63.99	53.81
Total	58.01	50.55

TABLE 6.4 – *Résultats obtenus, en termes de TEM, sans compensation et avec compensation du bruit additif.*

Les résultats exposés dans le tableau 6.4 montrent que la compensation du bruit additif nous permet de gagner plus de 7 % en absolu en termes de TEM. En nous appuyant sur ces résultats, nous pouvons conclure que la modélisation par analyse factorielle permet de trouver le sous-espace dans lequel le bruit additif est situé. Ce sous-espace est formé par les vecteurs colonnes de la matrice **U**. Ainsi, nous confirmons la possibilité de modéliser une grande partie de bruit additif présente dans le signal de la parole comme une composante additive $\mathbf{U}\mathbf{x}_{ph,b}$.

Dans les expériences suivantes, nous testons l'effet de deux facteurs sur l'estimation et la compensation du bruit additif : le niveau du bruit (rapport signal sur bruit) et le type du bruit. Les données de test utilisées dans ces expériences sont bruitées artificiellement.

6.6.3.2 Expérimentation sur des données bruitées artificiellement

Nous rappelons que les données d'apprentissage utilisées dans l'estimation des paramètres de la matrice **U** sont artificiellement bruitées. Dans ces expériences, les données de test sont aussi bruitées de la même façon. Le bruitage a été réalisé dans le domaine temporel en utilisant différents types de bruit : bruit de foule (brouhaha), bruit de voitures, bruit du supermarché et bruit de machines (moteur de bateau, ventilateur, machines d'usines). Le bruitage artificiel nous permet de contrôler le rapport signal sur bruit (*RSB*) dans les fichiers d'apprentissage et les fichiers de test. Nous allons exploiter ce contrôle pour répondre à la question suivante : est-ce que la correspondance entre le RSB des fichiers de test utilisés dans l'estimation de la matrice **U** et le RSB des fichiers de test sur lesquels est appliquée la compensation du bruit est nécessaire ?

Pour répondre à cette question, nous avons bruité notre corpus d'apprentissage de la parole propre avec quatre *RSB* différents. Nous obtenons quatre corpus d'apprentissage de parole bruitée *Train-N1*, *Train-N2*, *Train-N3* et *Train-N4* de RSB -8 dB, 1 dB, 8 dB et 16 dB respectivement. Nous utilisons chaque corpus pour estimer les paramètres du GMM-UBM, ainsi que les paramètres du modèle de bruit obtenus avec l'équation de l'analyse factorielle 6.20. Nous obtenons quatre matrices différentes, appelée $\mathbf{U}_{N1}$,$\mathbf{U}_{N2}$, $\mathbf{U}_{N3}$ et $\mathbf{U}_{N4}$.

De manière similaire à l'apprentissage, dans le test sont bruités les fichiers de test propres avec quatre RSB différents. Nous obtenons quatre corpus de test de la parole bruitée *Test-N1*, *Test-N2*, *Test-N3* et *Test-N4* de RSB -8 dB, 1 dB, 8 dB et 16 dB respectivement. Nous appliquons notre méthode de compensation de bruit sur chaque corpus, en utilisant chaque fois une des matrices $\mathbf{U}_{N1}$,$\mathbf{U}_{N2}$, $\mathbf{U}_{N3}$ et $\mathbf{U}_{N4}$. Une fois la compensation terminée, nous passons à l'étape de décodage. Les résultats obtenus sont toujours comparés avec la sortie du SRAP sur les fichiers bruités, mais sans appliquer la méthode de compensation.

Le tableau 6.5 présente les résultats obtenus en terme de TEM sur les quatre corpus de test artificiellement bruités *Test-N1*, *Test-N2*, *Test-N3* et *Test-N4*. La deuxième colonne montre le TEM obtenu sur les fichiers de test sans appliquer la méthode de compensation du bruit. Les quatre dernières colonnes exposent le TEM obtenu sur les mêmes fichiers, mais avec l'application de la compensation du bruit en utilisant à chaque fois une des quatre matrices $\mathbf{U}_{N1}$, $\mathbf{U}_{N2}$, $\mathbf{U}_{N3}$ et $\mathbf{U}_{N4}$.

	Sans compensation	$\mathbf{U}_{N1}$	$\mathbf{U}_{N2}$	$\mathbf{U}_{N3}$	$\mathbf{U}_{N4}$
Test-N1	38,10	34,69	**34,57**	34,91	35,37
Test-N2	41,80	37,03	**36,18**	36,46	36,36
Test-N3	45,90	44,03	43,18	**42,06**	42,96
Test-N4	49,40	47,05	46,10	44,54	**43,75**

TABLE 6.5 – *Résultats obtenus, en termes de TEM, sans compensation du bruit puis en appliquant la méthode de compensation avec des RSB différents.*

Les résultats exposés dans le tableau 6.5 montrent que le gain le plus élevé est obtenu lorsque nous utilisons dans la compensation une matrice estimée sur des données

ayant un RSB proche de celui des données de test. Cela montre que la correspondance entre le RSB des données d'apprentissage et des données de test est importante dans la méthode de compensation proposée.

