Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole.

Loïc Barrault

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

École Doctorale 166 I2S « Mathématiques et Informatique »
Laboratoire d'Informatique (EA 931)

Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole.

par
Loïc BARRAULT

Soutenue publiquement le 18 Juillet 2008 devant un jury composé de :

M.	Henri MÉLONI	Professeur, LIA, Avignon	Président du jury
M^{me}	Régine ANDRÉ-OBRECHT	Professeur, IRIT, Toulouse	Rapporteur
M.	Pietro LAFACE	Professeur, Politecnico de Torino	Rapporteur
M^{me}	Martine ADDA-DECKER	CR1-HDR, LIMSI/CNRS, Paris XI-Orsay	Examinateur
M.	Guillaume GRAVIER	CR, IRISA/CNRS, Rennes	Examinateur
M.	Renato DE MORI	Professeur, LIA, Avignon	Directeur de thèse
M.	Driss MATROUF	Maître de Conférences, LIA, Avignon	Co-Directeur de thèse

Laboratoire d'Informatique d'Avignon

Résumé

La Reconnaissance Automatique de la Parole (RAP) est affectée par les nombreuses variabilités présentes dans le signal de parole. En dépit de l'utilisation de techniques sophistiquées, un système de RAP seul n'est généralement pas en mesure de prendre en compte l'ensemble de ces variabilités. Nous proposons l'utilisation de diverses sources d'information acoustique pour augmenter la robustesse des systèmes de reconnaissance.

La combinaison de différents jeux de paramètres acoustiques repose sur l'idée que certaines caractéristiques du signal de parole sont davantage mises en avant par certains jeux de paramètres que par d'autres. L'intérêt est donc d'exploiter les points forts de chacun. Par ailleurs, les différentes partitions de l'espace acoustique opérées par les modèles acoustiques peuvent être mises à profit dans des techniques de combinaison bénéficiant de leur éventuelle complémentarité.

Le diagnostic est au cœur de ce travail. L'analyse des performances de chaque jeu de paramètres permet la mise en évidence de contextes spécifiques dans lesquels la prédiction du résultat de reconnaissance est possible. Nous présentons une architecture de diagnostic dans laquelle le système de RAP est vu comme un « canal de transmission » dont l'entrée correspond aux phonèmes contenus dans le signal de parole et la sortie au résultat de reconnaissance. Cette architecture permet de séparer les différentes sources d'ambiguïté à l'intérieur du système de reconnaissance. Les analyses ont permis d'intégrer des stratégies de combinaison post-décodage à un niveau segmental élevé (phonème ou mot).

Des techniques de combinaison des probabilités *a posteriori* des états d'un modèle de Markov caché connaissant un vecteur de paramètres acoustiques sont également proposées. Afin d'améliorer l'estimation des probabilités *a posteriori*, les probabilités obtenues avec différents modèles acoustiques sont fusionnées. Pour combiner les probabilités de manière cohérente, les modèles acoustiques doivent avoir la même topologie. Par conséquent, nous avons développé un protocole permettant d'entraîner des modèles de même topologie avec des paramètres acoustiques différents. Plusieurs méthodes pour estimer des facteurs de pondération et pour générer des modèles acoustiques complémentaires sont également présentées.

Abstract

Automatic Speech Recognition (ASR) is affected by many variabilities present in the speech signal. Despite sophisticated techniques, a single ASR system is usually incapable of considering all these variabilities. We propose to use various sources of acoustic information in order to increase precision and robustness.

Combination of various acoustic feature sets is motivated by the assumption that some characteristics that are de-emphasized by a particular feature set are emphasized by another. Therefore, the goal is to make the most of their strengths. In addition, acoustic models make different partition of the acoustic space so that they can be used in a combination scheme relying on their complementarity.

Diagnosis is at the core of this research. Performance analysis of each feature set brings out specific contexts where the prediction of the recognition result is possible. We propose a diagnosis architecture in which the ASR system is shown as a "channel model" which takes as input the phonemes present in the speech signal and outputs phoneme hypotheses given by the system. This architecture allows different sources of confusion to be separated within the recognition system. The performed analyses enable the introduction of post-decoding combination strategies at a high segmental level (word or phoneme).

Combination of a posteriori probabilities of states of a Hidden Markov Model (HMM) given a feature frame is also proposed. In order to better estimate such a posteriori probabilities, probabilities obtained with several acoustic models are fused. For the sake of consistency, the topology of the acoustic models has to be equivalent. In consequence, we propose a new fast, efficient protocol to train models having the same topology but using different acoustic feature sets. Several methods to estimate weighting factors and to generate complementary acoustic models for combination are also suggested.

Remerciements

Je tiens tout d'abord à remercier l'ensemble des membres de mon jury. Merci à Henri Méloni, président du jury, à mes rapporteurs Régine André-Obrecht et Pietro Laface et à mes examinateurs Martine Adda-Decker et Guillaume Gravier pour le temps qu'ils ont consacré à la lecture de mon document ainsi que pour leurs remarques judicieuses et leurs critiques constructives.

Mes plus profonds remerciements vont à mon directeur Renato De Mori. C'est un honneur d'avoir travaillé avec un chercheur de si grande envergure.

Je remercie spécialement Driss Matrouf, co-directeur de ma thèse. Son expérience, ses conseils pratiques et nos fameuses "drisscussions" ont été pour moi d'une grande aide tout au long de ce doctorat.

سكرن دريس مترف

Cette thèse achève mes 8 années de formation « dans le sud ». Je remercie chaleureusement les membres de l'IUP GMI et du Laboratoire Informatique d'Avignon pour les merveilleux moments qu'on a passé ensemble.

Je commence par les potes de l'IUP : Steph, Cissou, Christian, Thomthom et Nanou . . . *5 jeunes gens plein d'avenir ! !*
Merci à JP, Cathy, Mireille, Dom, Laurence, Philou, Pierrot, Thierry, Jef, Georges, Fred, Corinne, Gilloux, Nimaan, MJ, Eric, Alain, ML, Mathieu ainsi qu'à la nouvelle génération, TiFred, Stan, Nico . . . j'en passe et des meilleurs.
Et puis comment pourrais-je ne pas mentionner tous les *potos* du labo avec lesquels on a bien fait la fête ! Je pense à Laurianne et Thomas, Christian, Antho et Bérénice, William et Virginie, Nico et Nicole, Ben et Laure, Christophe et Charlotte, Alex.

Cette thèse aura également permis la naissance d'une grande amitié. Merci aux *Juju* : Stéphane, Nathalie et Sarah pour ces moments de rigolade, de balade et d'escalade toujours dans la simplicité et l'humour. Merci à Stéphane pour ces parties de Magic endiablées qui permettent de se vider la tête. Merci à Nathalie pour ces années de cohabitation et de collaboration, c'était un plaisir . . . *on a fait du bon boulot ! !* Je suis sûr que cela continuera . . . et même dans dix-7 ans ! !

Merci à toute ma famille, principalement à mes parents et à mes frères et soeurs, qui se sont toujours enquis de savoir comment avançait mon travail. Vous m'avez continuellement encouragé dans ce que je faisais. C'est très réconfortant de pouvoir compter sur vous !

Je dédie cette thèse à ma femme, **Nathalie**, et à mes deux filles, **Romane** et **Louisa**. Merci à Nath pour sa patience, ses encouragements, sa foi en moi et pour le plus merveilleux des cadeaux qu'elle a pu me faire. Merci aux poupettes pour leur joie de vivre. Rien de tel que leurs sourires pour débuter la journée en beauté !

Préambule

Cette thèse a été en grande partie financée par le projet européen DIVINES : Diagnostic and Intrinsic Variabilities In Natural Speech, initié par le 6th *Framework Programme* de la Commission Européenne. Le projet vise à proposer des alternatives aux techniques *état-de-l'art* d'extraction de paramètres acoustiques, de modélisation acoustique et de modélisation linguistique dans le but de combler l'écart entre les performances de reconnaissance de la parole humaine et automatique. Ces nouvelles techniques se basent sur le diagnostic de la nature et de la cause des erreurs produites par le système de reconnaissance.

Les informations complémentaires concernant ce projet peuvent être trouvées sur le site web dédié : http://divines-project.org.

Dans le cadre du partenariat, nous avons utilisé le système de reconnaissance de Loquendo. Ce système a été entraîné sur plusieurs langues, telles que l'Italien, l'Espagnol ou l'Allemand, ce qui explique pourquoi les corpus utilisés dans la première partie de mon travail sont les parties italienne et espagnole d'Aurora3.

Table des matières

Introduction

L'apparition de l'Internet moderne au début des années 1990 nous a propulsé dans l'ère de l'information et de la communication. Depuis, les technologies n'ont cessé de progresser, permettant de communiquer et d'accéder à l'information de n'importe quel endroit et de plus en plus facilement. Les interactions nécessaires entre l'homme et ces technologies requièrent une interface qui se doit d'être naturelle et utilisable. Les hommes n'ayant de meilleur moyen de communication que le mode oral, il est apparu évident de le transposer à la communication homme-machine. C'est dans ce sens que la communauté scientifique a fourni un grand effort de recherche depuis plus de 50 ans pouvant aujourd'hui être mis à profit dans des applications vocales pour lesquelles l'intérêt du public est croissant. Cependant, afin de rendre ce moyen de communication utilisable au quotidien, il est impératif d'avoir un système adapté à l'être humain. D'ailleurs, comme Bristow (1986) le fait remarquer :

> *"Speech recognition is about computers learning how to communicate with humans, rather than vice versa."*

Cela signifie que les contraintes d'utilisation doivent être limitées et que le système doit pendre en compte certains critères environnementaux dépendant du cadre d'utilisation de l'application. Un système peut être utilisé par n'importe qui et dans n'importe quel endroit. Il est donc confronté à de nombreux événements qui viennent perturber le signal de parole.

Variabilités du signal de parole

Les systèmes de reconnaissance de la parole actuels se basent sur l'estimation de l'enveloppe spectrale du signal de parole afin d'extraire l'information pertinente pour la reconnaissance. Or, le signal de parole est sujet à de nombreuses déformations ou variabilités qui entraînent des modifications de cette enveloppe spectrale. Les sources de variabilités ont des effets divers et variés sur le signal de parole.

L'environnement : le bruit extérieur capté par le microphone est totalement décorrélé du signal de parole et engendre une modification du spectre de ce dernier. Le phénomène de réverbération, correspondant à la réflexion de l'onde acoustique sur les objets de l'environnement, provoque une déformation du signal de parole capté par le microphone. Cette déformation est, quant-à elle, corrélée avec le signal de parole. La présence de plusieurs locuteurs discutant en même temps est une des problématiques concernant la robustesse des systèmes de RAP. Pour pouvoir communiquer dans un environnement très bruité, le locuteur augmente, généralement, la puissance de sa voix, ce qui a pour effet de modifier le spectre fréquentiel du signal de parole. Cet effet indirect du bruit environnant est appelé effet Lombard.

