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

THÈSE

présentée pour obtenir le grade de Docteur en Sciences
de l'Université d'Avignon et des Pays de Vaucluse

SPÉCIALITÉ : Informatique

Modèles acoustiques compacts pour les systèmes embarqués

par **Christophe LÉVY**

Soutenue le 30 novembre 2006 devant un jury composé de :

M.	Renato DeMori	Professeur, LIA, Avignon	Président du jury
M.	Paul Deléglise	Professeur, LIUM, Le Mans	Rapporteurs
M.	Jean-Paul Haton	Professeur, LORIA, Nancy 2	
M.	Laurent Besacier	MdC, CLIPS, Grenoble	Examinateurs
M.	Stéphane Pineau	Senior Software Engineer	
M.	Jean-François Bonastre	MdC HDR, LIA, Avignon	Directeurs de thèse
M.	Georges Linarès	MdC, LIA, Avignon	

École Doctorale Sciences et Agronomie
Laboratoire Informatique d'Avignon

à mon frère ...

Remerciements

Par ces quelques lignes, je tiens à remercier tous les membres de mon jury : Paul Deléglise et Jean-Paul Haton qui ont accepté la charge d'en être les rapporteurs ainsi que Renato De Mori, Stéphane Pineau (sans qui je n'aurais pas fait de thèse) et Laurent Besacier pour leur participation à ce jury en tant qu'examinateurs.

Je tiens aussi à remercier Jean-François Bonastre et Georges Linarès pour leur encadrement permanent durant ces travaux. Au passage, un petit conseil pour les actuels et futurs thésards de JF. . . Il faut toujours laisser gagner le chef au bowling... Pour ceux qui décident de se lancer avec Georges, je leur conseille soit de toujours refuser de faire du sport avec lui soit d'avoir une bonne paire de chaussures. . .

Un grand merci à Pascal Nocera qui m'a obtenu un financement pour ces travaux.

D'une façon plus générale, j'ai aussi apprécié les conditions de travail offertes par le LIA et l'IUP ; pour cela je remercie tous les membres de ces deux institutions.

Je remercie également tous les thésards, présents et passés, de l'équipe pour l'ambiance qu'ils ont entretenue : Domi (rectificatif par rapport à ton propre commentaire . . . tes blagues sont lourdes), Teva (grâce à toi j'aurai au moins une jolie figure dans ce manuel), Sylvain, Corinne (même si on parle ici de la pré-histoire), Yannick, Alex (au fait on court demain ?), Ben (1&2), Nico, Will, Anthony (grâce à qui j'aurais pu avoir une autre figure. . . que j'ai supprimée au dernier moment). . . et tous ceux que j'ai oubliés, qu'ils me pardonnent.

Enfin, je tiens à remercier l'ensemble des membres de ma famille qui m'ont tous apporté, à un moment ou un autre, un soutien pour en arriver là aujourd'hui. J'embrasse plus particulièrement mes parents qui m'ont toujours laissé faire mes choix, même quand ils n'étaient pas forcément bons. Et enfin, un grand merci à Charlotte qui me supporte depuis près de 10 ans ... mais encore

plus pendant ces derniers mois.

Résumé

Depuis le lancement des téléphones portables au milieu des années 90, leurs ventes n'ont cessé de progresser. Leur taille, comme celle de l'ensemble des systèmes embarqués (téléphone, GPS, PDA. . .), a constamment été réduite, quand, dans le même temps, le nombre de services offerts n'a fait qu'augmenter. D'une manière générale, la plupart des systèmes embarqués offre aujourd'hui une interface homme-machine complexe et peu conviviale. L'intégration d'un moteur de reconnaissance de la parole dans ces systèmes offre une voie intéressante pour améliorer leur ergonomie.

Cette thèse s'inscrit dans le cadre de la Reconnaissance Automatique de la Parole (RAP) intégrée dans les systèmes embarqués. Les ressources disponibles dans ces systèmes sont nettement inférieures à celles des ordinateurs généralement utilisés pour la RAP, tant du point de vue de la puissance de calcul que de la quantité de mémoire. Les travaux que nous présentons s'inscrivent dans cette problématique de la RAP en situation de ressources réduites et plus particulièrement dans le cadre de la réduction de la taille des modèles acoustiques.

En RAP les unités phonétiques sont, généralement, représentées par des modèles de Markov cachés gauche-droit à trois états. Afin d'améliorer les performances des systèmes, la tendance va vers l'utilisation de modèles contextuels et vers l'apprentissage de GMM complexes pour la modélisation acoustique. Cette approche nécessite une quantité de mémoire très importante qui n'est pas en adéquation avec les ressources disponibles dans les systèmes embarqués.

Dans ce travail, nous présentons une approche alternative dans laquelle une seule mixture de gaussiennes (le GMM général) représente l'ensemble de l'espace acoustique. Chaque état est ensuite estimé relativement au GMM général par une transformation, simple et compacte. Deux techniques sont proposées pour estimer les transformations permettant de caractériser les fonctions de densité de probabilité des différents états. Dans un premier temps, nous proposons de ré-estimer le poids de chacune des composantes du GMM général avec un critère maximisant la vraisemblance ou avec un critère discriminant.

Ensuite, nous présentons une seconde fonction de transformation combinant une transformation linéaire et globale du GMM général (par modification des moyennes et variances) et la ré-estimation des poids citée précédemment.

Cette approche permet un gain important en terme de compacité des modèles acoustiques en ne nécessitant que le stockage du GMM général et des différentes transformations. De plus, l'architecture présentée autorise une adaptation rapide de l'ensemble des modèles acoustiques à un nouvel environnement ou un nouveau locuteur, simplement par l'adaptation du GMM général (sans modifier les fonctions de transformations).

Nous évaluons nos méthodes sur deux tâches : la reconnaissance de chiffres isolés (environnement acoustique propre) et la reconnaissance de commandes vocales (environnement acoustique bruité). En comparant notre approche avec un système de référence, nous obtenons des performances significativement meilleures. L'architecture présentée permet une diminution du taux d'erreur quelle que soit la fonction de transformation utilisée (simple ré-estimation des poids ou transformation du GMM général puis ré-estimation des poids). La diminution relative du taux d'erreur peut atteindre les 55%. En ce qui concerne l'adaptation au locuteur ou à l'environnement acoustique par une simple adaptation du GMM général, les expériences réalisées montrent une diminution du taux d'erreur similaire à celle obtenue dans la littérature pour les systèmes de RAP classiques alors que notre méthode ne nécessite que peu de données d'adaptation et de ressources de calcul. Cette phase d'adaptation nous a permis de réduire de 9% à 12% (en relatif) les taux d'erreur obtenus avec la méthode présentée dans cette thèse.

Sommaire

Introduction

La téléphonie mobile est apparue au milieu des années 90. Depuis le début, la croissance de ce marché n'a jamais faibli, elle a même affiché des taux de progression impressionnants. Par exemple, le taux de croissance a atteint 83%[1] en 1999 et le marché a encore progressé (+4%[2]) au plus fort de la crise des nouvelles technologies de 2002.

Aujourd'hui, téléphoner au moyen d'un mobile est devenu un geste courant. Le marché des téléphones portables est toujours en plein essor : au 30 juin 2006, 81,3%[3] des français possédaient un téléphone portable. La croissance annuelle du marché de la téléphonie mobile, entre juin 2005 et juin 2006, dépasse 10%[4].

La nature mobile de ces appareils conduit les fabricants à réduire, sans cesse, leurs dimensions. Leur manipulation devient de plus en plus difficile, d'autant que le nombre croissant des applications disponibles (agenda, réveil, jeux...) rend l'interface moins intuitive.

Devant l'attention nécessaire pour l'utilisation de ces téléphones et devant la croissance de ce marché, le législateur a décidé de réglementer l'usage des mobiles au volant. Une loi a vu le jour en mars 2003 précisant que "l'usage d'un téléphone *tenu en main* par le conducteur d'un véhicule en circulation est interdit" (article R.412-6-1 du code de la route). Pour pallier ce problème, une généralisation de l'usage des kits "mains libres" (qui permettent l'utilisation des téléphones portables au volant) est observée. Un inconvénient subsiste : l'utilisateur a toujours besoin du clavier pour les fonctionnalités de base (numéroter/décrocher) comme pour les applications avancées. Intégrer une interface vocale en complément du traditionnel clavier représente une solution qui pourrait permettre de limiter (voir supprimer) l'usage du clavier.

De plus, l'intégration de la reconnaissance de la parole dans les mobiles per-

[1] source ARCEP - `http://www.art-telecom.fr`

[2] source ARCEP

[3] 49,07 millions de téléphones portables pour 60,32 millions de français (source ARCEP 30 juin 2006)

[4] en relatif (source ARCEP 30 juin 2006), soit 6,4% en absolu

mettrait d'ajouter de nouvelles fonctionnalités telles que :

- le "name dialing" : l'utilisateur n'a plus qu'à prononcer le nom d'une personne se trouvant dans son répertoire pour l'appeler ; la phase de recherche du nom disparaît alors ;
- la reconnaissance de mots clés : elle peut éviter à l'utilisateur l'usage du clavier et de l'écran pour décrocher lors d'un appel entrant ;
- la numérotation automatique : dans le cas où la personne à appeler n'appartient pas au carnet d'adresses, l'utilisateur peut dicter le numéro plutôt que de le composer à la main ;
- la reconnaissance de "commandes vocales" : l'utilisateur peut donner un certain nombre d'ordres pouvant être interprétés par le système (*e.g.* "composer numéro", "ajouter rendez-vous") ;
- ...

Ce travail traite de l'intégration d'un système de reconnaissance dans les systèmes embarqués. Cette problématique, aujourd'hui au cœur de tous les laboratoires de recherche des concepteurs de téléphones, est devenu un argument clé dans la compétition industrielle. Plusieurs firmes proposent déjà ce type de service mais, généralement, il ne fonctionne que dans de bonnes conditions acoustiques.

Cette thèse s'inscrit dans le cadre d'un partenariat (convention CIFRE) entre le Laboratoire Informatique d'Avignon[5] (LIA) et l'entreprise Stepmind SA[6].

Le domaine d'activité de Stepmind est la conception et le développement des circuits intégrés et des modules qui, grâce à des technologies de communication sans fil, permettent à toutes sortes d'objets d'échanger entre eux toutes sortes de données. Son ambition est, notamment, d'assurer la convergence entre les possibilités offertes par les réseaux GSM, poussés à leurs limites technologiques, et celles apportées par les nouveaux standards de réseaux locaux sans-fil. C'est pourquoi, depuis sa création, Stepmind mise sur les technologies d'accélération de la distribution de contenus basées sur les standards 802.11[a|b|g|i] et les technologies autour du GSM (notamment l'EDGE qui est présentée comme une alternative à l'UMTS). Un projet est en cours d'étude pour la création d'un chipset intégrant les fonctionnalités de traitement numérique et analogique du signal. L'objectif est d'intégrer un système de reconnaissance sur ce même chipset.

Le LIA pour sa part est un laboratoire de recherche de l'Université d'Avignon et des Pays de Vaucluse. Ses recherches s'orientent autour de 3 axes principaux : le traitement du langage naturel (dialogue, parole, texte), des réseaux et applica-

[5] http://www.lia.univ-avignon.fr
[6] http://www.stepmind.com

tions multimédias et enfin de la recherche opérationnelle. Dans le domaine oral, les compétences sont essentiellement orientées autour de la reconnaissance de la parole et de la reconnaissance du locuteur.

Cette thèse s'inscrit dans la reconnaissance automatique de la parole pour les systèmes embarqués. Une telle intégration est confrontée à trois problèmes majeurs :

- la puissance de calcul : globalement les systèmes embarqués disposent de peu de puissance ;
- la mémoire : les systèmes de RAP nécessitent généralement une grande quantité de mémoire quand celle-ci n'est disponible qu'en petite quantité dans les systèmes mobiles ;
- l'environnement acoustique variable : les systèmes mobiles sont utilisés dans des environnements acoustiques plutôt instables : le bruit de fond d'un bureau varie peu, alors qu'un système embarqué peut être utilisé dans la rue, au bureau ou dans une voiture.

Durant ce travail, nous avons pris le parti d'étudier essentiellement les contraintes mémoires.

Ce manuscrit contient quatre parties principales.

La première partie présente le contexte général. Dans un premier temps, le monde de l'embarqué et les contraintes qui lui sont propres sont présentés. A la suite de cela, nous détaillons le contexte général de la Reconnaissance Automatique de la Parole (RAP). Dans cette étude, nous avons choisi d'utiliser le téléphone portable comme exemple de système embarqué. Nous présentons donc les spécifications d'un système de RAP dans ce contexte. Cette partie se termine par un présentation des principales approches pour la réduction de la taille des modèles acoustiques.

La deuxième partie concerne le cœur de cette thèse. Après une rapide présentation des corpus utilisés, nous présentons plusieurs solutions pour réduire la taille des modèles acoustiques. L'hypothèse que nous mettons en avant dans ce chapitre est que les modèles acoustiques peuvent être définis les uns par rapport aux autres. En effet, nous proposons tout d'abord de modéliser l'espace acoustique dans son intégralité grâce à un seul et unique GMM. Ensuite, les modèles de chaque unité acoustique sont dérivés de ce GMM. Enfin, nous présentons une approche permettant d'adapter l'ensemble des modèles acoustiques par une simple adaptation du GMM initialement appris.

La troisième partie présente quelques conclusions et perspectives relatives au travail réalisé.

Enfin, en quatrième partie, le lecteur trouvera les annexes auxquelles il est

fait référence tout au long de ce document.

Première partie

Contexte

Chapitre 1

Le monde de l'embarqué

Sommaire

Résumé

Aujourd'hui, les systèmes embarqués (et notamment les téléphones portables) sont à la disposition de tous. Ces systèmes électroniques, bien que de plus en plus perfectionnés, ont des ressources très limitées en comparaison des ordinateurs classiques (même portables). Dans ce chapitre, nous détaillons les principales contraintes des systèmes embarqués.

Aujourd'hui, une très grande partie de la population, active ou non, possède un système embarqué. En effet, des très jeunes enfants aux seniors en passant par les adolescents, tout le monde possède un système embarqué. Dès leur plus jeune âge, parfois sans le savoir, les enfants peuvent utiliser de tels systèmes : jouets (par exemple, les nouvelles poupées qui parlent et répondent intègrent un module de reconnaissance vocale), téléphones portables, lecteurs MP3.... De même, la plupart des seniors possèdent également leur système embarqué : assistants personnels (type PDA), systèmes de positionnement par satellite (GPS - qu'il soit pour voiture, bateau, ou même pour des randonneurs) ou, tout simplement, téléphone portable.

Avec la course à la miniaturisation des téléphones (ou des systèmes embarqués dans leur ensemble), l'ajout de nouvelles fonctionnalités pose des problèmes en terme d'ergonomie. La reconnaissance vocale peut donc apparaître comme une solution envisageable (parmi d'autre) pour améliorer l'interface homme-machine. Cependant, les systèmes embarqués, qui sont par essence mobiles, sont utilisés dans des environnements acoustiques pouvant être très variables (ce qui n'est pas sans compliquer la reconnaissance comme nous le verrons plus tard). Enfin, les ressources disponibles dans un système embarqué ne sont pas (encore) les mêmes que celles dont disposent les serveurs de calcul classiquement utilisés pour la RAP.

Nous allons aborder ces trois points en détail dans les sections suivantes.

1.1 Ergonomie

Tous les systèmes embarqués montrent une tendance commune : la miniaturisation. La taille des téléphones portables, par exemple, a pratiquement été divisée par 2 (*cf.* figure 1.1), en près de 10 ans. La plupart des téléphones disponibles aujourd'hui avoisinent les 8 cm de long alors qu'il y a seulement deux ou trois ans ils mesuraient plus de 13 cm. Il en est de même pour leur poids : en 1992, les premiers téléphones pesaient près de 500 grammes alors que, maintenant, ils oscillent entre 70 et 100 grammes. Aujourd'hui, lors de l'achat d'un téléphone, il convient de vérifier que le clavier correspond à peu près à la taille des doigts de l'utilisateur. Cette course à la miniaturisation entraîne une manipulation de plus en plus délicate de ces appareils.

De plus, si les téléphones portables de première génération ne proposaient que peu de services (tels que la mémorisation de quelques numéros de téléphone), les nouvelles générations fournissent, pour leur part, beaucoup plus de fonctionnalités. Ces nouveaux appareils permettent de télécharger des son-

FIG. 1.1 – Evolution de la taille des téléphones

neries, de jouer, de gérer un agenda... Ces nouvelles fonctionnalités nécessitent souvent une interface complexe.

Depuis quelques mois, le marché fait émerger un nouveau type de systèmes embarqués qui concilie les fonctionnalités d'un téléphone et celle d'un PDA. Ceci permet aux téléphones de devenir de véritables outils de bureautique (traitement de texte, tableur, logiciel de présentation...). Le dernier-né de chez Hewlett-Packard est une parfaite illustration de cette évolution : l'iPAQ h6315 offre tous les services d'un PDA (agenda, rédaction/visualisation de documents, messagerie internet...) et d'un téléphone (*cf.* 1.2). Pour pouvoir rédiger un document, l'utilisation du stylet est peu envisageable, d'où l'intégration d'un clavier. Cependant, il faut noter que ce PDAphone ne mesure que 7cm de large et que 10 boutons ont été installés sur ces 7cm pour obtenir un clavier. Au final, ce clavier ne répond que très partiellement aux attentes des utilisateurs.

FIG. 1.2 – Exemple de PDA permettant de téléphoner : l'iPAQ h6315 de chez HP

Ces contraintes ergonomiques, amplifiées par la miniaturisation des téléphones, expliquent pourquoi le "name dialing"[1], la reconnaissance de chiffres ou le pilotage par la voix deviennent/deviendront des applications standards.

1.2 Environnement acoustique

Les performances des systèmes de RAP sont directement liées à l'environnement acoustique dans lequel ils sont utilisés ou, du moins, à leur capacité à s'y adapter. Ce point est démontré aisément au regard des évolutions des performances mesurées lors des différentes campagnes d'évaluation de systèmes de transcription de la parole.

En 1997, la première campagne d'évaluation nationale a été réalisée par l'AUPELF[2]. Elle a organisé des Actions de Recherche Concertées (ARC) pour évaluer les systèmes de reconnaissance de parole continue, grand vocabulaire (plus de 20 000 mots), indépendants du locuteur (ARC-B1 [Dolmazon 1997]). Les évaluations portaient sur de la parole lue, enregistrée dans de très bonnes conditions (salle insonorisée). Le taux d'erreur en mots (Word Error Rate) des systèmes de dictée vocale de l'époque (1997) se situait entre 40% et 50% ([Fohr 1997], [Savariaux 1997], un seul sortant du lot avec un WER proche de 12% ([Adda 1997]).

Aujourd'hui, les progrès scientifiques et technologiques permettent de s'attaquer à des conditions acoustiques plus difficiles. En janvier 2003, un groupement tripartite (l'AFCP[3], la DGA[4] et ELRA[5]) a lancé la seconde évaluation nationale : la campagne ESTER ([Gravier 2004]). Une des tâches de cette campagne est la transcription orthographique d'émissions radiophoniques. Les conditions acoustiques de cette nouvelle évaluation sont beaucoup plus difficiles que celle de 1997, mais beaucoup plus proches d'une application réelle de RAP. Dans le cadre d'ESTER, les résultats varient entre 11,9% et 61,9% [Galliano 2005].

Parallèlement, le NIST (National Institute of Standards and Technology) organise chaque année des évaluations dans le domaine de la reconnaissance de la parole[6] incluant des tâches variables comme la transcription enrichie (Rich Transcription) ou la recherche de mots clés (Spoken Term Detection).

[1] reconnaissance vocale de nom ou prénom
[2] http://www.auf.org/
[3] http://www.afcp-parole.org
[4] http://www.defense.gouv.fr/dga/
[5] http://www.elra.info
[6] http://www.nist.gov/speech/

Les avancées dans le domaine de la reconnaissance de la parole sont clairement visibles au vu de ces résultats. En effet, les performances lors des deux campagnes d'évaluations (ESTER et ARC B1) sont similaires en terme de WER alors que le contexte acoustique d'ESTER est nettement plus complexe (émissions radiophoniques, conversation téléphonique). Au regard de ces WER, relativement élevés, la reconnaissance embarquée semble une tâche bien complexe.

Notamment, dans le cadre de la RAP sur des systèmes embarqués, il n'est plus possible de parler d'un environnement acoustique unique ; plusieurs environnements acoustiques doivent être pris en compte. En effet, l'utilisateur doit pouvoir utiliser la reconnaissance quel que soit l'endroit où il se trouve : dans son bureau (où l'environnement peut être supposé plus calme), dans sa voiture, dans la rue (avec les différentes sources de bruits - voiture, métro, bus)...

Pour palier la multiplicité des environnements acoustiques, plusieurs solutions sont envisageables. La première solution, exhaustive, consiste à embarquer plusieurs modèles acoustiques : un pour chaque environnement. Une autre approche est concevable : obtenir un modèle adaptable très rapidement. La première approche semble plus complexe à intégrer dans les systèmes embarqués car elle nécessite plus d'espace mémoire afin de stocker un modèle par environnement.

1.3 Ressources

Une contrainte essentielle d'un système embarqué réside bien évidemment dans son autonomie. En effet, quel que soit le nombre de fonctionnalités que peuvent offrir de tels systèmes, si ceux-ci doivent être reliés à une prise électrique pour pouvoir fonctionner, ils perdent la notion de mobilité et donc une part importante de leur intérêt. L'autonomie dépend principalement de deux aspects : la batterie et les composants électroniques. Bien que d'importants progrès aient été réalisés dans le domaine des batteries, notamment depuis la généralisation des batteries au *lithium*, le choix des composants (ainsi que leurs performances) reste primordial.

Cette thèse est réalisée en collaboration avec une entreprise qui fabrique des puces électroniques de téléphone portable. Nous utiliserons donc les téléphones portables (principalement leurs contraintes) comme exemple d'intégration. Pour permettre au lecteur de mieux comprendre le fonctionnement d'un tel système embarqué, nous présentons ci-après (*cf.* figure 1.3) le schéma de principe d'un téléphone portable.

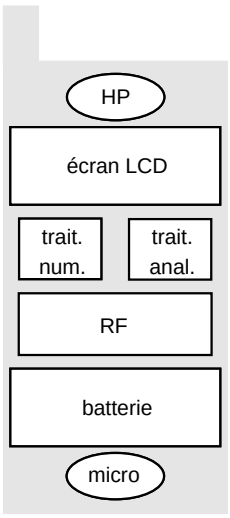

FIG. 1.3 – Architecture d'un téléphone portable

Les chiffres présentés dans 1.3.1 et 1.3.2 correspondent aux caractéristiques techniques des puces conçues par Stepmind (ou en projet) dans le cadre de ses activités. Ils ont pour unique but de donner un ordre de grandeur au lecteur.

1.3.1 Mémoires

Il existe différentes mémoires dans un téléphone portable :
- les mémoires dédiées au programme (les "Program ROM/RAM"). Le code est forcément exécuté depuis ce type de mémoire. Le bus "program" a souvent une bande passante inférieure à un bus de données. Donc, généralement, ces types de mémoires sont plutôt destinés à charger des segments de code ou à écrire quelques données d'initialisation. Elles peuvent aussi être utilisées pour les données, cependant leur emploi n'est pas très aisé ;
- les mémoires dédiées aux données (les "Data ROM/RAM"). Ces mémoires sont très simples d'accès et possèdent plusieurs modes d'adressage. Elles sont faites pour effectuer les calculs, sauvegarder les tableaux de coefficients pré-calculés pour les tables de *sinus*, *cosinus*, *logarithme*...

Dans chaque cas ("Data" ou "Program"), nous disposons de mémoire vive (RAM) et de mémoire morte (ROM). Les deux principales différences entre la RAM et la ROM sont :
- la nature de leur accès. En effet, les mémoires ROM ne sont pas ré-inscriptibles, *i.e.* elles sont écrites en usine une fois pour toutes. C'est, par exemple, dans les mémoires ROM que seront inscrites les tables de

cosinus ;
- la surface qu'elles occupent sur le circuit. En supposant que n Ko de RAM occupent une surface s, on peut estimer que la même quantité de ROM occupera une surface comprise entre $s/2$ et $s/3$.

Le tableau 1.1 indique les différentes quantités de mémoires disponibles sur la puce de Stepmind (en développement en 2004). Ces quantités de mémoire doivent permettre l'intégration de toutes les technologies nécessaires pour un téléphone, notamment les codecs de compression de parole, les algorithmes nécessaires à la transmission du signal de parole, les éventuels codecs pour la décompression MP3 et les algorithmes de décodage.

	quantité de mémoire
Program ROM	96 Ko
Program RAM	128 Ko
Data ROM	88 Ko
Data RAM	34 Ko

TAB. 1.1 – Mémoires disponibles sur la puce développée par Stepmind

La RAP étant la dernière technologie implémentée dans la puce de Stepmind, les quantités de mémoires disponibles sont relativement faibles. On peut estimer qu'il reste moins d'une quinzaine de Ko en ROM et autant en RAM ("Program" et "Data" confondus).

1.3.2 Processeurs

Comme évoqué en introduction (*cf.* 1.3), l'autonomie est l'une des caractéristiques principales d'un système embarqué. Elle est liée essentiellement à deux facteurs : la batterie et le choix des composants. Nous nous intéressons ici au second facteur (le choix des composants) et plus particulièrement au processeur du module de traitement numérique du signal.

Ce module est généralement composé non pas d'un seul processeur mais d'un processeur dédié aux calculs et d'un processeur dédié au traitement numérique du signal (DSP). Le DSP a pour avantage de posséder des instructions de base dédiées au traitement numérique du signal ainsi que des routines optimisées.

Plus ces processeurs vont être puissants, plus l'autonomie du système sera faible. L'exécution des instructions étant coûteuse en énergie, ceci explique que toutes les boucles et morceaux de code sont optimisés au maximum. Les concep-

teurs de chipset sont toujours à la recherche du moindre gain de MIPS.

Dans le cadre du projet de Stepmind, le processeur dédié au calcul pourrait être un ARM7[7] tournant à 100 MHz et le DSP un PalmDSP[8] ayant lui aussi à une fréquence voisine des 100 MHz.

1.4 Conclusion

Dans un premier temps, nous avons présenté le problème d'ergonomie des téléphones. En effet, depuis leur lancement la taille des téléphones n'a cessé de diminuer alors que dans le même temps le nombre de services offerts a nettement augmenté (gestion de mails, gestion de l'agenda, jeux. . .). La RAP apparaît comme une alternative permettant d'améliorer l'ergonomie de ces systèmes ; l'ajout d'un clavier plus complet permettrait aussi d'améliorer l'ergonomie mais augmenterait aussi la taille des téléphones.

Nous avons, ensuite, présenté les deux principales contraintes liées aux systèmes embarqués : un environnement acoustique variable et des ressources limitées.

En comparant les performances des systèmes de RAP lors d'évaluations nationales (ARC-B1 1997 et ESTER 2003), nous constatons que les taux d'erreurs sont similaires malgré les différences dans les tâches réalisées. En effet, alors que les enregistrements étaient réalisés dans une chambre sourde pour la campagne ARC-B1, ceux d'ESTER étaient des enregistrements d'émissions radiophoniques. Cette constatation montre les gains en performance réalisés durant l'intervalle de temps séparant ces deux campagnes. Dans le cadre de la RAP embarquée sur les téléphones portable, l'environnement acoustique peut être bruyant (lors d'une utilisation dans la rue par exemple) mais aussi changeant (passage de la rue à l'intérieur de la voiture) ; ce qui complexifie encore la reconnaissance, par comparaison avec ESTER.

Enfin, les ressources des systèmes embarqués sont très limitées. En prenant l'exemple de la puce développée par Stepmind, nous avons montré qu'un téléphone portable dispose de peu de ressources de calcul (un DSP et un CPU à 100MHz) et mémoire (moins de 200Ko). Les SRAP nécessitant généralement des ordinateurs de dernière génération (avec un processeur de plusieurs GHz et plusieurs Go de mémoire), les capacités d'un téléphone portable sont donc très éloignés de celles généralement requises par les SRAP classiques.

[7] http://www.arm.com/products/CPUs/families/ARM7Family.html
[8] http://www.dspg.com/technology/dsp_licensing_overview.html

Embarqué un SRAP n'est donc pas une tâche triviale et nécessite une importante refonte des algorithmes.

Chapitre 2

Reconnaissance automatique de la parole

Sommaire

Résumé

Dans ce chapitre, nous présentons le principe de la reconnaissance automatique de la parole. Pour cela, après quelques généralités, nous décrivons les principaux modules d'un système de RAP. Ensuite, nous détaillons les trois approches possibles pour faire de la reconnaissance dans un système embarqué. Pour finir, nous présentons un rapide état de l'art sur les modules d'un système de reconnaissance qui sont concernés par la reconnaissance de la parole intégrée dans un système embarqué.

Depuis le début des années 1950 et les premiers travaux dans le domaine de la reconnaissance automatique de la parole, les objectifs ont bien évolué.

En 1949, Jean Dreyfus-Graf étudiait la déviation électrique du spot d'un oscilloscope en fonction du signal de parole ([Dreyfus-Graf 1950]).

En 1952, K.H. Davis, R. Biddulph et S. Balashek ([Davis 1952]) proposaient le premier système de reconnaissance de chiffres dépendant du locuteur (de "zero" à "nine").

Aujourd'hui, la communauté de chercheurs travaille sur des systèmes de reconnaissance de parole continue, indépendants du locuteur ou de l'environnement acoustique, le tout en temps réel (tâche TTR d'ESTER [Gravier 2004]).

Les premiers travaux utilisant un ordinateur datent de 1959 (les travaux antérieurs utilisaient eux les moyens de l'électronique analogique) et sont l'œuvre de Jim et Karma Forgie au Lincoln Laboratory du MIT. Leur ordinateur de l'époque, le TX-0, occupait une salle entière (*cf.* figure 2.1).

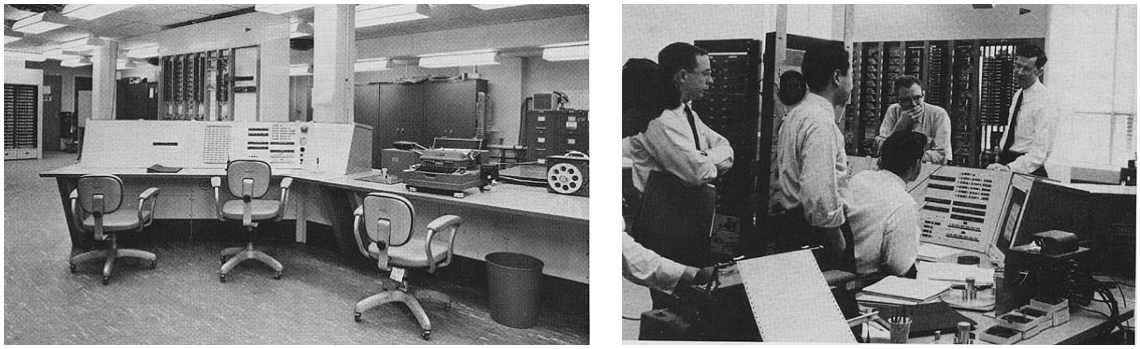

FIG. 2.1 – Le TX-0, premier "ordinateur" utilisé pour la reconnaissance vocale au laboratoire Lincolm du MIT en 1959.

Le volume des ordinateurs actuels a fortement réduit et leurs performances ont augmenté. Cependant, leur taille actuelle (même pour les ordinateurs portables) ne permet pas encore de les intégrer dans des systèmes embarqués grand public (téléphone, PDA, jouet. . .).

2.1 Introduction

La parole est pour l'homme le principal moyen de communication. C'est sans doute pour cela que Shannon s'attacha dès les années 1940 à en étudier les mécanismes. Dans son étude sur la théorie de la communication ([Shannon 1948]), il modélise la communication comme une information source

qui doit être envoyée à un destinataire (*cf.* figure 2.2). Pour cela, l'émetteur code son message avant de l'envoyer sur le canal de transmission. Ce message codé peut alors être perturbé par différentes sources de bruits (système de codage/décodage différents, bruits sur le canal de transmission, interférences...). Une fois le message arrivé à destination, le destinataire doit encore le décoder pour pouvoir en comprendre le contenu.

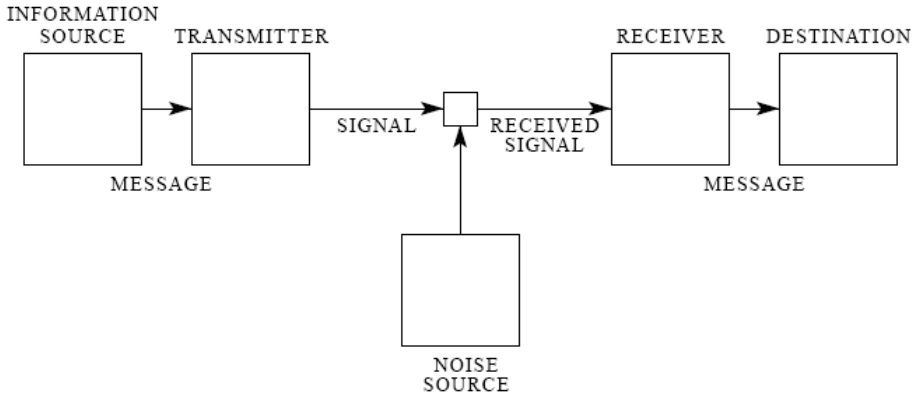

FIG. 2.2 – Schéma de principe pour la modélisation d'une communication (original extrait de [Shannon 1948])

2.2 Principaux modules d'un système de RAP

Les systèmes classiques de RAP sont composés essentiellement de cinq modules (*cf.* figure 2.3) :

- la paramétrisation du signal, qui doit permettre de ne garder que les informations pertinentes de ce dernier ;
- les modèles acoustiques, qui doivent représenter au mieux les unités acoustiques choisies (phonèmes, diphones, mots...) ;
- les modèles linguistiques, qui doivent être une représentation la plus vraisemblable possible du langage ;
- le dictionnaire, qui doit contenir l'ensemble des mots que l'on souhaite pouvoir reconnaître (dans certains cas le dictionnaire peut être spécifique à une application) ;
- le système de reconnaissance lui même.