Dans l'expérience suivante, nous allons tester l'importance de la correspondance entre le type de bruit qui affecte les fichiers d'apprentissage et le bruit qui affecte les fichiers de test. Nous avons utilisé quatre types de bruit dans l'étape de bruitage : bruit de foule (brouhaha), bruit de la rue, bruit du supermarché et bruit des machines (moteur de bateau, ventilateur, machine d'usines). Nous notons ces bruits $T1$, $T2$, $T3$ et $T4$ respectivement. Après le bruitage, nous obtenons quatre corpus d'apprentissage de la parole bruitée *Train-T1*, *Train-T2*, *Train-T3* et *Train-T4*. Nous utilisons chaque corpus pour estimer les paramètres du GMM-UBM, ainsi que les paramètres du modèle de bruit obtenus avec l'équation de l'analyse factorielle 6.20. Nous obtenons quatre matrices différentes, appelée $\mathbf{U}_{T1}$,$\mathbf{U}_{T2}$, $\mathbf{U}_{T3}$ et $\mathbf{U}_{T4}$. Ainsi, nous estimons un cinquième modèle en utilisant les quatre corpus *Train-T1*, *Train-T2*, *Train-T3* et *Train-T4* ensemble. Nous appelons la matrice obtenue (U_{T1234}). Les vecteurs colonnes de cette matrice forment une base d'un sous-espace dans laquelle les différents bruits sont situés.

Dans le test, nous bruitons les fichiers de test propres avec les quatre types de bruit $T1$, $T2$, $T3$ et $T4$. Nous obtenons quatre corpus de test de parole bruitée *Test-T1*, *Test-T2*, *Test-T3* et *Test-T4*. Nous appliquons notre méthode de compensation du bruit sur chaque corpus, en utilisant à chaque fois une des matrices $\mathbf{U}_{T1}$,$\mathbf{U}_{T2}$, $\mathbf{U}_{T3}$, $\mathbf{U}_{T4}$ et U_{T1234}. Une fois la compensation terminée, nous passons à l'étape de décodage. Les résultats obtenus sont toujours comparés avec la sortie du SRAP sur les fichiers bruités, mais sans appliquer la méthode de compensation.

Le tableau 6.6 présente les résultats obtenus, en termes de TEM, sur les quatre corpus de test artificiellement bruités *Test-T1*, *Test-T2*, *Test-T3* et *Test-T4*. La deuxième colonne montre le TEM obtenu sur les fichiers bruités, mais sans compensation. Les cinq dernières colonnes exposent le TEM obtenu sur les mêmes fichiers, mais avec l'application de la compensation du bruit en utilisant à chaque fois une des cinq matrices $\mathbf{U}_{T1}$,$\mathbf{U}_{T2}$, $\mathbf{U}_{T3}$, $\mathbf{U}_{T4}$ et U_{T1234}.

	Sans compensation	$\mathbf{U}_{T1}$	$\mathbf{U}_{T2}$	$\mathbf{U}_{T3}$	$\mathbf{U}_{T4}$	$\mathbf{U}_{T1234}$
Test-T1	41,80	*36,18*	37,95	37,51	42,02	**35,74**
Test-T2	43,50	42,78	***37,58***	40,17	39,14	37,95
Test-T3	44,70	42,17	42,87	*38,86*	44,66	**36,15**
Test-T4	42,90	42,77	42,26	41,95	*36,69*	**36,15**

TABLE 6.6 – *Résultats obtenus, en termes de TEM, sans compensation du bruit puis en appliquant la méthode de compensation incluant différents types de bruit.*

Les résultats exposés dans le tableau 6.6 montrent un gain significatif lorsque nous utilisons dans la compensation une matrice estimée sur des données bruitées par un bruit similaire à celui présent dans le fichier de test. Mais le gain le plus important est généralement observé dans le cas où est utilisée, dans la compensation, la matrice estimée sur toutes les données bruitées (la matrice $\mathbf{U}_{N1234}$).

6.7 Conclusion

Dans ce chapitre nous avons montré l'utilité de la modélisation par analyse factorielle pour la robustesse des SRAP face aux multiples variabilités nuisibles. Nous avons proposé une méthode simple de compensation de la variabilité locuteur, de la variabilité canal et du bruit additif, s'appuyant sur la modélisation d'analyse factorielle. Nous avons affirmé la possibilité de modéliser les différents types de variabilité acoustique nuisible comme une composante additive dans le domaine cepstral. Nous soustrayons cette composante des vecteurs cepstraux pour compenser ces variabilités pénalisantes pour la reconnaissance de la parole.