Le locuteur : une même personne ne prononcera jamais deux fois la même phrase de la même manière, cette variabilité intra-locuteur peut causer l'échec de la reconnaissance si elle n'est pas prise en compte. D'autre part, les caractéristiques physiologiques, sociales ou environnementales propres à chaque locuteur (âge, sexe, accent, ...) impliquent des différences entre les prononciations d'un même mot produisant une variabilité dite inter-locuteur. Ce type de variabilité est matérialisée dans le système par des vecteurs acoustiques très différents représentant la même unité acoustico-phonémique.

Le canal de transmission : les propriétés du canal (le type de microphone utilisé, la distance entre le microphone et le locuteur, ...) provoquent des variabilités au niveau du signal de parole.

Comme l'a fait remarquer Gauvain (2000) dans son état de l'art de la reconnaissance de la parole, l'amélioration des modèles acoustique, lexical, syntaxique et sémantique est nécessaire pour rendre les systèmes vocaux utilisables en conditions réelles.

Élements de robustesse actuels

La robustesse d'un système est définie par sa capacité à faire face à des événements nouveaux non prévus initialement. C'est un domaine de recherche très fertile et de nombreuses techniques ont été développées pour améliorer chaque composante du système. Pour cela, il est nécessaire d'intégrer des méthodes permettant de prendre en compte ou d'atténuer ces variabilités.

Les traitements possibles effectués à chaque niveau du système de reconnaissance sont présentés dans les paragraphes suivants. Les différents composants d'un système de reconnaissance automatique de la parole seront décrits en détail dans le chapitre 1.

Paramètres acoustiques : des traitements spécifiques peuvent être mis en œuvre pour rendre les paramètres acoustiques plus robustes au bruit. L'objectif est de normaliser l'espace des vecteurs acoustiques. Certaines techniques opèrent dans le domaine temporel, tel que le « soft-thresholding » (Donoho, 1995; Gemello et al., 2002). D'autres modifient le spectre du signal comme la soustraction spectrale (Boll, 1979; Martin, 1994). La normalisation du cepstre moyen est une technique couramment utilisée pour réduire l'influence du canal de transmission. Elle agit dans le domaine cepstral (Acero, 1990). La technique de normalisation de la longueur du conduit vocal (Cohen et al., 1995) compense les différences de longueur de conduit vocal entre les différents locuteurs.

Modèles acoustiques : une des principales contraintes pour le bon apprentissage des modèles acoustiques est la quantité de données disponible pour l'estimation des paramètres du modèle. Chaque unité phonétique doit être suffisamment représentée dans le corpus d'apprentissage. En outre, un problème de modélisation se pose lorsque les données d'apprentissage sont très différentes des données de la tâche ciblée. Les modèles acoustiques peuvent alors être adaptés afin de mieux faire correspondre leurs paramètres aux différentes prononciations des unités phonétiques pouvant être rencontrées.

On distingue deux approches principales pour augmenter la robustesse des modèles acoustiques. La première consiste à modifier les paramètres des modèles acoustiques. On peut citer l'adaptation Maximum *a Posteriori* (MAP, voir section 1.5.1) et ses dérivées qui ont été créées pour modifier les paramètres des modèles acoustiques en vue de prendre en compte les conditions d'enregistrement, le bruit environnant, le locuteur et ses caractéristiques (genre, âge, etc.) (Gauvain et Lee, 1994; Matrouf et al., 2001). Une alternative pour l'adaptation des modèles acoustiques est la *Maximum Likelihood Linear Regression* (MLLR, voir section 1.5.2) proposée par Leggetter et Woodland (1995). La seconde approche consiste à estimer un modèle du bruit et à le combiner avec un modèle de parole non bruitée. Un exemple de ce type d'approche est la combinaison parallèle de modèles (Gales et Young, 1996). Différents types d'adaptation de modèles acoustiques sont présentées dans Matrouf (1997).

Modèles de langage : Un critère de robustesse pour les modèles de langage est la taille du contexte utilisé. Lors de l'apprentissage, il est nécessaire d'avoir un jeu de données suffisamment conséquent pour estimer correctement les paramètres du modèle. La complexité de la modélisation linguistique peut-être réduite en diminuant la taille du vocabulaire. Par exemple, dans les serveurs d'appels téléphoniques actuels, un petit ensemble de mots clés est utilisé pour accéder aux services proposés. Ceci n'est bien sûr pas satisfaisant du point de

vue utilisabilité. Cette technique ne fait que se substituer aux commandes hiérarchisées par touches dédiées[1] qui sont très contraignantes pour l'utilisateur.

Hypothèses de reconnaissances : Les systèmes de reconnaissance de la parole génèrent en sortie la suite de mots la plus probable connaissant les modèles acoustique et linguistique. Cette suite de mots est issue du parcours du treillis de mots par un algorithme de recherche. Cet algorithme permet de trouver les N-meilleures séquences de mots en fonction de leurs scores acoustique et linguistique. À ce niveau, des techniques de réévaluation des scores peuvent être employées afin de générer de nouvelles hypothèses. Dans ce type d'approche, la robustesse se présente sous la forme de mesures qui tiennent compte de la qualité du signal, de la confiance accordée au système ou d'autres caractéristiques dynamiques du signal de parole. De nombreux travaux ont été effectués dans l'élaboration de mesures de confiance visant à estimer la qualité du système et répercuter cette information dans le processus de reconnaissance. Une partie d'entre elles est présentée dans la section 3.5 (p 76).

Malgré l'existence de toutes ces méthodes permettant d'augmenter la robustesse et la précision des systèmes, les performances restent encore insuffisantes lorsque la parole est spontanée, bruitée et/ou enregistrée dans des conditions différentes de l'apprentissage. Les techniques permettant d'augmenter la robustesse des différents composants d'un système de reconnaissance améliorent les résultats, mais des imperfections subsistent.

Travaux réalisés

Avec pour objectif d'accroître la robustesse des systèmes de reconnaissance de la parole, nous proposons de combiner plusieurs jeux de paramètres différents correspondant à différentes manières de représenter le signal de parole. La combinaison de systèmes, que ce soit avant ou après le décodage et quel que soit le niveau segmental, tire parti des avancées dans chaque domaine (paramétrisation et modélisation acoustique, stratégies de décodage, réévaluation des hypothèses de reconnaissance). L'important est d'établir une stratégie d'évaluation des performances des systèmes permettant d'identifier leurs différences, leurs points forts et leurs faiblesses afin de les combiner de manière intelligente. On peut naturellement se poser la question de savoir ce que sont des systèmes différents ? Comment mesurer cette différence ? Et surtout, comment profiter au maximum de ces différences afin de trouver la meilleure manière de les combiner ?

[1] Exemple de commandes par touches dédiées : pour les prévisions météo tapez 1, pour les horaires de cinéma tapez 2, ...

L'idée de combiner des systèmes n'est pas nouvelle, mais il apparaît que les techniques appliquant des règles de combinaison de manière systématique ne donnent pas totale satisfaction.

Afin d'effectuer une combinaison plus adaptée, moins systématique, il est important d'identifier et de comprendre les causes des échecs de la reconnaissance automatique de la parole. Il est intéressant d'identifier les points forts de chaque type de paramètres et de chaque modélisation acoustique. Le diagnostic des erreurs commises par les systèmes de reconnaissance afin de déterminer les causes est un domaine de recherche crucial (Kim et Rahim, 2004). La compréhension des raisons pour lesquelles telle ou telle erreur est commise permet de mettre en œuvre des méthodes visant à améliorer le système.

Notre travail consiste donc à exploiter diverses sources d'information acoustique (différents jeux de paramètres acoustiques) ainsi que plusieurs types de modélisations acoustiques : modèles de Markov cachés et réseaux de neurones et de les combiner de différentes manières. D'une part, la combinaison de différentes techniques d'analyse acoustique repose sur l'idée que certains jeux de paramètres capturent mieux certaines caractéristiques du signal de parole que d'autres et vice-versa. La diversité des paramètres acoustiques acoustiques utilisés a pour but de profiter des points forts de chacun d'entre eux pour améliorer les performances du système de reconnaissance. D'autre part, les modèles acoustiques ne partitionnent pas l'espace acoustique de la même manière. On peut donc présupposer que les erreurs ne seront pas commises pour les mêmes contextes phonétiques et que, si les modèles sont suffisamment complémentaires, les erreurs de l'un seront corrigées par l'autre et réciproquement.

Nous présentons également des architectures d'étude permettant de séparer l'influence des paramètres acoustiques de celle des modèles acoustiques sur le résultat de la reconnaissance. L'utilisation de mesures issues de la théorie de l'information nous permet de définir des contextes favorables ou défavorables au bon déroulement du processus de reconnaissance pour chaque système.

Organisation du document

Ce document est organisé en deux grandes parties. La première partie présente le contexte d'étude et les outils utilisés dans les systèmes de reconnaissance de la parole. La seconde correspond au travail réalisé durant cette thèse, à savoir, l'analyse et la combinaison de systèmes utilisant des paramètres et des modèles acoustiques différents.

Le premier chapitre présente les différents composants d'un système de reconnaissance de la parole, de la paramétrisation du signal audio jusqu'à la géné-

ration des hypothèses de mots, en passant par le modèle acoustique, le modèle de langage et le module de décodage.

Le deuxième chapitre détaille les différentes techniques d'analyse du signal de parole ainsi que les méthodes d'extraction de paramètres pertinents pour la reconnaissance de la parole. Un rapide survol des fondements de l'analyse du signal permet de comparer les paramètres cepstraux obtenus avec différents algorithmes d'extraction. Des techniques de filtrage permettant d'augmenter la robustesse des paramètres sont présentées. L'analyse en ondelettes multi-résolution ainsi que d'autres types de paramètres sont également présentés.

Dans le chapitre 3, une présentation des techniques « état de l'art » de combinaison de systèmes de reconnaissance de la parole est effectuée. Les différents types de combinaisons à tous les niveaux du système de reconnaissance sont présentés.

La première partie de mon travail est présentée dans le chapitre 4. Elle consiste à analyser et combiner des systèmes de RAP utilisant des jeux de paramètres et/ou des modèles acoustiques différents à un niveau segmental élevé (phonème ou mot). Différentes architectures de diagnostic permettant la comparaison et l'évaluation des systèmes sont présentées. Des techniques de combinaison aboutissant à une amélioration des performances ont été mises en œuvre à partir des résultats d'analyse obtenus.

La combinaison de systèmes au niveau de la trame est présentée dans le chapitre 5. Afin de combiner plusieurs modèles au niveau de la trame, il est important que les modèles soient synchrones afin de fusionner les probabilités qu'ils fournissent de manière cohérente. Pour cela, nous avons établi un protocole rapide et efficace qui permet de générer des modèles de même topologie, (*i.e.* faisant la même partition de l'espace acoustique). Nous avons ensuite introduit différents facteurs de pondération dans le processus de combinaison. Ces facteurs de pondération ont pour but d'estimer la qualité des modèles, et de modifier la combinaison en fonction de la confiance que l'on peut accorder à un modèle acoustique. La segmentation forcée (utilisée pour l'adaptation des modèles) a été analysée afin d'observer son influence sur le résultat de la reconnaissance. La combinaison est optimale lorsque les systèmes sont vraiment complémentaires, nous avons donc analysé le degré de similitude des systèmes par l'intermédiaire d'une mesure de concordance et observé le lien entre la concordance des système et les résultats obtenus.