Ces différentes composantes d'un système de RAP, bien que toutes nécessaires pour la reconnaissance de parole continue, sont relativement indépendantes les unes des autres.

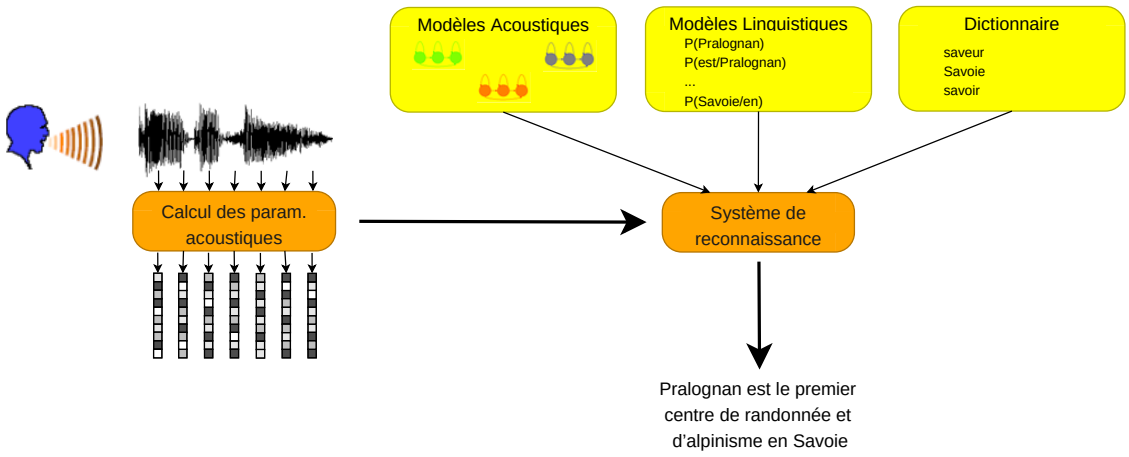

FIG. 2.3 – Schéma de principe d'un système de reconnaissance automatique de la parole

Comme nous l'avons expliqué dans 1.3, les systèmes embarqués (notamment les téléphones portables) ne disposent pas des mêmes ressources que les ordinateurs classiques. Il en résulte qu'un système embarqué ne pourra pas fournir les mêmes services qu'un système de RAP classique.

Les services présentés en introduction peuvent reposer sur un système capable de faire du décodage de mots isolés (voir de mots-connectés) plutôt que sur un système de reconnaissance de parole continue. De plus, au regard des ressources limitées dont dispose un système embarqué, il semble plus réaliste d'avoir pour objectif un système de reconnaissance de mots isolés.

Choisir un système de reconnaissance de mots isolés permet de s'affranchir du problème des modèles de langage (aspect très important pour les systèmes type dictée vocale) ; ceci explique pourquoi nous n'aborderons pas ce problème dans le chapitre suivant. Le lecteur pourra se référer à [Haton 2006] pour plus de détails sur cet aspect.

2.3 Rapide état de l'art de la RAP

Comme vu auparavant, un système de RAP comporte cinq modules principaux (*cf.* le chapitre 2.2). Cependant, la modélisation statistique du langage ne rentre pas dans le cadre fixé (la reconnaissance de chiffres isolés et la reconnaissance de commandes vocales). D'autre part, la gestion du lexique sera abordée dans la partie 3.2.1.

Ce rapide état de l'art ne concernera donc que les trois modules suivants :
- la paramétrisation ;
- le principe du décodage ;
- les modèles acoustiques.

Cet aperçu concerne la RAP d'un point de vue général. Une partie plus spécifique aux modèles acoustiques compacts, sur lesquels nous nous sommes focalisés, sera développée dans le chapitre 4.

2.3.1 Paramétrisation du signal

La parole apparaît comme une variation de la pression de l'air dans l'appareil phonatoire humain (pour plus de détails sur ces aspects, le lecteur se référera à [Tubach 1989] chapitre II). Les différents traits acoustiques du signal de parole sont notamment : sa fréquence fondamentale, son énergie, son spectre... Chacun de ces éléments étant lui-même intimement lié à une grandeur perceptible : pitch, intensité, timbre...

L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes. Un tel système prend un signal en entrée et retourne un vecteur de paramètres (appelé indifféremment *vecteur acoustique* ou encore *vecteur d'observations*). Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), discriminants (pour faciliter la reconnaissance) et robustes (aux différents bruits et/ou locuteurs).

Il existe un certain nombre d'approches pour la paramétrisation. Nous présentons ici celles utilisées le plus couramment dans la littérature :
- paramétrisation basée sur un modèle de production de la parole ;
- paramétrisation basée sur une analyse dans le domaine cepstral.

2.3.1.1 Paramétrisation basée sur un modèle de production de la parole

LPC (Linear Predictive Coefficients - [Tremain 1982])

Cette approche est basée sur les connaissances expertes en production de la parole. Le conduit vocal est modélisé comme un filtre auto-régressif (AR). Ceci permet d'approximer l'échantillon de l'instant n ($s(n)$) par une combinaison

linéaire des p échantillons précédents (P étant l'ordre du modèle).

$$\tilde{s}(n) = \sum_{i=1}^{P} a_i * s(n-i) \tag{2.1}$$

L'erreur de prédiction du modèle peut être estimée par :

$$e(n) = s(n) - \tilde{s}(n) \tag{2.2}$$

On peut donc estimer l'erreur quadratique moyenne par :

$$E_n = \sum_m e(m)^2 = \sum_m \left[s(m) - \sum_{i=1}^{P} a_i * s(m-i) \right]^2 \tag{2.3}$$

Minimiser cette erreur quadratique revient à annuler les dérivées partielles de E_n par rapport aux a_i. Pour cela, plusieurs approches sont présentées dans la littérature (méthode de covariance, méthode d'auto-corrélation). Pour plus de détails, le lecteur pourra se référer notamment à [Tubach 1989].

2.3.1.2 Paramétrisation basée sur une analyse dans le domaine cepstral

Comme nous l'avons précisé ci-dessus, le signal de parole (s_n) est le résultat de la convolution entre un signal excitateur g_n (la glotte) et le conduit vocal b_n :

$$s_n \ = \ g_n * b_n \tag{2.4}$$

Le passage, par homomorphisme, dans un domaine où l'opérateur de convolution est transformé en opérateur d'addition permet de décorreler les contributions de la source et du conduit du signal de parole. En pratique, l'utilisation de la transformée de Fourier donne les coefficients cepstraux :

$$\tilde{s}_n \ = \ \tilde{g}_n + \tilde{b}_n \tag{2.5}$$

où $\tilde{g}_n$ et $\tilde{b}_n$ sont les transposées dans le domaine quéfrentiel de g_n et b_n.

Plusieurs méthodes permettent d'obtenir des coefficients cepstraux :
- grâce à une récursion depuis les coefficients LPC, ce qui donne les coefficients LPCC ;
- par l'utilisation d'une FFT et d'une FFT inverse ; cette technique permet de calculer les coefficients MFCC, LFCC et PLP.

LPCC (Linear Predictive Cepstral Coefficients)

La méthode présentée dans [Miet 2001] permet de calculer les coefficients LPCC directement depuis les coefficients LPC.

$$LPCC_i = -LPC_i + \sum_{k=1}^{i-1}(1 - \frac{k}{i})LPC_k LPCC_{i-1} \tag{2.6}$$

Cette approche a pour but de modéliser davantage l'enveloppe du signal.

Le calcul des coefficients cepstraux (MFCC, LFCC et PLP) est souvent précédé d'une phase de pré-accentuation du signal, suivie d'un fenêtrage.

La pré-accentuation est définie de la manière suivante :

$$x[i] = x[i] - \alpha * x[i-1] \tag{2.7}$$

α est généralement compris entre 0,90 et 1 (la valeur classique de α est 0,97 - HTK, SPHINX).

Le fenêtrage appliqué généralement est celui de Hamming :

$$ham[i] = 0,54 - 0,46\cos(\frac{2 * \pi * i}{N}) \tag{2.8}$$

où N correspond à la longueur de la fenêtre. Dans la littérature d'autres fenêtrages sont parfois appliqués : fenêtrage de Hanning, fenêtrage de Blackman, fenêtrage de Kaiser...

Ces deux étapes sont préalables au calcul des coefficients MFCC, LFCC et PLP.

MFCC (Mel Frequency Cepstral Coefficient - [Davis 1980])

Afin de rapprocher l'analyse en banc de filtres de la perception humaine, les filtres ne sont généralement pas répartis de manière linéaire mais en fonction d'une échelle Mel. La correspondance entre une fréquence en Hz et en Mel se calcule de la manière suivante :

$$F_{mel} = 2595 * \log(1 + \frac{F_{Hz}}{700}) \tag{2.9}$$

Intuitivement, cela revient à utiliser une échelle linéaire en basse fréquence, puis logarithmique en haute fréquence.

La chaîne complète de calculs des coefficients MFCC est définie par la figure 2.4.

Généralement, seuls les 12 premiers coefficients cepstraux sont conservés et une vingtaine de filtres sont utilisés pour l'analyse en banc de filtres.

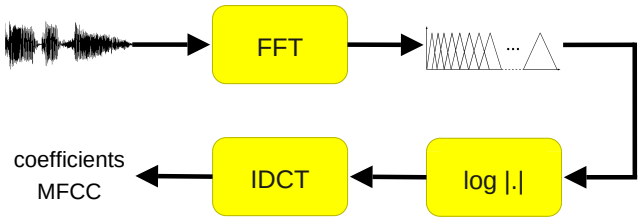

FIG. 2.4 – Chaîne de traitement pour obtenir les coefficients MFCC

LFCC (Linear Frequency Cepstral Coefficient)

Il s'agit d'une variante des MFCC. La différence vient de l'utilisation d'un banc de filtres linéaire, contrairement à l'échelle Mel des MFCC.

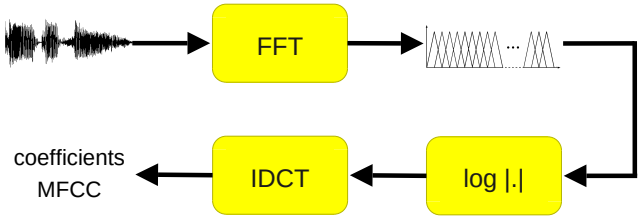

FIG. 2.5 – Chaîne de traitement pour obtenir les coefficients LFCC

PLP (Perceptual Linear Predictive)

Les coefficients PLP ont été présentés dans [Hermansky 1990] ; cependant l'implémentation utilisée dans ce travail se rapproche plus de la version présentée dans [Young 1995] (qui est aussi celle utilisée par le toolkit HTK - [Woodland 1993]). Cette extraction de paramètres est basée sur des connaissances expertes de l'appareil auditif humain.

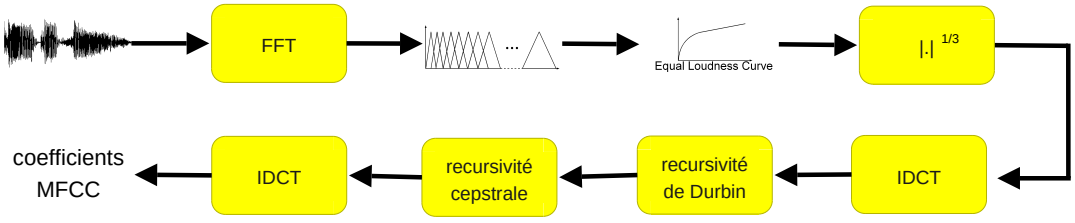

FIG. 2.6 – Chaîne de traitement pour obtenir les coefficients PLP

La paramétrisation du signal en coefficients PLP est finalement assez proche

d'une LPC ([Haton 2006]).

Energie

Généralement, l'énergie du signal est utilisée en complément des coefficients issus d'une paramétrisation basée sur une analyse dans le domaine cepstral. L'énergie correspond à la puissance du signal.

$$En = \sum_{n=0}^{N-1} s_n^2 \tag{2.10}$$

Le calcul de l'énergie se fait généralement sur des fenêtres glissantes de 25ms avec un décalage de 10ms (soit une valeur toutes les 10ms de signal).

2.3.1.3 Autres paramétrisations

Taux de passage par zéro

Le ZCR (Zero Crossing Rate) est un bon complément de l'énergie. Un taux de passage par zéro ([Kedem 1986]) faible et une énergie forte sont un bon indice d'un son voisé alors qu'un taux de passage par zéro élevé et une énergie plus faible caractérisent plutôt une zone non voisée.

$$Zcr = 0,5 * \sum_{n=0}^{N-1} |sign(x_n) - sign(x_{n-1)}| \tag{2.11}$$

Une évolution du ZCR est proposée par [Taboada 1994] ; il propose une bande d'amplitude autour de 0 pour limiter un certain nombre de phénomènes parasites qui provoquent de faibles oscillations aux alentours de 0.

Divers

Il existe de nombreuses références qui comparent les systèmes de paramétrisation : pour plus de détails le lecteur pourra se référer à [Psutka 2001] ou [Chen 2005]. Dans [Lévy 2003], nous avons proposé une telle comparaison dans le cadre des systèmes embarqués. A la lecture de ces différents travaux, les coefficients PLP apparaissent légèrement plus performants, particulièrement en conditions adverses.

Au-delà des principales approches pour la paramétrisation que nous venons de présenter, d'autres peuvent être trouvées dans la littérature telles que NPC (Neural Predictive Coding - extension non linéaire du codage LPC [Gas 2000]), LSF (Line Spectral Frequencies - les fréquences de raies spectrales, issue des LPC, [Paliwal 1993])...

2.3.1.4 Post traitement

Nous allons détailler deux compléments communément admis : la suppression de la moyenne cepstrale avec réduction de la variance et les paramètres dynamiques (dérivées premières et secondes).

Suppression de la moyenne cepstrale (CMS) et réduction de la variance

Généralement, après l'étape de paramétrisation proprement dite, une normalisation des paramètres est effectuée afin de rendre ces paramètres plus robustes au bruit ou au changement de canal. Cela revient généralement à soustraire la moyenne cepstrale puis à réduire la variance. La soustraction de la moyenne suit la règle suivante :

$$c_m^{'}(t) = c_m(t) - \sum_{\tau}^{signal} c_m(\tau) \tag{2.12}$$

la normalisation de la variance respecte :

$$\tilde{c}_m(t) = \frac{c_m^{'}(t)}{\sqrt{\sum_{\tau}^{signal} \left(c_m^{'}(\tau)\right)^2}} \tag{2.13}$$

où $c_m(t)$ représente les coefficients initiaux, $c_m^{'}(t)$ les coefficients centrés et $\tilde{c}_m(t)$ les coefficients centrés réduits.

La soustraction de la moyenne cepstrale et la normalisation de la variance sont réalisées soit sur l'intégralité du fichier à décoder soit sur une fenêtre glissante.

[Schwartz 1993] et [Chen 2002] montrent une amélioration non négligeable des performances grâce à cette normalisation.

Dérivées de premier et second ordre : Δ et $\Delta\Delta$

Pour enrichir la paramétrisation, les dérivées de premier et second ordre sont souvent utilisées. Cela permet d'ajouter de l'information concernant la dynamique du signal. Les coefficients Δ (dérivées du premier ordre) sont souvent estimés grâce au développement limité d'ordre 2 :

$$c_i'(t) = \frac{c_i(t+1) - c_i(t-1) + 2(c_i(t+2) - c_i(t-2))}{10} \tag{2.14}$$

où $c_i(t)$ correspond au $i^{ème}$ coefficient pour la trame t et $c_i'(t)$ sa dérivée.

Les $\Delta\Delta$ (coefficients du second ordre) sont estimés de la même manière à partir des coefficients du premier ordre.

2.3.2 Principe de décodage

Les premières recherches en reconnaissance automatique de la parole ont débuté à partir des années 1950. Dans les années 1960, une méthode appelée Dynamic Time Warping (DTW) est apparue. Elle repose sur les travaux de [Bellman 1957] et reste aujourd'hui une approche importante en reconnaissance de mots isolés.

Une seconde méthode a émergé dans les années 1975, avec les travaux de [Jelinek 1976]. Elle s'appuie sur l'utilisation des modèles de Markov cachés (*Hidden Markov Models* - HMM). Elle a permis de nombreuses avancées dans les domaines de la reconnaissance de la parole continue et de la reconnaissance multi-locuteurs, domaines dans lesquels la DTW était peu probante.

Enfin, à la fin des années 80, les réseaux de neurones sont venus offrir une nouvelle voie pour le traitement automatique de la parole ([Bourlard 1987]). Plusieurs types de réseaux de neurones coexistent mais, pour la reconnaissance de mots isolés, les approches les plus classiques sont le perceptron multi-couches et les TDNN (Time Delay Neural Network). Ces méthodes ne sont pas abordées dans ce document, mais le lecteur pourra se tourner vers [Bourlard 1990] ou [Waibel 1989] pour de plus amples informations sur ces approches.

2.3.2.1 Dynamic Time Warping : DTW

Principe général

L'idée directrice de la DTW consiste à estimer une mesure de similarité entre la représentation d'un mot référence et la représentation d'un mot inconnu afin d'évaluer l'écart entre ces deux mots.

Pour cela, nous disposons d'un ensemble de références R_x qui forment le dictionnaire (C) des n mots à reconnaître : $C = \{R_x\}_{1 \leqslant x \leqslant n}$. Muni d'une distance D, il devient alors possible de calculer un coût de déformation entre le mot inconnu (T) et une référence x (R_x). Le but est donc de réaliser un alignement temporel, le meilleur qui soit, entre une référence et un mot à tester. Le mot prononcé est trouvé par la résolution de :

$$t = \underset{R_x \in C}{\text{ArgMin}}\, D(T, R_x) \tag{2.15}$$

Soient R_x la référence d'un mot du dictionnaire et T le mot à reconnaître de longueurs respectives I et J. R est composée d'éléments r(1), r(2), r(3), ..., r(I) (respectivement pour T : t(1), t(2), t(3), ..., t(J)) qui représentent les vecteurs

acoustiques du signal à un instant donné. On appelle $d(r_i, t_j)$ la distance entre les vecteurs acoustiques r(i) et t(j). La figure 2.7 illustre ce principe.

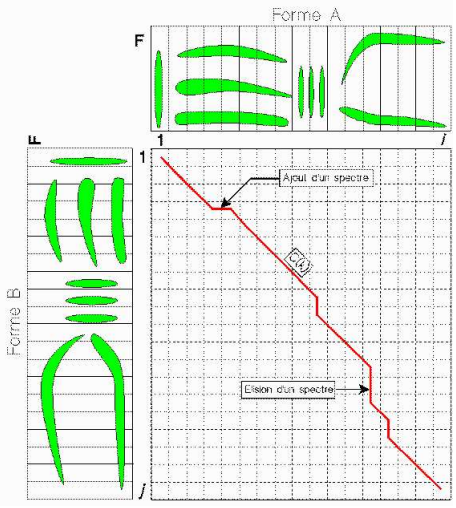

FIG. 2.7 – Principe de la DTW

Distances usuelles

Plusieurs distances[1] peuvent être utilisées en fonction des méthodes de paramétrisation retenues :

- la distance utilisant les normes L_n (n allant de 1 à l'∞) est plutôt utilisée dans les systèmes à base d'analyse cepstrale. La norme L_2 est la plus utilisée. Elle est plus connue sous le nom de distance euclidienne

$$d_n(r_i, t_j) = \left(\sum_{k=0}^{p} |r_k - t_k|^n \right)^{\frac{1}{n}} \tag{2.16}$$

- la mesure d'Itakura est plutôt utilisée dans le cadre d'une paramétrisation par prédiction linéaire

$$d_{it}(r_i, t_j) = \log \left[\frac{{r_i}^t R_b r_i}{{t_j}^t R_b t_j} \right], \tag{2.17}$$

  avec R_b représentant la matrice des coefficients d'autocorrélation évalués sur le segment t_j.

Contraintes locales

Afin de tenir compte des réalités physiques du mécanisme de production de la parole, les déplacements entre les vecteurs de paramètres sont limités

[1] au sens de *mesure de similarité*

(*contraintes locales*). Les contraintes locales les plus courantes sont représentées dans la figure 2.8.

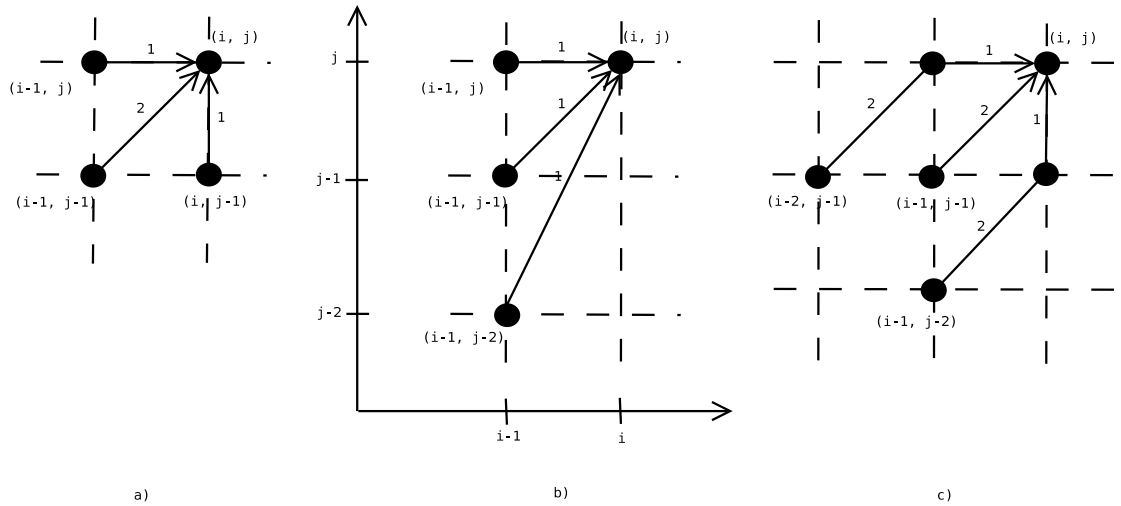

FIG. 2.8 – Contraintes locales utilisées dans la DTW

Le principe est donc de trouver le chemin d'alignement ayant un coût minimum. La distance cumulée g(i,j) est définie (en fonction de la contrainte locale choisie - ici 2.8.a) par :

$$g(i,j) = \min \begin{bmatrix} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2 * d(i,j) \\ g(i,j-1) + d(i,j) \end{bmatrix} \tag{2.18}$$

En normalisant par les longueurs de R et T, on obtient donc :

$$D(R,T) = \frac{g(I,J)}{I+J} \tag{2.19}$$

2.3.2.2 Hidden Markov Model : HMM

L'idée principale est de décomposer un mot en une suite de sous-unités lexicales (généralement les phonèmes) qui sont représentées par des modèles de Markov cachés.

La notation utilisée ici correspond à celle utilisée par L. Rabiner dans [Rabiner 1989].

Principe général

La reconnaissance consiste à retrouver le mot $\tilde{m}$ parmi un ensemble de mots possibles M ($M = \{m_i\}$), en fonction d'une suite d'observations O (suite de vec-

teurs acoustiques). Cette recherche est faite par maximisation de la probabilité d'émission d'un mot en fonction de la suite d'observations :

$$\tilde{m} = \underset{m}{\text{ArgMax}}\, P(m/O) \tag{2.20}$$

D'après le théorème de Bayes :

$$P(m/O) = \frac{P(O/m)P(m)}{P(O)} \tag{2.21}$$

$P(O)$ étant indépendant de m, on obtient :

$$\tilde{m} = \underset{m}{\text{ArgMax}}\, P(O/m)P(m) \tag{2.22}$$

avec $P(m)$ la probabilité d'apparition *a priori* du mot m, et $P(O/m)$ la probabilité *a posteriori* d'émission de la séquence O sachant le mot m.

Formalisme Markovien

Un modèle de Markov (illustré par la figure 2.9) est un automate fini qui change d'état à chaque unité de temps. De manière générale, on se limite aux HMM d'ordre 1, ce qui sous-entend que la possibilité d'être dans un état e_j au temps $t+1$ ne dépend que de l'état e_i dans lequel le système se trouvait à l'instant t (d'autres modèles existent voir [Haton 1994] qui propose des HMM d'ordre 2).

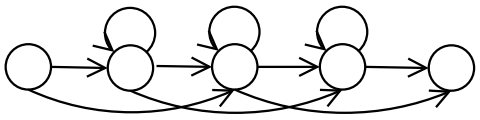

FIG. 2.9 – Exemple d'un HMM

De tels automates sont utilisés pour modéliser les sous-unités lexicales. Chaque unité est représentée par un HMM à n états, généralement trois. A ces trois états, on ajoute un état de début d et un état de fin f qui permettront l'enchaînement des HMM.

Des HMM modélisant des mots entiers ont également été utilisés ([Lee 1989]). Nous ne détaillons pas ce choix qui ne correspond pas aux objectifs de cette étude (*cf.* chapitre 3.2.1).

Chaque état est caractérisé par une fonction de densité de probabilité. Très souvent, ces fonctions sont des mixtures de gaussiennes (*Gaussian Mixture Model* : GMM). Les liaisons inter-états sont caractérisées par une matrice de transitions qui définit la probabilité de se déplacer d'un état i à un état j. Un HMM λ est donc caractérisé par :

- E : $E = \{1, 2, ..., N\}$, l'ensemble des N états du modèle ;
- A : $A = \{a_i j$, avec $1 \leqslant i, j \leqslant N\}$, la matrice de transition inter-états ;
- et B : $B = \{b_i$, avec $1 \leqslant i \leqslant N\}$, l'ensemble des fonctions de densité de probabilité associées à chacun des états.

La mise en oeuvre d'un système Markovien implique la résolution de trois problèmes :

- l'estimation de la probabilité d'une séquence d'observations : pour une suite d'observations O et un HMM λ, quelle est la probabilité $P(O|\lambda)$?
- le décodage d'une séquence d'observations : pour une suite d'observations O et un HMM λ, quelle est la séquence d'états qui correspond aux observations ?
- l'apprentissage des paramètres du modèle : comment estimer les paramètres λ ?

Une réponse à ces 3 questions est proposée dans les trois paragraphes suivants.

Estimation de la probabilité d'une séquence d'observations

La séquence d'observations $O = \{o_1, o_2, o_3, ..., o_t\}$, passant par la suite d'états $Q = \{q_0, q_1, q_2, ..., q_t\}$ aura pour probabilité :

$$P(O|Q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t, \lambda) \tag{2.23}$$

$$= b_{q_1}(o_1) b_{q_2}(o_2) b_{q_2}(o_3) \ ... \ b_{q_t}(o_T) \tag{2.24}$$

en supposant que les observations sont statistiquement indépendantes.

La probabilité du chemin peut, elle, être définie par :

$$P(Q|\lambda) = \pi_{q_1} \prod_{t=1}^{T} a_{q_t - 1 q_t} \tag{2.25}$$

$$= \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ ... \ a_{q_{t-1} q_t} \tag{2.26}$$

La probabilité conjointe des observations O et du chemin Q est donc :

$$P(O, Q|\lambda) = P(O|Q, \lambda) P(Q|\lambda) \tag{2.27}$$

Pour l'ensemble des chemins, cela donne :

$$P(O|\lambda) = \sum_{Q} P(O|Q, \lambda) P(Q|\lambda) \tag{2.28}$$

$$= \sum_{q_1, q_2, ..., q_n} \pi_{q_1} \prod_{t=1}^{T} a_{q_t - 1 q_t} \tag{2.29}$$

La complexité de ce calcul est très importante, de l'ordre de $2^T * N^T$ opérations avec T correspondant au nombre d'observations et N au nombre d'états.

L'utilisation de l'algorithme *Forward*, de programmation dynamique, permet de réduire ce coût de calcul tout en conservant une solution exacte. Cet algorithme comporte trois étapes :

1. initialisation :
$$\alpha_1(i) = \pi_i b_i(o_1) \qquad 1 \leq i \leq N$$
2. itération :
$$\alpha_{t+1}(j) = b_j(o_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij} \qquad 1 \leq j \leq N et 1 \leq t \leq T-1$$
3. conclusion :
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

avec $\alpha_t(i)$ qui correspond à la probabilité conjointe d'observer la séquence $o_1, o_2, ...o_t$ et l'état i au temps t. Cet algorithme est d'une complexité nettement inférieure, de l'ordre de N^2T.

Il possède un pendant appelé *backward* qui parcourt dans l'ordre inverse. Il comporte aussi trois étapes et peut être décrit de la manière suivante :

1. initialisation :
$$\beta_T(i) = 1 \qquad 1 \leq i \leq N$$
2. itération :
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \qquad 1 \leq i \leq N \; et \; t = T-1, T-2, ..., 1$$
3. conclusion :
$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$

Décodage d'une séquence d'observations

L'objectif est ici, connaissant la séquence d'observations $O = o_1, o_2, ..., o_T$ et les paramètres λ du HMM, de déterminer quel est le chemin, $S = q_1, q_2, ..., q_T$, le plus probable :

$$\tilde{S} = \underset{S}{\text{ArgMax}} \, P(O, S|\lambda) \tag{2.30}$$

De la même manière que pour l'évaluation de $P(O|\lambda)$, le test de l'ensemble des chemins possibles nécessite de l'ordre de N^T opérations. Un autre algorithme de programmation dynamique nous permet de réduire cette complexité : l'algorithme de Viterbi ([Viterbi 1967]). L'idée principale est d'utiliser la probabilité ($\delta_t(i)$) d'être dans l'état i, à l'instant t, pour la séquence $o_1, o_2, ..., o_t$ pour estimer la probabilité d'être dans l'état j avec la prochaine trame (o_{t+1}). Cette probabilité correspond au produit de $b_j(o_{t+1})$ par le max sur i des $\delta_t(i)a_{ij}$.

Cet algorithme récurrent est défini comme suit :

1. initialisation :

$$t = 0$$

$$r_0(m) = \pi_i b_i(o_1)$$

2. récurrence :

$$r_t(j) = \underset{i}{\text{Max}}\ r_{t-1}(i)a_{ij}b_j(o_t)$$

3. décision :

$$etatfinal = \underset{i}{\text{Max}}\ r_T(i)$$

$r_t(j)$ correspond à la probabilité maximale pour que les observations $O = o_1, o_2, ..., o_t$ aient été émises par λ en suivant un chemin arrivant en j.

L'ajout d'une variable de mémorisation du chemin parcouru (suite d'états) permet, par un retour arrière, de déterminer également la séquence d'état.

Apprentissage des paramètres du modèle

L'apprentissage des paramètres du modèle est certainement le problème le plus complexe des trois. Ces paramètres à estimer sont les transitions entre états (a_{ij}) et les probabilités d'émissions des états ($b_i(o)$). La topologie[2] du modèle ne fait pas partie des paramètres estimés, elle est, généralement, définie *a priori*.

Classiquement, la méthode utilisé pour la ré-estimation des paramètres λ cherche à maximiser la vraisemblance (critère du maximum de vraisemblance aussi appelé MLE pour Maximum Likelihood Estimation). L'objectif est donc, possédant une certaine quantité de données O, de ré-estimer les paramètres λ tels que :

$$\tilde{\lambda} = \underset{\lambda}{\text{ArgMax}} \prod_k P(O_k|\lambda) \tag{2.31}$$

[2]par topologie, nous entendons : nombre d'états des modèles, la possibilité ou non de passer d'un état à l'autre et enfin l'alphabet des symboles

D'autres critères ont été utilisés conjointement ou à la place de MLE, tels que le critère MAP (Maximum A Posteriori) ou le critère MMIE (Maximum Mutual Information Estimation).

La maximisation de la vraisemblance du critère MLE est obtenue par l'utilisation de l'algorithme *Baum-Welch* connu aussi sous le nom de *forward-backward*. Cet algorithme itératif converge vers un optimum local ; cependant, le temps nécessaire à cette convergence dépend directement de l'initialisation des paramètres. Cet algorithme nous assure que :

$$P(O|\tilde{\lambda}_n) \leq P(O|\tilde{\lambda}_{n+1}) \tag{2.32}$$

Définissons la grandeur ξ :

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O,\lambda) \tag{2.33}$$

qui correspond à la probabilité d'aller de l'état i à l'instant t à l'état j à l'instant $t+1$ sachant le modèle et les observations.