Dans les premières expériences de ce travail, nous nous sommes intéressés à la variabilité locuteur et à la variabilité canal, indépendamment l'une de l'autre. Ensuite, nous avons proposé un algorithme qui permet de compenser d'une manière conjointe les deux variabilités canal et locuteur. Les algorithmes que nous avons proposés sont fondés sur l'analyse factorielle. Les améliorations obtenues sont limitées mais encourageantes, avec un gain de 1,3 % observé.

Dans la seconde partie de travail, nous nous sommes intéressés à une autre type de source de bruit nuisible pour le SRAP : le bruit additif. Nous avons montré qu'avec la méthode proposée, il est possible de modéliser le bruit additif (ou une partie) comme composante additive dans le domaine cepstral. Nous avons présenté des résultats correspondants à des bruits réels affectant le signal de parole, mais aussi des résultats correspondant à des bruits ajoutés artificiellement. La compensation de cette composante (aussi dans le domaine cepstral) a permis d'améliorer la sortie du SRAP de 7,5 % en absolu en termes de TEM.

Chapitre 7

Conclusion et Perspectives

Dans cette thèse nous nous sommes intéressés à l'utilisation de méthodes provenant de l'analyse factorielle afin de tenter de résoudre trois problématiques liées au traitement de la parole : la modélisation acoustique compacte, la classification d'unités acoustiques et la compensation de variabilités nuisibles à la reconnaissance de la parole. Dans la suite nous résumons les travaux réalisés dans cette thèse ainsi que les résultats obtenus.

7.1 Travaux réalisés et résultats obtenus

7.1.1 Modélisation acoustique compacte (chapitre 4)

La première application de l'analyse factorielle dans cette thèse était destinée à la modélisation acoustique compacte. Dans cette application, nous avons réussi à réduire considérablement le nombre des paramètres du modèle acoustique tout en préservant de bonnes performances. La méthode utilisée consiste à représenter l'espace acoustique de la parole incluant tous les phonèmes par un seul modèle générique (GMM-UBM), puis à dériver les modèles des différents états des MMC depuis ce modèle générique, en mutualisant une large partie des modèles. Les dérivations des moyennes des gaussiennes sont obtenues en utilisant un modèle fondé sur l'analyse factorielle, alors que les moyennes ont été obtenues en utilisant le maximum de vraisemblance. Les variances sont restées inchangées par rapport au modèle du monde (UBM).

Dans nos premières expériences, nous avons cherché le point de fonctionnement optimal entre le taille et la performance des modèles. Nous avons fait varier la taille du modèle et le nombre des gaussiennes du GMM-UBM. Le meilleur modèle compact obtenu avec la modélisation d'Analyse Factorielle (FA) permet une réduction de la taille de 92 %, tout en maintenant des performances similaires à celles du modèle de base (nous rappelons que dans le modèle de base les GMM des états sont estimés indépendamment les uns des autres avec l'algorithme EM).

Nous avons montré que la modélisation FA permet de décomposer l'ensemble des paramètres du modèle acoustique compact en cinq sous-ensembles indépendants :

$$\pi = \{\pi_s\}, A = \{a_{s,s'}\}, UBM, \mathbf{U}, X = \{x_s\}, W = \{w_s\}$$

Où s indique l'état, π et A étant les probabilités initiales des états et les probabilités de transition. UBM est l'ensemble des paramètres du modèle du monde. X est l'ensemble des vecteurs spécifiques aux états s, et finalement, W représente l'ensemble des vecteurs des poids de gaussiennes spécifiques aux états s.

Cette décomposition permet une grande flexibilité dans l'adaptation du modèle acoustique aux locuteurs, aux genres ou aux nouvelles tâches. Grâce à cette décomposition, nous avons montré la possibilité d'adapter les modèles acoustiques en utilisant des données non transcrites [1]. En effet, l'estimation des paramètres de l'UBM ne nécessite pas de données transcrites. Dans notre modèle, tous les GMM des états sont obtenus à partir de l'UBM. L'adaptation de ce dernier à une tâche cible permet d'adapter le modèle acoustique dans sa globalité à la tâche en question sans avoir besoin de données transcrites. Dans nos expérimentations, nous avons utilisé des données enregistrées dans des conditions acoustiques bruitées (conférences, colloques, assemblées générales, etc.). L'absence de transcription de ces fichiers audio ne nous permet pas d'adapter le MMC avec les méthodes classiques. Mais l'adaptation de notre modèle à travers l'UBM nous a permis d'obtenir un gain de 5,93 % en absolu comparativement au modèle compact générique. Une seconde adaptation du modèle compact générique a été proposée dans nos expériences. Elle consiste à adapter les vecteurs d'états X. L'avantage de celle-ci est qu'elle peut être réalisée avec une quantité très limitée de données. En effet, l'estimation des vecteurs X ne demande que quelques dizaines de trames. Les expériences de cette adaptation sur les locuteurs nous a permis un gain de 1,02 % en absolu par rapport au modèle compact générique.