Finalement, un résumé des points clés de la thèse est présenté ainsi que quelques perspectives pour les travaux de recherche futurs.

Première partie

La Reconnaissance Automatique de la Parole

Chapitre 1

La Reconnaissance Automatique de la Parole

Sommaire

La Reconnaissance Automatique de la Parole (RAP) s'inscrit dans le cadre général de la reconnaissance de formes. Les formes peuvent être de nature très différentes comme par exemple un signal vocal ou une image. On distingue deux approches principales pour la reconnaissance de formes. La première consiste à décrire les formes comme des structures organisées de formes élémentaires. Dans ce cas, les classes sont décrites à partir de règles utilisant des formes primitives préalablement déterminées par des experts. Cette approche a donné lieu à peu de réalisations pratiques car elle pose le problème de l'inférence des règles ainsi que de l'extraction des formes primitives (Haton et al., 2006).

La seconde approche consiste à extraire des vecteurs de paramètres caractéristiques des formes et à utiliser une technique de classification permettant d'attribuer une classe à une forme donnée. C'est dans ce cadre que se situe le décodage statistique de la parole.

Le décodage statistique de la parole se base sur le propos suivant (Jelinek, 1998) : trouver la suite de mots W la plus vraisemblablement prononcée dans le signal de parole X. Le problème se résume donc à trouver la suite de mot $\hat{W}$ satisfaisant l'équation suivante (Bahl et al., 1990) :

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|X) \tag{1.1}$$

D'après le théorème de Bayes, cette probabilité peut se ré-écrire de la manière suivante :

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{1.2}$$

On distingue trois termes qui s'expliquent comme suit :

- $P(X|W)$ est la probabilité d'observer le signal de parole X étant donné que la suite de mot prononcée est W. Cette probabilité est calculée par le modèle acoustique.
- P(W) est la probabilité *a priori* de la suite de mot W. Cette probabilité est calculée par le modèle de langage.
- P(X) est la probabilité d'observer le signal de parole X. Cette probabilité est identique pour chaque suite de mots (elle ne dépend pas de W). Elle n'est pas utile pour déterminer la meilleure suite de mots, et peut donc être ignorée.

Un système de RAP est constitué de plusieurs composants, comme décrit dans la figure 1.1. Le signal en entrée est d'abord traité par le module de paramétrisation (*front-end*) afin de produire des vecteurs de paramètres acoustiques représentatifs de l'information nécessaire à la reconnaissance de la parole (section 2). Différents types de modèles acoustique et différentes techniques d'apprentissage et d'adaptation seront décrites dans ce chapitre. Les modèles acoustiques permettent de calculer la probabilité qu'un symbole (mot, phonème, etc.)

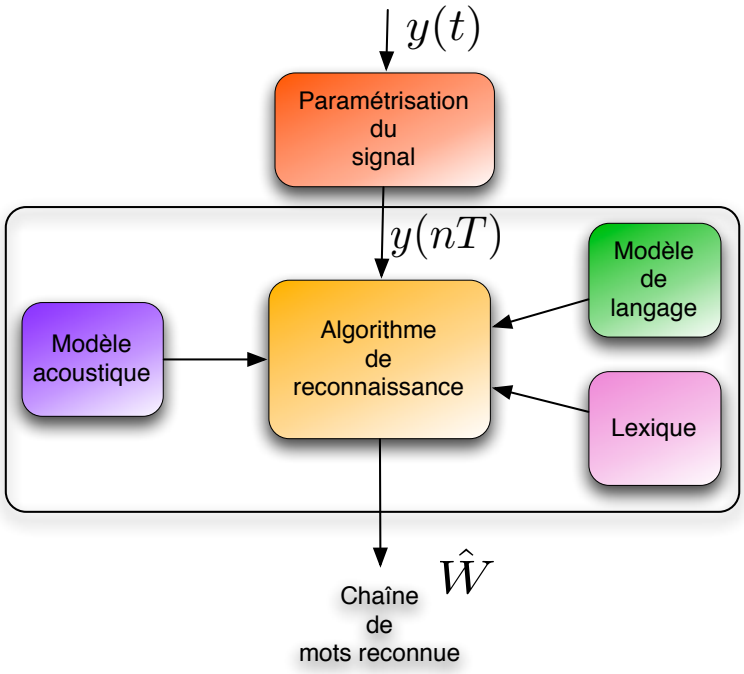

FIG. *1.1:* *Architecture d'un système de reconnaissance automatique de la parole.*

ait généré la trame de parole analysée. Les algorithmes permettant d'obtenir l'hypothèse finale du système de reconnaissance à partir des probabilités du modèle acoustique et du modèle de langage sont brièvement présentés en section 1.4).

1.1 Traitement du signal et paramètres acoustiques

Afin de pouvoir reconnaître le contenu d'un signal de parole correctement, il est nécessaire d'en extraire des paramètres caractéristiques et pertinents pour la reconnaissance. Pour ce faire, plusieurs techniques d'analyse du signal et d'extraction de paramètres peuvent être utilisées. L'extraction de paramètres acoustiques différents est un élément essentiel de cette thèse et, de ce fait, un chapitre leur est totalement dédié. J'invite donc le lecteur à se reporter au chapitre 2 pour de plus amples détails.

1.2 Les modèles acoustiques : modèles de Markov cachés

Les modèles acoustiques sont des modèles stochastiques qui sont utilisés conjointement à un modèle de langage afin de prendre des décisions quant-à la suite de mots contenue dans la phrase.

Le rôle du modèle acoustique est de calculer la probabilité qu'un événement linguistique (phonème, mot, ...) ait généré une séquence de vecteurs de paramètres extraits d'un signal de parole.

Quelques caractéristiques importantes des modèles acoustiques doivent être prises en compte. D'un point de vue utilisabilité, les modèles acoustiques doivent être robustes puisque les conditions acoustiques de la tâche de reconnaissance sont souvent différentes des conditions d'entraînement. En effet, le signal de parole possède de nombreuses variabilités qui ont pour conséquence d'augmenter la disparité entre la réalisation acoustique et le contenu linguistique. D'un point de vue pratique, les modèles acoustiques doivent être efficaces. Pour que leur utilisation soit acceptable, il est nécessaire qu'ils respectent certaines contraintes temporelles et donc proposer des temps de réponse relativement courts.

Les paramètres d'un modèle acoustique sont estimés à partir d'un corpus d'entraînement. Ce corpus d'entraînement est généralement transcrit manuellement. Cela permet d'identifier les segments de parole correspondant à chaque événement linguistique.

Actuellement, on distingue deux types de modèles acoustiques couramment utilisés : les modèles de Markov Cachés (*Hidden Markov Model* - HMM) utilisant des mixtures de gaussiennes (*Gaussian Mixture Models* - GMM, voir section 1.2.2), et les modèles hybrides HMM utilisant des réseaux de neurones (*Artificial Neural Network* - ANN, voir section 1.2.3). D'autres techniques (que je ne détaillerais pas) comme les machines à support vectoriel, ont récemment fait leur apparition.

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

Les sous-sections suivantes présentent les différents composants d'un HMM, ainsi que les techniques d'apprentissage et d'adaptation.

1.2.1 Structure d'un HMM

Un HMM (représenté dans la figure 1.2) est défini par :

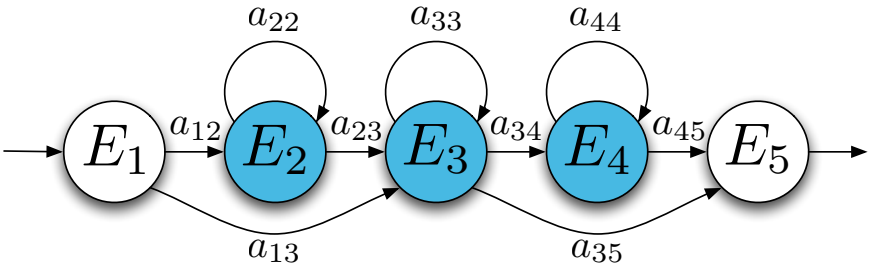

FIG. 1.2: *HMM à 5 états dont 3 émetteurs.*

- N : le nombre d'états composant le modèle.
- A : la matrice des probabilités a_{ij} de transition entre les états, de taille NxN. La somme des probabilités de transitions entre un état i et tous les autres états doit être égale à 1, *i.e.* $\forall i, \sum_{j=1}^{N} a_{ij} = 1$.
- π_i : la probabilité d'être dans l'état i à l'instant initial. La somme de ces probabilités doit également être égale à 1, *i.e.* $\sum_{i=1}^{N} \pi_i = 1$.
- b_i : la densité de probabilité de l'état i.

1.2.2 Les mixtures de gaussiennes

Un GMM est un mélange de distributions de probabilité qui suivent une loi gaussienne multivariée. Une fonction de densité de probabilité est estimée par la somme finie des gaussiennes composant le GMM.

La figure 1.3 montre un exemple de gaussienne bivariée (multivariée de dimension 2). Elle est définie par :

$$\mathcal{N}(\boldsymbol{\mu}, \Sigma, x) = \frac{1}{2\pi \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})} \tag{1.3}$$

avec $\boldsymbol{\mu}$ est la moyenne et Σ est la matrice de variance-covariance.

Les GMMs sont à la base des systèmes de reconnaissance HMMs les plus couramment utilisés. Une procédure itérative, détaillée dans la section 1.2.4, est utilisée pour leur apprentissage.

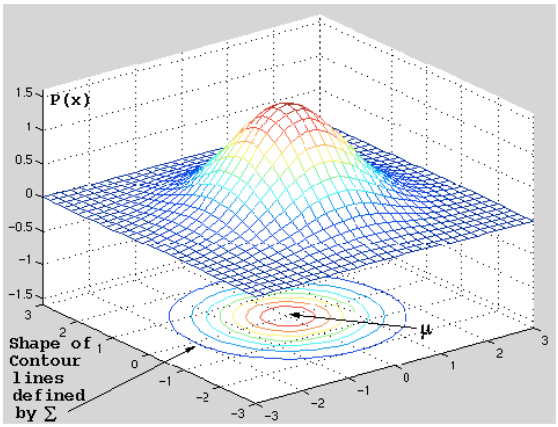

FIG. 1.3: *Exemple de densité de probabilité d'une gaussienne bivariée.*

1.2.3 Les réseaux de neurones artificiels

Les réseaux de neurones ou modèles neuromimétiques sont constitués de cellules élémentaires, appelées neurones, fortement connectées entre elles (Personnaz et Rivals, 2003). Ces neurones émettent en sortie une fonction non linéaire de la somme pondérée de leurs entrées. Les fonctions les plus utilisées sont les sigmoïdes ou les fonctions de Heavyside.