Et la grandeur γ

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{2.34}$$

qui correspond à la probabilité d'être dans l'état i à l'instant t sachant les observations O et les paramètres λ.

$$\begin{aligned}
\xi_t(i,j) &= P(q_t = i, q_{t+1} = j|O,\lambda) && (2.35)\\
&= \frac{P(q_t = i, q_{t+1} = j, O|\lambda)}{P(O|\lambda)} && (2.36)\\
&= \frac{P(q_t = i, o_1, o_2, ..., o_t|\lambda) * a_{ij} b_j(o_{t+1}) * P(o_{t+2}, ..., o_T|q_{t+1} = j, \lambda)}{P(O|\lambda)} && (2.37)
\end{aligned}$$

Notons que $P(q_t = i, o_1, o_2, ..., o_t|\lambda)$ correspond à $\alpha_t(i)$, défini durant la présentation de l'algorithme $forward$. $P(o_{t+2}, ..., o_T|q_{t+1} = j, \lambda)$ correspond au paramètre $\beta_{t+1}(j)$ de l'algorithme $backward$. Nous obtenons :

$$\xi_t(i,j) = \frac{\alpha_t(i) * a_{ij} b_j(o_{t+1}) * \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) * a_{ij} b_j(o_{t+1}) * \beta_{t+1}(j)} \tag{2.38}$$

La probabilité de transition entre l'état i et l'état j est ré-estimée par :

$$\begin{aligned}
\tilde{a}_{ij} &= \frac{\textit{espérance du nombre de transitions des états } i \textit{ vers } j}{\textit{espérance du nombre de passages en } i} && (2.39)\\
&= \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} && (2.40)
\end{aligned}$$

Il faut noter que le terme a_{ij} présent au numérateur de l'équation 2.38 assure qu'une transition non autorisée au départ reste non autorisée : γ_{ij} restant à 0.

La probabilité d'être dans l'état i au départ (π_i) correspond à :

$$\tilde{\pi}_i = \gamma_1(i) \tag{2.41}$$

La probabilité d'émission associée à un état est définie par :

$$\tilde{b}_j(k) = \frac{espérance\ du\ nombre\ d'émission\ du\ symbole\ v_k\ dans\ l'état\ i}{espérance\ du\ nombre\ de\ passages\ en\ i} \tag{2.42}$$

$$= \frac{\sum_{t=1,o_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2.43}$$

2.3.2.3 Systèmes experts

Les systèmes experts sont des systèmes cherchant à reproduire l'analyse faite par des experts humains (notamment des phonéticiens dans le cadre de la reconnaissance de la parole). Ces experts procèdent généralement en deux phases : une analyse visuelle du spectrogramme suivie d'un raisonnement contextuel avec les indices notés lors de la première phase.

De tels systèmes sont généralement composés de deux entités distinctes :
- la base de connaissance : contenant *les règles* et *les faits répertoriés* par un expert humain ;
- le moteur d'inférence : réalisant les déductions logiques à partir de la base de connaissances.

L'objectif principal de ces systèmes est de réaliser des raisonnements logiques comparables à ceux que feraient des experts humains de ce domaine.

Dans [Carbonell 1986], des exemples de règles et de faits sont présentés.

L'analyse visuelle d'un certain nombre de segments (x segments notés Sx) a permis d'extraire des faits :
- S1 : voyelle, des formants F1 à 800Hz, F2 à 1200Hz et F3 à 2500Hz ;
- S2 : plosive obtenant son maximum spectral à 3000Hz ;
- S3 : voyelle, des formants F1 à 300Hz, F2 à 1950Hz et F3 à 2500Hz ;
- S4 : . . .

Ils présentent aussi des exemples de règle :
- R1 : *SI* plosive *ET* 3000Hz $\leq$ maximum spectral $\leq$ 4500Hz *ET* le contexte droit est une labiale *ALORS* c'est un /t/ ;

- R2 : *SI* plosive *ET* 2500Hz $\leq$ maximum spectral $\leq$ 4500Hz *ET* le contexte droit est une voyelle non arrondie *ALORS* c'est un /k/ ;
- R3 : . . .

D'autres exemples de systèmes de "reconnaissance du français" basés sur des systèmes experts peuvent être trouvés dans [Bechet 1994], [Haton 1990], [Stern 1986] ou encore dans [Zue 1986] pour un système en anglais.

Ces systèmes ont progressivement été abandonnés avec la généralisation des systèmes Markoviens.

2.3.3 Modélisation du lexique

Les SRAP dont le système de décodage est basé sur la DTW ou sur les HMM nécessitent de posséder des modèles acoustiques. Ces modèles peuvent être des modèles d'unités sous-lexicales (phonèmes, di-phones, tri-phones, penta-phones. . .) ou des modèles de mots directement.

L'inconvénient majeur de l'utilisation de modèles de mot est la complexité engendrée par l'ajout d'un mot dans le lexique. Cette approche est généralement utilisée avec un système DTW. Il convient alors de faire un alignement entre la référence et le test. La référence est le mot complet (non découpé en sous-unités). A l'inverse, les approches se servant des HMM utilisent généralement une décomposition d'un mot comme une suite de phonèmes[3].

Les modèles de mots sont utilisés le plus souvent dans le cadre d'application ayant un lexique non évolutif et composé de peu d'entrées. La très grande majorité des systèmes de reconnaissance de parole continue grand vocabulaire utilise des sous-unités lexicales. En effet, ces systèmes ont un lexique comportant plus de $65\,000$ entrées, il est donc impossible de faire prononcer toutes les entrées du lexique par un utilisateur. De plus, le partage des sous-unités lexicales entre les différentes entrées du lexique permet aussi de limiter le nombre d'unités de base et d'augmenter, pour chaque unité, la quantité de données d'apprentissage.

[3] ou di-phones, tri-phones, penta-phones. . .

Chapitre 3

Contexte applicatif

Sommaire

Résumé

Dans ce chapitre, nous présentons les différentes architectures possibles pour l'intégration de services vocaux dans les systèmes embarqués. Nous rappelons ensuite les applications envisageables grâce à l'apport de la reconnaissance vocale avant de présenter les contraintes liées à ces applications et au contexte de l'embarqué. Différents choix liés aux architectures et aux contraintes sont également présentés.

Dans ce chapitre nous présentons, dans un premier temps, les différentes architectures permettant l'intégration de services vocaux dans les systèmes mobiles.

Dans la suite de ce chapitre, certaines de contraintes, notamment liées au choix de l'architecture, sont présentées.

3.1 Différentes architectures pour la reconnaissance dans des systèmes mobiles

Nous pouvons dénombrer trois grandes architectures pour la reconnaissance de la parole dans les systèmes embarqués :
- la reconnaissance déportée (*cf.* figure 3.1),
- la reconnaissance embarquée (*cf.* figure 3.2), et
- la reconnaissance répartie (*cf.* figure 3.3).

Ces différentes approches, que nous allons détailler dans les chapitres suivants, présentent toutes des avantages et des inconvénients.

3.1.1 La reconnaissance déportée

Le principe de la RAP déportée consiste à se servir du terminal uniquement comme un système d'acquisition du son. Une fois le signal enregistré, il est transmis via un réseau, généralement sans fil, à un serveur qui effectue le travail de reconnaissance.

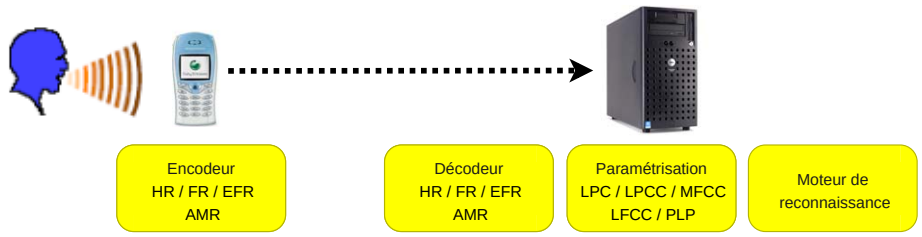

FIG. 3.1 – Architecture déportée

Cette approche permet d'envisager l'utilisation de serveurs beaucoup plus puissants et donc de fournir des services plus divers et généralement de meilleure qualité. La présence d'un serveur performant permet d'adjoindre d'autres modules à la reconnaissance vocale tels qu'un système de dialogue ou encore des vocabulaires dédiés à chaque application. Pour résumer, l'avantage

majeur de cette architecture est la levée des contraintes liées aux ressources limitées des systèmes embarqués.

Un des inconvénients de ce type d'architecture est la perte d'informations dûe à la transmission du signal. En effet, avant d'être envoyé au(x) serveur(s) de reconnaissance, le signal doit être encodé suivant un des codecs de l'opérateur (HR [ETSI-HR 2000], FR [ETSI-FR 2000], EFR [ETSI-EFR 2000], AMR [ETSI-AMR 2000]...). Ces codages (avec pertes) sont nécessaires car l'émission du signal sur le réseau a un coût non négligeable. De plus, au cours d'une transmission sur un réseau (sans fil ou non), différents phénomènes, comme la perte de paquets d'informations ou l'ajout de bruits viennent perturber le signal.

Parmi les inconvénients, nous pouvons encore citer le coût (tant sur le plan financier que sur l'aspect du trafic réseau) de transmission du signal au serveur et de la réponse du serveur au terminal.

Enfin, l'absence d'autonomie représente certainement l'inconvénient majeur de cette architecture. En effet, elle suppose que le téléphone soit connecté au réseau pour pouvoir fournir les services de reconnaissance, interdisant l'emploi de ceux-ci en local pour des applications comme les jeux, l'agenda...

Dans la littérature, les problèmes généraux de la reconnaissance déportée (pas seulement via des réseaux téléphoniques mobiles) sont abordés notamment dans [Lilly 1996] ou [Euler 1994]. Des exemples plus spécifiques aux réseaux GSM sont présentés dans [Gallardo-Antolin 1998], [Raj 2001] et [Kim 2001].

Comme exemple d'application, le lecteur pourra se tourner vers le projet "Multimedia Terminal Mobile" [Lefort 2002] (projet MTM IST 1999-11100) auquel a participé le LIA. Ce projet avait pour objectif de concevoir et de réaliser un système qui regrouperait les fonctionnalités d'un PDA, d'un téléphone portable de troisième génération (UMTS), d'un ordinateur portable et d'une caméra. Dans le cadre de ce projet, le LIA a développé un serveur permettant de faire de la reconnaissance déportée de mots isolés, ainsi que de la reconnaissance de locuteurs (un mode de reconnaissance répartie a également été présenté).

3.1.2 La reconnaissance embarquée

Avec cette approche, toutes les étapes de la reconnaissance sont effectuées dans le système embarqué. Aucun serveur à distance est nécessaire, le système est donc entièrement autonome : il réalise lui-même l'acquisition du son puis le décodage.

L'intérêt principal de cette approche est l'absence de contrainte : en effet, il

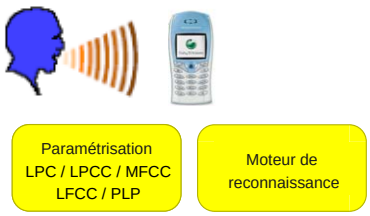

FIG. 3.2 – Architecture embarquée

n'est pas nécessaire de se limiter à l'utilisation des codecs de parole définis par l'ETSI[1], car aucune compatibilité avec un serveur n'est requise (contrairement à 3.1.3). De plus, les inconvénients de la reconnaissance déportée (*cf.* 3.1.1 - perte d'informations dûe à la transmission ou à la compression avant émission) sont résolus par l'absence de transmission.

Cette solution présente toutefois un inconvénient majeur : les ressources nécessaires. En effet, toutes les phases (paramétrisation du signal et reconnaissance) sont effectuées au sein du téléphone. Comme expliqué dans 2.2, les différentes phases de la reconnaissance nécessitent beaucoup de ressources comparées aux capacités d'un téléphone portable. De plus, l'augmentation des ressources d'un téléphone (mémoire et puissance de calcul), bien que possible, est très coûteuse.

Dans la littérature plusieurs implémentations dans des systèmes embarqués sont proposées. Les ressources disponibles dans ces systèmes sont variables suivant les projets. Trois grandes classes peuvent être distinguées :

- **très peu de ressources** : DSPFactory ([Cornu 2002]) propose un système nécessitant un DSP[2]/CPU à 4MHz. Un système présenté par Wang permet d'embarquer un système de reconnaissance de mots isolés sur un MCU[3] à 8MHz ([Wang 2004]) ;
- **peu de ressources** : le système d'IBM ([Deligne 2001]) recommande un DSP/CPU à 50 MHz et 1Mo de mémoire, celui développé par l'université de Graz ([Obermaier 1998]) un DSP/CPU à 30 MHz et ceux de Siemens ([Astrov 2003a] et [Astrov 2003b]) et Texas Instrument ([Gong 2000] - 128Ko de mémoire) exigent un DSP/CPU à 100 MHz ;
- **"système de demain"** : NEC ([Ishikawa 2006]) propose un système ayant besoin d'un processeur triple cœur (3 systèmes à 200MHz) et 6Mo de mémoire.

[1] Institut européen des normes de télécommunication - `http://www.etsi.org/`
[2] Processeur dédié au traitement numérique du signal
[3] Micro-contrôleur - CPU + mémoire + port de communication

Ce dernier système est beaucoup plus gourmand que les autres. Il semble plutôt destiné aux futures applications et paraît difficilement intégrable aujourd'hui au regard des ressources des systèmes embarqués disponibles actuellement.

3.1.3 La reconnaissance répartie

Ce mode de reconnaissance se situe entre la RAP déportée et la RAP embarquée. En effet, une partie du travail est effectuée sur le système embarqué et une autre partie sur un serveur distant. Généralement, la phase d'extraction des paramètres est réalisée sur le terminal (*cf.* [Maes 2000] , [Ramaswamy 1998], [Srinivasamurthy 2005]. . .) et la partie reconnaissance - à proprement parler - se trouve sur un serveur.

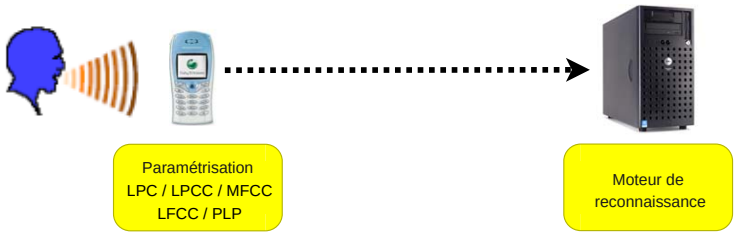

FIG. 3.3 – Architecture répartie

Cette approche permet de contourner les problèmes liés à la perte d'informations dûe au transport. Avec cette architecture, une seule phase "d'extraction de l'information" est nécessaire : les deux premières étapes (encodage et décodage) de l'architecture présentée dans 3.1.1 deviennent inutiles. Les pertes engendrées par l'encodage effectué sur le téléphone sont alors évitées (HR, FR, EFR et AMR sont tous des encodeurs avec perte).

Cette approche implique, par ailleurs, que le téléphone (et donc son fabricant) s'accorde avec l'opérateur désirant fournir le service de reconnaissance vocal, étant donné que l'extraction des paramètres est effectuée directement dans le téléphone. Ces paramètres sont ensuite utilisés par le système de reconnaissance durant la phase de décodage. Les codecs[4] classiques, définis par l'ETSI, sont plutôt basés sur les coefficients LPC alors que les systèmes actuels de RAP sont essentiellement basés sur les coefficients MFCC ou PLP ou encore sur des

[4] Logiciel permettant de convertir le signal analogique en signal numérique. Généralement, il permet aussi la compression du signal numérisé.

dérivés de l'une de ces deux paramétrisations, il est donc nécessaire d'ajouter des fonctions spécifiques de paramétrisation au sein des téléphones.

Enfin, cette approche (comme la RAP déportée) impose une connexion constante au réseau GSM pour bénéficier des services de RAP.

3.2 Choix applicatifs

Parmi les trois architectures présentées dans le chapitre précédent, celle qui offre le plus de liberté est certainement la reconnaissance embarquée. Cette option, bien que complexe à mettre en œuvre notamment à cause des ressources très limitées disponibles dans un système embarqué, a été retenue.

Les applications visées sont la reconnaissance de chiffres, le "name dialing", la commande vocale... Ces applications nécessitent uniquement de la reconnaissance de mots isolés. Pour nos travaux, nous avons décidé d'utiliser la reconnaissance de chiffres et la reconnaissance de commandes vocales comme critères de mesure. Ces applications imposent toutes l'utilisation d'un lexique dynamique, ce qui permet aussi au système d'être évolutif.

Ces exemples d'applications impliquent des contraintes. Les quatre principales que nous isolons concernent :
- le lexique ;
- la phase d'apprentissage ;
- l'indépendance ou non du système au locuteur ;
- l'environnement acoustique.

3.2.1 Lexique

Le lexique contient l'ensemble des mots que le système est susceptible de reconnaître. Les systèmes de RAP classiques utilisent généralement des lexiques très grands (de l'ordre de $65\,000$ mots). Certains utilisent même des lexiques appropriés à la nature des termes qu'ils vont reconnaître. Par exemple, dans leur travail quotidien, un médecin et un notaire n'emploient pas le même langage (vocabulaires différents). Pour rendre leurs systèmes plus performants, les laboratoires développent des versions spécifiques. Par exemple, Nuance, qui commercialise le système "Dragon NaturallySpeaking", propose une édition médicale et une édition notariale en plus de la version "classique".

Cette solution semble à l'heure actuelle être la meilleure en terme de performance. Cependant, elle est totalement inenvisageable pour les téléphones por-

tables. En effet, comme présenté dans 1.3.1, la quantité de mémoire disponible dans un téléphone est très limitée.

Par ailleurs, pour les applications envisagées (”name dialing”, reconnaissance de mots-clés ou numérotation automatique), il est difficile de prévoir tout le vocabulaire. Si pour les deux derniers exemples, les mots peuvent être connus (les chiffres par exemple), ce n'est pas le cas du ”name dialing”, pour lequel le problème est beaucoup plus complexe. En effet, l'utilisateur doit pouvoir choisir lui-même les mots dont il a besoin : un nom, un prénom, un surnom...

La solution retenue dans ce travail de thèse consiste à construire dynamiquement le dictionnaire. Cette solution permet de minimiser au maximum la taille du lexique sans limiter le nombre de ”mots” dont l'utilisateur peut se servir. Elle permet également de proposer une solution technique générique autorisant de nombreuses applications.

3.2.2 Apprentissage

La méthode du lexique dynamique implique une phase d'apprentissage pour chaque entrée du lexique. Chaque mot doit être modélisé (quel que soit le type du moteur de reconnaissance - HMM, DTW, réseau de neurones...) pour être inséré dans le lexique. Un lexique très complet, tel que ceux utilisés en reconnaissance de parole continue et qui contiennent généralement de l'ordre de 60k mots, est une solution envisageable. Cependant les ressources mémoire disponibles ne permettent pas de stocker un tel dictionnaire (en outre, la plupart des mots ne serviraient pas). De plus, dans le cadre du ”name dialing”, les lexiques génériques ne contiendraient pas les noms propres et les entrées par surnoms de l'utilisateur.

Dans la majorité des applications visées, ajouter des mots à partir d'une transcription orthographique est également possible mais cette solution n'a pas été étudiée dans un premier temps. Chaque fois que l'utilisateur voudra ajouter un mot, il devra le prononcer. Pour éviter une phase d'apprentissage trop longue, nous nous sommes imposés une contrainte forte : l'utilisateur ne devra prononcer le mot qu'une seule et unique fois pour l'insérer dans le dictionnaire.

Cette contrainte (très forte) ne permet pas d'envisager l'utilisation de modèle de mots. En effet, il n'est pas concevable d'apprendre un modèle HMM avec une seule occurrence de mot ; l'unité acoustique de base utilisée sera donc le phonème et la phonétisation des mots ajoutés dans le lexique sera le résultat d'une phase de décodage acoustico-phonétique ou d'une phonétisation par règles.

Durant nos expériences, comme nous n'avions pas forcément plusieurs occurrences du même mot pour un locuteur donné, nous avons utilisé une transcription phonétique exacte. Ceci ne peut avoir qu'un effet négatif, car la phase de phonétisation automatique que nous désirions utiliser aurait rendu le système "tolérant" à ses propres erreurs. En effet, s'il reconnaît un "in" à la place d'un "an" lors de la construction du lexique il est fort possible qu'il fasse la même erreur lors du décodage ; alors que la transcription exacte que nous avons simulée marquera bien un "an".

3.2.3 Système multi-locuteurs

Ce type de système doit être indépendant du locuteur. En effet, l'utilisateur final n'est pas forcément la personne qui a ajouté les numéros. Par exemple, le téléphone peut faire partie d'une flotte de téléphones portables partagée par un groupe de commerciaux.

Pour ces raison, le système de base présenté dans le chapitre 6 doit être un système indépendant du locuteur ; cependant dans le chapitre suivant (7) nous présentons des solutions pour adapter notre système à l'environnement et/ou au locuteur. L'approche présentée, bien que se voulant indépendante du locuteur, permet aussi de faire de la RAP dépendante du locuteur.

3.2.4 Environnement acoustique

L'environnement acoustique dans lequel un utilisateur peut utiliser un système embarqué est nettement plus variable que pour un système classique utilisé, généralement, seulement dans un bureau. Les systèmes auxquels nous faisons référence ici correspondent plus à des PDA, des téléphones portables ou des lecteurs MP3.

L'utilisateur sera donc amené à utiliser son système dans la rue, au bureau, dans la voiture. . . Ce système doit donc être très robuste au bruit et à la variété des environnements acoustiques.

3.3 Conclusion

Après avoir présenté les trois architectures envisageables (reconnaissance déportée, reconnaissance embarquée et reconnaissance répartie) pour la RAP dans les systèmes embarqués, le choix d'une reconnaissance embarquée a été

fait car cette architecture est la seule qui permette une utilisation entièrement autonome des fonctions de reconnaissance vocale. En effet, la reconnaisance répartie et la reconnaissance déportée font toutes deux appel à des serveurs de reconnaissance externes.

Différents choix applicatifs ont ensuite été présentés et justifiés, notamment concernant la constitution du lexique et la robustesse du système de reconnaissance.

Le lexique d'un SRAP doit contenir l'ensemble des mots à reconnaître, amenant les concepteurs de tels systèmes à proposer des lexiques de très grande taille pour minimiser le taux de mots hors vocabulaire. Cependant, dans le cadre de la RAP sur des systèmes embarqués (nous avons vu que la quantité de mémoire est très limitée) il n'est donc pas possible d'augmenter indéfiniment la taille du lexique. Pour contourner cette limitation , nous avons opté pour un lexique dynamique que l'utilisateur enrichit (via un décodage acoustico-phonétique ou une phonétisation à base de règle) au fur et à mesure des besoins.

Enfin, nous avons montré que les systèmes de RAP pour les téléphones (par exemple) devaient être robustes à l'environnement acoustique (bruyant et changeant). Dans le cadre d'appareils partagés (flotte de téléphones d'entreprise), le système de reconnaissance doit aussi être indépendant du locuteur.

Chapitre 4

Les modèles acoustiques compacts

Sommaire

Résumé

Ce chapitre est dédié aux modèles acoustiques compacts. Nous faisons un tour d'horizon des différentes approches permettant de réduire la taille des modèles acoustiques. Ces techniques s'orientent autour de deux grands axes : la réduction de la dimension de l'espace de représentation et la réduction du nombre de paramètres des modèles.

La réduction de la taille des modèles acoustiques n'est pas un problème récent. En effet, même si l'objectif n'était pas forcément l'intégration d'un SRAP dans un système embarqué, une grande quantité d'approches ont déjà été abordées dans la littérature.

Certaines techniques permettant la réduction des modèles acoustiques avaient pour but initial la mutualisation des paramètres, plutôt qu'une réelle recherche de diminution de la taille de ces modèles. Le tying et les modèles semi-continus (SCHMM) en sont la parfaite illustration. L'idée de ces approches était plus de palier un manque de données d'apprentissage que de diminuer l'empreinte mémoire des modèles.

La diminution du nombre de méta-paramètres d'un modèle acoustique permet aussi un gain en terme de temps de calcul. L'illustration la plus simple, de ce point, est la réduction du nombre de gaussiennes d'un GMM modélisant la fonction de probabilité d'émission d'un état. La vraisemblance devient alors nettement plus rapide à calculer pour un état modélisé avec 2 gaussiennes qu'un état représenté par une mixture de gaussiennes à 128 composantes (*i.e.* il suffit d'estimer la vraisemblance de 2 composantes au lieu des 128 initiales).

Les approches principales permettant, directement ou indirectement, une réduction de la taille des modèles sont abordées ci-après. Dans un premier temps, les techniques de réduction de l'espace de représentation sont présentées avant d'aborder les méthodes permettant de diminuer le nombre de paramètres des modèles.

4.1 Réduction de la dimension de l'espace de représentation

La réduction de la dimension de l'espace acoustique peut se faire principalement de deux manières :

- en supprimant simplement une partie des paramètres, comme les paramètres dynamiques, ou
- en projectant les paramètres dans un nouvel espace, de dimension inférieure.

Après une rapide discussion sur l'utilisation des paramètres dynamiques, nous présentons les deux principales méthodes de réduction de l'espace acoustique par projection vers un sous-espace : l'Analyse en Composante Principale (ACP) et l'Analyse Linéaire Discriminante (ALD).

4.1.1 Suppression des paramètres dynamiques

L'apport des paramètres dynamiques (Δ et $\Delta\Delta$) n'est plus à démontrer. Les différentes études de [Furui 1986], [Schwartz 1989], [Lee 1990] ou encore [Junqua 1993] montrent que l'utilisation des paramètres dynamiques permet une diminution significative du taux d'erreur que ce soit en reconnaissance de mots isolés, pour les travaux de Furui et de Junqua, ou en parole continue avec BYBLOS le système de BBN (travaux de Schwartz) ou SPHINX, le système de CMU (travaux de Lee).

Cependant, l'utilisation de ces paramètres multiplie par 3 la taille du vecteur acoustique ; en conséquence la taille de modèles acoustiques est elle aussi multipliée par ce même facteur. C'est pourquoi une manière rapide de réduire la taille des modèles acoustiques consiste à supprimer ces paramètres dynamiques. Le chapitre 6.1.2 présente plus en détail l'influence des paramètres dynamiques dans le cadre d'une application embarquée.

D'autres méthodes de réduction du nombre de paramètres (sans distinction de nature statique/dynamique) ont également été proposées, comme le "Knock-out" ([Sambur 1975]).

4.1.2 Méthodes d'analyse de données

Les méthodes d'analyse de données telles que l'ACP (Analyse en Composante Principale) et l'ALD (Analyse Linéaire Discriminante) sont couramment utilisées pour réduire l'espace de représentation des modèles acoustiques. Les vecteurs acoustiques sont composés d'un grand nombre de paramètres : classiquement, 39 coefficients correspondant aux 12 premiers coefficients cepstraux (MFCC ou PLP) plus l'énergie, auxquels sont ajoutées les dérivées de premier et second ordre.

Ces deux méthodes d'analyse de données ont pour but de partir d'un espace de dimensions N (ici 39) et de se projeter dans un espace de dimension inférieure R tel que $R \ll N$. Cependant, l'intention de la projection est différente pour l'ALD et l'ACP. L'ACP tente pour sa part de trouver un sous-espace permettant de maximiser la variance des données, alors que l'ALD cherche un sous-espace qui augmente la capacité discriminante de la représentation.

4.1.2.1 Analyse en Composantes Principales

L'ACP, notamment, utilisée dans le domaine de la parole par [Jankowski 1995], [Glass 1996], [Nouza 1996], a pour but, partant d'un espace initial de représentation, de se projeter dans un sous-espace, de dimension inférieure, dans lequel les données seront représentées de manière compacte et dont les axes sont décorrélés. Les axes (orthogonaux entre eux) du nouvel espace sont déterminés de telle manière qu'ils maximisent la variabilité des données. La variabilité suivant l'axe i est supérieure à celle de l'axe $i+1$ de part leur orthogonalité. Ce principe est illustré par la figure 4.1 pour un espace à deux dimensions.

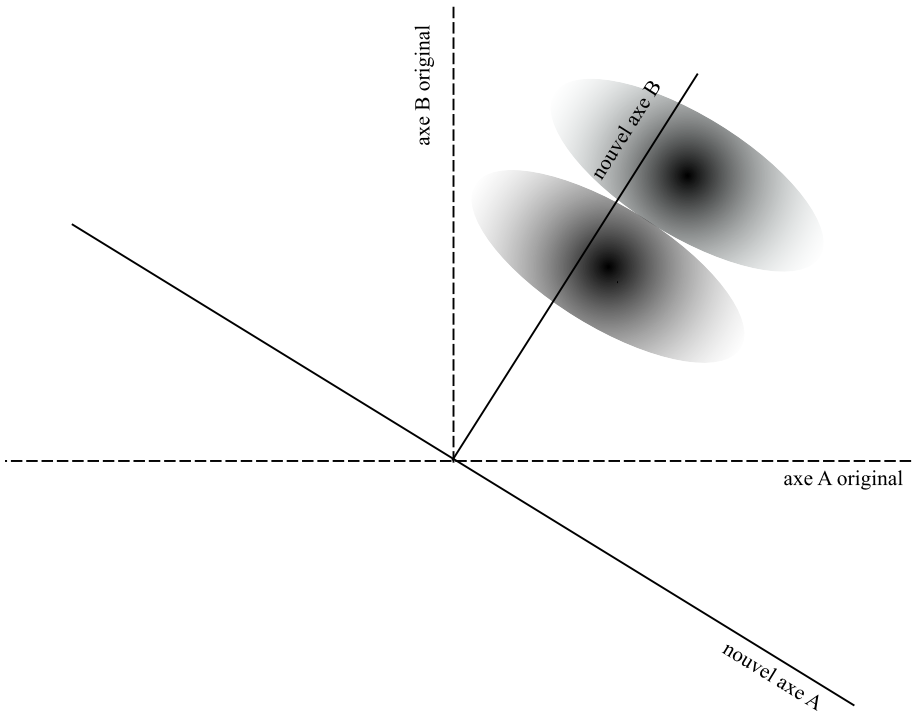

FIG. 4.1 – Exemple d'une Analyse en Composantes Principales d'un espace à deux dimensions.

Les axes principaux sont déterminés à partir de Σ, la matrice de co-variance des données.

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)' \tag{4.1}$$

où N correspond au nombre de vecteurs acoustiques (x_i) disponibles et μ est

le vecteur moyenne estimé grâce à :

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.2}$$

Les directions maximisant la variablité sont déterminées grâce aux valeurs propres (λ) et aux vecteurs propres associés (v_i). La rotation est donc définie par la matrice R :

$$R = [v_1 v_2 ... v_n] \tag{4.3}$$

A ce stade, le nouvel espace de représentation est de la même taille que l'espace inital. Ce nouvel espace permet juste de décorréler statistiquement les axes entre eux. Afin de réduire la taille du vecteur acoustique, seuls les axes comportant le maximum d'informations (déterminés par les plus grandes valeurs propres) sont conservés.

L'inconvénient principal de l'ACP est qu'elle est focalisée sur la recherche d'axes maximisant la variance des données sans tenir compte de la capacité discriminante des données.

4.1.2.2 Analyse Linéaire Discriminante

L'objectif principal de l'ALD (initialement présentée, pour le contexte de la RAP, par [Hunt 1989]), contrairement à l'ACP, est de séparer l'espace de manière à diminuer la variance intra-classe tout en augmentant la variance inter-classe : le choix des classes est donc un paramètre important à déterminer. Ceci impose également de connaître la classe à laquelle appartient chaque vecteur de paramètres.

La figure 4.2 illustre le fonctionnement de l'analyse discriminante avec des données en deux dimensions. On note que les deux distributions sont très peu discriminées, que ce soit suivant l'axe original A ou le B. Le premier axe obtenu avec l'ALD permet, lui, une classification nettement plus précise en discriminant bien les deux classes.

La nouvelle base est obtenue grâce aux vecteurs propres du produit : $\Sigma_{ec} * \Sigma_{ic}^{-1}$. Avec Σ_{ec} qui correspond à la variance entre-classe et Σ_{ic} la variance intra-classe.

La variance intra-classe, Σ_{ic}, que l'on cherche à diminuer, est estimée comme la somme pondérée de la matrice de covariance de toutes les classes :

$$\Sigma_{ic} = \frac{1}{N}\sum_{j=1}^{J} N_j \Sigma^j \tag{4.4}$$

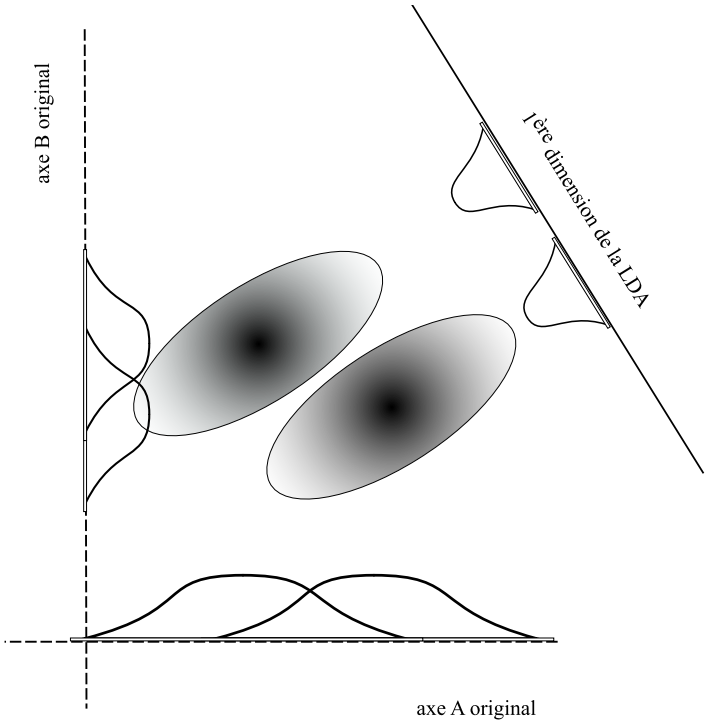

FIG. 4.2 – Exemple d'une Analyse Linéaire Discriminante d'un espace à deux dimensions.

où J est le nombre de classes, N_j le nombre de vecteurs associés à la classe j, N le nombre total de vecteurs et, enfin, Σ^j l'estimation de la matrice de covariance de la classe j définie par :

$$\Sigma^j = \frac{1}{N_j} \sum_{i=1}^{N_j} (x_i^j - \mu^j)(x_i^j - \mu^j)' \tag{4.5}$$

x_i^j correspond au $i^{ième}$ vecteur de la classe j. μ^j est la moyenne estimée de la classe j :

$$\mu^j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i \tag{4.6}$$

La variance entre-classe, Σ_{ec}, représente la variance à augmenter pour discriminer au mieux les différentes classes. Elle est estimée par :

$$\Sigma_{ec} = \frac{1}{N} \sum_{j=1}^{J} N_j(\mu^j - \mu)(\mu^j - \mu)' \tag{4.7}$$

$$= \Sigma - \Sigma_{ic} \tag{4.8}$$

μ est définie par l'équation 4.2.