7.1.2 Classification phonétique (chapitre 5)

Dans la modélisation acoustique compacte basée sur l'analyse factorielle (chapitre 4), tous les GMM associés aux états des MMC sont dérivés à partir d'un seul modèle générique appelé modèle du monde (UBM). Nous avons montré, dans nos expériences, que les modèles compacts obtenus ont des performances similaires à celles obtenues avec le modèle de base. Ces résultats prouvent que les *facteurs d'états* $\mathbf{x}_e$, avec un nombre limité de paramètres, caractérisent bien les états des MMC.

En s'appuyant sur ces résultats, nous avons proposé d'adopter ces vecteurs comme une représentation vectorielle d'états des MMC. Un intérêt indéniable de cette représentation vectorielle est la possibilité de traiter les états par des techniques d'*analyse de données*, telles que l'*analyse factorielle des correspondances*. Le fait de représenter les états par des vecteurs permet également de mesurer efficacement la similarité entre ces états,

1. Habituellement, afin de pouvoir réaliser le processus d'adaptation du MMC, une transcription des données audio doit être fournie (transcription de référence ou sortie des systèmes de RAP). Ces méthodes d'adaptation sont dites supervisées.

en utilisant des distances adaptées (distance euclidienne ou distance de Mahalanobis). Il est important de noter que cette tâche (mesure de similarité) était très complexe (voire très approximative) lorsque les états ne possédaient pour représentants que les GMM.

Dans la modélisation acoustique, nous avons proposé l'utilisation des facteurs d'états pour réaliser la procédure de partage d'états. Notre proposition consiste à s'appuyer sur les facteurs d'états pour classifier les états en utilisant l'algorithme de classification non-supervisée *k-means*.

Dans nos expériences, nous avons appliqué cette méthode sur la langue française. Nous avons observé une amélioration des performances du système (plus de 1 % en absolu en termes de taux d'erreur mot). L'avantage de la méthode proposée est qu'elle s'appuie uniquement sur l'information portée par les facteurs d'états pour déterminer les états acoustiquement proches. Ainsi, nous n'avons pas besoin de connaissances phonétiques habituellement nécessaires dans la réalisation de ce travail. En outre, nous pouvons correctement estimer les facteurs d'états sur une quantité limitée de données d'apprentissage. Nous avons décidé d'exploiter ces avantages dans la modélisation acoustique pour les langues peu dotées. La modélisation acoustique pour cette catégorie de langues souffre du manque de ressources informatiques et linguistiques. Dans les expériences, la méthode proposée est appliquée dans un premier temps sur les langues vietnamienne et berbère, où nous avons obtenu un gain relatif de 23 % et 20 %, respectivement, comparativement au système de référence.

En outre, nous avons montré l'intérêt de la représentation vectorielle des états pour une éventuelle visualisation graphique. De telles représentations graphiques peuvent être utiles dans différents domaines de recherche sur le traitement de la parole : la phonétique, la phonétique clinique, etc.

7.1.3 Compensation de la variabilité nuisible (chapitre 6)

De nombreuses techniques ont été proposées pour augmenter la robustesse des SRAP, en particulier leur résistance face aux bruits. L'objectif de ces techniques est de compenser les différences entre les conditions d'apprentissage et les conditions d'utilisation du système. Dans cette partie du travail, nous avons proposé une nouvelle méthode de compensation des multiples variabilités nuisibles aux systèmes de RAP. L'estimation et l'isolation du bruit se fait en utilisant l'analyse factorielle dans l'espace des super-vecteurs formé par la concaténation des moyennes des gaussiennes composant le GMM-UBM. Nous avons supposé que les variabilités nuisibles sont additives dans ce domaine et qu'elles sont situées dans un sous-espace de très faible dimension (comparativement à la dimension du super-vecteur).

Nous nous sommes intéressés à la variabilité locuteur et la variabilité canal, ainsi qu'au bruit additif. Nous avons étudié plusieurs scénarios d'estimation et d'isolation de ces variabilités. Dans un premier temps, nous avons testé l'estimation et la compensation indépendante des deux variabilités locuteur et canal. Le gain apporté par cette stratégie ne dépasse pas 0,8 % en absolu. Pour faire face à la présence simultanée de

plusieurs types de variabilités, nous avons proposé une extension du modèle de l'analyse factorielle afin d'estimer et de compenser conjointement la variabilité locuteur et la variabilité canal. Cette méthode a permis d'améliorer la qualité des transcriptions jusqu'à 1,3 % en absolu en termes de taux d'erreur mot par rapport à notre système de base.

Dans un second temps, nous avons orienté notre travail vers le bruit additif. Malgré l'effet non-linéraire du bruit additif sur le signal de parole dans le domaine cepstral, nous avons montré la possibilité de modéliser ce bruit comme une composante additive dans le domaine cepstral. La compensation de cette composante, réalisée également dans le domaine cepstral, a permis d'améliorer la sortie des SRAP de 7,5 % en absolu en termes de taux d'erreur mot.