Une des formes les plus répandues de réseau de neurones est le perceptron multicouches. Un perceptron est un réseau sans contre-réaction, ce qui signifie que les sorties des neurones de la couche i forment les entrées des neurones de la couche i+1. La figure 1.4 montre un perceptron à trois couches dont une cachée, permettant de reconnaître N symboles (phonèmes ou autres).

L'ANN est alimenté avec des paramètres acoustiques dont il se chargera de trouver la combinaison optimale permettant d'obtenir la meilleure classification.

Les scores en sortie du réseau de neurones sont généralement normalisés par une fonction appelée « softmax » définie par :

$$pp_i = \frac{e^{S_i}}{\sum_{j=1}^{C} e^{S_j}} \tag{1.4}$$

Les probabilités pp_i peuvent être interprétées comme des probabilités *a posteriori* des classes (ou symboles) i connaissant le signal de parole d'entrée. Ces probabilités sont ensuite normalisées avec la probabilité *a priori* des symboles pour obtenir des vraisemblances normalisées (*scaled likelihoods*) pouvant être exploitées comme probabilités d'émission des états d'un HMM classique.

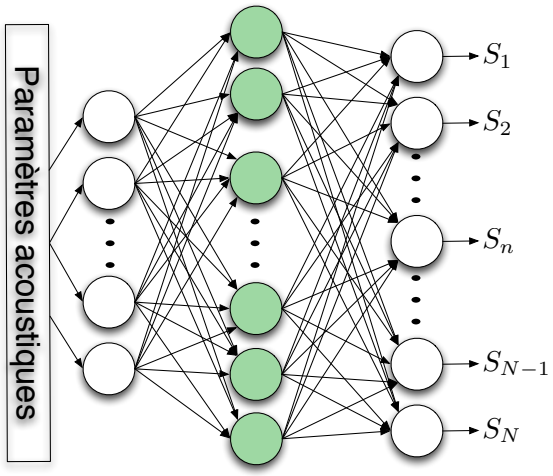

FIG. *1.4:* *Architecture d'un perceptron à trois couches dont une cachée (en gris) permettant de reconnaître N symboles.*

D'un point de vue philosophique, les ANNs sont très différents des GMMs dans le sens où ils n'estiment pas uniquement la probabilité d'un symbole indépendamment des autres. Ils intègrent plutôt une approche discriminante qui vise non seulement à donner une grande probabilité pour le symbole réellement émis, mais aussi à donner une faible probabilité aux autres classes. L'apprentissage des ANNs se fait par un processus itératif dit de « rétropropagation du gradient d'erreur » qui modifie les paramètres des neurones des couches cachées en fonction d'un critère (par exemple les moindres carrés).

Les expériences de la première partie de nos travaux utilisent un système hybride HMM/ANN acceptant deux types de paramètres acoustiques en entrée comme décrit dans Gemello et al. (1999).

1.2.4 Apprentissage d'un HMM

Afin d'utiliser un HMM pour la reconnaissance de la parole, il est nécessaire d'entraîner un modèle, ce qui signifie déterminer les paramètres optimaux des modèles de phonèmes constituant le HMM. Chaque état correspond à un GMM dont on doit estimer les paramètres (moyenne et matrice de covariance). Ceci est fait grâce à un corpus d'apprentissage dont on connaît la suite de mots prononcée pour chaque phrase.

L'apprentissage du HMM consiste à déterminer les paramètres $\widehat{\Theta} = \{N, A, \{\pi_i\}, \{b_i\}\}$ optimaux selon un critère de maximum de vraisemblance (*Maximum Likelihood* - ML). Le critère de ML correspond à trouver le

$\widehat{\Theta}$ qui maximise la fonction de vraisemblance comme suit :

$$\widehat{\Theta}_{ML}(Y) = \underset{\Theta}{\operatorname{argmax}}\, f(Y|\Theta) \tag{1.5}$$

Y étant l'ensemble des données d'apprentissage.

La complexité du problème d'optimisation est grande en raison des données incomplètes Y à notre disposition. Les données du corpus d'entraînement sont appelées données incomplètes car elles ne contiennent pas l'information concernant le GMM qui les a générées. Il est donc nécessaire d'utiliser l'approche itérative « Expectation and Maximisation » (**EM**), présentée dans Rabiner (1989) permettant de converger vers le modèle optimal. Généralement, le processus est interrompu lorsque le gain en vraisemblance est inférieur à un seuil prédéfini, ou lorsque le nombre d'itérations désiré a été atteint.

Comme son nom l'indique l'approche **EM** comporte deux phases :
- **E ou *expectation*** : dans cette étape, il faut compléter les données incomplètes Y en leur attribuant des données manquantes en fonction du modèle courant Θ^i.
- **M ou *maximisation*** : trouver le nouvel ensemble de paramètres Θ^{i+1} qui maximise la vraisemblance des données complètes Z_i connaissant le modèle Θ^i :

$$\Theta^{i+1} = \underset{\Theta}{\operatorname{argmax}}\, P(Z_i|\Theta) \tag{1.6}$$

Cet algorithme nécessite un modèle initial Θ^0, à partir duquel on effectue plusieurs itérations des étapes **E** et **M**. À partir de ce modèle initial, **EM** converge vers un maximum local, il en découle que le choix des paramètres du modèle initial va influencer grandement la convergence de l'algorithme.

Dans la pratique, on utilise un système de reconnaissance existant pour effectuer l'alignement forcé par rapport aux états (avec l'algorithme de Viterbi par exemple). Á partir de cette segmentation forcée, on applique l'algorithme **EM** au niveau de chaque état pour déterminer les paramètres du GMM correspondant.

Plusieurs solutions sont possibles pour initialiser chaque GMM. La première consiste à estimer une première gaussienne (moyenne et variance). Ensuite, deux gaussiennes sont créées à partir de la première en faisant varier la moyenne de $\pm\epsilon$ (déterminé en fonction de la variance de chaque paramètres). Puis on réestime les paramètres de chaque gaussienne avec les données qui ont le plus de vraisemblance avec elle. On réitère plusieurs fois pour atteindre le nombre de gaussiennes désiré. Une autre solution est d'utiliser l'algorithme des k-means (ou k-moyennes). Cet algorithme a pour but de partager l'espace en k parties en essayant de trouver les centres naturels de ces parties. L'objectif

est double : minimiser la variance intra-classe ou l'erreur quadratique[1] tout en maximisant la variance inter-classe.

Une fois le modèle Θ^0 déterminé, s'ensuivent plusieurs itérations d'**EM**. Les différents algorithmes se différencient par la manière de compléter les données incomplètes (étape **E**). Dans Baum et al. (1970), on cherche à trouver le modèle Θ qui maximise la log-vraisemblance $L(Z, \Theta)$ des données complètes étant donné le modèle courant Θ^i.

$$\Theta^{i+1} = \underset{\Theta}{\operatorname{argmax}}\, E\left[\log(L(Z,\Theta))|\Theta^i\right] \tag{1.7}$$

Cet algorithme utilise une procédure «avant-arrière» pour affecter les valeurs aux données manquantes et compléter les données incomplètes, d'où son autre nom de *Forward-Backward* Une autre manière de compléter les données incomplètes consiste à leur affecter leur valeur la plus probable comme dans l'algorithme de Viterbi (Forney, 1973). Le lecteur pourra se référer à Baum et al. (1966); Rabiner (1989); Rabiner et Juang (1993) pour les informations complémentaires à ce sujet.

1.2.5 Limitations des HMMs

Les modèles de Markov cachés reposent sur un ensemble d'hypothèses simplificatrices. Tout d'abord, les données à l'entrée d'un HMM sont supposées être statistiquement indépendantes, et donc la probabilité qu'un vecteur soit émis au temps t ne dépend pas des vecteurs précédemment émis. Cette hypothèse est irréaliste. En effet, les vecteurs de paramètres acoustiques sont calculés sur des portions de signal d'une durée très petite (en général 30 ms.), il est donc incorrect de penser que deux trames successives ne possèdent aucune corrélation statistique. L'utilisation des dérivées premières et secondes des paramètres acoustiques comblent en partie cette imprécision, mais à l'heure actuelle, les systèmes n'intègrent pas complètement la corrélation entre les trames de manière efficace. Les modèles segmentaux comme présentés dans Russell (1993) sont des modèles tentant de prendre en compte cette dépendance de manière intrinsèque lors des calculs des probabilités. Cependant, ils n'ont jamais montré de gain en performance probant et irrévocable au point de remplacer les modèles classiques.

Une autre limitation réside dans la modélisation de la durée, qui est implicite dans un HMM. Elle est déterminée par le critère visant a maximiser la pro-

[1] L'erreur quadratique E est définie par : $E = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2$, où μ_i est le centroïde des points x appartenant à la $i^{\text{ème}}$ classe S_i (au nombre de k).

babilité *a posteriori*. L'utilisation de modèles de trajectoire permettant de rendre compte de l'évolution temporelle du signal de parole à été proposé par Gong et Haton (1994).

L'utilisation de HMMs de premier ordre[2] repose sur l'hypothèse que la parole est également un processus de Markov de premier ordre. Des modèles d'ordre supérieurs ont été considérés (Mari et Haton, 1994; Mari et al., 1996), mais le compromis entre coût de calcul et gain en performance n'est pas évident.

1.3 Les modèles de langage

Le langage est la faculté de mettre en œuvre un système de symboles linguistiques (qui constituent la langue) permettant la communication et l'expression de la pensée. Cette faculté peut être mise en œuvre, notamment, par des moyens vocaux (parole), graphiques (écriture) et/ou gestuels (langue des signes).

Les modèles de langages ont pour objectif de décrire un langage. Deux types de modèles sont principalement utilisés. Les premiers sont à base de grammaires formelles mises au point par des experts en linguistique. Les autres sont des modèles stochastiques qui utilisent un corpus pour estimer des probabilités d'une suite de mots d'un langage de manière automatique. La génération d'un ensemble de règles décrivant un langage est un processus long et difficile, c'est pourquoi les modèles probabilistes sont privilégiés dans les systèmes de reconnaissance automatique de la parole.

Les modèles de langage stochastiques les plus utilisés sont les modèles N-grammes permettant d'estimer calculer la probabilité *a priori* (voir équation 1.8) d'une suite de mots W_k.

$$\begin{aligned} \hat{P}(W_k) &= \prod_{i=1}^{k} P(w_i|h_i) \text{ avec} \\ h_i &= \{w_1, ..., w_{i-1}\} \text{ pour } i > 2 \\ h_i &= \{w_1\} \text{ pour } i = 2 \\ h_i &= \{\emptyset\} \text{ pour } i = 1 \end{aligned} \tag{1.8}$$

Cette probabilité est estimée à partir de la probabilité conditionnelle de chaque mot composant la séquence étant donné un historique de taille réduite et fixe. En général, un historique de taille 1 (bi-grammes) ou 2 (tri-grammes) est utilisé, mais on peut étendre celui-ci à 3 (quadri-grammes) ou plus si la taille

[2]Pour les HMMs d'ordre 1, la probabilité d'être dans un état ne dépend que de l'état précédent, et non de l'ensemble des états traversés.

des données d'apprentissage est suffisamment grande pour obtenir une bonne estimation des probabilités.