De la même manière que pour l'ACP, les valeurs propres de $\Sigma_{ec} * \Sigma_{ic}^{-1}$ sont ordonnées pour le choix des axes (définis par les vecteurs associés aux valeurs propres).

[Eisele 1996] propose un comparatif entre des paramétrisations basées sur l'ALD, dans le domaine spectral et le domaine cepstral ; [Welling 1997] présente une autre approche pour l'utilisation de l'ALD : le vecteur t est concaténé avec les vecteurs $t-1$ et $t+1$ avant de faire une ALD pour revenir à la dimension initiale.

La HLDA (Heteroscedastic Linear Discriminant Analysis - [Kumar 1998]) correspond à une généralisation de l'ALD.

4.2 Réduction du nombre de paramètres des modèles

De nombreuses approches de réduction du nombre de paramètres des modèles sont proposées dans la littérature. L'approche la plus simple consiste à réduire le nombre de composantes du GMM associé à chaque état. La quantification vectorielle permet également une réduction de l'espace acoustique dans le cadre de la DTW. Des approches plus complexes sont basées sur le partage de paramètres tels que le tying ou les HMM semi-continus (SCHMM). Enfin, nous présentons le Subspace Distribution clustering HMM.

4.2.1 Réduction du nombre de gaussiennes

De la même manière que la réduction du nombre des paramètres par suppression des paramètres dynamiques, une réduction du nombre de paramètres peut être obtenue par une simple diminution du nombre de gaussiennes des GMM servant à modéliser les probabilités d'émission associées à chaque état des unités acoustiques.

Dans [Barras 1996] (chapitre 2.4, "Modèles indépendants du contexte"), le lecteur trouvera une étude complète sur les différentes combinaisons possibles entre réduction de la taille du vecteur acoustique et réduction du nombre de gaussiennes par état. Dans ces travaux, il est montré qu'une réduction du nombre de gaussiennes par état est généralement préférable à la suppression des paramètres dynamiques.

Cette approche est détaillée dans le chapitre 6.1.1 pour notre cadre applicatif.

4.2.2 La quantification vectorielle

L'idée initiale de la quantification vectorielle (Vector Quantization - VQ) était de diminuer le débit nécessaire lors de la transmission d'informations en utilisant un dictionnaire commun au niveau de l'émetteur et du destinataire afin de ne plus faire transiter l'information mais un code correspondant à cette information. Cette méthode permet effectivement de réduire le débit mais entraîne aussi une perte d'informations (généralement appelée *distorsion*).

En 1983, [Shore 1983] propose un système de reconnaissance basé sur la VQ qui s'affranchit de l'aspect temporel du signal de parole. Cela permet aussi de réduire l'espace mémoire nécessaire par comparaison à une approche DTW car il permet de ne retenir qu'un certain nombre de vecteurs (codewords dans la terminologie de Shore et Burton) pour une référence (codebook). La DTW nécessite, pour sa part, de garder l'ensemble des trames associées à une référence. La VQ apportait donc pour les systèmes DTW du début des années 80 un certain gain en terme d'occupation mémoire.

[Billi 1982] présente une intégration de la VQ dans des systèmes basés sur des HMM. En effet, au niveau de chacun des états, les GMM sont remplacés par un ensemble de codewords issus d'une VQ. Cette approche ne permet plus vraiment de gain en terme d'occupation mémoire. Cependant, un nouvel intérêt émerge : le calcul rapide de la vraisemblance. Comme présenté par [Bocchieri 1993] le GMM d'un état est partitionné en plusieurs classes (grâce à la VQ) et un représentant est estimé pour chaque classe. Lors du calcul de la vraisemblance d'une suite d'observations sachant un état, une première estimation est faite avec les représentants des classes. La vraisemblance exacte est ensuite calculée seulement pour les classes ayant une vraisemblance non négligeable.

En conclusion, la quantification vectorielle ne conduit pas nécessairement à une représentation acoustique très compacte. Par contre, elle autorise un gain important de temps de calcul.

4.2.3 Le tying

Le tying est basé sur le partage de paramètres entre modèles ou parties de modèles. Cette approche possède deux avantages principaux :

- une réduction de la quantité de données nécessaires pour un apprentissage robuste (elles sont mutualisées entre les états liés).
- un gain important en coût de calcul et/ou en coût de stockage. Au lieu de stocker N gaussiennes, il n'y en a plus que M à stocker (avec $M \ll N$) ; de même, lors du calcul de la vraisemblance, seules les M vraisemblances

élémentaires sont à calculer.

La première approche proposée par [Bellegarda 1990] consiste à partager les paramètres au niveau des mixtures de gaussiennes. [Young 1992] étend ce partage à l'ensemble des paramètres d'un système basé sur les HMM : les modèles, les états, les matrices de transition, les mixtures, les gaussiennes, les moyennes et les variances. La figure 4.3, reprise de [Young 1992], illustre cette proposition. [Young 1994] présente un comparatif entre du tying au niveau du modèle et du tying au niveau de l'état pour la construction de tri-phones.

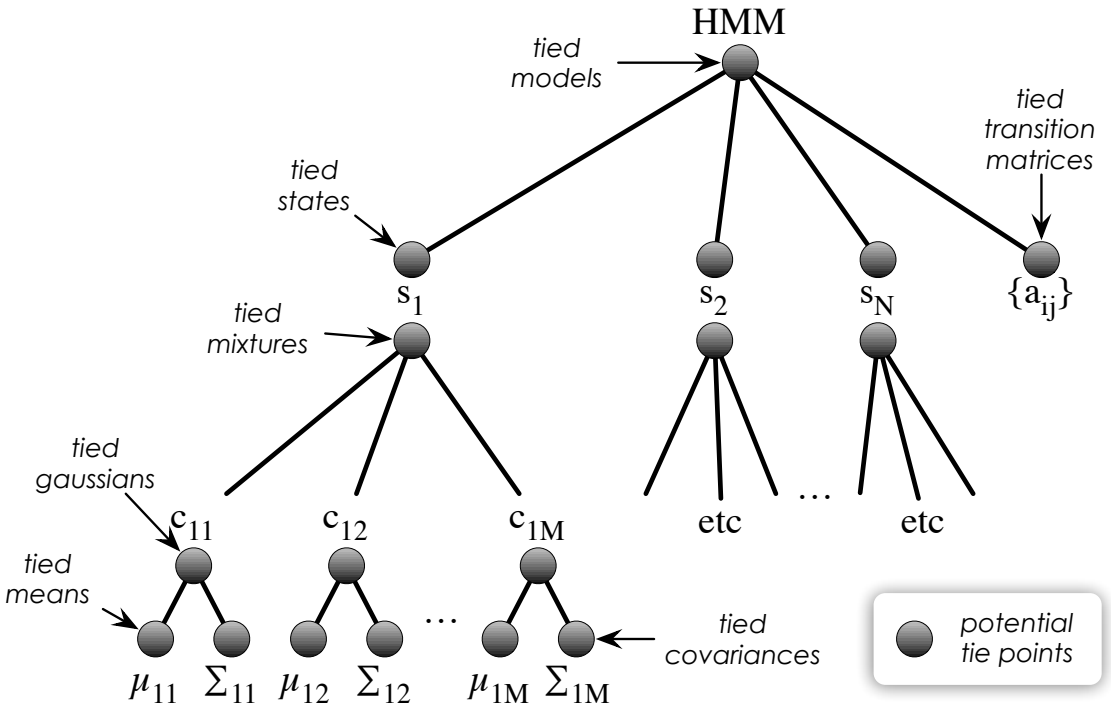

FIG. 4.3 – Représentation des différents niveaux possibles pour le partage de paramètres proposée par Young dans [Young 1992]

Une des problématiques de l'apprentissage de modèles acoustiques 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, il est généralement recommandé d'augmenter le nombre de gaussiennes des mixtures afin d'obtenir des modèles plus précis ; cependant il faut faire attention à disposer de suffisament de données pour apprendre correctement les mixtures. Ceci est encore plus vrai lors de l'utilisation de modèles contextuels car le nombre d'états devient nettement plus important. Young propose une méthode pour l'apprentissage de modèles de triphones sans avoir de données étiquetées "tri-phones", uniquement par partage de paramètres (*cf.* [Young 1992]). Cette approche per-

met aussi de modéliser correctement des phonèmes contextuels pour lesquels aucune donnée d'apprentissage n'est disponible. Dans les systèmes récents, le partage est fait au niveau des états modélisants un même phonème en contextes. Ce cloisonnement limite la taille minimale qui pourrait être atteinte par le partage d'état.

Le partage systématique de toutes les mixtures de gaussiennes entre les états porte un nom particulier : les HMM semi-continus (SCHMM). Les SCHMM peuvent être vus comme un cas particulier du tying.

4.2.4 Les modèles de Markov semi-continus (SCHMM)

Les modèles semi-continus[1] ont été abordés très tôt dans la littérature. Huang propose plusieurs articles dès le début des années 1990 ([Huang 1989], [Huang 1992]) bien avant l'apparition de la problématique de la reconnaissance embarquée. L'utilisation des HMM semi-continus (SCHMM) permettait de réduire le nombre de paramètres (comparé aux HMM continus) tout en conservant une modélisation suffisamment détaillée (comparé aux HMM discrets).

Contrairement aux HMM continus qui apprennent, état par état, une mixture de gaussiennes, tous les états des HMM semi-continus partagent le même ensemble de gaussiennes. Les états sont ensuite différenciés entre eux par un simple vecteur de poids. Ces poids sont généralement ré-estimés, avec des données propres à chaque état, suivant la formule de l'algorithme *Baum_Welch* (le lecteur pourra se référer au chapitre 2.3.2.2 et notamment à l'équation 2.42 pour plus de détails).

Dans le cadre des modèles semi-continus, la vraisemblance d'une trame x pour un état i s'exprime par :

$$P_i^{SCHMM}(x) \;=\; \sum_{m=1}^{M} c_{im}\, \mathcal{N}(x, \mu_m, \sigma_m) \tag{4.9}$$

où M correspond au nombre de composantes du GMM, c_{im} au poids de la $m^{ième}$ gaussienne de l'état i et $\mathcal{N}(x, \mu_m, \sigma_m)$ à la $m^{ième}$ loi gaussienne du GMM de moyenne μ_m et de variance σ_m.

Dans la suite de ce document, au chapitre 6.2.2.2, nous présentons une alternative à cette approche, incluant une ré-estimation des poids selon un critère discriminant.

[1] Dans la littérature, plusieurs angliscismes sont utilisés pour les modèles semi-continus. On trouve aussi bien "semi-continuous HMM" que "tied-mixture".

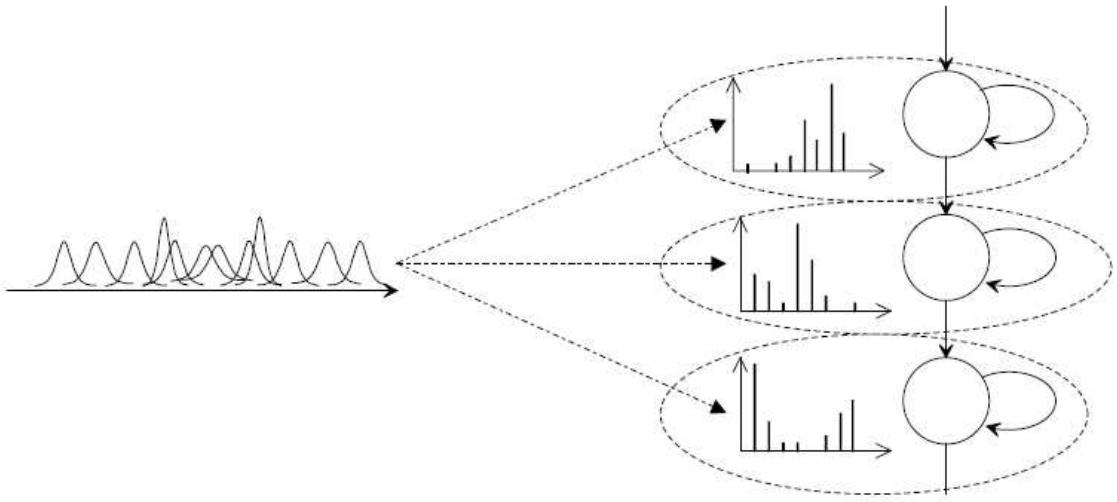

FIG. 4.4 – HMM semi-continu : les états sont différenciés entre eux uniquement par un vecteur de poids.

La construction du pool de gaussiennes était initialement basée soit sur un codebook issu d'une VQ dont les paramètres (moyennes ou moyennes et variances) sont ré-estimés pour permettre notamment le recouvrement entre composantes (impossible avec la VQ), soit sur un clustering des gaussiennes issues d'un HMM classique. Dans le chapitre 6.2.1 nous proposons une variante en transformant ce pool de gaussiennes en véritable GMM (par ré-estimation des moyennes et des poids de chaque composante).

Un des objectifs des SCHMM est d'obtenir des modèles avec peu de paramètres et ce spécialement dans le cadre de la reconnaissance de la parole embarquée dans des systèmes mobiles. [Duchateau 1998] présente différents critères pour ne pas stocker un poids pour chacune des gaussiennes du GMM mais uniquement pour un sous-ensemble de ces gaussiennes (suffisant pour représenter correctement un état donné). Quatre règles sont proposées :

- nombre de gaussiennes fixé : les N gaussiennes de poids le plus élévé sont alors gardées. Ce N est indépendant de l'état.
- seuil sur le poids : les gaussiennes possédant un poids inférieur à un seuil fixé *a priori* ne sont pas conservées. Le nombre de gaussiennes varie alors suivant les états.
- seuil sur le pourcentage des poids : les gaussiennes sont ordonnées suivant leur poids. Puis seules les N gaussiennes permettant d'atteindre $X\%$ des poids sont conservées. Chaque état peut avoir un nombre de gaussiennes différent.
- seuil sur l'occupation : si l'occupation d'une gaussienne (pour des données) pour un état est inférieure à un seuil fixé *a priori* alors cette gaussienne

n'est pas conservée pour cet état. Dans ce cas aussi, le nombre de gaussiennes par état peut différer.

L'utilisation de ce dernier critère donne les meilleurs résultats.

[Vaich 1999] présente un comparatif entre les CDHMM et les SCHMM dans le cadre de la reconnaissance de chiffres isolés (avec le corpus TIDIGITS). Les taux d'erreur sont affichés pour différentes tailles de modèles acoustiques. Nous pouvons noter que pour les modèles compacts le WER d'un système CDHMM est supérieur à 20% alors que le WER d'un système SCHMM est proche de 5%[2]. Les SCHMM semblent nettement plus performants dans le cadre de modèles acoustiques compacts (pour les gros modèles, les WER sont similaires).

Enfin, [Macas-Guarasa 1996] propose une amélioration pour les SCHMM en utilisant plusieurs GMM. Des GMM dépendant du sexe sont utilisés et permettent un gain significatif dans le cadre de la reconnaissance de mots isolés dans un contexte téléphonique.

4.2.5 Subspace Distribution Clustering HMM (SDCHMM)

Dans [Bocchieri 1997] et [Bocchieri 2001], Bocchieri et Mak proposent une nouvelle approche pour les systèmes de reconnaissance basés sur les HMM appelée Subspace Distribution Clustering HMM (SDCHMM). L'idée principale est de partir d'un modèle HMM classique (utilisant le tying ou non d'ailleurs) puis de séparer les vecteurs acoustiques en différents flux (streams ou subspaces) et, enfin, de faire un clustering pour réduire le nombre de gaussiennes de ces flux. La figure 4.5, tirée de [Bocchieri 2001], illustre cette idée.

Du point de vue du calcul de la probabilité d'émission d'une trame x pour un état i ; cela donne, pour le CDHMM :

$$P_i^{CDHMM}(x) \;=\; \sum_{m=1}^{M} c_{im}\, \mathcal{N}(x, \mu_{im}, \sigma_{im}) \tag{4.10}$$

où c_{im} correspond au poids de la $m^{ième}$ gaussienne de l'état i et $\mathcal{N}(x, \mu_{im}, \sigma_{im})$ à la gaussienne m de l'état i. Après la création des flux, nous obtenons :

$$P_i^{CDHMM}(x) \;=\; \sum_{m=1}^{M} c_{im} \left(\prod_{k=1}^{K} \mathcal{N}(x_k, \mu_{imk}, \sigma_{imk}) \right) \tag{4.11}$$

Enfin le clustering de chacun des flux donne :

$$P_i^{SDCHMM}(x) \;=\; \sum_{m=1}^{M} c_{im} \left(\prod_{k=1}^{K} \mathcal{N}^{quan}(x_k, \mu_{imk}, \sigma_{imk}) \right) \tag{4.12}$$

[2] les auteurs ne précisent pas si la taille des différents modèles est équivalente

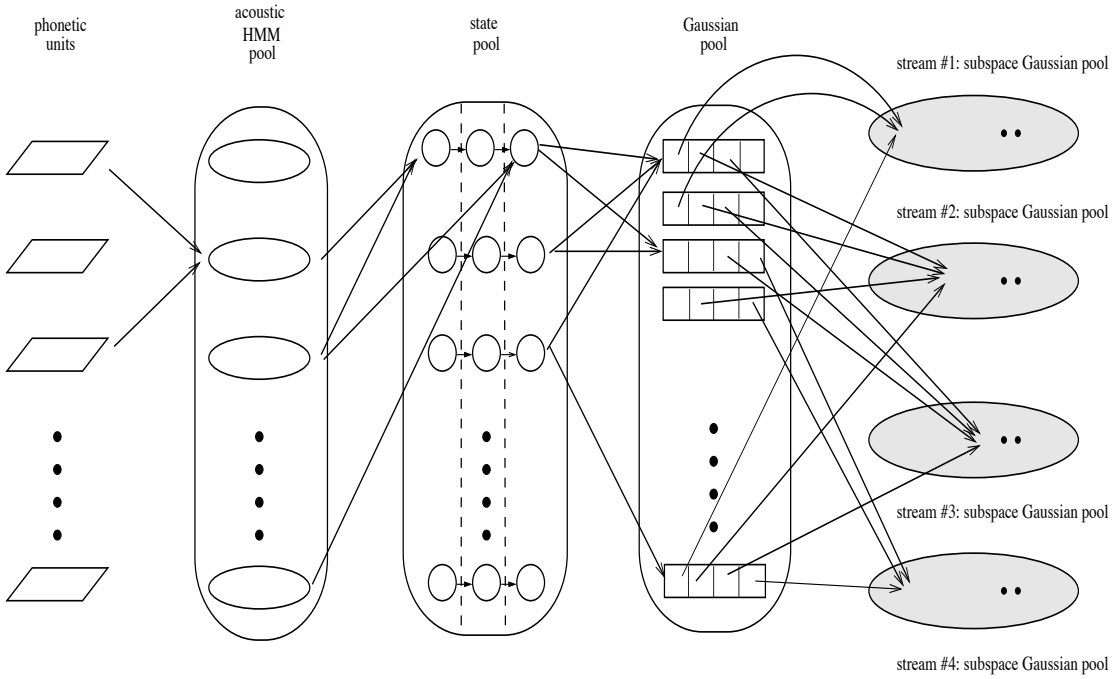

FIG. 4.5 – Exemple d'un "Subspace Distribution Clutering HMM" à 4 flux

L'utilisation des SDCHMM permet un gain du point de vue du stockage des modèles acoustiques grâce au partage des paramètres à tous les niveaux (HMMs, états, gaussiennes). Les SDCHMM apportent aussi un gain en terme de calcul. Ce gain est dû d'une part, au clustering qui autorise une réduction du nombre de composantes et d'autre part au partage des gaussiennes[3] qui permet de pré-calculer la vraisemblance des "clusters" en amont afin d'estimer la vraisemblance des états par une simple somme de ces vraisemblances pré-calculées.

Cette approche semble similaire à l'utilisation des SCHMM avec des flux. Cependant, en regardant l'expression de la vraisemblance d'un état i pour un frame x utilisant K flux (*cf.* l'équation 4.13), on peut noter que les opérateurs $\sum$ et $\prod$ sont inversés.

$$P_i^{SCHMM\_stream}(x) \;=\; \prod_{k=1}^{K} \sum_{m=1}^{M} c_{imk}\, \mathcal{N}^{quan}(x_k, \mu_{mk}, \sigma_{mk}) \tag{4.13}$$

Les SCHMM supposent une indépendance entre les flux, alors que les SDCHMM supposent que les flux sont dépendants.

Plusieurs remarques peuvent être faites à propos des SDCHMM :

[3] comme pour les SCHMM

- si K=1, un seul flux utilisé, alors le modèle SDCHMM n'est autre que CDHMM classique.
- si chacun des flux comprend un seul paramètre alors l'approche SDCHMM correspond au feature-tying présenté dans [Sagayama 1995].

Dans la version originale, Bocchieri et Mak proposaient de dériver les paramètres du SDCHMM depuis un CDHMM. Dans [Mak 2001], ils proposent d'apprendre les paramètres directement sans passer par l'étape du CDHMM.

Enfin, Cho propose dans [Cho 2004] un modèle hybride combinant les SCHMM et les SDCHMM. Cette approche permet de réduire un peu plus la taille du modèle acoustique.

4.3 Conclusion

Un grand nombre de méthodes permettant de réduire la taille des modèles acoustiques sont présentées dans la littérature.

Les méthodes basées sur la réduction de nombre de paramètres[4] ont généralement été pensées pour répondre au problème de manque de données durant la phase d'apprentissage des modèles acoustiques.

D'autres approches se focalisent sur la réduction de la taille des modèles acoustiques, en particulier par la réduction de la dimension de l'espace de représentation[5].

Ces approches permettent toutes d'obtenir directement ou indirectement une réduction de la taille des modèles acoustiques.

[4] quantification vectorielle, tying, Semi-Continuous HMM ou encore Subspace Distribution Clustering HMM

[5] notamment différentes méthodes d'analyse de données

Deuxième partie

Travail réalisé

Chapitre 5

Corpus et contexte expérimental

Sommaire

Résumé

Ce chapitre présente les trois corpus utilisés pour valider les méthodes proposées : BREF, BDSON et VODIS. Le premier a été uniquement utilisé pour l'apprentissage initial des modèles acoustiques. Les deux derniers correspondent à des conditions acoustiques différentes : BDSON étant un corpus enregistré en laboratoire et VODIS un corpus enregistré dans des voitures. La seconde partie de ce chapitre présente le cadre expérimental utilisé durant ce travail.

Notre étude se limite au domaine acoustique de la reconnaissance de la parole. Dans les chapitres 6 et 7, nous proposons plusieurs approches pour réduire l'espace nécessaire au modèle acoustique ainsi que pour améliorer sa représentation.

Afin d'évaluer l'influence de toutes ces approches, il est nécessaire de définir un critère objectif de mesure de qualité ; nous évaluerons chaque méthode en terme de reconnaissance de mot. Pour cela, nous définissons deux taux d'erreur, le DER (Digit Error Rate) et le CER (Command Error Rate), de la manière suivante :

$$DER = \frac{\textit{nombre de chiffres mal reconnus}}{\textit{nombre de chiffres à reconnaître}}$$

$$CER = \frac{\textit{nombre de commandes vocales mal reconnues}}{\textit{nombre de commandes vocales à reconnaître}}$$

Dans les deux sections suivantes, nous présenterons les différents corpus utilisés ainsi que le protocole expérimental mis en place pour l'évaluation des méthodes.

5.1 Corpus

Pour modéliser l'espace acoustique avec une approche stochastique, un grand ensemble de données est nécessaire. Pour cela, nous utilisons trois corpus répartis en deux groupes :

- apprentissage des modèles acoustiques : **BREF**, et
- adaptation et évaluation des approches proposées : **BDSON** et **VODIS**.

5.1.1 BREF

Ce corpus a été développé en 1991 ([Lamel 1991]) par le LIMSI afin de fournir un ensemble de données françaises permettant le développement et l'évaluation de systèmes de dictée vocale. Son développement a été financé par plusieurs partenaires : le GDR-PRC Communication Homme/Machine, la CEE (projet ESPRIT Polyglot) et l'Aupelf-Uref.

L'ensemble des données *TRAIN* représente plus de 100 heures de paroles lues (non spontanées) prononcées par un ensemble de 120 locuteurs (65 femmes et 55 hommes). Les conditions d'enregistrement sont optimales, ceux-ci ayant été effectués dans une chambre sourde. Les phrases sont toutes extraites du journal *Le Monde*. Il s'agit donc d'un style journalistique.

Ce corpus a été utilisé uniquement pour apprendre les modèles acoustiques initiaux (avant adaptation), car les deux autres corpus utilisés (BDSON et VODIS) ne disposaient pas d'assez de données pour apprendre de manière satisfaisante les modèles acoustiques. De plus, cette répartition des corpus correspond à une situation réaliste dans laquelle on ne dispose que d'une faible quantité de données spécifiques aux conditions réelles d'utilisation des systèmes de reconnaissance.

5.1.2 BDSON

Le corpus BDSON ([Carré 1984]) est composé de phrases phonétiquement équilibrées, de suites CVC, de logatomes, de chiffres... Les séquences ont été prononcées par 32 locuteurs (16 hommes et 16 femmes).

Seul le sous-corpus contenant les chiffres isolés a été retenu (composé d'enregistrements prononcés par 15 hommes et 15 femmes). Ce sous-corpus a été divisé en deux sous-ensembles :

- un pour l'adaptation (BADAPT). Il contient 700 occurrences de chiffres prononcées par 7 locuteurs (4 hommes et 3 femmes). Ce sous-ensemble a été utilisé pour adapter le modèle acoustique générique. Il a également été utilisé pour construire les modèles spécifiques aux phonèmes.
- un pour l'évaluation (BTEST). Il est composé de 2 300 occurrences de chiffres prononcées par 23 locuteurs (11 hommes et 12 femmes).

Les locuteurs du corpus BDSON sont différents de ceux du corpus BREF. Les locuteurs des corpus BADAPT et BTEST sont aussi différents. Les tests sont réalisés en mode indépendant du locuteur (sauf mention contraire explicite), les locuteurs de test n'apparaissant jamais durant les phases d'apprentissage.

Le corpus de test étant limité à des chiffres isolés, les résultats sont exprimés en DER (Digit Error Rate - taux d'erreur de reconnaissance de chiffres).

5.1.3 VODIS

VODIS[1] [Geutner 2000] est un corpus français dédié aux applications embarquées dans une voiture. Il a été réalisé dans le cadre du projet européen VODIS (Voice Operated Driver Information System) de la branche Telematics. Cette branche s'intéresse à l'intégration d'un système complexe de navigation par satellites à l'intérieur d'un véhicule et à sa commande par la voix. Le but de

[1] Nous souhaitons remercier le LORIA qui nous a fourni ce corpus.

ce projet est le développement d'une interface vocale pour un système de navigation de la marque Blaupunkt : le Berlin RCM202A. L'interface vocale interagit avec le système de navigation, le téléphone cellulaire et l'autoradio (lecteur CD et radio).

Dans le cadre de ce projet, une certaine quantité de données a été collectée. Ce corpus est composé d'enregistrements effectués par 200 personnes dans deux voitures : Peugeot 406 et Renault Safrane. Il contient une grande variété de données : lettres, chiffres, commandes vocales, mots épelés, phrases phonétiquement équilibrées...

Ces enregistrements sont réalisés avec plusieurs microphones (close-talk et far-talk - micro plus ou moins distant du locuteur). L'environnement acoustique varie suivant les différentes sessions d'enregistrement (les fenêtres sont ouvertes ou non, la radio est allumée ou non, la climatisation est activée ou pas...).

Seules les parties contenant les commandes isolées enregistrées en condition close-talk (près de la bouche) ainsi que les phrases phonétiquement équilibrées ont été utilisées.

Nous avons divisé le corpus VODIS en deux sous-ensembles :
- le premier, pour l'adaptation au contexte (VADAPT) contient 2 712 commandes vocales prononcées par 39 personnes ;
- le second (VTEST), pour l'évaluation contient 11 136 commandes. Elles sont prononcées par 160 locuteurs qui prononcent les 70 commandes différentes (en moyenne).

De même qu'avec BDSON, les locuteurs de VODIS sont différents de ceux de BREF et de BDSON. Ils sont aussi différents dans VADAPT et VTEST.

La liste des commandes peut être trouvée dans l'annexe B.

Les phrases phonétiquement équilibrées sont utilisées pour réaliser l'adaptation au locuteur dans le chapitre 7.

5.2 Contexte expérimental

Les applications les plus intéressantes pour la reconnaissance vocale embarquée sont très certainement la reconnaissance de nom ("name dialing") ou la reconnaissance de numéro de téléphone. Ces deux orientations concernent la reconnaissance de mots isolés, ce qui, compte-tenu des ressources disponibles d'un téléphone (par exemple), semble être la seule orientation possible à court terme (les SRAP classiques exigent beaucoup trop de ressources pour être

intégrés dans un téléphone).

Les corpus disponibles permettent de répondre partiellement aux applications envisagées. En effet, nous disposons de deux corpus :

- BDSON contient des chiffres isolés dans un environnement acoustique idéal.
- VODIS est composé de commandes vocales destinées à commander un téléphone, un GPS et un autoradio. Les commandes ont été enregistrées au sein même de voitures, donc dans un environnement réaliste (bruité).

Nous allons effectuer de la reconnaissance de chiffres isolés et de la reconnaissance de commandes vocales.

Le système de reconnaissance de mots isolés que nous utilisons est issu du TOOLKIT de reconnaissance de parole du LIA (SPEERAL [Nocera 2002]). Notre moteur de reconnaissance utilise un simple Viterbi qui donne le mot du lexique ayant la vraisemblance maximale.

Chapitre 6

Réduction de la taille des modèles acoustiques

Sommaire

Résumé

Dans ce chapitre, nous présentons deux méthodes de réduction de la taille des modèles acoustiques. L'objectif visé consiste à intégrer le système de reconnaissance dans des systèmes embarqués. Nous commençons par présenter une approche simple permettant d'établir un système de référence. La seconde approche constitue le cœur de ce travail de thèse. Elle repose sur l'utilisation d'un modèle GMM général que nous dérivons, par transformation simple, pour obtenir les modèles acoustiques d'état.

Comme décrit auparavant, une des contraintes principales des systèmes embarqués réside dans la faiblesse des ressources mémoires disponibles (*cf.* 1.3.1).

Les systèmes de RAP actuels sont, en majorité, basés sur une méthode Markovienne. Comme nous l'avons présenté précédemment (*cf.* 2.3.2.2), cette approche nécessite de disposer de modèles pour chaque unité acoustique (mots, phonèmes, di-phones, tri-phones, penta-phones...). Ces besoins ne sont pas en adéquation avec les capacités des systèmes embarqués. Ce point constitue le c œur de ce travail de thèse : nous allons essayer d'apporter quelques éléments de réponse concernant la réduction de la taille des modèles acoustiques.

La problématique principale de ce document concerne donc l'espace mémoire nécessaire au stockage des modèles acoustiques des systèmes de RAP. Pour un modèle acoustique HMM classique, cette quantité peut être estimée de la manière suivante :

$$nb\_comp * (\underbrace{2 * nb\_param * taille\_param}_{\text{param. d'une composante}} + \underbrace{taille\_param}_{\text{poids d'une composante}}) \quad (6.1)$$

où nb_comp correspond au nombre de composantes (généralement des lois gaussiennes) utilisées pour modéliser l'environnement acoustique (généralement égal à la somme du nombre de gaussiennes des GMM de chaque état émetteur des HMM), où nb_param correspond au nombre de paramètres de chaque composante (taille du vecteur acoustique) et où $taille\_param$ représente la taille mémoire nécessaire pour stocker un paramètre (double, float, int...).

Dans un souci de simplification, le problème de l'optimisation du codage des paramètres n'a pas été abordé dans ce document. Le système actuel de RAP du LIA (SPEERAL) utilisant des doubles, nous avons fait de même : le paramètre $taille\_param$ vaut donc 8 octets. Cependant, pour une implémentation réelle dans un système embarqué, l'utilisation du codage en nombre entier est obligatoire (*cf.* [Huggins-Daines 2006] pour PocketSphinx ou [Franco 2002]).