7.2 Perspectives

La modélisation acoustique compacte fondée sur l'analyse factorielle nous permet d'obtenir un gain en termes de taille des modèles. En se basant sur le modèle acoustique obtenu, nous pouvons également améliorer le temps de calcul lors du processus de décodage. Comme les GMM des états sont dérivés à partir d'un seul GMM-UBM, le calcul des probabilités pour chaque couple trame/état peut être limité à un sous-ensemble de gaussiennes sélectionnées à partir du modèle du monde. Habituellement, pour réaliser cette étape, nous classifions les gaussiennes de tous les états dans le MMC (des dizaines de milliers de gaussiennes), ensuite nous sélectionnons les classes qui possèdent les poids les plus importants.

Dans la deuxième partie expérimentale de cette thèse, nous avons proposé d'adopter les facteurs d'états estimés par l'analyse factorielle, comme représentation vectorielle des états des MMC. Nous nous sommes appuyés sur ces facteurs pour réaliser la procédure de partage des états des MMC dans le cadre de la modélisation acoustique pour les langues peu dotées. Les bons résultats obtenus nous encouragent, dans des travaux futurs, à utiliser les facteurs d'états dans l'étape de *bootstraping* (Osterholtz et al., 1992) précédant l'étape de partage d'états. Le *bootstraping* consiste à obtenir un tableau de correspondances phonétiques (*phone mapping*) entre une ou plusieurs langues sources et la langue cible (la langue peu dotée). Les modèles acoustiques des phonèmes indépendants du contexte de la langue source sont dupliqués afin d'obtenir des modèles acoustiques indépendants du contexte de la langue peu dotée. Dans la littérature, les méthodes les plus utilisées pour réaliser le *bootstraping* sont les méthodes manuelles à base de connaissances linguistiques et phonétiques (Le et Besacier, 2009; Schultz et Waibel, 1998).

L'efficacité due à l'utilisation des facteurs d'états dans la classification des états des MMC, nous motive à utiliser les facteurs d'états dans la construction de tableaux de correspondance entre plusieurs langues sources et la langue peu dotée. La réalisation de ce test nécessite l'estimation des facteurs x_{ph} sur chaque phonème ph pour chaque langue (source et destination). Le fait de représenter les phonèmes par des vecteurs per-

met de mesurer efficacement la similarité entre eux, en utilisant des distances adaptées (distance euclidienne ou distance de Mahalanobis). De cette manière, nous pouvons construire automatiquement la table de correspondance sans disposer nécessairement de connaissances linguistiques et phonétiques.

Dans nos travaux, l'utilisation des facteurs d'états a été restreinte à la procédure de partage dans le cadre de la modélisation acoustique pour les systèmes de RAP. Cependant, cette représentation vectorielle peut avoir des applications dans tous les domaines du traitement de la parole : reconnaissance de la parole, reconnaissance du locuteur, reconnaissance de la langue, synthèse de la parole, etc.

Dans la troisième partie, nous avons proposé une méthode de compensation du bruit additif du signal de la parole dans le domaine cepstral. Dans des prochains travaux, nous proposons de tester notre méthode dans un domaine où le bruit ambiant est additif avec le signal de la parole (*i.e.* le domaine spectral).

Dans l'approche de compensation que nous avons proposée, la variabilité due à l'effet de Lombard n'est pas prise ne compte explicitement. Dans les applications réelles, l'effet de Lombard a toute son importance, notamment dans le cas de RSB faibles (Rajasekaran et al., 1986). Il est donc nécessaire de tenir compte de cette variabilité lors de nos prochains travaux. Néanmoins, cette approche de compensation a été testée dans le cas de la parole bruitée naturellement (donc présence de l'effet Lombard) et a donné de bons résultats.

Bibliographie personnelle

– M. Bouallegue, D. Matrouf, et G. Linarès, 2011a. A simplified subspace gaussian mixture to compact acoustic models for speech recognition. Dans les actes de ICASSP, 4896-4899.

– M. Bouallegue, D. Matrouf, M. Rouvier, et G. Linarès, 2011b. Subspace gaussian mixture models for vectorial hmm-states representation. Dans les actes de ASRU, 512-516.

– M. Rouvier, M. Bouallegue, D. Matrouf, et G. Linarès, 2011. Factor analysis based session variability compensation for automatic speech recognition. Dans les actes de ASRU, 141-145.

– M. Bouallegue, D. Matrouf, M. Rouvier, et G. Linarès, 2012. Subspace gaussian mixture models based on noise compensation for speech recognition. Dans les actes de INTERSPEECH.

– M. Bouallegue, E. Ferreira, D. Matrouf, G. Linarès, M. Goudi, et P. Nocera, 2012. Acoustic modeling for under-resourced languages based on vectorial hmm-states representation using subspace gaussian mixture models. Dans les actes de SLT, 330-335.