Dans le cas d'un modèle tri-gramme, l'équation 1.8 peut être réécrite de la manière suivante :

$$\hat{P}(W_k) = P(w_1) \cdot P(w_2|w_1) \cdot \prod_{i=3}^{k} P(w_i|w_{i-2}, w_{i-1}) \tag{1.9}$$

La probabilité conditionnelle d'un mot connaissant son historique est estimée par un critère de maximum de vraisemblance (Maximum Likelihood - ML) défini par :

$$p(w_i|w_{i-n+1}, \ldots, w_{i-1}) = \frac{\mathcal{O}(w_{i-n+1}, \ldots, w_{i-1}, w_i)}{\mathcal{O}(w_{i-n+1}, \ldots, w_{i-1})} \tag{1.10}$$

où $\mathcal{O}(w_j, \ldots, w_J)$ correspond au nombre d'occurrences de la séquence de mot $w_j, \ldots, w_J$ dans le corpus d'apprentissage.

Les limitations de ce genre de modèle viennent de la taille et de la diversité du corpus d'apprentissage. Toutes les séquences de mots n'apparaissent pas dans le corpus et par conséquent une probabilité nulle peut leur être attribuée, ce qui pose un problème pour le calcul de la probabilité de l'équation 1.2. La technique la plus usuelle est celle dite de repli (*back-off*) qui calcule la probabilité d'une séquence de mots non observée lors de l'apprentissage en utilisant un modèle de langage plus général.

1.4 Décodage de la parole avec un HMM

La reconnaissance de la parole utilisant des modèles stochastiques est effectuée par un décodage statistique qui consiste à sélectionner, parmi l'ensemble des événements linguistiques, celui qui correspond aux données observées avec la plus grande probabilité.

En reconnaissance automatique de la parole, une séquence d'observations correspond généralement à une phrase constituée de plusieurs mots qui sont eux-mêmes constitués de plusieurs phonèmes. Si les langues avaient un nombre limité de phrases possibles, alors on pourrait raisonnablement avoir un modèle pour chaque phrase. Ce n'est pas le cas dans la pratique, et on doit donc considérer une approche différente lorsque l'ensemble des événements linguistiques est grand. Typiquement, les événements linguistiques sont représentés en concaténant des unités plus petites dont l'ensemble est de taille raisonnable. Chacune

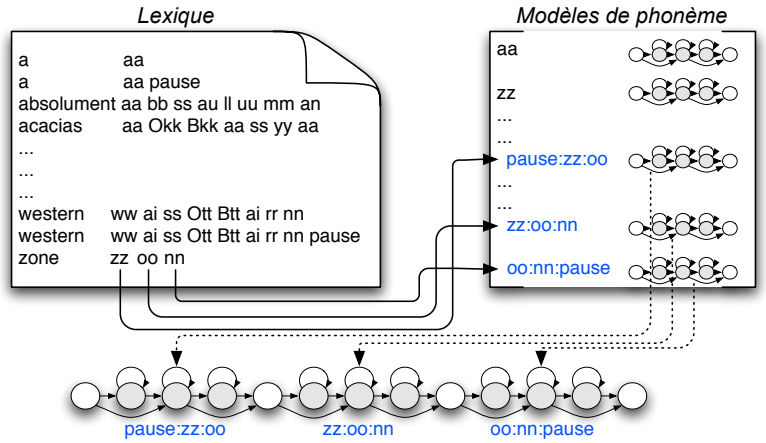

FIG. 1.5: *Concaténation de modèles de phonème afin d'obtenir un modèle du mot "zone". À noter qu'un même mot peut avoir différents modèles permettant de prendre en compte des variantes de prononciation.*

de ces unités aura son propre modèle permettant par assemblage d'obtenir des modèles pour les unités plus grandes.

En pratique, on utilise des modèles correspondant à des unités linguistiques élémentaires plus petites que le mot comme les phonèmes (plus rarement les syllabes). On obtient les modèles des mots en suivant les règles de composition des modèles de phonèmes décrites dans le lexique[3]. Ensuite, on forme les phrases en concaténant les modèles de mots. Dans le HMM, chaque unité est décomposée en états, correspondants à différentes parties de cette unité. En général, un modèle de phonème est composé de 3 états émetteurs servant à décrire la forme (*pattern*) du phonème à reconnaître. Pour la reconnaissance de la parole, on considère généralement que les phonèmes (contextuels ou non contextuels) possèdent trois parties, chacune correspondant à un état :

- un contexte gauche correspondant à la transition entre le phonème précédent et le phonème courant
- une partie centrale considérée comme stationnaire
- un contexte droit correspondant à la transition entre le phonème courant et le suivant.

Cette architecture convient parfaitement à la reconnaissance de la parole puisque la production de la parole est basée sur l'émission d'une séquence de sons basiques provenant d'une liste de taille limitée (les phonèmes). De plus, la structure d'un HMM permet une concaténation facile des modèles de phonèmes. En pratique, on modélise les phonèmes et/ou les transitions entre phonèmes que l'on compose pour former des mots et ensuite des phrases (voir fi-

[3]Le lexique est l'ensemble des règles permettant de décomposer un mot en sous-unités (phonèmes, phonèmes en contexte, syllabes).

gure 1.5).

À partir des probabilités générées par le modèle acoustique et des probabilités des séquences de mots obtenues avec le modèle de langage et le lexique, l'algorithme de décodage permet de produire la meilleure hypothèse de phrase. Certains systèmes proposent une liste des N-meilleures hypothèses voire même un treillis dont un exemple est présenté dans la figure 1.6. Différents types d'algorithme sont envisageables. Les plus célèbres sont l'algorithme de Viterbi (aussi appelé *Beam Search*) présenté dans Viterbi (1967) et l'algorithme $A*$ présenté entre autre dans Nocera et al. (2002). Il est important de souligner que ces algorithmes trouvent le chemin optimal permettant de maximiser la probabilité de la suite d'observations étant donné le modèle utilisé sans pour autant générer la totalité du treillis, ce qui serait trop coûteux en temps et en mémoire.

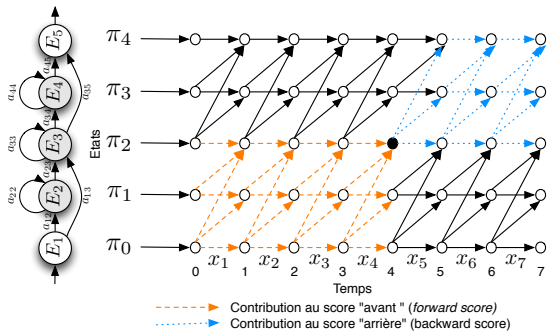

FIG. 1.6: *Treillis états/temps. Les colonnes du treillis correspondent aux états parcourus à un temps donné. Les arcs, correspondant aux transitions entre états, sont pondérés par le produit de la probabilité de transition et la probabilité d'émission calculée avec la distribution associée à l'état.*

1.5 Adaptation des modèles acoustiques

L'apprentissage des modèles acoustiques est effectué avec un corpus d'entraînement qui, souvent, ne correspond pas à la tâche pour laquelle le modèle est construit. Cela est en partie dû au fait que la construction d'un corpus d'apprentissage est un processus long et fastidieux étant donné qu'il faut retranscrire et aligner manuellement chaque phrase. De plus, les paramètres acoustiques varient en fonction de l'environnement d'enregistrement. Il est donc utile de prendre en compte ces différences en modifiant les paramètres des modèles acoustiques. Afin de pallier ce genre de problème, des techniques d'adaptation des modèles acoustiques ont été développées. Elles consistent à modifier les paramètres du modèle acoustique afin qu'ils correspondent à une tâche détermi-

née. Par exemple, un modèle adapté à la voix d'un homme (ou d'une femme) fournira de meilleures performances pour reconnaître un message vocal prononcé par un homme (ou une femme).

1.5.1 Maximum *a Posteriori*

L'adaptation Maximum *a Posteriori* (MAP) proposée par Gauvain et Lee (1994), est un processus permettant de modifier les paramètres d'un modèle acoustique initial afin de le rapprocher d'un corpus de données spécifique. Les données de ce corpus sont généralement plus proches des données de test que celles du corpus d'apprentissage, ce qui permet d'augmenter les performances du modèle sans pour autant relancer tout le processus d'entraînement depuis le début.

L'estimation par maximum *a posteriori* consiste à utiliser des probabilités *a priori* dans une approche par maximum de vraisemblance (*Maximum Likelihood Expectation* - MLE). Soit θ, les paramètres du HMM que l'on doit estimer à partir des observations x avec la fonction de densité de probabilité (*probability density function* - p.d.f.) $f(x|\theta)$). Si on considère que θ a une p.d.f. *a priori* g, la p.d.f. *a posteriori* θ_{MAP} est définie comme suit :

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, g(\theta|x) \tag{1.11}$$

$$= \underset{\theta}{\operatorname{argmax}}\, f(x|\theta)g(\theta) \tag{1.12}$$

La méthode MAP estime donc le θ comme le mode (valeur qui maximise la vraisemblance) de la p.d.f *a posteriori* :

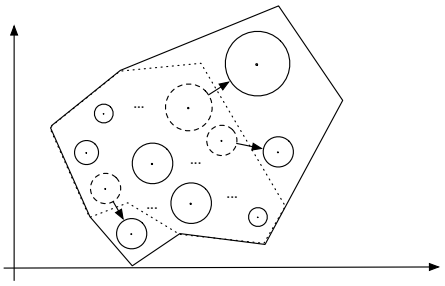

FIG. 1.7: *Adaptation par maximum a posteriori - MAP.*

Les systèmes de reconnaissance de la parole tirent de gros bénéfices de l'adaptation MAP. Dans Gauvain et Lee (1994), une amélioration très significative des performances a été obtenue en comparaison à une approche basée sur le maximum de vraisemblance.

Cependant, cette technique comporte quelques inconvénients qu'il est utile de préciser. Tout d'abord, elle nécessite une quantité de données diversifiées suffisamment importante pour contenir des observations concernant chaque état du modèle acoustique. En effet, cette technique modifie localement toutes les gaussiennes des GMMs présents dans les états du HMM, comme représenté dans la figure 1.7 (Bellot, 2006). L'amélioration attendue ne sera pas aussi grande si certaines composantes du modèle ne sont pas adaptées. Cela peut être le cas si le corpus d'adaptation n'est pas suffisamment grand ou s'il est incomplet.