Les systèmes classiques de reconnaissance de parole continue grand vocabulaire sont généralement composés de plusieurs millions de paramètres (voir plusieurs dizaines de millions). Par exemple, le système SPEERAL du LIA supporte plus de $10\ 000$ modèles de tri-phones représentés par $3\ 600$ états émetteurs[1] eux-mêmes caractérisés par une mixture de 64 gaussiennes. Le nombre total de paramètres dépasse alors les 13 millions de paramètres ($3\ 600 * 64 * (2 * 39 + 1)$). Le signal est représenté par un vecteur acoustique à 39 coefficients (12 et l'énergie + leurs dérivées de premier et second ordre - les Δ et $\Delta\Delta$). L'empreinte mémoire d'un tel modèle acoustique complet peut être estimée à environ 300Mo. Nous

[1] L'utilisation du tying permet d'obtenir $10\ 000$ tri-phones avec seulement $3\ 600$ états.

avons vu qu'un téléphone portable standard dispose d'une petite centaine de Ko pour les DATA ROM/RAM, ce qui ne permet évidemment pas d'intégrer les modèles acoustiques standards.

Dans la suite de ce document, afin d'estimer la quantité d'espace mémoire requise nous proposons d'utiliser le nombre de paramètres comme critère d'évaluation. En effet, la taille (en Mo) du fichier du modèle acoustique n'a que peu de sens étant donné que nous n'abordons pas le problème du codage des paramètres (tel que le codage en nombre entier). Ce qui nous donne pour un modèle HMM classique :

$$nb\_comp * (\underbrace{2 * nb\_param}_{\text{param. d'une composante}} + \underbrace{1}_{\text{poids d'une composante}}) \tag{6.2}$$

Un modèle acoustique avec des phonèmes non-contextuels, 128 composantes gaussiennes par état et un vecteur acoustique de 39 coefficients comprend plus de 1 million de paramètres ($13\,824 * (2 * 39 + 1)$), ce qui ne l'autorise toujours pas à être intégré dans un système embarqué.

Dans le but d'obtenir un modèle respectant les contraintes mémoires d'un téléphone portable, une première approche, simple, consistant à réduire le nombre de paramètres pour modéliser les états du HMM a été mise en place. Dans un premier temps, nous étudions l'influence du nombre de gaussiennes par état sur le CER[2] et le DER[3]. Dans un second temps, pour diminuer significativement la taille des modèles acoustiques, nous avons réduit la taille du vecteur acoustique. En effet, nous avons opté pour une paramétrisation PLP (qui semble la plus performante comparée aux MFCC, LPCC, LPC - *cf.* [Lévy 2004], [Chen 2005] ou [Psutka 2001]) avec 12 coefficients statiques ainsi que l'énergie. Nous n'avons donc pas utilisé les coefficients dynamiques (Δ et $\Delta\Delta$). Cette approche n'est pas optimale ; elle a pour objectif d'établir des références (*baselines*) pour comparer nos approches.

Enfin, une méthode basée sur la modélisation de l'espace acoustique par un GMM général sera présentée. Elle est développée en deux étapes dont une est optionnelle. La première étape correspond à une ré-estimation des poids des gaussiennes du GMM général pour décrire chaque état. Elle est similaire aux approches des HMM semi-continus présentés dans 4.2.4. Cependant, la méthode de ré-estimation des poids peut être différente (nous présentons une méthode originale pour la ré-estimation des poids à l'aide d'un critère discriminant) ainsi que l'approche pour la construction du dictionnaire de gaussiennes initial. L'autre étape que nous proposons peut être considérée comme

[2]CER - Command Error Rate - pourcentage de commandes mal reconnues
[3]DER - Digit Error Rate - pourcentage de chiffres mal reconnus

un préalable à la ré-estimation des poids. Elle est caractérisée par la transformation du GMM général en un GMM local via une simple transformation linéaire. La même transformation est effectuée sur toutes les composantes du GMM.

6.1 Système de référence

Afin de pouvoir comparer notre approche avec un système de référence (que l'on dénommera *baseline*) nous partons d'un modèle de HMM classique et nous essayons de réduire simplement sa taille sans toucher à sa structure. Les techniques de réduction de l'espace acoustique (LDA, ACP) ou de partage de paramètres (tying) ne sont pas abordées. Nous nous sommes intéressés uniquement aux macro-paramètres. Pour cela, nous effectuons deux étapes successives :

- la réduction du nombre de gaussiennes par état ;
- la réduction du nombre de paramètres du vecteur acoustique.

Il s'agit des deux facteurs principaux qui interviennent pour estimer la taille d'un modèle acoustique. Un troisième facteur existe : le nombre d'états émetteurs ; cependant la mise en place d'un système de tying d'états émetteurs (*cf.* 4.2.3) n'a pas été évaluée.

6.1.1 Réduction du nombre de gaussiennes par état

La réduction du nombre de gaussiennes par état permet de donner des indications sur l'évolution du taux d'erreur (CER et DER - se référer au chapitre 5 pour la définition de ces deux indicateurs) face à cette réduction drastique de la taille du modèle acoustique. Comme nous l'avons énoncé précédemment, l'objectif est de construire simplement un modèle acoustique capable d'être embarqué dans un téléphone portable sans chercher une solution optimale. Cette réduction du nombre de gaussiennes ne réduit pas suffisamment la taille du modèle acoustique pour les systèmes embarqués et ne correspond qu'à une première étape - les premiers modèles respectant les contraintes matérielles sont présentés dans la section suivante.

Le tableau 6.1 présente l'évolution du CER et de la taille des modèles acoustiques en fonction du nombre de gaussiennes de chaque état émetteur. Les HMM sont appris avec le corpus BREF puis une phase d'adaptation MAP (poids, moyenne et variance - [Gauvain 1994]) est effectuée avec la sous-partie VADAPT de VODIS.

Nous pouvons noter qu'avec le modèle acoustique le plus performant (utilisant 128 gaussiennes par état - non-contextuel - et un vecteur acoustique de 39 coefficients) le CER est de 1,80%. Cependant, ce modèle acoustique ne peut pas être intégré du fait de son trop grand nombre de paramètres (plus de 1 million). Les modèles à 2 et 4 gaussiennes par état (correspondant respectivement à 17 064 et 34 128 paramètres) représentent de bons compromis performance/taille du modèle. Les CER de ces modèles sont respectivement de 5,48% et de 3,40%.

# gauss/état	# paramètres	CER
2	17 064	5,48%
4	34 128	3,40%
128	1 092 096	1,80%

TAB. 6.1 – Evolution du CER et de la taille du modèle acoustique en fonction du nombre de gaussiennes des états émetteurs (non-contextuels). Le vecteur acoustique comprend 39 coefficients (12 PLP plus l'énergie avec leurs Δ et leurs $\Delta\Delta$). 11 136 tests effectués sur le corpus VODIS (commandes vocales / bruité).

Cette réduction de la taille du modèle acoustique par un facteur de plus de 30 (pour le modèle à 4 gaussiennes par état) entraîne une augmentation absolue du CER de moins de 1,60%. Le modèle à 2 gaussiennes par état, pour sa part, permet une réduction de l'espace mémoire par un facteur supérieur à 60.

Une deuxième série d'expériences a été effectuée sur un corpus de chiffres enregistrés dans de bonnes conditions : BDSON. Les modèles acoustiques sont appris de la même manière que pour les expériences avec VODIS : première phase d'apprentissage avec les données de BREF puis adaptation avec le sous-corpus BADAPT de BDSON. Les tests sont effectués sur BTEST (2 300 tests).

# gauss/état	# paramètres	DER
2	17 064	1,48%
4	34 128	0,96%
128	1 092 096	0,96%

TAB. 6.2 – Evolution du DER et de la taille du modèle acoustique en fonction du nombre de gaussiennes des états émetteurs. Le vecteur acoustique comprend 39 coefficients (12 PLP plus l'énergie avec leurs Δ et leurs $\Delta\Delta$). 2 300 tests effectués sur le corpus BDSON (chiffres / non bruité).

Les résultats sont similaires à ceux obtenus sur VODIS (*cf.* tableau 6.2). Le DER pour le modèle acoustique le plus complet est de 0,96%. Les modèles avec 4 gaussiennes et 2 gaussiennes ont respectivement un DER à 0,96% et 1,48% (ce qui correspond à une augmentation absolue de DER de 0,52% pour le dernier

modèle).

6.1.2 Réduction de la taille du vecteur acoustique

Associée à l'approche précédemment décrite, la réduction de la taille du vecteur acoustique de 39 coefficients à 13 coefficients (suppression des Δ et $\Delta\Delta$) permet d'obtenir la *baseline* d'un modèle acoustique dont la taille correspond aux capacités d'un système embarqué.

# gauss/état	# paramètres	CER
2	5 832	5,80%
4	11 664	4,80%
128	373 248	3,94%
full[4]	1 092 096	1,80%

TAB. 6.3 – Evolution du CER et de la taille du modèle acoustique en fonction du nombre de gaussiennes par état émetteur. Le vecteur acoustique comprend 13 coefficients (12 PLP plus l'énergie). 11 136 tests effectués sur le corpus VODIS (commandes vocales / bruité).

La suppression de l'information dynamique amène une nouvelle augmentation du CER (*cf.* tableau 6.3). Pour les modèles à 2 gaussiennes par état ce CER passe de 5,48% à 5,80%. Cela représente une augmentation relative du CER de 6% alors que le nombre de paramètres a été divisé par 3. De même, pour le modèle à 4 gaussiennes, le CER passe de 3,40% à 4,80% (pour la même réduction du nombre de paramètres).

La même technique appliquée sur le corpus BDSON donne des résultats similaires. En effet, les DER (*cf.* tableau 6.4) passent respectivement de 1,48% à 4,96% et de 0,96% à 4,43%.

Le DER du modèle à 128 gaussiennes est très légèrement supérieur au DER obtenu avec le modèle à 4 gaussiennes. Ce résultat *a priori* non attendu s'explique par le fait que, dès le modèle à 4 gaussiennes, nous atteignons les limites de performance du système (*i.e.* la capacité du reconnaisseur est suffisante au vu des données d'apprentissage). De plus, l'intervalle de confiance est d'environ 0,9% pour ces 2 modèles. Nous pouvons donc considérer que cet écart n'est pas significatif.

[4]full : pour rappel, cela correspond au modèle composé de 128 gaussiennes par état et d'un vecteur acoustique de 39 coefficients.

[5]full : pour rappel, cela correspond au modèle composé de 128 gaussiennes par état et d'un vecteur acoustique de 39 coefficients.

# gauss/état	# paramètres	DER
2	5 832	4,96%
4	11 664	4,43%
128	373 248	4,52%
full[5]	1 092 096	0,96%

TAB. 6.4 – Evolution du DER et de la taille du modèle acoustique en fonction du nombre de gaussiennes par état émetteur. Le vecteur acoustique comprend 13 coefficients (12 PLP plus l'énergie). 2 300 tests effectués sur le corpus BDSON (chiffres / non bruité).

Il est à noter que l'influence des paramètres dynamiques semble plus importante pour les modèles acoustiques complexes. Le passage du CER de 5,48% à 5,80% ne représente qu'une augmentation relative de 6% alors que pour le modèle à 128 gaussiennes, le CER passe de 1,80% à 3,94%, soit une augmentation relative de 119%.

6.1.3 Conclusion sur la réduction "simple" du modèle acoustique

Pour conclure ce premier chapitre, nous avons obtenu des modèles acoustiques dont la taille permet une intégration dans un système embarqué. En effet, les modèles à 2 et 4 gaussiennes par état (comprenant respectivement 5 832 et 11 664 paramètres) associés à une paramétrisation réduite (12 coefficients PLP plus l'énergie) sont suffisamment compacts vis-à-vis des contraintes matérielles. Nous obtenons donc les *baselines* suivantes :

- VODIS / modèle très compact : 5 832 paramètres pour un CER de 5,80% ;
- VODIS / modèle compact : 11 664 paramètres pour un CER de 4,80% ;
- BDSON / modèle très compact : 5 832 paramètres pour un DER de 4,96% ;
- BDSON / modèle compact : 11 664 paramètres pour un DER de 4,43%.

Comme attendu, les modèles moins compacts sont plus robustes au bruit. En effet, comparativement, l'augmentation relative du CER est de 322% pour le modèle compact VODIS (corpus le plus bruité) alors que le DER augmente de 516% pour le même modèle avec le corpus BDSON. L'augmentation du nombre de paramètres permet donc de renforcer la robustesse au bruit des modèles acoustiques.

En terme de CER, une augmentation de 1,80% (modèle à 128 gaussiennes par état) à 4,80% pour le "gros" modèle VODIS peut être notée. Le CER monte même à 5,80% pour le modèle VODIS le plus compact. Cette augmentation, bien qu'importante, permet néanmoins une réduction drastique de la taille du modèle acoustique (par des facteurs de 90 et 180 respectivement). Cette étape est

nécessaire pour respecter les contraintes matérielles d'intégration d'un système embarqué.

6.2 Représentation acoustique basée sur un GMM général

L'idée principale présentée dans cette thèse est l'utilisation d'un seul GMM pour représenter l'ensemble de l'espace acoustique comme dans le cas des SCHMM. Contrairement aux approches classiques pour les systèmes basés sur des HMM (hormis le cas des HMM semi-continus), les GMM des états ne sont pas appris indépendamment les uns des autres avec des données propres aux états. L'ensemble de l'espace acoustique est modélisé par un unique GMM et les modèles d'état sont dérivés depuis ce GMM.

Dans notre approche, nous n'utilisons pas un simple codebook de gaussiennes comme cela est réalisé généralement pour les SCHMM mais un véritable GMM. Cette différence permet l'utilisation d'algorithmes classiques comme MMIE (*cf.* 6.2.2.2) et MLE (*cf.* 6.2.2.1).

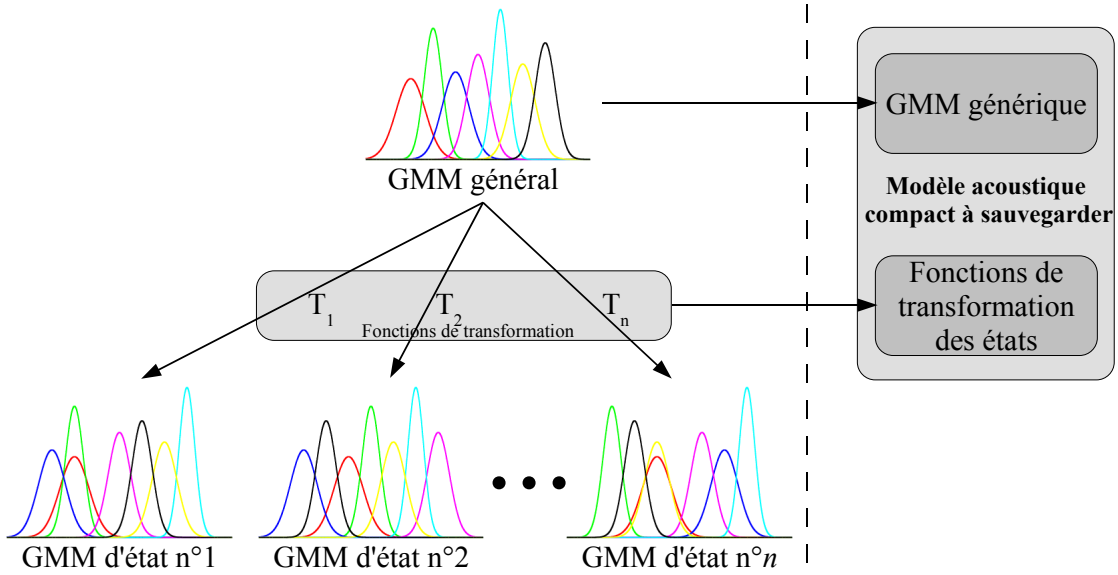

FIG. 6.1 – Principe général de l'approche proposée

Les états des HMM sont ensuite différenciés les uns des autres grâce à une fonction de transformation estimée avec peu de données. Nous proposons plusieurs approches pour cette fonction de transformation :

- la première est très proche des modèles semi-continus, la différence se situe au niveau de la ré-estimation des poids. Deux critères discriminants sont proposés, MMIE et FMMIE (méthode rapide pour la ré-estimation discriminante des poids) ;
- la seconde est basée sur une ré-estimation des paramètres (poids, moyenne et variance) du GMM général.

La figure 6.1 présente une vue générale de notre approche. Elle nécessite uniquement le stockage du GMM générique (appelé aussi GMM général) et de la fonction de transformation propre à chaque état.

Cette approche permet de réduire l'espace mémoire nécessaire au stockage des modèles acoustiques, mais elle permet aussi de s'affranchir d'un second problème : la quantité de données nécessaire pour l'apprentissage de modèles statistiques. En effet, les approches classiques ont besoin d'une quantité de données importante pour l'apprentissage des GMM d'états. Cependant, dans le cadre de la reconnaissance de parole embarquée dans les téléphones, très peu de données sont disponibles. Les corpus actuels sont généralement enregistrés dans les centraux téléphoniques ; le signal a alors été transformé par la phase de codage/décodage (pour diminuer la quantité de données transmises) et détérioré par la transmission à travers le canal (filaire ou non). L'enregistrement du signal au sein même du téléphone est plus compliqué, donc plus rare. Ce qui explique la rareté et la faible taille des corpus enregistrés directement dans un téléphone[6].

De plus, le regroupement de l'ensemble des gaussiennes permet de supprimer une part de redondance d'informations. Deux unités acoustiquement proches peuvent partager un grand nombre de gaussiennes. L'approche basée sur un GMM général permettra de ne stocker qu'une seule fois la composante alors que l'approche classique (sans tying) imposera de la stocker pour chaque état.

Notre méthode permet d'utiliser l'ensemble des données disponibles pour apprendre le GMM général, les états sont ensuite dérivés de ce GMM avec le peu de données propres aux états dont nous disposons.

Dans les sections suivantes, nous présenterons la construction du GMM général avant de détailler nos deux approches.

[6]nous n'en avons d'ailleurs pas trouvés. VODIS est celui qui se rapproche le plus de la réalité. Le microphone close-talk que nous avons utilisé n'a pas les mêmes contraintes qu'un micro intégré à l'intérieur du téléphone, notamment les perturbations mécaniques et électromagnétiques.

6.2.1 Construction du GMM général

Le GMM général utilisé dans notre approche se situe à mi-chemin entre un GMM classique (tel que ceux servant à la modélisation d'un état) et un codebook de gaussiennes. En effet, plutôt que d'apprendre un GMM directement depuis un ensemble de données, nous dérivons notre GMM général d'un HMM par regroupement des gaussiennes de chaque état du HMM, par fusion des gaussiennes (afin d'atteindre le nombre de gaussiennes désiré) et par une ré-estimation des paramètres. cette procédure est illustrée par le schéma 6.2.

FIG. 6.2 – Schéma de principe de la construction du GMM général

La première étape consiste à apprendre un HMM classique pour l'ensemble des modèles acoustiques du français. Ces HMM sont ensuite adaptés avec les données propres aux états avec le corpus d'adaptation[7] adéquat.

Dans un second temps, nous créons un codebook de gaussiennes en regroupant toutes les gaussiennes de tous les états précédemment appris (le poids de chaque composante est équiprobable). Afin d'obtenir le nombre de composantes désiré (pour garder un nombre constant de paramètres durant l'évaluation de nos approches), nous fusionnons deux à deux les gaussiennes les plus proches (en terme de perte de vraisemblance).

La distance entre deux gaussiennes ($\mathcal{N}_1(\mu_1, \Sigma_1, c_1)$ et $\mathcal{N}_2(\mu_2, \Sigma_2, c_2)$) utilisée est définie par :

$$D(\mathcal{N}_1, \mathcal{N}_2) = \frac{c_1}{c_1 + c_2} \log(\frac{\sqrt{\Sigma}}{\sqrt{\Sigma_1}}) + \frac{c_2}{c_1 + c_2} \log(\frac{\sqrt{\Sigma}}{\sqrt{\Sigma_2}}) \tag{6.3}$$

[7] sous-corpus BADAPT pour le corpus BDSON ou VADAPT pour VODIS

où Σ correspond à la variance de la gaussienne résultant de la fusion de $\mathcal{N}_1$ et $\mathcal{N}_2$ telle que définie par l'équation 6.6. Pour une description plus complète de cette distance le lecteur pourra, par exemple, se référer à [Bellot 2006].

La gaussienne $g'(c', \mu', \Sigma')$, résultante de la fusion des gaussiennes $g_i(c_i, \mu_i, \Sigma_i)$ et $g_j(c_j, \mu_j, \Sigma_j)$, est définie par :

$$c' = c_i + c_j \tag{6.4}$$

$$\mu' = \frac{c_i * \mu_i + c_j * \mu_j}{c_i + c_j} \tag{6.5}$$

$$\Sigma' = \frac{c_i}{c_i + c_j}\Sigma_i + \frac{c_j}{c_i + c_j}\Sigma_j + \frac{c_i * c_j}{(c_i + c_j)^2}(\mu_i - \mu_j)(\mu_i - \mu_j)^{tr} \tag{6.6}$$

La dernière étape consiste en une ré-estimation des paramètres, poids et moyenne, de chacune des composantes afin d'obtenir une véritable mixture de gaussiennes (GMM) et non seulement, un codebook. Cette ré-estimation est faite par une passe d'EM, maximisant la vraisemblance (MLE). L'utilisation de ce GMM permet d'appliquer des traitements statistiques classiques (utilisés par exemple pour l'adaptation, *cf.* chapitre 7).

6.2.2 Ré-estimation des poids

Les états sont discriminés entre eux uniquement par le poids de chacune des composantes. Cette méthode nécessite de stocker l'ensemble initial de gaussiennes et, pour chaque état émetteur ($etem$), un vecteur de poids dont la taille correspond au nombre de gaussiennes du GMM général. Pour estimer la taille du modèle, nous obtenons l'équation suivante :

$$\underbrace{nb\_gauss * 2 * nb\_param}_{\text{GMM général}} + \underbrace{nb\_etem * nb\_gauss}_{\text{poids des gaussiennes}} \tag{6.7}$$

où nb_gauss est le nombre de gaussiennes du GMM général, nb_param la taille du vecteur acoustique et nb_etem le nombre d'états émetteurs du HMM.

Les résultats présentés dans le tableau 6.5 montrent que l'ensemble des gaussiennes du GMM général n'est pas nécessaire pour représenter correctement un état. Nous proposons donc de ne garder que les N meilleures gaussiennes ([Park 2004], [Huggins-Daines 2006]) (au sens du maximum de vraisemblance) pour caractériser un état du HMM. Ceci nous permet alors, pour un nombre constant de paramètres, d'augmenter le nombre de gaussiennes du GMM général. Le nombre de paramètres du modèle est maintenant estimé par :

$$\underbrace{nb\_gauss * 2 * nb\_param}_{\text{GMM général}} + \underbrace{nb\_etem * nb\_gauss\_sel}_{\text{poids des gaussiennes}} \tag{6.8}$$

où nb_gauss_sel correspond au nombre de gaussiennes que nous gardons pour modéliser un état.

Les résultats du tableau 6.5 montrent que la coupure sur les modèles très compacts améliore très légèrement le CER ; cependant, cette amélioration n'est pas significative car inférieure à l'intervalle de confiance. Cette diminution peut s'expliquer comme un filtrage sur les composantes non caractéristiques de l'état, tout en renforçant le poids des gaussiennes les plus importantes.

	sans coupure	avec coupure
modèle très compact	6,37%	6,05%
modèle compact	5,07%	5,15%

TAB. 6.5 – CER obtenu avec l'ensemble des gaussiennes et avec la coupure dynamique. La modélisation acoustique repose sur un SCHMM. 11 136 tests effectués sur le corpus VODIS.

Cette sélection des gaussiennes a été faite de la manière suivante :
- les gaussiennes sont ordonnées en fonction de leur poids

$$c_i > c_{i+1} \quad \forall i \in 0, 1, ..., N-1 \tag{6.9}$$

- pour un seuil α fixé, seul le nombre minimum de gaussiennes permettant d'atteindre ce seuil est conservé

$$\sum_{i=0}^{N} c_i > \alpha \quad | \quad \sum_{i=0}^{N-1} c_i < \alpha \tag{6.10}$$

- variation de α afin d'obtenir une empreinte mémoire constante pour toutes les approches. α est fixé de manière à obtenir une moyenne de 20 gaussiennes par état pour les modèles très compacts et une moyenne de 30 gaussiennes par état pour les modèles compacts.

Les deux paragraphes suivants présentent des méthodes de ré-estimation des poids. La première consiste en une ré-estimation des poids basée sur le principe du maximum de vraisemblance (nous sommes donc proche du cas classique des HMM semi-continus). La seconde approche utilise un critère discriminant MMIE (Maximum Mutual Information Estimation). Nous présentons tout d'abord l'approche classique proposée par [Bahl 1986] pour l'estimation des paramètres discriminants puis une approximation rapide qui ne nécessite que très peu de calculs.

6.2.2.1 Ré-estimation suivant le maximum de vraisemblance : WRE-MLE

Cette approche consiste à estimer le vecteur de poids dépendant de l'état. Ceci permet alors de définir sa fonction de densité de probabilité par le modèle

GMM général et son vecteur de poids spécifique. Le système résultant est très proche des HMM semi-continus (SCHMM).

Le poids de la $i^{ème}$ gaussienne est ré-estimé suivant le critère du maximum de vraisemblance. La fonction de ré-estimation utilisée est donc définie par :

$$\tilde{c}_i = \frac{c_i * L(tr|G_i)}{\sum_{j=1}^{nb_g} c_j * L(tr|G_j)} \tag{6.11}$$

où $L(tr|g_x)$ correspond à la vraisemblance des données relatives à l'état (tr) pour la $x^{ième}$ gaussienne (G_x) et c_x à son poids initial.

6.2.2.2 Ré-estimation suivant un critère discriminant : WRE-MMIE et WRE-FMMIE

L'apprentissage des HMM en utilisant un critère discriminant maximisant l'information mutuelle (Maximum Mutual Information Estimation - MMIE) a déjà été largement étudié ces dernières années, notamment dans [Bahl 1986], [Valtchev 1996] ou [Woodland 2000]. L'objectif est de minimiser le risque d'erreur en maximisant l'écart de vraisemblance entre la bonne transcription et toutes les autres. Ce critère a souvent été utilisé dans le cadre des HMM continus mais plus rarement dans celui des HMM semi-continus.

La recherche des paramètres λ améliorant la capacité discriminante des modèles est réalisée par des algorithmes d'optimisation maximisant la fonction objective :

$$F_{mmie}(\lambda) = \sum_{r=1}^{R} log\left(\frac{P_\lambda(O_r|M_{w_r})P(w_r)}{\sum_{\tilde{w}} P_\lambda(O_r|M\tilde{w}) * P(\tilde{w})}\right) \tag{6.12}$$

où w_r est la transcription correcte, M_x la séquence de modèles correspondant à l'hypothèse x, $P(x)$ la probabilité linguistique et O_r une suite d'observations. Le dénominateur est constitué de la somme du produit des probabilités acoustique ($P_\lambda(O_r|M_x)$) et linguistique ($P(x)$) sur toutes les hypothèses possibles.

Une des difficultés majeures rencontrées pour l'estimation des paramètres optimaux réside dans la complexité de la fonction objective qui intègre, dans son dénominateur, l'ensemble des chemins incorrects susceptibles d'être empruntés (ainsi que les séquences de modèles associées). Pour atteindre une complexité acceptable, il faut généralement limiter le nombre de ces chemins, par exemple en ne gardant que les n meilleurs issus d'un treillis de mots ou de phonèmes (*cf.* [Valtchev 1997]). L'estimation des modèles par MMIE reste néanmoins bien

plus coûteuse que par MLE qui ne nécessite pas l'évaluation d'hypothèses incorrectes.

Dans le cas particulier des modèles semis-continus, seule la ré-estimation des poids est nécessaire. Par ailleurs, le partage massif des gaussiennes permet de réduire considérablement la complexité de l'estimation des paramètres. En effet, le calcul des vraisemblances est limité au nombre réduit des gaussiennes du dictionnaire. D'autre part, la présence des mêmes composantes dans tous les états permet une sélection directe des composantes discriminantes. C'est ce que nous mettons en évidence en développant la formule de ré-estimation des poids proposée dans [Povey 1999]. Les poids $\tilde{c}_{jm}$ maximisant la fonction objective peuvent être obtenus par maximisation de l'expression suivante :

$$\sum_{j,m} \left[\gamma_{jm}^{num} \log(\tilde{c}_{jm}) - \frac{\gamma_{jm}^{den}}{c_{jm}} \tilde{c_{jm}} \right] \tag{6.13}$$

où γ_{jm}^{num} et γ_{jm}^{dem} sont respectivement les taux d'occupation estimés sur les exemples corrects (num) et incorrects (den), c_{jm} le poids de la $m^{ième}$ composante du $j^{ième}$ état à l'itération précédente et $\tilde{c}_{jm}$ le nouveau poids.

En optimisant chaque terme de cette somme pour un ensemble de poids fixé, la convergence peut être obtenue après quelques itérations. Chacun de ces termes étant convexe, la formule de ré-estimation des poids se déduit directement de l'équation précédente :

$$\tilde{c_{jm}} = \frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} c_{jm} \tag{6.14}$$

où γ_{jm}^{k} est la probabilité d'être dans la $m^{ième}$ composante du $j^{ième}$ état estimée sur l'ensemble Ω_k qui regroupe les trames associées au $k^{ième}$ état. Ce taux d'occupation peut s'exprimer en fonction des vraisemblances $L()$:

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} \left[\frac{L(X|S_j)}{\sum_i L(X|S_i)} * \frac{c_{jm} L(X|G_{jm})}{L(X|S_j)} \right] \tag{6.15}$$

Ce qui donne :

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} \left[c_{jm} * \frac{L(X|G_{jm})}{\sum_i L(X|S_i)} \right] \tag{6.16}$$

En isolant, dans le dénominateur, la vraisemblance de la trame X sachant l'état S_k, on obtient :

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} \left[c_{jm} * \frac{L(X|G_{jm})}{L(X|S_k) + \sum_{i \neq k} L(X|S_i)} \right] \tag{6.17}$$

Dans des modèles semi-continus, les composantes gaussiennes G_{jm} sont indépendantes de l'état j. Dans ce cas, les taux d'occupation peuvent s'écrire :

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} \left[c_{jm} * \frac{L(X|G_{km})}{L(X|S_k) + \sum_{i \neq k} L(X|S_i)} \right] \tag{6.18}$$

En notant

$$\epsilon_k = \sum_{i \neq k} L(X|S_i) \tag{6.19}$$

Le rapport des taux d'occupation devient :

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{\sum_{X \in \Omega^j} \frac{L(X|G_{jm})}{L(X|S_j)+\epsilon_i}}{\sum_l \sum_{X \in \Omega^l} \frac{L(X|G_{lm})}{L(X|S_l)+\epsilon_l}} \tag{6.20}$$

En supposant ϵ_x négligeable devant les autres termes, nous obtenons l'approximation suivante :

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{c_{jm}}{\sum_l c_{lm}} \tag{6.21}$$

Enfin, en appliquant cette approximation dans l'équation 6.14, on obtient la formule de ré-estimation des poids suivante :

$$\tilde{c}_{jm} = \frac{{c_{jm}}^2}{\sum_l c_{lm}} \tag{6.22}$$

où $\tilde{c}_{jm}$ correspond au poids ré-estimé de la $m^{ième}$ gaussienne de l'état j, $\tilde{c}_{jm}$ au poids initial et $\sum_l c_{lm}$ à la somme sur tous les états l des poids de la $m^{ième}$ gaussienne.

Après ré-estimation, les vecteurs de poids de chaque état sont re-normalisés.

Cette fonction de mise à jour des poids à chaque itération ne nécessite pas de ré-estimation des vraisemblances sur le corpus d'apprentissage. En terme de temps de calcul, son coût est comparable à celui de l'apprentissage du modèle MLE initial, le poids MMIE se déduisant directement des poids MLE.

6.2.2.3 Comparaison des différents critères de ré-estimation des poids

Les expériences effectuées sur le corpus VODIS (corpus très bruité) sont résumées dans le tableau 6.6. Elles montrent que la ré-estimation des poids avec le critère du maximum de vraisemblance (MLE) et avec le critère discriminant classique (MMIE) semblent équivalentes entre elles et plus performantes

(*cf.* 6.6) que notre méthode de ré-estimation rapide (FMMIE). En effet, le CER est équivalent entre l'approche MLE et l'approche MMIE pour le modèle très compact (6,05% et 5,99%) et il est identique pour le modèle compact (5,15%).

L'approche FMMIE (ré-estimation rapide des poids suivant un critère discriminant), que nous proposons, semble bien moins performante pour le corpus VODIS. Le CER est nettement supérieur, environ 2% à 2,5% en variation absolue, pour les deux tailles de modèle acoustique.

	WRE		
	MLE	FMMIE	MMIE
modèle très compact	6,05%	8,54%	5,99%
modèle compact	5,15%	7,50%	5,15%

TAB. 6.6 – CER selon la fonction de transformation utilisée : MLE (utilisation du critère de maximum de vraisemblance), FMMIE (notre approche pour ré-estimation rapide de poids avec un critère discriminant) et MMIE (critère discriminant classique présenté par [Povey 1999]). 11 136 tests effectués sur le corpus VODIS (commandes vocales / bruité).