Liste des illustrations

Liste des tableaux

Bibliographie

(Acero, 1990) A. Acero, 1990. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Thèse de doctorat, Department of Electrical and Computer Engineering Carnegie Mellon University, Pittsburgh, Pennsylvania.

(Anderson et al., 1994) O. Anderson, P. Dalsgaard, et W. Barry, 1994. On the use of data-driven clustering techniques for language identification of poly and mono-phonemes for four european languages. 121–124. ICASSP, Adelaide.

(Berment, 2004) V. Berment, 2004. *Méthodes pour informatiser des langues et des groupes de langues peu dotées*. Thèse de doctorat, Université J. Fourier - Grenoble I, Grenoble, France.

(Beulen et Ney, 1998) K. Beulen et H. Ney, 1998. Automatic question generation for decision tree based state tying. Dans les actes de *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, USA, 805–809.

(Beyerlein, 1998) P. Beyerlein, 1998. Discriminative model combination. Dans les actes de *ICASSP*, Volume 1. Institute of Electrical Engineers (IEE).

(Bocchieri et Mak, 2001) E. Bocchieri et B. K.-W. Mak, 2001. Subspace distribution clustering hidden markov model. *IEEE Transactions on Speech and Audio Processing* *9*(3), 264–275.

(Chaker, 1984) S. Chaker, 1984. Textes en linguistique berbère : introduction au domaine berbère. 291.

(Chen et Goodman, 1996) S. F. Chen et J. Goodman, 1996. An empirical study of smoothing techniques for language modeling. Dans les actes de *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, Stroudsburg, PA, USA, 310–318. Association for Computational Linguistics.

(Constantinescu et Chollet, 1997) A. Constantinescu et G. Chollet, 1997. On cross-language experiments and data-driven units for alisp. 606–613. Automatic Speech Recognition and Understanding ASRU.

(Davis et Mermelstein, 1980) S. Davis et P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing* *28*(4), 357–366.

(de Mareüil et al., 2000) P. B. de Mareüil, C. Corredor-Ardoy, D. Matrouf, et M. Adda-Decker, 2000. Classement automatique de phonèmes dans un cadre multilingue et application à l'identification de la langue. Dans les actes de *In Proc. ICASSP*.

(Dempster et al., 1977) A. P. Dempster, N. M. Laird, et D. B. Rubin, 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B 39*, 1–38.

(Demuynck et al., 1996) K. Demuynck, J. Duchateau, et D. V. Compernolle, 1996. Reduced semi-continuous models for large vocabulary continuous speech recognition in dutch. Dans les actes de *ICSLP*.

(Dendrinos et al., 1991) M. Dendrinos, S. Bakamidis, et G. Carayannis, 1991. Speech enhancement from noise : A regenerative approach. *Speech Communication 10*(1), 45–57.

(Ephraim et Trees, 1995) Y. Ephraim et H. L. V. Trees, 1995. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing 3*(4), 251–266.

(Fletcher et Munson, 1933) H. Fletcher et W. A. Munson, 1933. Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America 5*(2), 82–108.

(Flitti, 2005) F. Flitti, 2005. *Techniques de réduction de données et analyse d'images multispectrales astronomiques par arbres de Markov*. Thèse de doctorat, Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection,, Strasbourg, France.

(Forney, 1973) G. D. Forney, 1973. The viterbi algorithm. *Proceedings of IEEE 61(3)*, 268–278.

(Furui, 1986) S. Furui, 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing 34*(1), 52–59.

(Gales, 1992) M. J. F. Gales, 1992. An improved approach to the hidden Markov model decomposition of speech and noise. Dans les actes de *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume 1, San Francisco, 233–236.

(Gales et Young, 1996) M. J. F. Gales et S. J. Young, 1996. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing 4*(5), 352–359.

(Galliano et al., 2005) S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, et G. Gravier, 2005. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *INTERSPEECH*, 1149–1152. ISCA.

(Ghani et al., 2005) R. Ghani, R. Jones, et D. Mladenic, 2005. Building minority language corpora by learning to generate web search queries. Volume 7, New York, NY, USA, 56–83. Springer-Verlag New York, Inc.

(Gong et Haton, 1994) Y. Gong et J. Haton, 1994. Stochastic trajectory modeling for speech recognition. Volume 1, 57–60. Dans les actes de IEEE International Conference on Acoustics, Speech and Language Processing, Adelaide, SA, Australia.

(Hermansky, 1990) H. Hermansky, 1990. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am. 57*(4), 1738–52.

(Hermansky et Jr., 1991) H. Hermansky et L. A. C. Jr., 1991. Perceptual linear predictive (plp) analysis-resynthesis technique. Dans les actes de *EUROSPEECH*. ISCA.

(Hermansky et al., 1992) H. Hermansky, N. Morgan, A. Bayya, et P. Kohn, 1992. Rasta-plp speech analysis technique. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on 1*, 121–124.

(Holmes et Sedgwick, 1986) J. N. Holmes et N. C. Sedgwick, 1986. Noise compensation for speech recognition using probabilistic models. Dans les actes de *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 741–744.