1.5.2 Régression linéaire

L'adaptation par régression linéaire MLLR consiste à estimer une transformation qui modélise les différences, supposées linéaires, entre les conditions d'entraînement et celles de test (voir figure 1.8 (Bellot, 2006)). L'estimation de

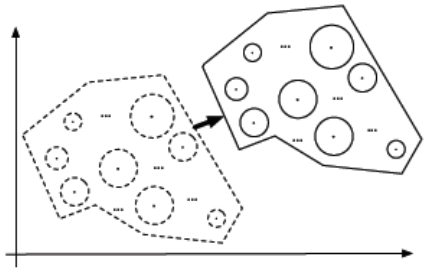

FIG. *1.8:* *Adaptation par régression linéaire à maximum de vraisemblance - MLLR.*

la transformation s'effectue au sens du critère de maximum de vraisemblance. Contrairement à l'adaptation MAP, l'ensemble des données d'adaptation est utilisé pour modifier l'ensemble des paramètres du modèle sans distinction et dans une seule et même direction . De ce fait, elle nécessite beaucoup moins de données d'adaptation que MAP. Des classes des régression peuvent être définies, comme par exemple le regroupement des fricatives, afin d'affiner la transformation et ainsi améliorer les performances de l'adaptation.

1.6 Mesures d'évaluation

En reconnaissance automatique de la parole, la mesure d'évaluation la plus répandue est le taux d'erreur mot (*Word Error Rate* - WER). Le WER consiste à comparer la phrase reconnue et la phrase de référence (celle qui a effectivement

été prononcée). Il est défini comme suit :

$$WER = \frac{I + D + S}{W} * 100 \tag{1.13}$$

avec I le nombre d'insertions, D le nombre de suppressions, S le nombre de substitutions et W le nombre de mots dans la référence.

Une autre mesure d'évaluation que nous utiliserons est l'entropie conditionnelle moyenne ou *"equivocation"* (Shannon, 1948) apportée par le système. En considérant le système comme un canal de transmission ayant une source S, produisant des symboles f à son entrée, et un récepteur R, recevant des symboles g en sortie, alors l'equivocation $H_R(S)$ est définie par :

$$H_R(S) = -\sum_{f,g} P(f,g) \log_2 P(g|f) \tag{1.14}$$

Cette mesure permet d'évaluer le degré de confusion apporté par le système de reconnaissance.

1.7 Conclusion

Dans ce chapitre, les principes de fonctionnement et les différents constituants d'un système de reconnaissance automatique de la parole ont été présentés.

Le signal de parole est sujet à des perturbations qui provoquent des variabilités dans les paramètres acoustiques. Il en résulte plusieurs problèmes. Deux trames d'observations acoustiques semblables peuvent représenter des symboles différents, et donc induire le classifieur en erreur. De plus, les modèles acoustiques sont générés à partir de ces paramètres acoustiques variables. La modélisation de l'espace acoustique est optimisée de manière globale, et la répartition des symboles dans l'espace acoustique ne sera donc pas forcément correcte.

Les méthodes visant à augmenter la robustesse des modules de paramétrisation et de modélisation acoustique permettent de réduire ou de compenser les variabilités sans toutefois atteindre les performances escomptées pour que le système soit utilisable de manière naturelle.

Le travail réalisé dans cette thèse consiste à utiliser diverses sources d'informations, matérialisées par différents jeux de paramètres acoustiques, afin de les combiner à plusieurs niveau. Une grande partie du travail réalisé dans cette thèse consiste à analyser les différences entre deux jeux de paramètres acoustiques et à exploiter leur complémentarité lors de la combinaison. Ces jeux de

paramètres sont issus de méthodes d'analyse fondées sur différentes approches afin que les caractéristiques du signal de parole extraites soient les plus hétérogènes possible, *a priori*. En outre, certains paramètres intègrent des méthodes permettant d'augmenter leur robustesse face au bruit environnant.

Les modèles acoustiques utilisés dans les expériences sont de deux types : modèles hybrides HMM/ANN et modèles HMM/GMM. Des techniques d'adaptation ont été employées afin d'augmenter leur robustesse.

Dans la partie II, plusieurs architectures exploitant différents jeux de paramètres sont présentées. Le diagnostic effectué sur les sorties des systèmes permettent d'identifier des zones dans lesquelles un jeu de paramètres est meilleur que l'autre et d'adapter la combinaison pour privilégier le type de paramètres ayant la plus grande confiance dans un contexte donné.

Le premier type de combinaison présenté dans le chapitre 4 consiste à exploiter les sorties des systèmes de reconnaissance de la parole utilisant le même type de modélisation mais des paramètres acoustiques différents. Ces systèmes sont ensuite combinés à un niveau segmental plus ou moins élevé (mot ou phonème).

Le second type de combinaison consiste à combiner les probabilités *a posteriori* générées par des modèles acoustiques utilisant des paramètres différents au niveau de la trame (voir chapitre 5).

Le chapitre suivant présente les différentes manières de traiter le signal de parole dans le but de générer des paramètres acoustiques robustes exploitables dans un système de reconnaissance de la parole.

Chapitre 2

Les paramètres acoustiques

Sommaire

Le signal de parole est trop redondant et variable pour être utilisé directement dans un système de reconnaissance automatique de la parole. Il est donc nécessaire d'en extraire l'information pertinente afin de caractériser et d'identifier le contenu linguistique. Le signal de parole est représenté, en général, dans le domaine fréquentiel montrant l'évolution temporelle de son spectre. Ce domaine est approprié pour la reconnaissance puisque l'on peut raisonnablement considérer que les propriétés du spectre restent stationnaires durant des intervalles de temps d'environ une dizaine de ms (valeur adoptée de manière classique).

Les systèmes de reconnaissance intègrent un module de paramétrisation dont le rôle est de créer des vecteurs de paramètres acoustiques résultant de l'analyse spectrale du signal de parole. La plupart des techniques de paramétrisation consistent à décrire l'enveloppe du spectre à court terme dans le domaine fréquentiel. D'autres techniques peuvent être utilisées comme l'analyse en ondelette.

2.1 Coefficients cepstraux de prédiction linéaire

La prédiction linéaire est une technique issue de l'analyse de la production de la parole permettant d'obtenir des coefficients de prédiction linéaire (*Linear Prediction Coefficients* - LPC). Des paramètres cepstraux LPCC (*Linear Prediction Cepstral Coefficients*) sont ensuite calculés à partir de ces coefficients.

Dans ce cadre d'analyse, le signal de parole s est considéré comme la conséquence de l'excitation du conduit vocal par un signal provenant des cordes vocales. La prédiction s'appuie sur le fait que les échantillons de parole adjacents sont fortement corrélés, et que par conséquent, l'échantillon s_n peut être estimé en fonction des p échantillons précédents.

Par prédiction linéaire, on obtient donc une estimation du signal :

$$\hat{s}_n = \sum_{i=1}^{p} a_i \, s_{n-i} \tag{2.1}$$

où les a_i sont des coefficients constants sur une fenêtre d'analyse.

La définition devient exacte si on inclut un terme d'excitation :

$$s_n = \sum_{i=1}^{p} a_i \, s_{n-i} + G \, u_n \tag{2.2}$$

où u est le signal d'excitation normalisé et G le gain de l'excitation. La transformée en Z de cette égalité donne :

$$GU(z) = \left(1 - \sum_{i=1}^{p} a_i \, z^{-i}\right) S(z) \tag{2.3}$$

d'où :

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i \, z^{-i}} = \frac{G}{A(z)} \tag{2.4}$$

L'équation 2.4 peut être interprétée comme suit : le signal s est le résultat de l'excitation du filtre tout pôle $H(z) = \dfrac{G}{A(z)}$ par le signal d'excitation u.
Les coefficients a_i sont les coefficients qui minimisent l'erreur quadratique moyenne :

$$E_n = \sum_m e_{n+m}^2 = \sum_m \left[s_{n+m} - \sum_{i=1}^{p} a_i \, s_{n+m-i} \right]^2 \tag{2.5}$$

À partir de ces échantillons prédis, on peut calculer les paramètres cepstraux comme définis dans Markel et Gray Jr. (1976). Le cepstre est le résultat de la transformée de Fourier inverse appliquée au logarithme de la transformée de Fourier du signal de parole. Les paramètres cepstraux c_i sont les coefficients du développement de Taylor du logarithme du filtre tout pôle :

$$\ln\left[\frac{G^2}{|A(z)|^2}\right] = \sum_{-\infty}^{+\infty} c_i \, z^{-i} \tag{2.6}$$

ce qui donne :

$$\begin{cases} c_0 = \ln(G^2) \\ c_m = a_m + \sum_{k=i-p}^{i-1} \frac{k}{i} \, c_k \, a_{i-k} \text{ avec } i = p, \ldots, N_c \\ c_{-m} = c_m \end{cases}$$

Les paramètres cepstraux ont l'avantage d'être peu corrélés entre eux. Cela permet d'utiliser des matrices de covariances diagonales pour leur moment de second ordre, et ainsi gagner beaucoup de temps lors du décodage. Les différentes étapes de l'analyse LPCC sont détaillées dans la figure 2.1.

Comme dit précédemment, ce modèle provient de l'analyse de la production de la parole. D'autres formes d'analyses qui tiennent compte du mode de perception auditive de la parole plutôt que du mode de production sont présentées dans les sections suivantes.

2.2 L'analyse en banc de filtres

L'analyse par banc de filtres est une technique initialement utilisée pour le codage du signal de parole. Elle produit des paramètres cepstraux «*Mel-Frequency Cepstral Coefficients*» (MFCC). Le signal de parole est analysé à l'aide de filtres passe-bande permettant d'estimer l'enveloppe spectrale en calculant l'énergie dans les bandes de fréquences considérées.

Les bandes de fréquences des filtres sont espacées logarithmiquement selon une échelle perceptive afin de simuler le fonctionnement du système auditif humain. Les échelles perceptives les plus utilisées sont celles de Mel et de Bark. Plus la fréquence centrale du filtre est basse, plus la bande passante du filtre est étroite. Augmenter la résolution pour les basses fréquences permet d'extraire plus d'information dans ces zones où elle est plus dense.