Cette approche ne semble pas apporter de gain, en terme de CER, comparée aux *baselines* obtenues précédemment (*cf.* tableau 6.3). Les CER peuvent être considérés comme équivalents en tenant compte de l'intervalle de confiance. Le modèle très compact a un CER de 6,05% pour l'approche MLE et 5,99% pour l'approche MMIE alors que le CER de référence est à 5,80% (l'intervalle de confiance est de 0,45%).

Des expériences ont aussi été effectuées sur le corpus BDSON (ce corpus est enregistré dans des conditions laboratoire et n'est donc pas bruité). Le tableau 6.7 présente les résultats obtenus dans ces conditions : les conclusions sont différentes.

	WRE		
	MLE	FMMIE	MMIE
modèle très compact	3,35%	2,78%	3,13%
modèle compact	2,83%	2,17%	2,48%

TAB. 6.7 – DER selon la fonction de transformation utilisée : MLE (utilisation du critère de maximum de vraisemblance), FMMIE (notre approche pour la ré-estimation rapide de poids avec un critère discriminant) et MMIE (critère discriminant classique présenté par [Povey 1999]). 2 300 tests effectués sur le corpus BDSON (chiffres / non bruité).

Dans le cas de données non bruitées, l'approche FMMIE est meilleure que l'approche MMIE classique. En effet, le DER du modèle compact est de 2,17% pour FMMIE et de 2,48% pour MMIE. De même, pour le modèle très compact le

DER est lui aussi inférieur pour la méthode FMMIE (2,78% contre 3,13%).

Nous pouvons également noter que les approches discriminantes (FMMIE et MMIE) sont plus performantes que l'approche MLE.

Pour terminer, il faut remarquer que dans ce cadre-là (expérience sur BD-SON) l'approche proposée sur la ré-estimation des poids s'avère nettement plus performante que les *baselines*. Le DER des *baselines* était de 4,96% pour le modèle très compact et de 4,43% pour le modèle compact alors que, dans ce cas, nous obtenons des DER respectifs à 2,79% (modèle très compact - FMMIE) et 2,17% (modèle compact - FMMIE).

En conclusion, nous pouvons retenir plusieurs éléments :
- FMMIE est rapide à calculer ;
- FMMIE semble peu robuste au bruit ;
- l'apport de l'approche basée sur le GMM général est mitigé en terme de performance pour des données bruitées ;
- diminution relative du DER entre 32% et 55% lors de l'utilisation de données propres.

Ces points sont détaillés dans les paragraphes suivants.

La méthode discriminante de ré-estimation des poids (FMMIE) que nous avons abordée présente un intérêt : son très faible coût de calcul. En effet, l'approche classique de ré-estimation, telle que présentée par [Povey 1999], nécessite le calcul de la vraisemblance de l'ensemble des chemins possibles (bons comme mauvais). Notre approche est basée uniquement sur la valeur initiale des poids. Le poids ré-estimé d'une gaussienne m pour l'état j correspond à ce poids élevé au carré et divisé par la somme des poids de la $m^{ième}$ composante sur tous les états.

L'approche FMMIE apporte un gain important dans le cadre de données propres, cependant elle semble peu résistante au bruit : sa comparaison avec MMIE sur un corpus bruité donne des résultats nettement inférieurs. Le CER de l'approche FMMIE est supérieur d'environ 40%, en relatif, pour des données bruitées (8,54% au lieu de 5,99% pour le modèle très compact) et ce, quelle que soit la taille du modèle utilisé.

L'utilisation d'un GMM général pour représenter l'ensemble de l'espace acoustique puis la dérivation des HMM d'état depuis ce GMM semble intéressante dans le cas de données propres ou du moins peu bruitées. Dans le cas où les données sont très bruitées, une étape de débruitage du signal s'avère impérative.

Enfin, cette approche avec le GMM général permet une importante réduction

du temps de calcul nécessaire pour l'estimation de la vraisemblance de chaque état. En effet, il suffit de pré-calculer la vraisemblance de chaque composante du GMM général puis d'effectuer une somme pondérée de ces vraisemblances :

$$L(X|S_j) = \sum_{i}^{nb_g} c_{i,j} * L(X|G_i) \tag{6.23}$$

où $L(X|G_i)$ est la vraisemblance des données X pour la gaussienne i (pré-calculée une seule fois pour tous les états) et $c_{i,j}$ son poids dans l'état j.

Le gain, pour l'estimation de la vraisemblance des états, est très important. Bien que cette thèse soit plutôt orientée sur l'occupation mémoire, il est important de noter que l'approche avec GMM général et ré-estimation des poids permet un gain important en temps de calcul.

Des expériences ([Linarès 2006]) ont aussi été menées dans un contexte de petits modèles pour la reconnaissance de parole continue (sur une heure issue du corpus ESTER). Ces modèles acoustiques, basés sur des HMM semi-continus, utilisent aussi l'approche WRE_{MLE} et WRE_{FMMIE} et donnent de bons résultats avec la ré-estimation discriminante rapide ($FMMIE$).

6.2.3 Transformation linéaire depuis le GMM général

Dans ce chapitre, nous proposons une étape intermédiaire avant l'application de WRE. Elle consiste à adapter globalement le GMM à l'aide d'une transformation linéaire simple. Pour les mêmes raisons que précédemment, nous ne gardons que les N meilleures gaussiennes. L'espace mémoire nécessaire au stockage de ce modèle est donc défini par :

$$\underbrace{nb\_gauss * (2 * nb\_param)}_{\text{GMM général}} + \underbrace{nb\_etem * (2 * nb\_param + nb\_gauss\_sel)}_{\text{transfo. linéaire et poids}} \tag{6.24}$$

où nb_param correspond à la taille du vecteur acoustique, nb_gauss au nombre de gaussiennes du GMM général, nb_etem au nombre d'états émetteurs et nb_gauss_sel au nombre de gaussiennes mémorisées.

La méthode LIAMAP présentée dans [Matrouf 2003] et [Lévy 2007] permet d'adapter globalement un GMM en utilisant uniquement une transformation globale, simple (maximisant la vraisemblance), appliquée sur toutes les composantes. La transformation étant commune à toutes les gaussiennes, l'ensemble des données est utilisé pour apprendre cette transformation ; ce qui nécessite une moins grande quantité de données spécifiques à l'état (et non plus à la gaussienne d'un état). Cette technique permet d'adapter le GMM général ($GMMgnl$)

aux données propres à chaque état et non plus aux données de chaque gaussienne de chaque état. La transformation porte sur les paramètres de moyenne et de variance. La forme générale de cette transformation est donnée ci-dessous :

$$\mu_{GMMetat} = \alpha * \mu_{GMMgnl} + \beta \tag{6.25}$$

$$\Sigma_{GMMetat} = \alpha^2 * \Sigma_{GMMgnl} \tag{6.26}$$

avec α (commun pour $\mu_{GMMetat}$ et $\Sigma_{GMMetat}$) et β définis ci-après.

L'idée principale de cette adaptation (*cf.* figure 6.3) est d'estimer une transformation entre les gaussiennes (μ, Σ) et $(\tilde{\mu}, \tilde{\Sigma})$. Elle sont obtenues de la manière suivante :

1. en fusionnant toutes les gaussiennes du GMM générique afin d'obtenir (μ, Σ) ;
2. en adaptant (avec MAP) le GMM générique avec les données spécifiques de l'état et en fusionnant ensuite toutes les gaussiennes de ce nouveau GMM afin d'obtenir $\widetilde{\mu}$ et $\widetilde{\Sigma}$.

Chaque gaussienne finale (définie par sa moyenne μ'_m et sa matrice de covariance Σ'_m) est calculée de la manière suivante :

$$\mu'_m = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}(\mu_m - \mu) + \widetilde{\mu} \tag{6.27}$$

$$\Sigma'_m = \widetilde{\Sigma}\Sigma^{-1}\Sigma_m \tag{6.28}$$

L'équation 6.27 peut être développée en :

$$\mu'_m = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu_m - \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu + \widetilde{\mu} \tag{6.29}$$

Posons

$$\alpha = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2} \tag{6.30}$$

et

$$\beta = -\widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu + \widetilde{\mu} \tag{6.31}$$

Les équations 6.27 et 6.28 deviennent :

$$\mu'_m = \alpha\mu_m + \beta \tag{6.32}$$

et

$$\Sigma'_m = \alpha^2\Sigma_m \tag{6.33}$$

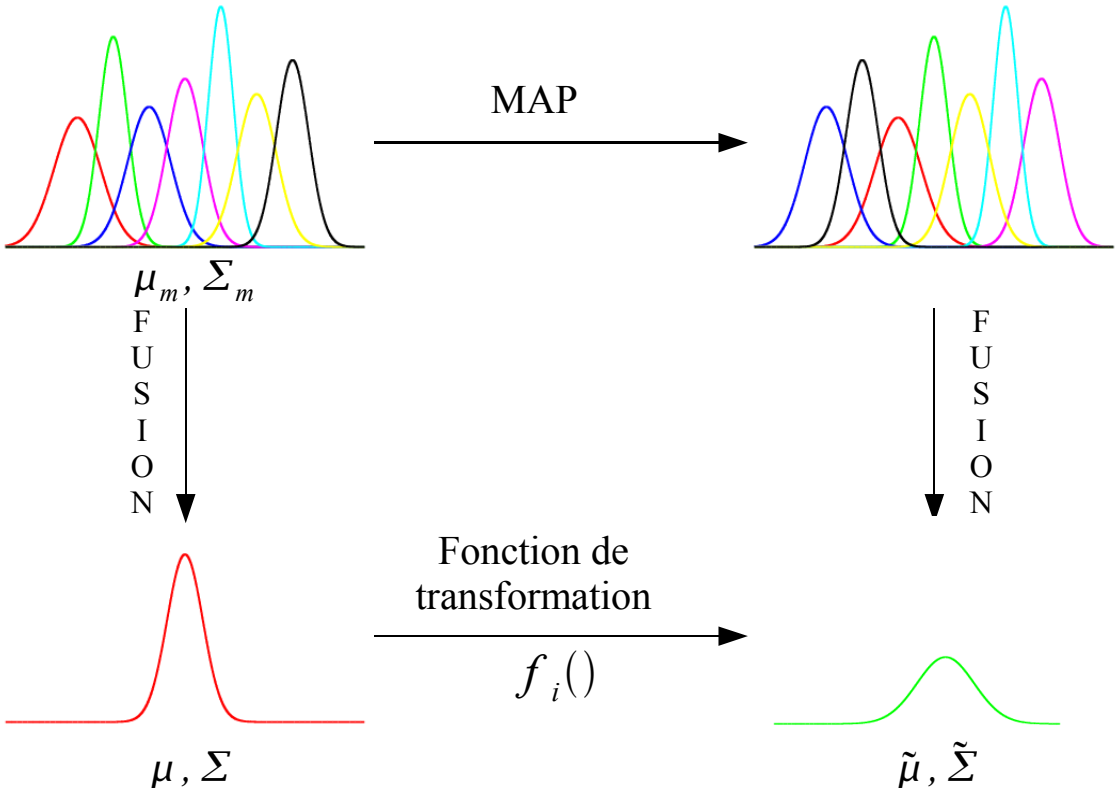

FIG. 6.3 – Principe général de la transformation LIAMAP

Les équations 6.32 et 6.33 correspondent à une simple transformation linéaire définie uniquement par les vecteurs α et β (la transformation est commune pour l'ensemble des gaussiennes du GMM).

ULT est ici présentée comme une étape préalable (optionnelle) à la réestimation des poids. L'étape WRE (*cf.* 6.2.2) est toujours employé - que l'on applique ou non l'ULT. La figure 6.4 illustre le processus complet.

L'ajout de cette étape nécessite un temps de calcul supplémentaire par rapport à l'approche WRE. En effet, durant le test, nous devons estimer dans un premier temps le GMM de chaque état avant de pouvoir calculer sa vraisemblance ; seuls le GMM général et les paramètres α et β de la transformation sont stockés. L'approche ULT+WRE impose aussi de calculer la vraisemblance de chaque gaussienne de chaque état contrairement à une approche WRE simple (sans ULT préalable) qui permet de déduire la vraisemblance de chaque état par une somme pondérée de la vraisemblance de chaque gaussienne (préalablement calculée).

Les expériences effectuées en milieu bruité (corpus vodis - *cf.* 6.8) montrent

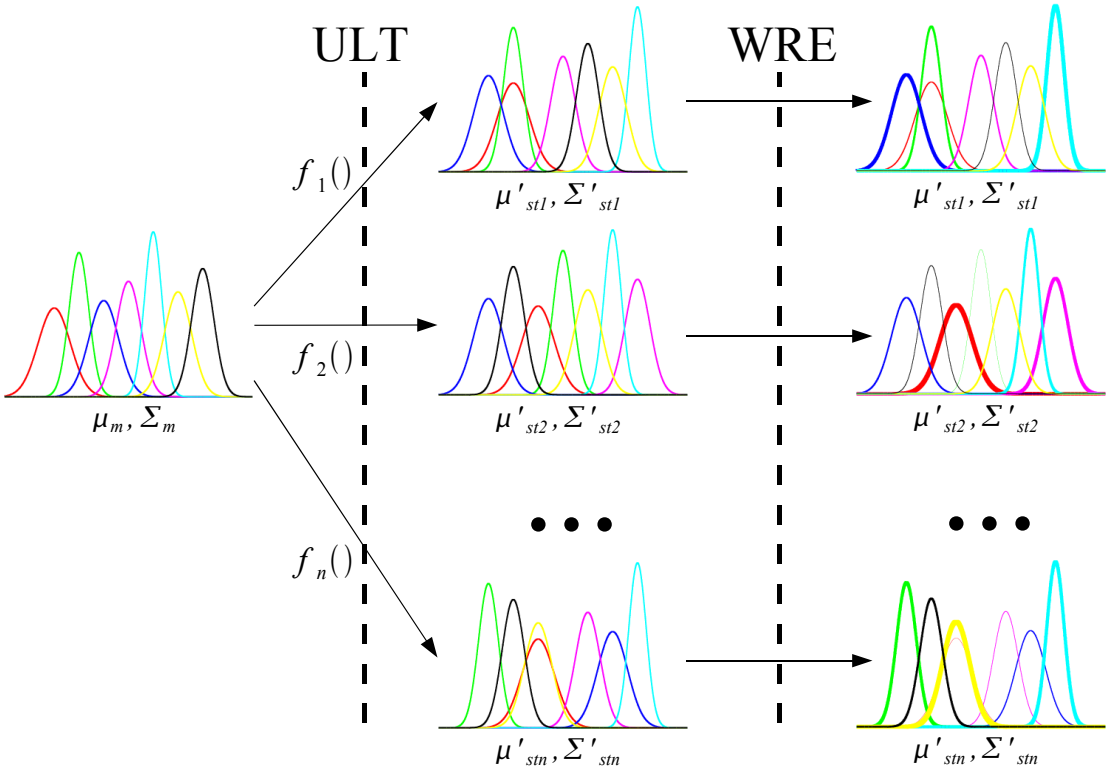

FIG. 6.4 – Transformation permettant de passer du GMM général, indépendant des états, au GMM dépendant des états

un gain intéressant en terme de CER. En effet, le CER est proche de 4% pour le modèle compact et proche de 5% pour le modèle très compact.

	ULT+WRE	
	MLE	MMIE
modèle très compact	5,25%	5,11%
modèle compact	4,01%	4,27%

TAB. 6.8 – CER selon l'application de l'ULT et de la fonction de transformation utilisée : MLE et MMIE. 11 136 tests effectués sur le corpus VODIS.

Il est intéressant de noter que l'étape ULT modifie le GMM général de telle manière que le choix de la méthode de ré-estimation des poids semble nettement moins déterminant que dans le cas de l'approche WRE. Le CER ne varie qu'entre 5,11% (MMIE) et 5,25% (MLE) pour le modèle très compact et entre 4,01% (MLE) et 4,27% (MMIE) pour le modèle supérieur. En terme de nombre d'erreurs de décodage de commande vocale, cela correspond à 16 (respectivement 29) commandes bien reconnues par un système et mal reconnues par l'autre sur

un total de 11 136 tests.

Les expériences effectuées avec BDSON sont présentées dans le tableau 6.9. Elles donnent, elles aussi, de bons résultats avec des taux d'erreur relativement faibles. Le DER est à 3,04% (avec la ré-estimation MLE) pour le modèle très compact et à 2,26% (ré-estimation MMIE) pour le modèle compact.

	ULT+WRE	
	MLE	MMIE
modèle très compact	3,04%	3,39%
modèle compact	2,78%	2,26%

TAB. 6.9 – DER selon l'application de l'ULT et de la fonction de transformation utilisée : MLE et MMIE. 2 300 tests effectués sur le corpus BDSON.

De même que pour les expériences bruitées, le critère de ré-estimation des poids (MLE ou MMIE) a peu d'influence sur les performances (même dans le cas du gros modèle où la différence peut apparaître comme significative, elle reste encore inférieure à l'intervalle de confiance).

6.3 Conclusion

L'approche globale (telle que présentée par le schéma 6.4) permet de générer des modèles acoustiques d'état à partir d'une représentation de l'ensemble de l'espace acoustique. Ce processus est composé de deux étapes (ULT et WRE) dont la première est optionnelle.

Le tableau 6.10 présente la synthèse de ces approches novatrices et les compare, en terme de reconnaissance de commande vocale, avec la *baseline* présentée dans 6.1 sur le corpus VODIS.

Les modèles acoustiques *baselines* ont des CER à 5,80% pour le modèle très compact et à 4,80% pour le modèle compact. Notre approche (WRE ou WRE+ULT) permet d'obtenir un CER minimum à 5,11% (ULT+WRE avec ré-estimation discriminante des poids - MMIE) pour le modèle très compact : ceci représente une réduction relative de plus de 12%. La contrainte de taille supérieure permet, quant à elle, d'obtenir un CER de 4,01% (ULT+WRE et ré-estimation des poids par MLE) contre une *baseline* à 4,80%. Cette réduction absolue de 0,8% représente une baisse relative de plus de 16%.

Nous pouvons voir que l'approche WRE seule ne permet pas d'amélioration significative du CER. Le meilleur taux d'erreur est de 5,99% pour la ré-

	WRE			ULT+WRE		*baseline*
	MLE	FMMIE	MMIE	MLE	MMIE	
modèle très compact	6,05%	8,54%	5,99%	5,25%	5,11%	5,80%
modèle compact	5,15%	7,50%	5,15%	4,01%	4,27%	4,80%

TAB. 6.10 – Tableau récapitulatif des résultats des approches WRE, ULT+WRE et de la *baseline*. CER en fonction de l'application de l'ULT ou non et de la fonction de transformation utilisée : MLE, FMMIE et MMIE. 11 136 tests effectués sur le corpus VODIS.

estimation avec MMIE alors que la *baseline* est à 5,80% pour le modèle très compact.

Notre approche complète (ULT+WRE) permet, elle, une réduction relative de CER entre 12% et 16% sans augmentation de l'espace nécessaire au stockage du modèle acoustique[8].

Pour des données non bruitées (tableau 6.11), l'approche WRE semble donner des résultats légèrement meilleurs que l'approche WRE+ULT. Le DER minimal, pour le modèle très compact, est obtenu avec l'approche WRE et ré-estimation par FMMIE ; le DER est de 2,78%. Ceci représente une baisse relative de plus de 40% (le DER de la *baseline* étant de 4,96%). De même, le modèle compact passe d'un DER à 4,32% à 2,17% soit une diminution relative de près de 50%.

	WRE			ULT+WRE		*baseline*
	MLE	FMMIE	MMIE	MLE	MMIE	
modèle très compact	3,35%	2,78%	3,13%	3,04%	3,39%	4,96%
modèle compact	2,83%	2,17%	2,48%	2,78%	2,26%	4,32%

TAB. 6.11 – Tableau récapitulatif des résultats des approches WRE, ULT+WRE et de la *baseline*. DER en fonction de l'application de l'ULT ou non et de la fonction de transformation utilisée : MLE, FMMIE et MMIE. 2 300 tests effectués sur le corpus BDSON.

Pour conclure, l'approche que nous proposons apporte une amélioration très importante dans le cadre de données propres (BDSON), que l'ULT soit utilisée ou non et quelle que soit la méthode de ré-estimation des poids. La *baseline* pour le modèle très compact était de 4,96% et les cinq variantes de notre approche ont un DER entre 2,78% et 3,39%. De même pour le modèle compact, la *baseline* est à 4,32% et le DER varie entre 2,17% et 2,83%. Soit une diminution relative du DER entre 30% et 50%.

Lorsque les données sont très bruitées (VODIS), l'approche WRE seule ne

[8]En effet, pour l'approche ULT+WRE, le nombre de composantes du GMM est réduit afin d'obtenir un nombre de paramètres total équivalent à l'approche WRE simple.

suffit pas. La *baseline* montre un CER inférieur aux résultats obtenus par WRE. Les meilleures configurations obtiennent un CER supérieur d'au minimum 0,2% (pour le modèle très compact avec ré-estimation MMIE). L'ULT permet, pour sa part, des gains de CER entre 10% et 15%. L'ULT semble donc être nettement plus résistante aux bruits.

De même, l'utilisation de FMMIE (notre approche de ré-estimation, discriminante, rapide des poids) semble assez peu résistante au bruit. Avec le corpus VODIS, la ré-estimation des poids avec les critères MLE ou MMIE est toujours plus performante qu'avec FMMIE. Cependant, elle reste très performante avec les données peu bruitées.

Chapitre 7

Adaptation du GMM général

Sommaire

Résumé

Dans ce chapitre nous cherchons à montrer qu'un modèle acoustique comporte deux types d'informations : celles relatives à l'unité acoustique (un phonème par exemple) et celles concernant les locuteurs et/ou l'environnement. Pour cela, nous allons essayer de tirer parti de l'architecture proposée dans le chapitre précédent en adaptant le GMM général (afin de le spécialiser à un locuteur ou à un environnement donné) tout en conservant les transformations apprises initialement. Après une rapide présentation des méthodes d'adaptation classiques en RAP, nous présentons l'adaptation du GMM général défini dans le chapitre précédent.

A l'origine, les Systèmes de Reconnaissance Automatique de Parole (SRAP) étaient dépendants du locuteur ([Davis 1952]). Les limites de ces systèmes ont très vite été perçues, ce qui a suscité un certain nombre de travaux de recherche pour les rendre, dans un premier temps, multi-locuteurs. Les avancées scientifiques et les progrès techniques ont enfin permis de faire des systèmes indépendants du locuteur[1]. Actuellement, la plupart des SRAP sont initialement indépendants du locuteur puis procèdent à une adaptation au locuteur et à l'environnement, généralement "en ligne". Dans [Bellot 2006], le chapitre 3.3 présente une série d'expériences montrant les effets de l'adaptation du modèle acoustique sur l'évolution du WER. Pour l'évaluation AUPELF ARC B1, le système de référence indépendant du locuteur obtient un WER de l'ordre de 20%. Un premier modèle adapté au genre du locuteur atteint 19% et, enfin, le WER des modèles adaptés au locuteur de test est de 17,4% ce qui représente un gain relatif de l'ordre de 14%.

Pour les systèmes avec approche statistique, les modèles acoustiques sont généralement appris indépendamment du locuteur/utilisateur final, principalement pour deux raisons :

- l'identité du locuteur n'est généralement pas connue lors de la phase d'apprentissage des modèles acoustiques ;
- les modèles acoustiques sont appris avec une très grande quantité de données (plusieurs dizaines, voire centaines d'heures de parole) et il paraît impensable de demander à la même personne de parler aussi longtemps pour apprendre un modèle.

L'apprentissage de modèles acoustiques dépendants du locuteur est donc difficilement envisageable. C'est pourquoi des méthodes d'adaptation de ces modèles indépendants ont été étudiées. Les deux familles principales sont décrites dans la section suivante.

Après cette présentation succincte des méthodes classiques d'adaptation, nous étudions l'intégration d'une phase d'adaptation dans le cadre de notre approche de modélisation compacte de l'espace acoustique. Nous souhaitons adapter le GMM général avec le peu de données propres à un locuteur dont nous disposons (quelques phrases) sans avoir à ré-apprendre les transformations (WRE ou ULT+WRE). Cela permettrait, d'une part, d'adapter l'ensemble des états directement via le GMM général et autoriserait, d'autre part, l'adaptation des états pour lesquels nous ne disposons pas de données.

[1] L'utilisation des méthodes statistiques et de corpus de grandes tailles y sont pour beaucoup.

7.1 Méthodes d'adaptation classiques

Les techniques d'adaptation des modèles acoustiques ont toutes pour objectif de rapprocher un modèle acoustique appris dans certaines conditions d'un jeu de données proches des conditions de test. Ce jeu de données est généralement relativement restreint, ce qui explique que le modèle n'ait pas été appris directement avec ces données. Il est la plupart du temps caractéristique d'un nouvel environnement acoustique ou d'un nouveau locuteur.

L'adaptation peut être supervisée ou non supervisée. Par méthode supervisée, nous entendons les méthodes utilisant une transcription exacte des données d'adaptation, contrairement aux méthodes dites non supervisées qui ne pré-supposent aucune connaissance *a priori* sur le contenu linguistique des données d'adaptation.

Pour l'adaptation supervisée, il est nécessaire de distinguer deux cas de figure selon que l'utilisateur final fasse partie du corpus d'apprentissage ou non. S'il est inclus dans ce corpus, le système utilise alors les données propres à ce locuteur pour adapter les modèles acoustiques. Dans le cas contraire (qui correspond au cas le plus courant dans le cadre des systèmes commercialisés), le système demande à l'utilisateur de prononcer un certain nombre de phrases. Le SRAP réalise ensuite un alignement temporel afin d'affecter chaque trame de signal à un état donné, voire à une composante précise de cet état. Une fois cette étape terminée, l'adaptation des modèles peut être réalisée.

Classiquement, l'adaptation non supervisée se déroule de la manière suivante : l'utilisateur prononce des phrases quelconques, le SRAP réalise un décodage (avec alignement temporel) de ces phrases afin d'associer chaque trame à un état (voir à une composante du GMM d'un état). Le système possède des données étiquetées qu'il peut utiliser pour faire de l'adaptation. Si le décodage initial contient trop d'erreurs, l'adaptation effectuée sera de moindre qualité. Les erreurs qui interviennent durant le décodage expliquent aisément pourquoi les méthodes supervisées sont généralement supérieures aux méthodes non supervisées.

Les techniques d'adaptation classiques peuvent se répartir en deux familles principales :

- l'adaptation par régression linéaire : MLLR[2] ;
- l'adaptation par *Maximum A Posteriori* : MAP.

Ces deux méthodes sont présentées, succinctement, dans les paragraphes suivants.

[2] Maximum Likelihood Linear Regression

7.1.1 MAP : *Maximum A Posteriori*

L'adaptation MAP présentée par [Gauvain 1994] ou [Reynolds 2000] permet de modifier le modèle initial grâce à un jeu de données spécifiques, généralement plus proches des données de test que celles qui ont permis l'apprentissage des modèles acoustiques.

Cette adaptation peut être perçue comme étant à mi-chemin entre l'adaptation et l'apprentissage. En effet, l'adaptation MAP peut être vue comme une seconde phase d'apprentissage contrainte par des connaissances *a priori* (la valeur initiale des paramètres et/ou leurs distributions). L'objectif de cette adaptation est de rapprocher les paramètres du modèle initial vers des données spécifiques (données d'adaptation). Comme illustré par la figure 7.1 [Bellot 2006], l'idée est de déplacer une partie des paramètres par des transformations spécifiques appliquées indépendamment à chaque paramètre.

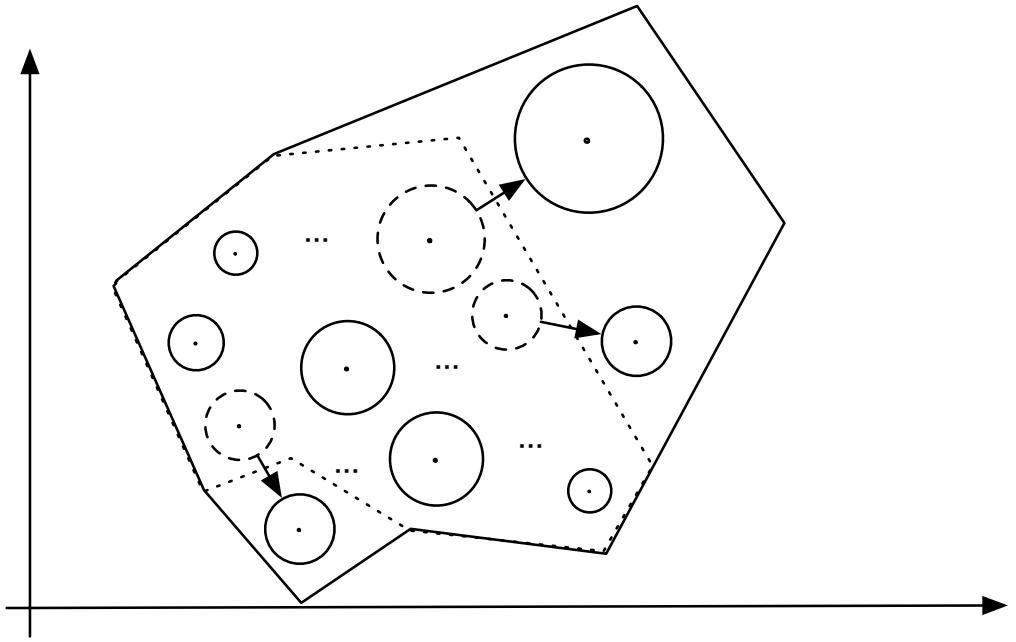

FIG. 7.1 – Méthode d'adaptation MAP. Cette méthode d'adaptation modifie chaque paramètre (moyenne et écart type) du GMM, composante par composante.

L'inconvénient majeur de cette méthode d'adaptation réside dans la nécessité de disposer de suffisamment de données d'adaptation car les paramètres de chaque composante du GMM sont considérés comme indépendants les uns des autres. Ceci implique aussi de disposer de données pour toutes les composantes des GMM que l'on souhaite adapter. MAP peut donc être considérée comme une méthode d'adaptation locale, du fait de l'adaptation indépendante de chaque

composante du GMM.

Le coût de calcul requis par MAP peut également être considéré comme un inconvénient non négligeable dans le cadre de la reconnaissance/adaptation embarquée dans un système mobile. Une phase d'apprentissage avec les nouvelles données s'avère nécessaire avant de pouvoir effectuer l'adaptation des paramètres initiaux, et ceci pour chaque modèle acoustique.

MAP est appliquable sur l'ensemble des paramètres du HMM : poids, moyenne et variance de chaque gaussienne. La définition de [Gauvain 1994] est couramment utilisée en reconnaissance de la parole et celle de [Reynolds 2000] en reconnaissance du locuteur ; la différence entre les deux est faible. Le contexte adaptatif utilisé dans la section suivante étant plus proche de la reconnaissance du locuteur[3], nous utiliserons la définition de [Reynolds 2000]. La ré-estimation des paramètres est définie par :

$$\tilde{w}_i = [\alpha_i^w n_i/T + (1-\alpha_i^w)w_i]\,\gamma \tag{7.1}$$

$$\tilde{\mu}_i = \alpha_i^\mu E_i(x) + (1-\alpha_i^\mu)\mu_i \tag{7.2}$$

$$\tilde{\sigma}_i^2 = \alpha_i^\sigma E_i(x^2) + (1-\alpha_i^\sigma)(\sigma_i^2+\mu_i^2) - \tilde{\mu}_i^2 \tag{7.3}$$

où γ est un facteur permettant que la somme des poids des composantes soit égale à 1, n_i représente l'occupation de la gaussienne i, T correspond au nombre total de trames, $E_i(x)$ est l'estimation de la moyenne des données d'adaptation et $E_i(x^2)$ l'estimation de la variance avec les mêmes données. α_i^ρ est définie par :

$$\alpha_i^\rho = \frac{n_i}{n_i + r^\rho} \tag{7.4}$$

avec r^ρ correspondant à un facteur de régulation[4].

L'apport de l'adaptation MAP est incontestable pour les systèmes de reconnaissance grand vocabulaire classiques. [Gauvain 1994] présente des améliorations très significatives du WER[5] dans le cadre de l'apprentissage de modèles dépendants du locuteur (comparé à une approche MLE). Cette adaptation nécessite cependant une quantité de données plus importante que l'approche MLLR décrite dans le paragraphe suivant.

[3] La modélisation de l'ensemble de l'espace par un GMM puis sa spécialisation vers un contexte donné ressemblent fortement à l'apprentissage d'un modèle du monde que l'on adapte pour obtenir un modèle de locuteur.

[4] Le facteur de régulation correspond au nombre de données nécessaires pour accorder la même confiance au modèle initial et au modèle estimé avec les données d'adaptation.

[5] Word Error Rate - Taux d'erreur mot

7.1.2 MLLR : Maximum Likelihood Linear Regression

Cette méthode d'adaptation, présentée initialement par [Leggetter 1994], propose, dans sa version originale, d'apprendre une transformation globale et linéaire pour tous les modèles acoustiques via une régression linéaire utilisant le critère de maximum de vraisemblance (comme illustré par la figure 7.2, [Bellot 2006]).