(Huang et al., 2001) C. Huang, T. Chen, S. Z. Li, E. Chang, et J.-L. Zhou, 2001. Analysis of speaker variability. Dans les actes de *INTERSPEECH*, 1377–1380. ISCA.

(Huang et al., 1989) X. D. Huang, H. W. Hon, et K. F. Lee, 1989. Large-vocabulary speaker-independent continuous speech recognition with semi-continuous hidden markov models. Dans les actes de *in Proc. European Conference on Speech Communication and Technology*, 163–166.

(IPA, 1999) IPA, 1999. *Handbook of the International Phonetic Association : A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

(Jelinek, 1976) F. Jelinek, 1976. Continuous speech recognition by statistical methods. Volume 4, 536–556. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP.

(Jimenez et Landgrebe, 1995) L. O. Jimenez et D. Landgrebe, 1995. High dimensional feature reduction via projection pursuit. Rapport technique.

(Jolliffe, 1986) I. T. Jolliffe, 1986. *Principal component analysis*. New York : Springer.

(Kenny, 2006) P. Kenny, 2006. Joint factor analysis of speaker and session variability : Theory and algorithms. Rapport technique.

(Kenny et al., 2005a) P. Kenny, G. Boulianne, et P. Dumouchel, 2005a. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing 13*(3), 345–354.

(Kenny et al., 2005b) P. Kenny, G. Boulianne, P. Ouellet, et P. Dumouchel, 2005b. Factor Analysis Simplified. Dans les actes de *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005)*, Volume 1, 637–640.

(Kneser et Ney, 1995) R. Kneser et H. Ney, 1995. Improved backing-off for m-gram language modeling. Dans les actes de *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume I, Detroit, Michigan, 181–184.

(Le et Besacier, 2009) V. B. Le et L. Besacier, 2009. Automatic speech recognition for under-resourced languages : Application to vietnamese language. *IEEE Transactions on Audio, Speech and Language Processing 17*(8), 1471–1482.

(Lee et al., 1990) K. Lee, X. D. Huang, et H. Hon, 1990. On semicontinuous hidden markov modeling. Volume Vol. 1, 689–692.

(LEVY, 2006) C. LEVY, 2006. *Modèles acoustiques compacts pour les systèmes embarqués*. Thèse de doctorat, l'Université d'Avignon et des Pays de Vaucluse, Avignon, France.

(Lim, 1978) J. Lim, 1978. Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. *IEEE Trans. Acoustic Speech Signal Processing ASSP*, 471–472.

(Liu et Fung, 2005) Y. Liu et P. Fung, 2005. Acoustic and phonetic confusions in accented speech recognition. Dans les actes de *INTERSPEECH*, 3033–3036. ISCA.

(Mak et Barnard, 1996) B. Mak et E. Barnard, 1996. Phone clustering using the bhattacharyya distance. Dans les actes de *ICSLP*. ISCA.

(Mari et al., 1997) J.-F. Mari, J. P. Haton, et A. Kriouile, 1997. Automatic word recognition based on second-order hidden markov models. *IEEE Transactions on Speech and Audio Processing 5*(1), 22–25.

(Markel et Gray, Jr., 1976) J. D. Markel et A. H. Gray, Jr., 1976. *Linear prediction of speech*. Berlin-Heidelberg-New York : Springer-Verlag.

(Matrouf et al., 2007) D. Matrouf, N. Scheffer, B. G. B. Fauve, et J.-F. Bonastre, 2007. A straightforward and efficient implementation of the factor analysis model for speaker verification. Dans les actes de *INTERSPEECH*, 1242–1245.

(Morii et al., 1990) S. Morii, T. Morii, M. Hoshimi, S. Hiraoka, T. Watanabe, et K. Niyada, 1990. Noise robustness in speaker independent speech recognition. Dans les actes de *ICSLP*.

(N. Thi, 2006) M. H. N. Thi, 2006. *Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens*. Thèse de doctora, Université Henri Poincaré, Nancy, France.

(Ney et al., 1992) H. Ney, R. Haeb-Umbach, B.-H. Tran, et M. Oerder, 1992. Improvements in beam search for 10000-word continuous speech recognition. Dans les actes de *IEEE ICASSP-92*, San Francisco, CA, I.9–12. IEEE.

(Nocera et al., 2004) P. Nocera, C. Fredouille, G. Linarès, D. Matrouf, J.-F. Bonastre, , et F. Béchet, 2004. Lia's french broadcast news transcription system. Dans les actes de *SWIM : Lectures by Masters in Speech Processin*, Maui, Hawaii, 14–17.

(Nocera et al., 2002) P. Nocera, G. Linares, D. Massonié, et L. Lefort, 2002. Phoneme lattice based A* search algorithm for speech recognition. Dans les actes de *TSD*, 301–308.