Il est possible d'utiliser directement les coefficients obtenus à la sortie des filtres pour la reconnaissance de la parole, cependant, d'autres coefficients plus

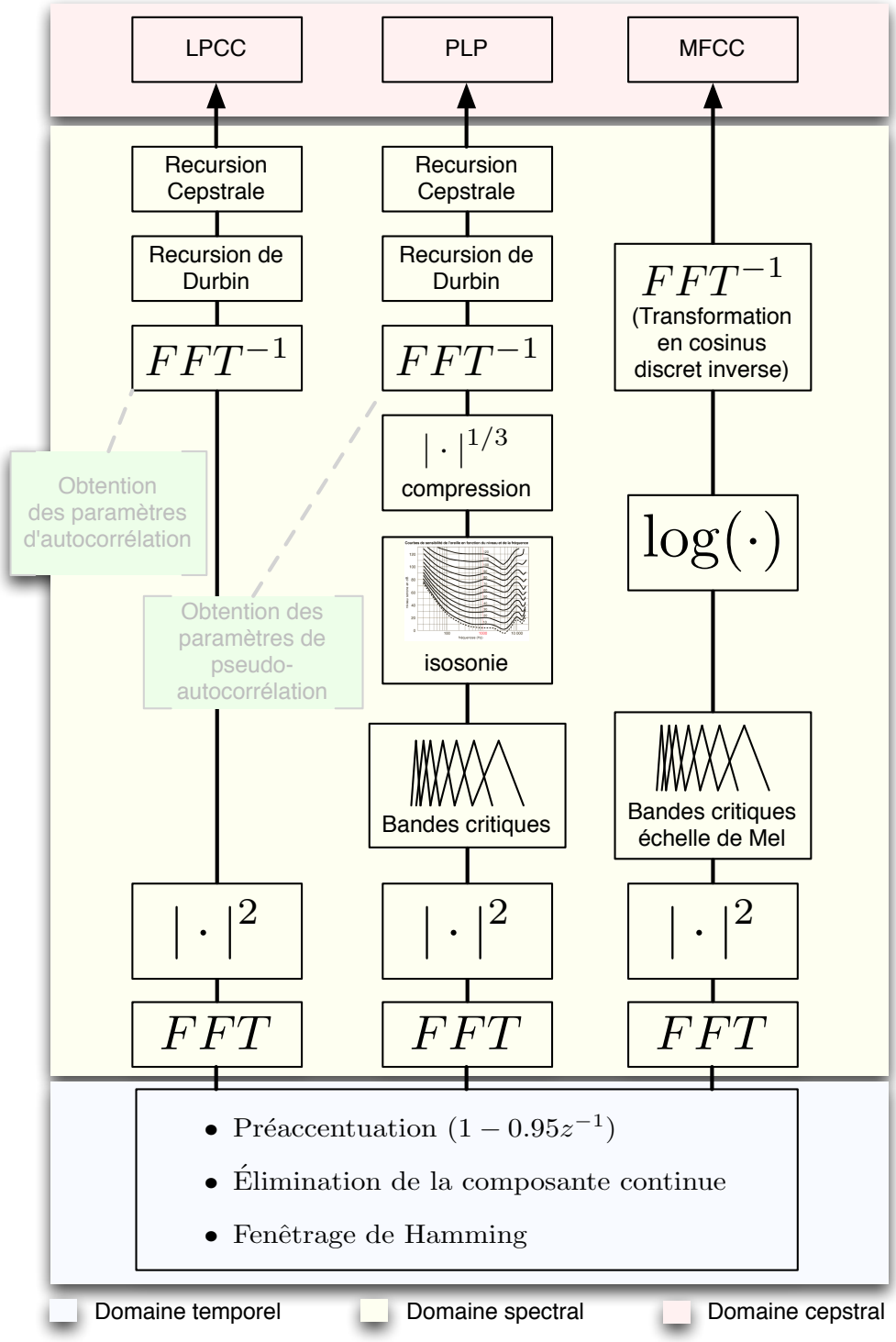

FIG. 2.1: *Comparaison de trois analyses du signal : LPCC, PLP et MFCC.*

discriminants, plus robustes au bruit ambiant et surtout décorrélés entre eux sont préférés : les coefficients cepstraux. Un ensemble de M coefficients cepstraux, généralement entre 10 et 15, sont calculés en effectuant un liftrage (filtrage dans le domaine cepstral) du spectre en puissance d'un signal selon la transformée en cosinus discret (*Discrete Cosinus Transform* - DCT) :

$$c_i = \sum_{m=0}^{M} S_m \cos\left(\frac{\pi}{N_f}\,(m+\frac{1}{2})\,i\right) \text{ pour } i = 0, ..., M-1 \tag{2.7}$$

où N_f est le nombre de filtres utilisé.

Le coefficient c_0 correspond à l'énergie moyenne de la trame. De manière générale, on ne le prend pas en compte afin de rendre les MFCC peu sensibles à la puissance acoustique du signal de parole.

Les différentes étapes de l'analyse MFCC sont détaillées dans la figure 2.1.

2.3 Analyse par prédiction linéaire perceptuelle

L'analyse par Prédiction Linéaire Perceptuelle (*Perceptual Linear Prediction* - PLP) repose sur un modèle de perception de la parole. Les différentes étapes de l'analyse PLP sont détaillées dans la figure 2.1.

Elle est basée sur le même principe que l'analyse prédictive et intègre trois caractéristiques de la perception (Hermansky, 1990) :

1. **Intégration des bandes critiques** : la prédiction linéaire produit la même estimation de l'enveloppe spectrale pour toute la zone de fréquences utiles, ce qui est en contradiction avec le fonctionnement de l'appareil perceptif humain. En effet, l'oreille humaine a la faculté d'intégrer certaines zones de fréquences en bande appelées bandes critiques. Les bandes critiques sont réparties selon l'échelle de Bark, dont la relation avec la fréquence est définie par :

   $$f = 600\ sinh(z/6) \tag{2.8}$$

   avec f la fréquence en Hertz et z la fréquence en Bark. La nouvelle densité spectrale est échantillonnée selon cette nouvelle échelle, ce qui augmente la résolution pour les basses fréquences.

2. **Préaccentuation pas courbe d'isosonie** : cette caractéristique provient de la psychoaccoustique qui a montré que l'intensité sonore d'un son pur perçue par l'appareil auditif varie avec la fréquence de ce son. Ainsi, dans

l'analyse PLP, afin de prendre en compte la manière dont l'appareil auditif perçoit les sons, la densité spectrale doit être multipliée par une fonction de pondération non linéaire. Cette fonction peut être estimée en utilisant l'abaque sur laquelle sont reportées les lignes isosoniques (figure 2.2). Ces lignes correspondent à la trajectoire d'égale intensité sonore pour différentes fréquences d'un son pur. En pratique, cette préaccentuation est remplacée par l'application du filtre passe-haut dont la transformée en Z est $(1 - 0.95z^{-1})$.

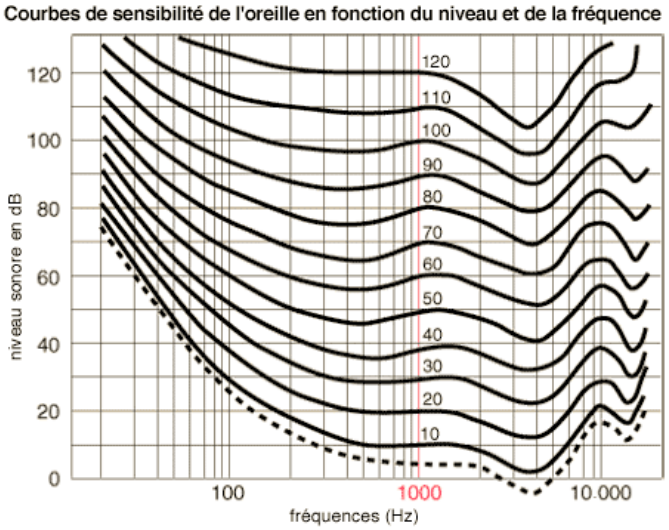

FIG. 2.2: *Abaque des courbes d'égale intensité sonore (Stevens, 1957).*

3. **Loi de Stevens** : l'intégration des bandes critiques et la préaccentutation ne suffisent pas à faire correspondre l'intensité mesurée et l'intensité subjective (appelée sonie). La loi de Stevens donne la relation entre ces deux mesures :

$$\text{sonie} = (\text{intensité})^{0.33} \tag{2.9}$$

Les PLP sont basés sur le spectre à court terme du signal de parole, comme les coefficients LPC. Cela signifie que le signal est analysé sur une fenêtre glissante de courte durée. En général, on utilise une fenêtre de longueur 10 à 30 ms. que l'on décale de 10 ms pour chaque trame.

2.4 RASTA PLP et J-RASTA PLP

Afin d'augmenter la robustesse des paramètre PLP, on peut envisager l'analyse spectrale relative RASTA (*RelAtive SpecTrAl*), présentée par Hermansky et

Morgan (1994) comme une façon de simuler l'insensibilité de l'appareil auditif humain aux stimuli à variation temporelle lente. Cette technique traite les composantes de parole non linguistiques, qui varient lentement dans le temps, dues au bruit convolutif (log-RASTA) et au bruit additif (J-RASTA). En pratique, RASTA effectue un filtrage passe-bande sur le spectre logarithmique ou sur le spectre compressé par une fonction non linéaire. L'idée principale est de supprimer les facteurs constants dans chaque composante du spectre à court-terme avant l'estimation du modèle tout-pôle. L'analyse RASTA est souvent utilisée en combinaison avec les paramètres PLP (Hermansky et Morgan, 1994). Les étapes d'une analyse RASTA-PLP sont décrites dans la figure 2.3. Les étapes

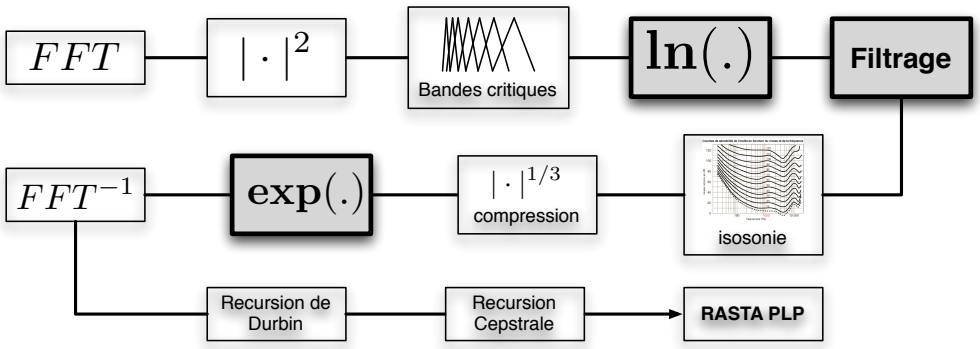

FIG. 2.3: *Analyse RASTA PLP.*

grisées sont celles qui font la spécificité du traitement RASTA. La différence entre RASTA et J-RASTA se situe au niveau du logarithme ($4^{\text{ème}}$ étape) : $ln(x)$ pour RASTA et $ln(1 + Jx)$ pour J-RASTA.

2.5 Analyse à résolution multiple

L'analyse à résolution multiple (*Multi Resolution Analysis* - MRA), décrite dans Perogaro (2000) et Gemello et al. (2006), effectue une analyse en ondelettes d'une fenêtre de signal audio. Cela consiste à faire passer le signal dans un arbre de filtres passe-bas et passe-haut, à la sortie desquels l'énergie à court terme est calculée (voir figure 2.4). À chaque niveau de l'arbre, le signal est entièrement décrit, mais dans une résolution fréquentielle et temporelle différente.

Comme on peut le constater, la disposition des filtres n'est pas intuitive, car il faut prendre en compte le phénomène de repliement spectral qui recopie dans les basses fréquences le signal haute fréquence inversé. Ensuite, il faut regrouper les énergies calculées aux feuilles de l'arbre pour former les trames qui seront utilisées dans le système de reconnaissance de la parole.