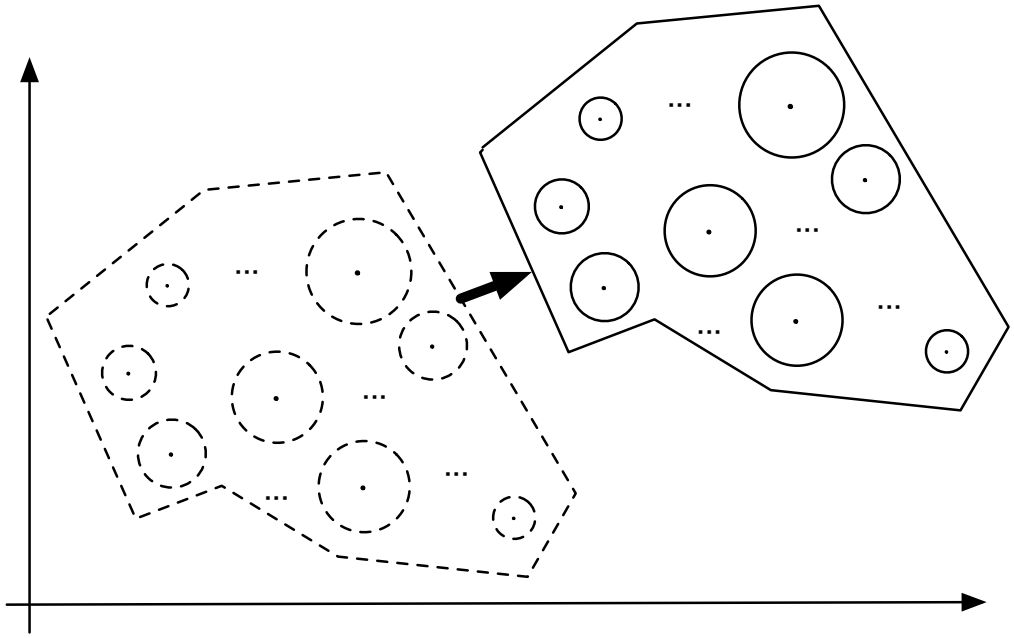

FIG. 7.2 – Méthode d'adaptation MLLR. Cette méthode adapte de la même manière l'ensemble des paramètres de chaque composante de tous les GMM.

Cette approche nécessite généralement moins de données d'adaptation que MAP (et ses dérivées) car l'ensemble des données est utilisé pour la régression. Avec la MLLR, une trame sert à l'adaptation de l'ensemble des paramètres alors que pour MAP seules les données associées à une composante servent à son adaptation.

Dans sa version originale, la MLLR s'applique uniquement sur les moyennes des gaussiennes et est définie de la manière suivante :

$$\tilde{\mu}_j = W_j \, \nu_j \tag{7.5}$$

avec

$$\nu_j = [1, \mu_{j1}, \mu_{j2}, ..., \mu_{jn}]' \tag{7.6}$$

W_j est appris sur les données d'adaptation.

La densité de probabilité initialement définie par :

$$b_j, m(o) = \mathcal{N}(o, \mu_{j,m}, \Sigma_{j,m}) \quad (7.7)$$

$$= \frac{1}{(2\pi)^{n/2}\Sigma^{1/2}} e^{-\frac{1}{2}(o-\mu_{j,m})'\Sigma^{-1}(o-\mu_{j,m})} \quad (7.8)$$

devient :

$$b_j, m(o) = \frac{1}{(2\pi)^{n/2}\Sigma^{1/2}} e^{-\frac{1}{2}(o-W_j\,\nu_j)'\Sigma^{-1}(o-W_j\,\nu_j)} \quad (7.9)$$

Le HMM est alors défini non plus par $\lambda = (\pi, A, B)$ mais par $\lambda = (\pi, A, B, W)$. W est estimé suivant le critère de maximum de vraisemblance :

$$\tilde{W} = \underset{W}{\text{ArgMax}}\, P(o/\lambda) \quad (7.10)$$

Afin d'améliorer les performances de cette adaptation, plusieurs classes de régression peuvent être utilisées (regroupement des fricatives par exemple) ce qui permet d'affiner la transformation. Gales propose une autre amélioration dans [Gales 1996] en adaptant aussi la variance et non plus seulement la moyenne.

7.1.3 Généralités sur les méthodes d'adaptation

Il est reconnu que les adaptations MAP et MLLR permettent, toutes deux, des gains importants en terme de taux de reconnaissance. Cependant, ces méthodes présentent des inconvénients majeurs dans le cadre applicatif présenté dans ce document. En effet, ces techniques nécessitent une quantité de données d'adaptation non négligeable. Cette quantité varie selon la méthode utilisée (MAP ou MLLR). MAP nécessite, généralement, plus de données car toutes les gaussiennes de tous les états sont adaptées indépendamment les unes des autres. La MLLR, pour sa part, apprend une transformation unique pour l'ensemble des modèles, ce qui lui permet d'utiliser toutes les données disponibles ; cependant cette quantité est tout de même variable en fonction du nombre de classes de régression utilisées.

L'adaptation de chacun des états a un coût non négligeable du point de vue du calcul. Dans notre cadre, la reconnaissance embarquée dans les systèmes mobiles, cet aspect est très important et constitue une contrainte majeure.

7.2 Adaptation du GMM général

Dans le chapitre 6, nous avons proposé une nouvelle approche pour la modélisation des unités acoustiques : l'ensemble de l'espace est représenté par

un seul GMM et les unités sont uniquement différenciées par une fonction de transformation simple. Nous avons montré que cette approche permet, en respectant les contraintes mémoires, d'apporter une amélioration en terme de taux de reconnaissance et qu'elle peut aussi permettre un gain important en terme de ressource de calcul.

Cette approche présente les états de manière relative (par la fonction de transformation) à une référence commune (le GMM général). Nous proposons de vérifier si le déplacement de la référence, vers un locuteur par exemple, permet toujours de différencier les états entres eux par une simple transformation. Ceci revient à vérifier si les transformations appliquées au GMM, pour obtenir les modèles des différents états, sont invariantes à une modification du dit GMM.

Pour cela, nous proposons d'adapter le GMM général avec les données propres à un locuteur ; tout comme les modèles de locuteurs peuvent être issus d'un modèle du monde en reconnaissance du locuteur. Le GMM de chaque état serait donc généré par l'application de sa transformation sur ce GMM adapté[6]. Ce principe est illustré par la figure 7.3.

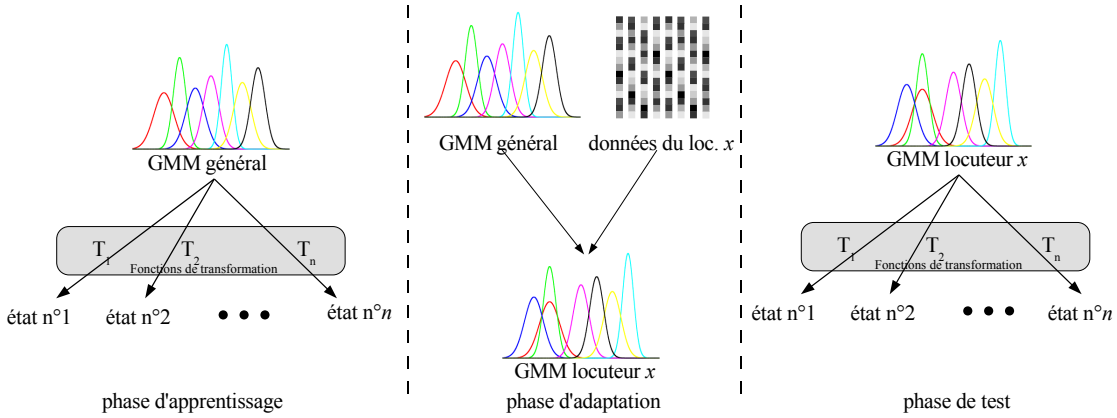

FIG. 7.3 – Principe de l'adaptation du GMM général : phase d'apprentissage des transformations avec un GMM général indépendant du locuteur, phase d'adaptation du GMM général avec des données propres au locuteur, phase d'utilisation avec remplacement du GMM général par le GMM du locuteur.

Cette adaptation globale du GMM, qui représente l'ensemble des modèles acoustiques, s'appuie sur le postulat suivant : si un décalage est constaté entre une unité acoustique indépendante du locuteur et une autre dépendante du locuteur, alors ce même décalage existe probablement entre toutes les unités acoustiques. C'est ce que nous souhaitons mettre en évidence en adaptant le GMM général sans modifier les fonctions de transformation.

[6] et non plus sur le GMM général

Le corpus VODIS contient un sous-ensemble de phrases phonétiquement équilibrées. Chaque locuteur prononce cinq phrases. Ces phrases ont été utilisées pour adapter le GMM général afin d'obtenir des GMM généraux propres à chaque locuteur (appelés *GMM locuteur x* dans la figure 7.3). Ces phrases sont totalement différentes des commandes à décoder. Cette adaptation MAP a été réalisée suivant la définition de Reynolds sur les moyennes des gaussiennes (une adaptation des variances aurait nécessité plus de données que celles dont nous disposions[7]). Le facteur de régulation (ρ) a été fixé à 14 (valeur classique en reconnaissance automatique du locuteur). Dans cette série d'expériences, nous n'avons pas fait d'adaptation en ligne du GMM général ; seule une adaptation du GMM appliquée avant les tests et avec des phrases phonétiquement équilibrées a été réalisée.

Le graphique 7.4 présente l'évolution du CER du modèle très compact en fonction du nombre de phrases utilisées pour l'adaptation du GMM général.

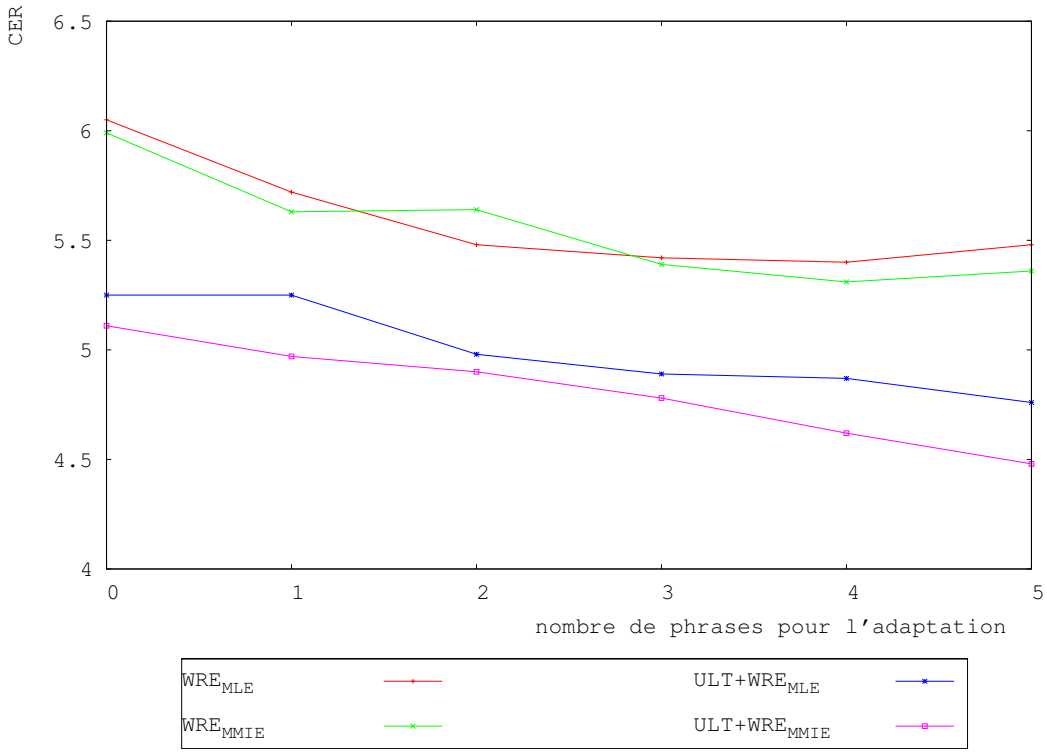

FIG. 7.4 – Evolution du CER en fonction du nombre de phrases phonétiquement équilibrées utilisées pour l'adaptation du GMM général (modèle très compact). L'indice 0 correspond au modèle initial sans adaptation. 11 136 tests effectués sur le corpus VODIS.

[7]En outre, en reconnaissance du locuteur, les systèmes "état de l'art" ne réalisent qu'une adaptation de la moyenne.

Une diminution régulière du CER peut être constatée pour l'ensemble des fonctions de transformation présentées dans ce manuel ; le graphique 7.4 illustre cela. L'approche WRE_{FMMIE} n'est pas présentée sur ce graphique car elle n'est pas sensible à l'adaptation du GMM général ; son CER reste voisin des 8,5% (*cf.* tableau récapitulatif de l'adaptation : tableau 7.1). Ceci peut s'expliquer par le caractère très sélectif de notre méthode discriminante de ré-estimation des poids. L'approximation du ϵ qui est faite avec notre méthode (*cf.* l'équation 6.19 page 83) n'est plus valable. En effet, l'adaptation MAP effectuée sur le GMM général a pour but de rapprocher ce GMM des données de test ; ce rapprochement tend à rendre cet ϵ difficile à négliger dans l'équation 6.19.

Le graphique 7.5 présente, pour sa part, les différentes valeurs de CER pour le modèle compact.

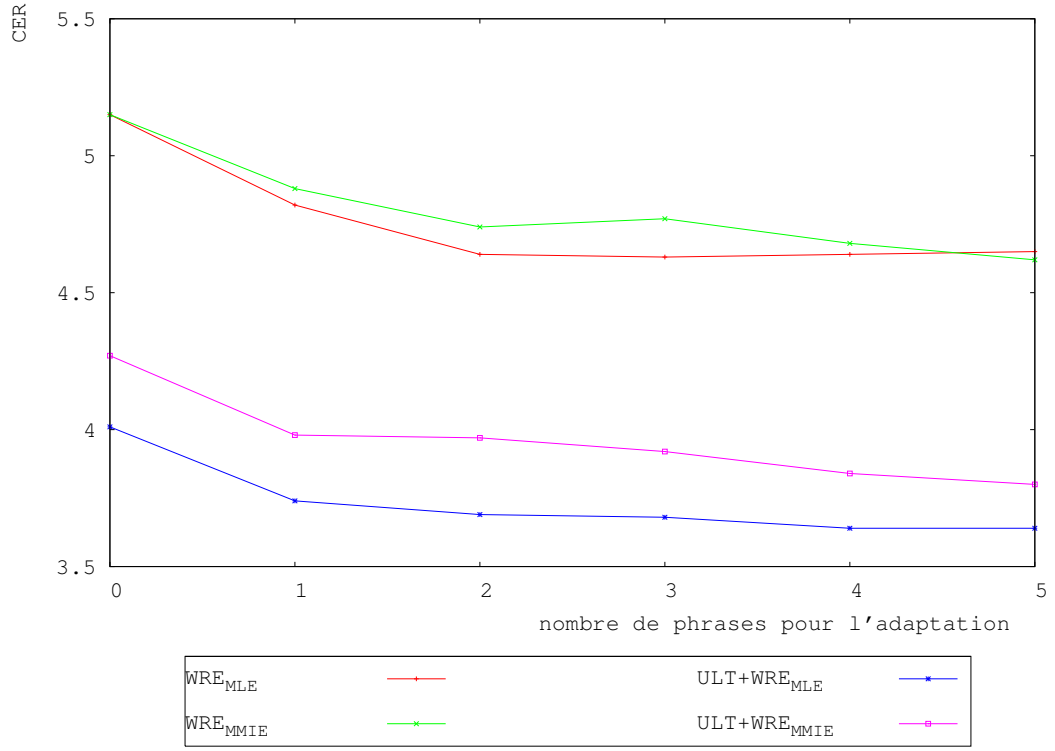

FIG. 7.5 – Evolution du CER en fonction du nombre de phrases phonétiquement équilibrées utilisées pour l'adaptation du GMM général (modèle compact). L'indice 0 correspond au modèle initial sans adaptation. 11 136 tests effectués sur le corpus VODIS.

L'évolution du CER est comparable pour les modèles compacts et les modèles très compacts. Le phénomène de non-adaptabilité du GMM avec la ré-adaptation des poids par FMMIE apparaît également (ce qui explique aussi son absence sur ce graphique).

Nous pouvons constater que, comme attendu, plus le nombre de phrases d'adaptation disponibles est important, plus l'adaptation est performante.

Les tableaux 7.1 présentent les résultats obtenus pour l'adaptation du GMM général avec l'approche WRE (7.0(a)) et avec l'approche ULT+WRE (7.0(b)) en adaptant le GMM général avec les cinq phrases phonétiquement équilibrées dont nous disposons. Une amélioration significative du CER peut être notée et ce quelle que soit l'approche utilisée.

(a) approche WRE

	sans adaptation			avec adaptation		
	MLE	FMMIE	MMIE	MLE	FMMIE	MMIE
modèle très compact (6 k.)	6,05%	8,54%	5,99%	5,48%	8,67%	5,36%
modèle compact (11 k.)	5,15%	7,50%	5,15%	4,67%	7,28%	4,63%

(b) approche ULT+WRE

	sans adaptation		avec adaptation	
	MLE	MMIE	MLE	MMIE
modèle très compact (6 k.)	5,25%	5,11%	4,76%	4,48%
modèle compact (11 k.)	4,01%	4,27%	3,64%	3,80%

TAB. 7.1 – CER obtenu pour l'approche WRE (7.0(a)) et l'approche ULT+WRE (7.0(b)) sans adaptation au locuteur et avec adaptation (en utilisant 5 phrases phonétiquement équilibrées). 11 136 tests effectués sur le corpus VODIS.

En effet, l'approche WRE seule (tableau 7.0(a)) permet un gain relatif jusqu'à 10%. Le CER du modèle très compact, avec ré-estimation des poids par MMIE, passe de 5,99% à 5,36%, ce qui représente un gain relatif de 10,52%. Le modèle compact obtient des gains similaires (gain relatif de 10,1% pour le modèle compact avec ré-estimation des poids avec MMIE).

Les modèles dont les poids sont ré-estimés avec la méthode FMMIE ne supportent pas l'adaptation car cette approche néglige la vraisemblance des gaussiennes ne servant théoriquement pas à modéliser un état (le fameux ϵ). En effet, l'adaptation rapproche les gaussiennes du GMM général vers les données d'adaptation ; nous en arrivons donc à négliger des gaussiennes dont la vraisemblance, après adaptation, n'est plus négligeable.

Au regard du tableau 7.0(b), le gain relatif (en terme de CER) est légèrement meilleur lors de l'utilisation de l'approche ULT+WRE. L'amélioration du CER se situe entre 9% et 12% (en relatif). Le modèle compact avec ré-estimation MLE obtient le gain le moins élevé : 9,22%, le CER passant de 4,01% à 3,64%. Le modèle très compact avec la même ré-estimation des poids (MMIE) obtient, pour sa part, une réduction relative du CER de 12,33%.

7.3 Conclusion sur l'adaptation du GMM général

L'approche que nous proposons dans le chapitre 6 repose sur la modélisation de l'ensemble de l'espace acoustique avec un seul GMM global. Les différentes unités acoustiques sont ensuite représentées de manière relative par une simple transformation de ce GMM.

Dans ce chapitre, nous avons émis l'hypothèse que cette représentation relative des états était orthogonale à la transformation du GMM général. En utilisant une approche comparable à celle utilisée en reconnaissance du locuteur, nous avons adapté le GMM général qui représente l'ensemble de l'espace acoustique et l'ensemble des locuteurs (d'un point de vue statistique) avec des données propres à un locuteur, en émettant l'hypothèse que les transformations permettraient toujours de distinguer les différentes unités acoustiques.

Les résultats présentés dans les tableaux 7.0(a) et 7.0(b) valident l'hypothèse émise sur l'orthogonalité des transformations dépendantes du locuteur et des transformations dépendantes du phonème. En effet, hormis pour l'approche WRE_{MMIE}, nous constatons une diminution relative du CER entre 9% et 12%. Le meilleur taux d'erreur étant obtenu avec la méthode $ULT + WRE_{MLE}$.

Un intérêt majeur de notre approche est de ne pas nécessiter de phase d'alignement étant donné qu'un seul GMM est adapté : on ne parle donc plus d'adaptation supervisée ou non supervisée.

Les gains présentés semblent cohérents avec l'état de l'art de l'adaptation ([Bellot 2006] observe des gains de l'ordre de 14%). Cependant, d'autres approches, notamment pour l'adaptation en ligne du GMM général (adaptation à partir des données du test en cours), sont envisageable.

Troisième partie

Conclusions et perspectives

Conclusions et perspectives

Conclusion

Alors qu'une course effrénée à la miniaturisation des téléphones se poursuit depuis leur apparition, ceux-ci intègrent, dans le même temps, de plus en plus de services. Ces deux phénomènes complexifient l'usage des téléphones, rendant les interfaces utilisateur peu conviviales, notamment à cause d'un recours fréquent à un clavier peu maniable. L'ajout d'une modalité vocale à ces interfaces offre une alternative intéressante en terme d'ergonomie.

La généralisation attendue du mode "mains libres" (au volant d'un véhicule par exemple) renforce l'intérêt d'un mode vocal et plus particulièrement de la reconnaissance de la parole. Une commande vocale de l'interface des téléphones permettrait de décrocher la ligne, de numéroter, d'accéder à son agenda ou à sa messagerie sans manipuler le téléphone.

Si l'intérêt d'une reconnaissance vocale intégrée à un téléphone semble peu discutable, le faible niveau de ressources (mémoire et puissance de calcul) de tels systèmes mobiles ainsi que la variabilité de l'environnement d'utilisation représentent un verrou technologique important. En effet, les systèmes classiques de reconnaissance de la parole nécessitent une quantité de ressources importante, correspondant à une station de travail de dernière génération, alors qu'un téléphone portable ne dispose par exemple que de quelques centaines de Ko de mémoire.

Les travaux présentés dans ce document abordent ce problème de la reconnaissance automatique de la parole en situation de ressources (très) limitées. Nous nous sommes spécifiquement intéressés à la problématique de la modélisation acoustique, en cherchant à réduire l'empreinte mémoire des modèles acoustiques, tout en favorisant les capacités d'adaptation de ces modèles.

Pour évaluer notre approche, nous nous sommes confrontés à deux tâches :

la reconnaissance de chiffres isolés (DER) et la reconnaissance de commandes vocales (CER) qui couvrent bien le panel des applications visées.

Après avoir rapidement présenté les principales méthodes permettant d'obtenir des modèles acoustiques compacts, nous avons détaillé une nouvelle approche pour modéliser les unités acoustiques en conservant le formalisme markovien. Dans cette approche, l'ensemble de l'espace acoustique est représenté par un unique GMM (appelé GMM général) et les modèles acoustiques associés à chaque état du HMM sont obtenus de manière relative, par une transformation de ce GMM. Cette solution, proche des HMM semi-continus, offre potentiellement trois avantages majeurs : une réduction importante du nombre de paramètres des modèles, une réduction de la quantité de calcul et une grande capacité d'adaptation.

Dans le but de constituer un comparatif entre une approche HMM classique et notre solution, nous avons cherché à diminuer le nombre de méta-paramètres des modèles HMM classiques afin de respecter les contraintes mémoire typiques d'un téléphone portable. Pour cela, nous avons réduit la taille du vecteur acoustique (sans utiliser pas les paramètres dynamiques - nous avons donc seulement 13 coefficients) et diminué le nombre de composantes du GMM caractérisant chaque état (2 ou 4 gaussiennes selon la taille globale souhaitée). Nous avons alors obtenu deux références (*baselines*) nommées "modèle compact" et "modèle très compact". Ces deux modèles offrent un taux de réduction du nombre de paramètres proche de 90 pour le premier et de 180 pour le second. Dans le même temps, le taux d'erreur de reconnaissance de chiffres isolés (DER) obtenu à l'aide du modèle très compact est passé de 0,96% (modèle complet avec plus d'un million de paramètres) à 4,96% (4,43% pour le modèle compact) et le taux d'erreur de reconnaissance de commandes (CER) du modèle très compact a augmenté de 1,80% (modèle complet) à 5,80% (4,80% pour le modèle compact).

L'approche que nous avons présentée, dans ce document, comporte deux étapes.

La première étape, appelée WRE, consiste à ré-estimer pour chaque état le vecteur de poids du GMM général. Dans un premier temps, nous avons utilisé un critère classique qui maximise la vraisemblance du GMM pour les données relatives à l'état : le critère MLE (Maximum Likelihood Estimation). Ensuite, nous avons proposé l'utilisation d'un critère discriminant (MMIE - Maximum Mutual Information Estimation). Une quantité de calculs très importante est nécessaire pour le critère MMIE, ce qui limite généralement son utilisation. Afin de s'affranchir de cette contrainte, nous avons proposé une approximation du critère MMIE dans le cadre des HMM semi-continus que nous appelons FMMIE. Cette approximation permet de ré-estimer les poids avec un critère discriminant tout

en conservant un coût de calcul similaire à celui du critère MLE. Cette première étape apporte des résultats mitigés en fonction de la qualité des données. Avec des données bruitées, les taux d'erreur sont similaires entre les *baselines* et l'approche WRE (hormis pour l'approche WRE_{FMMIE}). Cependant, si les données sont propres, le gain relatif apporté par WRE se situe entre 32% et 55%, ce qui permet d'obtenir un DER à 2,17% alors que le DER du système de référence est à 4,43%. En outre, WRE permet une réduction importante du coût de calcul, en pré-calculant une seule fois la vraisemblance de chaque gaussienne du GMM général, puis en estimant la vraisemblance des états à l'aide d'une simple somme pondérée des vraisemblances pré-calculées.

Dans une seconde étape, afin d'améliorer les performances de notre approche lors d'une utilisation en milieu bruité, nous avons présenté la méthode ULT qui réalise une adaptation du GMM général avant la ré-estimation des poids. Cette approche complète est appelée $ULT + WRE$. Cette adaptation du GMM général repose sur une transformation linéaire globale appliquée à l'ensemble des composantes de ce GMM. L'intérêt de cette approche est qu'elle apprend, pour un état donné, une transformation commune à toutes les composantes à partir de l'ensemble des données d'apprentissage relatives à cet état. Cette approche nécessite une quantité réduite de données. De plus, l'occupation mémoire d'une telle transformation reste très réduite. Avec des données propres, le gain apporté par $ULT + WRE$ est similaire à celui obtenu par WRE seule (entre 31% et 54% de diminution relative du taux d'erreur). Cependant, avec des données bruitées, cette approche permet une diminution relative du taux d'erreur oscillant entre 9% et 16% (toujours comparativement aux références) alors que l'approche WRE obtenait seulement des performances comparables aux *baselines* (voire légèrement inférieures). $ULT + WRE$ est donc plus résistante aux bruits que ne l'est WRE. Néanmoins, elle a un inconvénient non négligeable : elle nécessite une quantité de calculs supérieure à l'approche WRE. En effet, l'estimation de la vraisemblance d'un état nécessite de calculer, dans un premier temps, le GMM associé à cet état (en appliquant la transformation linéaire sur le GMM général) avant de pouvoir estimer la vraisemblance de l'état.

La plupart des SRAP "état-de-l'art" utilisent des méthodes d'adaptation pour rapprocher les modèles acoustiques initiaux d'un nouvel environnement et/ou d'un nouveau locuteur. Dans la littérature, l'adaptation au locuteur des modèles acoustiques permet un gain relatif compris entre 5% et 15%.

L'architecture que nous avons proposée permet d'envisager une adaptation de tous les modèles d'état en adaptant uniquement le GMM général. Nous formulons l'hypothèse qu'une adaptation du GMM général (celui qui sert de base aux transformations définissant les modèles des états) rapprocherait celui-ci de l'environnement de test tout en conservant inchangées les transformations.

Nous avons expérimenté cette idée dans le cadre d'une adaptation au locuteur (réalisée par MAP). Les résultats expérimentaux valident cette hypothèse. Nous observons une réduction relative du taux d'erreur compris entre 9% et 12%, ceci correspond globalement aux taux obtenus classiquement dans la littérature pour la phase d'adaptation au locuteur. De plus, cette approche de l'adaptation, contrairement aux approches classiques en RAP, ne nécessite pas la passe de décodage intermédiaire, généralement nécessaire pour l'adaptation non supervisée, étant donné que nous n'avons qu'un seul modèle à adapter (toutes les données d'adaptation peuvent être liées à ce modèle). Enfin, le dernier avantage de notre approche est qu'elle permet d'adapter l'ensemble des modèles d'état uniquement par modification du GMM général ; l'ensemble des modèles est alors adapté, même ceux pour lesquels aucune donnée spécifique n'est disponible.

En conclusion, nous avons proposé une approche permettant d'obtenir des modèles très compacts (moins de 5k et 11k paramètres). Ces modèles autorisent un gain important en terme de taux d'erreur comparativement à une approche classique de réduction du nombre de méta-paramètres d'un HMM. Cette architecture offre par ailleurs une solution efficace et peu coûteuse pour adapter l'ensemble du modèle acoustique aux conditions de test. Enfin, même si notre objectif principal consistait à réduire la taille mémoire des modèles acoustiques, notre approche permet également une forte réduction du coût de calcul lorsque la technique WRE est employée.

Perspectives

Plusieurs axes de perspectives peuvent être proposés :
- adaptation classique du GMM général ;
- adaptation en ligne du GMM général ;
- recherche d'autres fonctions de transformation ;
- réduction du temps de calcul pour le décodage.

Dans le chapitre 7, nous avons montré qu'une adaptation du GMM général suffisait à rapprocher l'ensemble des modèles acoustiques des conditions de test. Une solution très simple (MAP), issue de la reconnaissance du locuteur a été employée dans ce cadre. D'autres techniques, comme un apprentissage du modèle général par SAT (Speaker Adaptative Training) ou une adaptation de celui-ci par SMAPGM (Structural adaptation using MAP and Gaussians Merging), peuvent s'avérer plus performantes. De même, notre approche facilite grandement l'adaptation en ligne des modèles, toujours par le fait qu'un seul GMM doit être déplacé. Ajouter, en temps réel, la prise en compte de l'environnement de test devrait apporter un gain additionnel en terme de performance

(comparativement à une simple adaptation au locuteur).

La technique $ULT + WRE$ présentée dans 6.2.3 ajoute une transformation, linéaire et globale, du GMM général préalablement à la ré-estimation du vecteur de poids. Cette transformation relativement simple - et peut-être peu appropriée pour des GMM avec un grand nombre de composantes - a été employée avec ML comme critère d'optimisation. Nous souhaitons étudier d'autres transformations pour cette étape ULT, en utilisant des critères d'optimisation discriminants et/ou des transformations structurelles ou multi-classes. Une représentation structurelle du GMM initial offrirait potentiellement de nombreux avantages que ce soit pour la modélisation de l'espace acoustique ou pour les transformations/adaptations du GMM général.

Toujours dans le cadre de $ULT + WRE$, une autre voie d'amélioration consisterait à éviter le surcoût en terme de complexité de calcul ajouté par ULT, comparativement à WRE. En effet, WRE permet un calcul simplifié des vraisemblances : seules les vraisemblances des composantes du GMM général sont vraiment calculées, ce qui n'est plus le cas avec ULT qui, en transformant le GMM différemment pour chaque état du modèle, impose de recalculer la vraisemblance des GMM pour chacun des états. Une perspective intéressante consisterait à appliquer une transformation de type ULT directement sur la vraisemblance issue des composantes du GMM général plutôt que sur le modèle lui-même, ce qui associerait les gains de WRE (en terme de complexité de calcul) à ceux d'ULT (en terme de performance).

Pour finir, un exemple concret d'implémentation de l'architecture que nous avons présentée est en cours de développement au LIA dans le cadre du projet RNRT (Réseau National de Recherche en Télécommunication) BIOBIMO (BIOmétrie BImodale sur MObile). L'objectif de ce projet est l'identification biométrique des personnes sur des systèmes mobiles. Au niveau vocal, l'identification biométrique comprend notamment de la reconnaissance du locuteur et de la vérification de mot de passe, ce que permet aisément l'architecture présentée dans ce document.

Quatrième partie

Annexes

Annexe A

Bibliographie personnelle

Chapitre de livre

Lévy C., Linarès G., Nocera P. et Bonastre J.-F., Embedded mobile phone digit-recognition, *chapitre 7 in Advances for In-Vehicle and Mobile Systems*, H. Abut, J.H.L. Hansen and K. Takeda (Eds.), Springer Science, 2007.

Conférences internationales

Lévy C., Linarès G. et Bonastre J.-F., Gmm-based acoustic modeling for embedded speech recognition, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP'2006)*, Pittsburgh, Pennsylvania, USA, 2006.

Lévy C., Linarès G., Nocera P. et Bonastre J.-F., Reducing computational and memory cost for cellular phone embedded speech recognition system, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2004)*, pages 309-312, Montreal, Canada, 2004.

Conférences nationales

Lévy C., Linarès G. et Bonastre J.-F., Représentation acoustique compacte pour un système de reconnaissance de la parole embarquée, dans *XXVIèmes Journées d'Etudes sur la Parole JEP'2006*, pages 131-134, Dinard, France, 2006.

Linarès G., Lévy C. et Plagniol J.-C., Estimation rapide de modèles semi-continus discriminants, dans *XXVIèmes Journées d'Etudes sur la Parole JEP'2006*, pages 131-134, Dinard, France, 2006.