(Osterholtz et al., 1992) L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, et A. Waibel, 1992. Testing generality in janus : A multilingual speech translation system.

(Povey et al., 2010) D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, et S. Thomas, 2010. Subspace gaussian mixture models for speech recognition. Dans les actes de *ICASSP*, 4330–4333.

(R. Beaufort, 2006) A. R. R. Beaufort, 2006. système de synthèse de la parole à orientation linguistique. 509–512. Proceedings of JEP 2006.

(Rajasekaran et al., 1986) P. Rajasekaran, G. Doddington, et J. Picone, 1986. Recognition of speech under and in noise. *ICASSP*, 733–736.

(Reichl et Chou, 1998) W. Reichl et W. Chou, 1998. Decision tree state tying based on segmental clustering for acoustic modeling. Dans les actes de *icassp*, Volume 2, 801–804.

(Roe, 1987) D. B. Roe, 1987. Speech Recognition with a Noise-Adapting Codebook. Dans les actes de *Proceedings of the IEEE ICASSP-87*, 1139–1142.

(Russell, 1993) M. Russell, 1993. A segmental hmm for speech pattern modelling. Dans les actes de *Proc. ICASSP*, Volume II, 499–502.

(San-segundo et al., 2001) R. San-segundo, B. Pellom, K. Hacioglu, W. Ward, et J. Pardo, 2001. Confidence measures for spoken dialogue systems. Volume 1, 393–396. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP.

(Sarikaya et al., 2005) R. Sarikaya, Y. Gao, M. Picheny, et H. Erdogan, 2005. Semantic confidence measurement for spoken dialog systems. *IEEE Transactions on Speech and Audio Processing 13*(4), 534–545.

(Scannell, 2003) K.-P. Scannell, 2003. Automatic thesaurus generation for minority languages : an irish example. Volume Vol. 2, 11–14. TALN, Batz-sur-Mer, France.

(Schultz et Waibel, 1998) T. Schultz et A. Waibel, 1998. Multilingual and crosslingual speech recognition. Dans les actes de *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 259–262.

(Schultz et Waibel, 2001) T. Schultz et A. Waibel, 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication 35*(1-2), 31–51.

(Senay, 2011) G. Senay, 2011. *Approches semi-automatiques pour la recherche d'information dans les documents audios*. Thèse de doctorat, l'Université d'Avignon et des Pays de Vaucluse, Avignon, France.

(Singh et al., 1999) R. Singh, B. Raj, et R. M.Stern, 1999. Automatic clustering and generation of contextual questions for tied states in hidden markov models. Volume vol. 1, 117–120. International Conference on Spoken Language Processing (ICSLP).

(Soong et Huang, 1991) F. K. Soong et E.-F. Huang, 1991. A tree-trellis based fast search for finding the n-best sentence hypotheses in continuous speech recognition. Dans les actes de *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume 1, Toronto, 705–708.

(Suaudeau et Andé-Obrecht, 1994) N. Suaudeau et R. Andé-Obrecht, 1994. An efficient combination of acoustic and suprasegmental informations in a speech recognition system. Volume 1, 65–68. ICASSP, Adelaïde.

(Tachbelie et al., 2012) M. Tachbelie, S. Teferra Abate, et L. Besacier, 2012. Using different acoustic, lexical and language modeling units for ASR of an under-resourced language - Amharic. *Speech Communication Journal Vol. 56 - Special Issue on Processing Under-Resourced Languages*, 181–194. (Impact-F 1.28 estim. in 2012).

(Tipping et Bishop, 1999) M. E. Tipping et C. M. Bishop, 1999. Mixtures of probabilistic principal component analyzers. *Neural Comput. 11*(2), 443–482.

(Tran, 2003) D.-D. Tran, 2003. Building a large vietnamese speech database. Rapport de master tic, Vietnam.

(Vaseghi et Milner, 1992) S. Vaseghi et B. P. Milner, 1992. Speech recognition in noisy environments. Dans les actes de *ICSLP*. ISCA.

(Viterbi, 1967) A. Viterbi, 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on 13*(2), 260 –269.

(Wellekens, 1987) C. Wellekens, 1987. Explicite time correlation in hidden markov models for speech recognition. 384–386. Proc. ICASSP, Dallas.

(Woszczyna, 1998) M. Woszczyna, 1998. *Fast speaker independant large vocabulary continuous speech recognition*. Thèse de doctorat, de l'université de Karlsruhe, Allemagne.

(Young, 1992) S. J. Young, 1992. The general use of tying in phoneme-based hmm speech recognisers. Dans les actes de *ICASSP*, 569–572.

(Young et al., 1994) S. J. Young, J. J. Odell, et P. C. Woodland, 1994. Tree-based state tying for high accuracy modelling. Dans les actes de *HLT*.

(Young et Woodland, 1993) S. J. Young et P. C. Woodland, 1993. The use of state tying in continuous speech recognition. Dans les actes de *EUROSPEECH*. ISCA.