Lévy C., Linarès G., Nocera P. et Bonastre J.-F., Reconnaissance de chiffres isolés embarquée dans un téléphone portable, dans *XXVèmes Journées d'Etudes sur la Parole JEP'2004*, Fez, Maroc, 2004.

Auran C., Bouzon C., Hirst D., Lévy C. et Nocera P., Algorithme de prédiction d'élisions de phonèmes et influence sur l'alignement automatique dans le cadre du projet Aix-MARSEC, dans *XXVèmes Journées d'Etudes sur la Parole JEP'2004*, Fez, Maroc, 2004.

Workshops internationaux

Lévy C., Linarès G. et Bonastre J.-F., Mobile phone embedded digit-recognition, dans *Workshop on DSP in Mobile and Vehicular Systems*, Sesimbra, Portugal, 2005.

Lévy C., Linarès G. et Nocera P., Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems, dans *Workshop on DSP in Mobile and Vehicular Systems*, Nagoya, Japan, 2003.

Annexe B

Lexiques

VODIS

- abandonner : aa bb an dd oo nn ei
- aéroport : aa ei rr oo Opp Bpp oo rr
- affichez la carte : aa ff ii ch ei ll aa Okk Bkk aa rr Ott Btt
- aigus : ei gg uu
- ailleurs : aa yy oe rr
- autres destinations : au Ott Btt rr dd ai ss Ott Btt ii nn aa ss yy on
- bis : bb ii ss
- carnet d'adresses : Okk Bkk aa rr nn ai dd aa dd rr ai ss
- carrefour : Okk Bkk aa rr eu ff ou rr
- cassette : Okk Bkk aa ss ai Ott Btt
- CD : ss ei dd ei
- centre ville : ss an Ott Btt rr eu vv ii ll
- c'est bon : ss ei bb on
- changer : ch an jj ei
- choisir destination : ch ww aa zz ii rr dd ai ss Ott Btt ii nn aa ss yy on
- commencer guidage : Okk Bkk au mm an ss ei gg ii dd aa jj
- correction : Okk Bkk au rr ai Okk Bkk ss yy on
- couper le son : Okk Bkk ou Opp Bpp ei ll eu ss on
- culture : Okk Bkk uu ll Ott Btt uu rr
- curiosité : Okk Bkk uu rr yy au zz ii Ott Btt ei
- décrocher : dd ei Okk Bkk rr oo ch ei
- destination : dd ai ss Ott Btt ii nn aa ss yy on
- distance : dd ii ss Ott Btt an ss
- distraction : dd ii ss Ott Btt rr aa Okk Bkk ss yy on
- échangeur : ei ch an jj eu rr
- enlever la carte : an ll eu vv ei ll aa Okk Bkk aa rr Ott Btt

- faux : ff au
- frontière : ff rr on Ott Btt yy ai rr
- garage : gg aa rr aa jj
- gare : gg aa rr
- graves : gg rr aa vv
- grossir : gg rr au ss ii rr
- guidage silencieux : gg ii dd aa jj ss ii ll an ss yy eu
- guidage vocal : gg ii dd aa jj vv au Okk Bkk aa ll
- hôtel : au Ott Btt ai ll
- info : in ff au
- information : in ff oo rr mm aa ss yy on
- itinéraire : ii Ott Btt ii nn ei rr ai rr
- j'épelle : jj ei Opp Bpp ai ll
- loisirs : ll ww aa zz ii rr
- mémoriser : mm ei mm au rr ii zz ei
- menu principal : mm eu nn uu Opp Bpp rr in ss ii Opp Bpp aa ll
- mode automatique : mm oo dd au Ott Btt au mm aa Ott Btt ii Okk Bkk
- mode manuel : mm oo dd mm aa nn uu ai ll
- montrer la destination : mm on Ott Btt rr ei ll aa dd ai ss Ott Btt ii nn aa ss yy on
- navigation : nn aa vv ii gg aa ss yy on
- non : nn on
- numéro : nn uu mm ei rr au
- OK : au Okk Bkk ei
- oui : ou ii
- page précédente : Opp Bpp aa jj Opp Bpp rr ei ss ei dd an Ott Btt
- page suivante : Opp Bpp aa jj ss uy ii vv an Ott Btt
- position actuelle : Opp Bpp au zz ii ss yy on aa Okk Bkk Ott Btt uu ai ll
- raccrocher : rr aa Okk Bkk rr au ch ei
- radio : rr aa dd yy au
- rappel automatique : rr aa Opp Bpp ai ll au Ott Btt au mm aa Ott Btt ii Okk Bkk
- réduire : rr ei dd uy ii rr
- réessayer : rr ei ei ss ai yy ei
- remettre le son : rr eu mm ai Ott Btt rr ll eu ss on
- répertoire : rr ei Opp Bpp ai rr Ott Btt ww aa rr
- répéter : rr ei Opp Bpp ei Ott Btt ei
- répondre : rr ei Opp Bpp on dd rr
- restaurant : rr ai ss Ott Btt au rr an
- rue : rr uu
- sortie d'autoroute : ss oo rr Ott Btt ii dd au Ott Btt au rr ou Ott Btt
- station service : ss Ott Btt aa ss yy on ss ai rr vv ii ss

- station service autoroutière : ss Ott Btt aa ss yy on ss ai rr vv ii ss au Ott Btt au rr ou Ott Btt ii ai rr
- téléphone : Ott Btt ei ll ei ff oo nn
- ville : vv ii ll

BDSON

- zéro : zz ei rr au
- un : un
- deux : dd eu
- trois : Ott Btt rr ww aa
- quatre : Okk Bkk aa Ott Btt rr
- cinq : ss in Okk Bkk
- six : ss ii ss
- sept : ss ai Ott Btt
- huit : uy ii Ott Btt
- neuf : nn oe ff

Annexe C

Nombre de paramètres

• Pour les **HMM classiques** dont le nombre de paramètres est définie par l'équation 6.2, nous obtenons donc :

– modèles très compacts : 2 gaussiennes par état, 108 états émetteurs et 13 coefficients par vecteur acoustique soit

$$2 * 108 * (2 * 13 + 1) \ = \ 5\,832 \ \text{paramètres}$$

– modèles compacts : 4 gaussiennes par état, 108 états émetteurs et 13 coefficients par vecteur acoustique soit

$$4 * 108 * (2 * 13 + 1) \ = \ 11\,664 \ \text{paramètres}$$

• Pour **l'approche WRE** l'estimation du nombre de paramètres est définie par l'équation 6.8, soit :

– modèles très compacts : 141 gaussiennes pour le GMM général, 20 gaussiennes sélectionnées par état, 108 états émetteurs et 13 coefficients par vecteur acoustique soit

$$141 * 2 * 13 + 108 * 20 \ = \ 5\,826 \ \text{paramètres}$$

– modèles compacts : 324 gaussiennes pour le GMM général, 30 gaussiennes sélectionnées par état, 108 états émetteurs et 13 coefficients par vecteur acoustique soit

$$324 * 2 * 13 + 108 * 30 \ = \ 11\,664 \ \text{paramètres}$$

• Pour **l'approche ULT+WRE** l'estimation du nombre de paramètres est définie par l'équation 6.24, soit :

– modèles très compacts : 33 gaussiennes pour le GMM général, 20 gaussiennes sélectionnées par état, 108 états émetteurs et 13 coefficients par vecteur acoustique soit

$$33 * 2 * 13 + 108 * (2 * 13 + 20) \ = \ 5\ 826 \ \text{paramètres}$$

– modèles compacts : 216 gaussiennes pour le GMM général, 30 gaussiennes sélectionnées par état, 108 états émetteurs et 13 coefficients par vecteur acoustique soit

$$216 * 2 * 13 + 108 * (2 * 13 + 30) \ = \ 11\ 664 \ \text{paramètres}$$

Liste des figures

Liste des tableaux

Liste des acronymes

AFCP	Association Francophone de la Communication Parlée
ACP	Analyse en Composante Principale (PCA en anglais)
ADL	Analyse Discriminante Linéaire (LDA en anglais)
AMR	Adaptative Multi Rate
ARCEP	Autorité de Régulation des Communications Électroniques et des Postes
CDHMM	Continuous Density Hidden Markov Model - Modèle de Markov Caché avec densité de probabilité continue
CPU	Central Processing Unit - Unité de calcul centrale, généralement le processeur principal
DGA	Délégation Générale pour l'Armement
DSP	Digital Signal Processor - processeur de signal numérique
DHMM	Discrete Hidden Markov Model - Modèle de Markov Caché discret
DTW	Dynamic Time Warping - alignement temporel dynamique
EDGE	Enhanced Data rate for GSM Evolution - norme pour la téléphonie de 2.5 génération
EFR	Enhanced Full Rate
ELRA	European Language Resources Association
ETSI	European Telecommunications Standards Institute - Institut européen des normes de télécommunication
FFT	Fast Fourier Transform - Transformation de Fourier rapide
GPS	Global Positioning System - système mondial de positionnement
GSM	Global System for Mobile communications - nom du standard de la technologie de téléphone cellulaire déployée tout d'abord en Europe
FR	Full Rate
HMM	Hidden Markov Model - Modèle de Markov Caché
HR	Half Rate
LAR	Logarithm Area Ratios

LFCC	Linear Frequency Cepstral Coefficients
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
MAP	*Maximum A Posteriori*
MCU	Micro Control Unit - micro-contrôleur
MFCC	Mel Frequency Cepstral Coefficients
MIPS	Million d'Instructions Par Seconde
MLE	Maximum Likelihood Estimation
MLLR	Maximum Likelihood Linear Regression
MMC	Modèle de Markov Caché
MMIE	Maximum Mutual Information Estimation
MTM	Multimedia Terminal Mobile - terminal mobil multimédia
PDA	Personal Digital Assistant - assistant personnel
PLP	Perceptual Linear Predictive
RAM	Random Access Memory - mémoire vive
RAP	Reconnaissance Automatique de la Parole
ROM	Read Only Memory - mémoire morte
SCHMM	Semi-Coninuous Hidden Markov Model - Modèle de Markov Caché semi-continu
SCDHMM	Subspace Distribution Clustering Hidden Markov Model
SRAP	Système de Reconnaissance Automatique de la Parole
UMTS	Universal Mobile Telecommunications System - Système universel de télécommunications mobiles
WER	Word Error Rate - taux d'erreur en mots
WLAN	Wireless Local Area Network - réseau local sans fil

Bibliographie

[Adda 1997] Adda G., Adda-Decker M., Gauvin J.-L. et Lamel L.-F., Le système de dictée du LIMSI pour l'évaluation AUPELF'97, *Journées Scientifiques et Techniques*, pages 31–34, 1997.

[Astrov 2003a] Astrov S. et Andrassy B., Large vocabulary speaker independent isolated word recognition for embedded systems, dans *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, pages 1137–1140, Geneva, Switzerland, 2003.

[Astrov 2003b] Astrov S., Bauer J.-G. et Stan S., High performance speaker and vocabulary independent ASR technology for mobile phones, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2003)*, pages 281–284, Hong Kong, 2003.

[Bahl 1986] Bahl L.-R., Brown P. F., de Souza P.-V. et Mercer R.-L., Maximum Mutual Information Estimation of Hidden Markov Model parameters for speech recognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, pages 49–52, Tokyo, Japan, 1986.

[Barras 1996] Barras C., *Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés.*, Thèse de doctorat, Université de Paris 6, 1996.

[Bechet 1994] Bechet F., *Système de traitement de connaissances phonétiques et lexicales : application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu*, Thèse de doctorat, université d'Avignon, LIA, 1994.

[Bellegarda 1990] Bellegarda J.-R. et Nahamoo D., Tied mixture continuous parameter modeling for speech recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(12) :2033–2045, 1990.

[Bellman 1957] Bellman R., *Dynamic programming*, Princeton University Press, 1957.

[Bellot 2006] Bellot O., *Adaptation au locuteur des modèles acoustiques dans le cadre de la reconnaissance automatique de la parole*, Thèse de doctorat, université d'Avignon, LIA, 2006.

[Billi 1982] Billi R., Vector quantization and markov source models applied to speech recognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1982)*, tome 7, pages 574–577, 1982.

[Bocchieri 1993] Bocchieri E., Vector quantization for the efficient computation of continuous density likelihoods, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1993)*, tome 2, pages 692–695, Minneapolis, Minnesota, USA, 1993.

[Bocchieri 1997] Bocchieri E. et Mak B., Subspace distribution clustering for continuous observation density hidden markov models, dans *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'1997)*, pages 107–110, Rhodes, Greece, 1997.

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

[Bourlard 1987] Bourlard H. et Wellekens C.-J., Multi-layer perceptrons and Automatic Speech Recognition, dans *Proceedings of the IEEE First Annual International Conference on Neural Networks*, pages IV :407–416, San Diego, 1987.

[Bourlard 1990] Bourlard H. et Wellekens C.-J., Links between Markov models and multi-layer perceptrons, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 :1167–1178, 1990.

[Carbonell 1986] Carbonell N., Damestoy J.-P., Fohr D., Haton J.-P. et Lonchamp F., APHODEX, design and implementation of an acoustic-phonetic decoding expert system, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, tome 11, pages 1201– 1204, Tokyo, Japan, 1986.

[Carré 1984] Carré R., Descout R., Eskénazi M., Mariani J. et Rossi M., The French language database : defining, planning and recording a large database, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1984)*, pages 324–327, San Diego, California, USA, 1984.

[Chen 2005] Chen C.-P., Bilmes J. et Ellis D.-P.-W., Speech feature smoothing for robust ASR, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2005)*, pages 525–528, Philadelphia, Pennsylvania, USA, 2005.

[Chen 2002] Chen C.-P., Bilmes J. et Kirchhoff K., Low-resource noise-robust feature post-processing on Aurora 2.0, dans *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'2002)*, pages 2445–2448, Denver, Colorado, USA, 2002.

[Cho 2004] Cho Y., Kim S. et Yook D., Hybrid model using Subspace Distribution Clustering Hidden Markov Models and Semi-Continuous Hidden Markov Models for embedded speech recognizers, dans *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP'2004)*, pages 669–672, Jeju Island, Korea, 2004.

[Cornu 2002] Cornu E., Destrez N., Dufaux A., Sheikhzadeh H. et R.Brennan, An ultra low power, ultra miniature voice command system based on hidden markov models, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2002)*, tome 4, pages 3800–3803, Orlando, Florida, USA, 2002.

[Davis 1952] Davis K., Biddulph R. et Balashek S., Automatic recognition of spoken digits, dans *Journal of the Acoustical Society of America*, tome 24-6, pages 637–642, 1952.

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

[Deligne 2001] Deligne S., Eide E., Gopinath R.-A., Kanevksy D., Maison B., Olsen P., Printz H. et Sedivy J., Low-resource speech recognition of 500-word vocabularies, dans *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech'2001)*, pages 1833–1836, Aalborg, Denmark, 2001.

[Dolmazon 1997] Dolmazon J.-M., Bimbot F., Adda G., El-Bèze M., Caerou J.-C., Zeiliger J. et Adda-Decker M., ARC B1 - organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale, dans *Journées Scientifiques et Techniques*, pages 13–18, Avignon, France, 1997.

[Dreyfus-Graf 1950] Dreyfus-Graf J., Sonograph and sound mechanics, dans *Journal of the Acoustical Society of America*, tome 22-6, pages 731–739, 1950.

[Duchateau 1998] Duchateau J., Demuynck K., Compernolle D. V. et Wambacq P., Improved parameter tying for efficient acoustic model evaluation in large vocabulary continuous speech recognition, dans *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'1998)*, tome 5, pages 2215–2218, Sydney, Australia, 1998.

[Eisele 1996] Eisele T., Haeb-Umbach R. et Langmann D., A comparative study of linear feature transformation techniques for automatic speech recog-

nition, dans *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'1996)*, tome 1, pages 252–255, Philadelphia, Pennsylvania, USA, 1996.

[ETSI-AMR 2000] ETSI-AMR, Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90 version 7.2.1), 2000, URL `http://www.3gpp.org/ftp/Specs/archive/06_series/06.90/0690-721.zip`.

[ETSI-EFR 2000] ETSI-EFR, Enhanced Full Rate (EFR) speech transcoding (GSM 06.60 version 8.0.1), 2000, URL `http://www.3gpp.org/ftp/Specs/archive/06_series/06.60/0660-801.zip`.

[ETSI-FR 2000] ETSI-FR, Full Rate (FR) speech transcoding (GSM 06.10 version 8.2.0), 2000, URL `http://www.3gpp.org/ftp/Specs/archive/06_series/06.10/0610-820.zip`.

[ETSI-HR 2000] ETSI-HR, Half Rate (HR) speech transcoding (GSM 06.20 version 8.0.1), 2000, URL `http://www.3gpp.org/ftp/Specs/archive/06_series/06.20/0620-801.zip`.

[Euler 1994] Euler S. et Zinke J., The influence of speech coding algorithms on automatic speech recognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1994)*, tome 1, pages 621–624, Adelaïde, Australia, 1994.

[Fohr 1997] Fohr D., Haton J.-P., Mari J.-F., Smaïli K. et Zitouni I., MAUD : un prototype de machine à dicter vocale, dans *Journées Scientifiques et Techniques*, pages 25–30, Avignon, France, 1997.

[Franco 2002] Franco H., Zheng J., Butzberger J., Cesari F., Frandsen M., Arnold J., Gadde V.-R. R., Stolcke A. et Abrash V., Dynaspeak : Sri's scalable speech recognizer for embedded and mobile systems, dans *Proceedings of the ARPA workshop on Human Language Technology*, San Diego, CA, USA, 2002.

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

[Gales 1996] Gales M., Pye D. et Woodland P., Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation, dans *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'1996)*, tome 3, pages 1832–1835, Philadelphia, Pennsylvania, USA, 1996.

[Gallardo-Antolin 1998] Gallardo-Antolin A., de Maria F. D. et Valverde-Albacete F., Recognition from GSM digital speech, dans *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'1998)*, pages 584–587, Sydney, Australia, 1998.

[Galliano 2005] Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J.-F. et Gravier G., The ESTER Phase II evaluation campaign for the rich transcription of french broadcast news, dans *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech'2005)*, pages 1149–1152, Lisboa, Portugal, 2005.

[Gas 2000] Gas B., Zarader J.-L. et Chavy C., A new approach to speech coding : the neural predictive coding, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 4(1) :120–127, 2000.

[Gauvain 1994] Gauvain J.-L. et Lee C.-H., Maximum A Posteriori estimation for multivariate gaussian mixture observations of Markov chains, dans *IEEE Transactions on Speech and Audio Processing*, tome 2-2, pages 291–298, 1994.

[Geutner 2000] Geutner P., Arevalo L. et Breuninger J., VODIS - voice-operated driver information systems : a usability study on advanced speech technologies for car environments, dans *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000)*, pages 378–382, Beijing, China, 2000.

[Glass 1996] Glass J., Chang J. et McCandless M., A probabilistic framework for feature-based speech recognition, dans *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'1996)*, tome 4, pages 2277–2280, Philadelphia, Pennsylvania, USA, 1996.

[Gong 2000] Gong Y. et Kao Y.-H., Implementing a hich accuracy speaker-independent continuous speech recognize on a fixed-point, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2000)*, tome 6, pages 3686–3689, Istanbul, Turkey, 2000.

[Gravier 2004] Gravier G., Bonastre J.-F., Geoffrois E., Galliano S., McTait K. et Choukri K., ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français, dans *XXVèmes Journées d'Etudes sur la Parole JEP'2004*, pages 253–256, Fès, Morocco, 2004.

[Haton 1994] Haton J.-F. M. A.-P., Automatic word recognition based on second-order hidden markov models, dans *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP'1994)*, pages 247–250, Yokohama, Japan, 1994.

[Haton 1990] Haton J.-P., Bonneau A., Fohr D., Laprie Y., Gong Y. et Pierrel J.-M., Décodage acoustico-phonétique : problèmes et éléments de solution, *Traitement du Signal*, 7(4) :293–313, 1990.

[Haton 2006] Haton J.-P., Cerisara C., Fohr D., Laprie Y. et Smaïli K., *Reconnaissance automatique de la parole*, Dunod, 2006.

[Hermansky 1990] Hermansky H., Perceptual Linear Predictive (PLP) analysis of speech, *Journal of the Acoustical Society of America*, 87-4 :1738–1752, 1990.

[Huang 1992] Huang X., Alleva F., Hon H.-W., H. M.-Y. et Rosenfeld R., The SPHINX-II speech recognition system : an overview, *Computer Speech and Language*, 7(2) :137–148, 1992.

[Huang 1989] Huang X. et M.Jack, Large-vocabulary speaker-independent continuous speech recognition with Semi-Continiuous Hidden Markov Models, dans *Proceedings of the 1st European Conference on Speech Communication and Technology (Eurospeech'1989)*, pages 1163–1166, Paris, France, 1989.

[Huggins-Daines 2006] Huggins-Daines D., Kumar M., Chan A., Black A.-W., Ravishankar M. et Rudnicky A., Pocketsphinx : a free, real-time continuous speech recognition system for hand-held devices, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2006)*, pages 185–188, Toulouse, France, 2006.

[Hunt 1989] Hunt M.-J. et Lefebvre C., A comparison of several acoustic representations for speech recognition with degraded and undegraded speech, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1989)*, tome 1, pages 262–265, Glasgow, UK, 1989.

[Ishikawa 2006] Ishikawa S., Yamabana K., Isotani R. et Okumura A., Parallel LVCSR algorithm for cellphone-oriented multicore processor, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2006)*, pages 177–180, Toulouse, France, 2006.

[Jankowski 1995] Jankowski C.-R., Vo H.-D.-H. et Lippmann R.-P., A comparison of signal processing front ends for automatic word-recognition, dans *IEEE Transactions on Speech and Audio Processing*, tome 3(4), pages 286–293, 1995.

[Jelinek 1976] Jelinek F., Continuous speech recognition by statistical methods, dans *Proceedings of the IEEE*, tome 64-4, pages 532–556, 1976.

[Junqua 1993] Junqua J.-C., Wakita H. et Hermansky H., Evaluation and optimization of perceptually-based asr front-end, *IEEE Transactions on Speech and Audio Processing*, 1(1) :39–48, 1993.

[Kedem 1986] Kedem B., Spectral analysis and discrimination by zero-crossings, *Proceedings of the IEEE*, 74(11) :1477– 1493, 1986.

[Kim 2001] Kim H. et Cox R.-V., A bitstream-based front-end for wireless speech recognition on IS-260 communication system, *IEEE Transactions on Speech and Audio Processing*, 9 :558–568, 2001.

[Kumar 1998] Kumar N. et Andreou A.-G., Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, dans *Speech Communication*, tome 26, pages 283–297, 1998.

[Lamel 1991] Lamel L.-F., Gauvain J.-L. et Eskénazi M., BREF, a large vocabulary spoken corpus for French, dans *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech'1991)*, pages 505–508, Gênes, Italie, 1991.

[Lee 1989] Lee C.-H., Juang B.-H., Soong F.-K. et Rabiner L.-R., Word recognition using whole word and subword models, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1989)*, tome 1, pages 683–686, Glasgow, UK, 1989.

[Lee 1990] Lee K.-F., Hon H.-W. et Reddy R., An overview of the SPHINX speech recognition system, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1) :35–45, 1990.

[Lefort 2002] Lefort L., Merlin T., Bonastre J.-F. et Nocera P., Le projet MTM reconnaissance de la parole et du locuteur sur une plateforme embarquée, dans *XXIVèmes Journées d'Etudes sur la Parole JEP'2002*, pages 265–268, Nancy, France, 2002.

[Leggetter 1994] Leggetter C.-J. et Woodland P.-C., Speaker adaptation of continuous density hmms using multivariate linear regression, dans *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP'1994)*, pages 451–454, Yokohama, Japan, 1994.

[Lévy 2003] Lévy C., Linarès G. et Nocera P., Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems, dans *Workshop on DSP in Mobile and Vehicular Systems*, Nagoya, Japan, 2003.

[Lévy 2004] Lévy C., Linarès G., Nocera P. et Bonastre J.-F., Reducing computational and memory cost for cellular phone embedded speech recognition system, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2004)*, pages 309–312, Montreal, Canada, 2004.

[Lévy 2007] Lévy C., Linarès G., Nocera P. et Bonastre J.-F., *Embedded mobile phone digit-recognition*, chapitre 7 in Advances for In-Vehicle and Mobile Systems, H. Abut, J.H.L. Hansen and K. Takeda (Eds.), Springer Science, 2007.

[Lilly 1996] Lilly B. et Paliwal K., Effect of speech coders on speech recognition performance, dans *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'1996)*, tome 4, pages 2344–2347, Philadelphia, Pennsylvania, USA, 1996.

[Linarès 2006] Linarès G., Lévy C. et Plagniol J.-C., Estimation rapide de modèles semi-continus discriminants, dans *XXVIèmes Journées d'Etudes sur la Parole JEP'2006*, pages 131–134, Dinard, France, 2006.

[Macas-Guarasa 1996] Macas-Guarasa J., Gallardo A., Ferreiros J., Pardo J. et Villarrubia L., Initial evaluation of a preselection module for a flexible large vocabulary, dans *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'1996)*, pages 1343–1346, Philadelphia, Pennsylvania, USA, 1996.

[Maes 2000] Maes S.-H., Cohen G., Hoory R. et Chazan D., Conversational networking : Conversational protocols for transport, coding and control, dans *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000)*, tome 2, pages 198–201, Beijing, China, 2000.

[Mak 2001] Mak B.-K. et Bocchieri E., Direct training of subspace distribution clustering hidden markov model, *IEEE Transactions on Speech and Audio Processing*, 9(4) :378–387, 2001.

[Matrouf 2003] Matrouf D., Bellot O., Nocera P., Linarès et Bonastre J.-F., Structural linear model-space transformations for speaker adaptation, dans *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech'2003)*, pages 1625–1628, Geneva, Switzerland, 2003.

[Miet 2001] Miet G., *Towards wideband speech by narrowband speech bandwitdth extension : magic effect or wideband recovery ?*, Thèse de doctorat, University of Maine, 2001.

[Nocera 2002] Nocera P., Linares G., Massonié D. et Lefort L., Phoneme lattice based a* search algorithm for speech recognition, dans *Proceedings of the Fifth International Conference on Text, Speech, Dialogue*, pages 300–308, Brno, Czech Republic, 2002.

[Nouza 1996] Nouza J., Feature selection methods for hidden markov model-based speech recognition, dans *Proceedings of the International Conference on Pattern Recognition*, tome 2, pages 186–190, Vienna, Austria, 1996.

[Obermaier 1998] Obermaier B. et Rinner B., A tms320c40 based speech recognition system for embedded applications, dans *The 2nd European DSP Education & Research Conference*, Paris, France, 1998.

[Paliwal 1993] Paliwal K.-K. et Atal B.-S., Efficient vector quantization of lpc parameters at 24 bits/frame, *IEEE Transactions on Speech and Audio Processing*, 1(1) :3–14, 1993.

[Park 2004] Park J. et Ko H., Compact acoustic model for embedded implementation, dans *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP'2004)*, pages 693–696, Jeju Island, Korea, 2004.

[Povey 1999] Povey D. et Woodland P.-C., Frame discrimination training of hmms for large vocabulary speech recognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1999)*, pages 333–336, Phoenix, Arizona, USA, 1999.

[Psutka 2001] Psutka J., Müller L. et Psutka J.-V., Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task, dans *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech'2001)*, pages 1813–1816, Aalborg, Danemark, 2001.

[Rabiner 1989] Rabiner L.-R., A tutorial on hidden markov models and selected applications in speech recognition, dans *IEEE Transactions on Speech and Audio Processing*, tome 77-2, pages 257–285, 1989.

[Raj 2001] Raj B., Migdal J. et Singh R., Distributed speech recognition with codec parameters, dans *Automatic Speech Recognition and Understanding Workshop, IEEE, ASRU'2001*, pages 127–30, Madonna di Campiglio, Trento, Italy, 2001.

[Ramaswamy 1998] Ramaswamy G.-N. et Gopalakrishnan P.-S., Compression of acoustic features for speech recognition in network environments, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1998)*, tome 2, pages 977–980, Seattle, Washington, USA, 1998.

[Reynolds 2000] Reynolds D.-A., Quatieri T.-F. et Dunn R.-B., Speaker verification using adapted gaussian mixture models, *Digital Signal Processing*, 10 :19–41, 2000.

[Sagayama 1995] Sagayama S. T. A. S., Four-level tied-structure for efficient representation of acoustic modeling, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1995)*, tome 1, pages 520–523, Detroit, Michigan, USA, 1995.

[Sambur 1975] Sambur M., Selection of acoustic features for speaker identification, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(2) :176–182, 1975.

[Savariaux 1997] Savariaux C., Farhat A., Héon M., O'Shaghnessy D. et Lee C.-Z., Nouvelles avancées en reconnaissance de la parole continue grand vocabulaire du français basées sur le système de reconnaissance de l'INRS-Télécommunications, dans *Journées Scientifiques et Techniques*, Avignon, France, 1997.

[Schwartz 1993] Schwartz R., Anastasakos T., Kubala F., MAkhoul J., Nguyen L. et Zavaliagkos G., Comparative experiments on large vocabulary speech recognition, dans *Proceedings of the ARPA workshop on Human Language Technology*, pages 75–80, Princeton, NJ, USA, 1993.

[Schwartz 1989] Schwartz R., Barry C., Chow Y.-L., Derr A., Feng M.-W., Kimball O., Kubala F., Makhoul J. et Vandegrift J., The BBN BYBLOS continuous speech recognition system, dans *Proceedings of the ARPA workshop on Human Language Technology*, pages 94–99, Philadelphia, Pennsylvania, USA, 1989.

[Shannon 1948] Shannon C.-E., A mathematical theory of communication, dans *The Bell System Technical Journal*, tome 27, pages 379–423, 623–656, 1948.

[Shore 1983] Shore J. et Burton D., Discrete utterance speech recognition without time alignment, *IEEE Transactions on Information theory*, 29(4) :473–491, 1983.

[Srinivasamurthy 2005] Srinivasamurthy N., Ortega A. et Narayanan S., Efficient scalable encoding for distributed speech recognition, *IEEE Transactions on Speech and Audio Processing*, 2005.

[Stern 1986] Stern P.-E., Eskenazi M. et Memmi D., An expert system for speech spectrogram reading, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, tome 11, pages 1193– 1196, Tokyo, Japan, 1986.

[Taboada 1994] Taboada J., Feijoo S., Balsa R. et Hernandez C., Explicit estimation of speech boundaries, *Proceedings of the IEEE Science, Measurement and Technology*, 141(3) :153–159, 1994.

[Tremain 1982] Tremain T.-E., The government standard Linear Predictive Coding algorithm : LPC10, dans *Speech Technology Magazine*, tome 1-2, pages 40–49, 1982.

[Tubach 1989] Tubach J., rédacteur, *La parole et son traitement automatique*, Masson, collection technique et scientifique des télécommunications édition, 1989.

[Vaich 1999] Vaich T. et Cohen A., Comparison of Continuous-Density and Semi-Continuous HMM in isolated words recognition systems, dans *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech'1999)*, pages 1515–1518, Budapest, Hungary, 1999.

[Valtchev 1997] Valtchev V., Odell J., Woodland P. et Young S., MMIE training of large vocabulary recognition systems, dans *Speech Communication*, tome 22, pages 303–314, 1997.

[Valtchev 1996] Valtchev V., Odell J.-J., Woodland P.-C. et Young S.-J., Lattice-based discriminative training for large vocabulary speechrecognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1996)*, tome 2, pages 605–608, Atlanta, GA, USA, 1996.

[Viterbi 1967] Viterbi A., Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, dans *IEEE Transactions on Information Theory*, tome 2-13, pages 260–269, 1967.

[Waibel 1989] Waibel A., Hanazawa T., Hinton G., Shikano K. et Lang K., Phoneme recognition using Time-Delay Neural Networks, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37 :328–339, 1989.

[Wang 2004] Wang D., Zhang L., Liu J. et Liu R., Emnedded speech recognition system on 8 bits MCU core, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2004)*, tome 5, pages 301–304, Montreal, Canada, 2004.

[Welling 1997] Welling L. et Ney N. H. A., Acoustic front-end optimization for large vocabulary speech recognition, dans *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'1997)*, pages 2099–2102, Rhodes, Greece, 1997.

[Woodland 2000] Woodland P.-C. et Povey D., Large scale discriminative training for speech recognition, In ISCA ITRW Automatic Speech Recognition : Challenges for the Millenium, pages 7-16, Paris, 2000.

[Woodland 1993] Woodland P.-C. et Young S.-J., The htk tied-state continuous speech recogniser, dans *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech'1993)*, pages 2207–2210, Berlin, Germany, 1993.

[Young 1995] Young S., Large vocabulary continuous speech recognition : A review, dans *Automatic Speech Recognition and Understanding Workshop, IEEE, ASRU'1995*, pages 3–28, Snowbird, Utah, USA, 1995.

[Young 1994] Young S., Odell J. et Woodland P., Tree-based state tying for high accuracy acoustic modelling, dans *Proceedings of the ARPA workshop on Human Language Technology*, pages 307–312, Plainsboro, New Jersey, USA, 1994.

[Young 1992] Young S.-J., The general use of tying in phoneme-based HMM speech recognisers, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1992)*, pages 569–572, San Francisco, California, USA, 1992.

[Zue 1986] Zue V. et Lamel L., An expert spectrogram reader : A knowledge-based approach to speech recognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, tome 11, pages 1197– 1200, Tokyo, Japan, 1986.