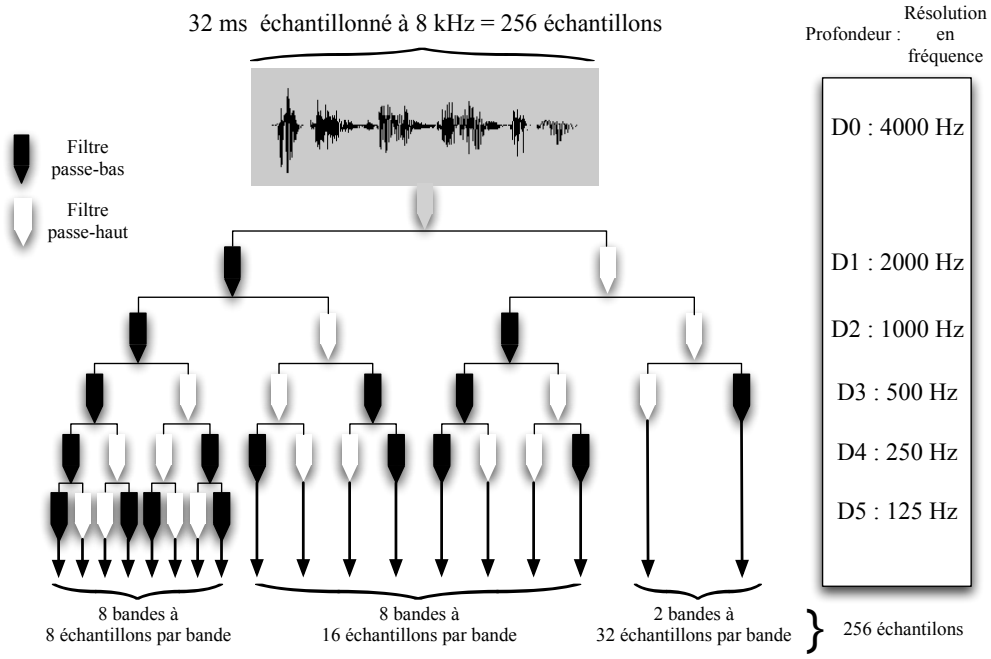

FIG. 2.4: *Principe de l'analyse à résolution multiple. Les filtres passe-haut sont en noir, les passe-bas en blanc. À chaque profondeur Dx, la résolution en fréquence est divisée par deux (cf. colonne de droite), mais la résolution temporelle est doublée.*

Considérons une fenêtre de taille N échantillons, qui se déplace de M échantillons. Pour MRA, les valeurs utilisées pour N sont 256 (32 ms) ou 384 (48 ms), et M est fixé à 80 échantillons (soit 10 ms). À noter que ce front-end a été développé pour des applications téléphoniques. Le nombre d'échantillons obtenus dans les noeuds de l'arbre diminue quand on descend dans l'arbre, mais l'intervalle temporel associé aux échantillons filtrés reste inchangé.

Selon le principe d'indétermination d'Heisenberg, il existe une relation entre la résolution temporelle et la résolution fréquentielle des échantillons dans les différentes sous-bandes. Sur la base de ce principe, le produit de la résolution en temps et celle en fréquence ne doit pas être inférieur à un certain seuil. Étant donné qu'à chaque niveau de l'arbre, la résolution fréquentielle est divisée par deux (cf. figure 2.4), on peut considérer des intervalles temporels d'intégration différents pour chaque niveau de l'arbre. Pour cela, on utilise l'extracteur de paramètres sur le même nombre d'échantillons à chaque niveau, ce qui a pour conséquence de diviser l'intervalle temporel par deux.

Pour les 8 premières bandes (de 0 à 1 kHz) on utilise les 8 échantillons disponibles. Pour les 8 bandes suivantes (de 1kHz à 3 kHz) on n'utilise que les 8 échantillons centraux sur les 16 disponibles. Enfin, pour les deux dernières bandes (de 3 kHz à 4 kHz) on utilise seulement 10 échantillons sur les 32 dispo-

nibles. Tout ceci est détaillé dans le tableau 2.1.

Niveau de l'arbre	Résolution en fréq. [Hz]	Intervalle temporel en ms (éch.)	
		32 (N=256)	48 (N=384)
1	4000	10 (80)	10 (80)
2	2000	10 (40)	10 (40)
3	1000	10 (20)	10 (20)
4	500	10 (10)	12 (12)
5	250	16 (8)	24 (12)
6	125	32 (8)	48 (12)

TAB. 2.1: *Correspondance entre résolution fréquencielle et temporelle pour l'analyse MRA.*

À la sortie de ces filtres, on doit appliquer une opération d'extraction de paramètres acoustiques sur les échantillons filtrés obtenus. Notons c_i les échantillons d'un noeud de l'arbre, et N leur nombre. Cette opération est appelée *intégration*. Les opérateurs disponibles pour l'intégration sont nombreux, les plus utilisés sont :

- L'énergie moyenne par échantillon :

$$E = \frac{1}{N}\sum_{i=1}^{N} c_i^2 \tag{2.10}$$

- La norme p :

$$X = \frac{1}{N}\sum_{i=1}^{N} |c_i|^p \text{ avec } p = 1, 2, 3. \tag{2.11}$$

- L'entropie moyenne :

$$H = \frac{1}{N}\sum_{i=1}^{N} c_i^2 \cdot \log c_i^2 \tag{2.12}$$

- L'opérateur *teager* :

$$T = \frac{1}{N-2}\sum_{i=2}^{N-1} (c_i^2 - c_{i-1} \cdot c_{i+1}) \tag{2.13}$$

- La dimension théorique (combinaison de l'entropie moyenne et de l'énergie moyenne) :

$$TD = E \cdot \exp^{-\frac{H}{E}} \tag{2.14}$$

Les paramètres MRA ont la particularité de ne pas décrire l'enveloppe spectrale du signal, mais plutôt de représenter le signal en terme d'énergie présente dans chaque bande de fréquences et d'utiliser la redondance de représentation de ce signal de parole à chaque niveau de l'arbre. L'intérêt de considérer de tels paramètres est qu'on peut supposer que l'information qu'ils contiennent sera différente de celle fournie par les représentations cepstrales.

2.6 Paramètres acoustiques *Tandem*

Les paramètres **tandem** (*Tandem Features*), tels que présentés dans Hermansky et al. (2000), sont calculés à partir de paramètres discriminants obtenus à l'aide d'un réseau de neurones.

Les systèmes de reconnaissance automatique de la parole utilisent en général des modèles à base de GMMs pour estimer les distributions de vecteurs de paramètres décorrélés qui correspondent à des unités acoustiques de courte durée (syllabes, phonèmes, phonèmes en contexte, ...). En comparaison, les systèmes hybrides ANN/HMM utilisent des réseaux de neurones entraînés de manière discriminante pour estimer les distributions de probabilité des unités étant donné les observations acoustiques.

L'approche **tandem** consiste à combiner des paramètres discriminants issus d'un réseau de neurones avec une modélisation des distributions par GMMs. Le réseau de neurones génère les probabilités postérieures des unités qui sont ensuite transformés pour être utilisés comme paramètres d'entrée pour le modèle HMM/GMM qui est alors appris de manière conventionnelle.

Les transformations sur les distributions de probabilité sont de différentes sortes. Les réseaux de neurones produisent directement des probabilités *a posteriori* contrairement aux mixtures de gaussiennes. Étant donné que les probabilités postérieures ont une distribution très biaisée, il est avantageux de les transformer en prenant leur logarithme par exemple. Une alternative à cela est d'omettre la dernière non-linéarité à la sortie du réseau de neurones. Cette non-linéarité, le *softmax*, correspond à normaliser les exponentiels (ce qui est très proche de prendre le logarithme des probabilités). Les vecteurs de probabilités postérieures ont tendance à posséder une valeur élevée, correspondant au phonème prononcé, et les autres basses.

Les réseaux de neurones n'ont pas la contrainte d'utiliser des paramètres acoustiques décorrélés comme les HMMs. Cependant, il s'avère que la transformation de Karhunen-Loeve, plus connue sous le nom d'analyse en composante principale (*Principal Component Analysis* - PCA) est utile pour décorréler les paramètres, vraisemblablement parce qu'elle augmente la correspondance entre les paramètres et les modèles à base de mixture de gaussiennes.

Les principaux résultats obtenus avec ce genre de technique sont présentés dans Hermansky et al. (2000) et Morgan et al. (2004).

2.7 Autres paramètres acoustiques

Beaucoup d'autres paramètres acoustiques ont été développés afin, le plus souvent, de compléter les paramètres existants. La plupart d'entre eux ne sont pas suffisant, lorsqu'ils sont utilisés seuls, pour créer des modèles acoustiques performants. Ainsi, dans Vaseghi et al. (1997), l'utilisation de caractéristiques modélisant les segments phonétiques de la parole avec des paramètres spectro-temporels multi-résolution est proposée. Ces paramètres de corrélation décrivent la trajectoire de la parole sur la durée d'une unité phonétique.

L'ajout de paramètres apportant de l'information différente a été considéré. Une caractéristique prosodique (le voisement) utilisée conjointement aux paramètres LPCC fournie une amélioration significative des résultats (Thomson et Chengalvarayan, 1998). Le paramètre de voisement est dérivé du signal temporel sous deux formes différentes : la périodicité (structure périodique du signal) et le *jitter* (petites fluctuations des cycles de la glotte). Des paramètres acoustiques représentant le voisement ont également été proposés dans Zolnay et al. (2002). Ces paramètres sont fondés sur l'analyse de la largeur et de la longueur des pics du spectre harmonique du signal de parole.

Dans Kamal Omar et Hasegawa-Jonhson (2002), plusieurs aspects du signal de parole sont considérés afin d'être sélectionnés pour former un nouveau vecteur d'observations. Ces caractéristiques comprennent : le voisement (voisé, non voisé, silence), la manière d'articulation (voyelle, nasale, fricative, stop, glide, silence), la position d'articulation (avant, latérale, basse, haute, arrière, ...) et la durée (tendue/strident, relachée/non strident, réduite/agitée). Elles sont toutes issues des traits distinctifs donnés par Stevens (1998). Ces traits phonologiques X sont sélectionnés selon un critère d'information mutuelle maximum avec les paramètres acoustiques Y (MFCC ou PLP) défini comme suit :

$$I(X,Y) = \int \sum_{i=1}^{N} P(y|x_i) \log \frac{P(y|x_i)}{P(y)} dy \tag{2.15}$$

où N correspond à la taille du vecteur de traits phonologiques, x_i à la $i^{\text{ème}}$ valeur de ce vecteur. $P(x_i)$ est calculée en utilisant le corpus d'entraînement et $P(y|x_i)$ est modélisée par une fonction de densité de probabilité dans un GMM.

D'autres techniques modifient le protocole de calcul de paramètres standards afin d'améliorer les paramètres. Dans Pujol et al. (2005), une technique de filtrage de fréquences a été employée pour décorréler les paramètres MFCC. Ce jeu de paramètres a montré de bonnes performances, seul ou en combinaison dans un système multi-flux avec les paramètres J-RASTAPLP, pour diverses tâches de reconnaissance plus ou moins bruitées. Dans Hariharan et al.

(2001), une approche multi-résolution et multi-bandes permet d'obtenir des paramètres acoustiques plus robustes au bruit.

Comme on peut le constater, ces autres paramètres prennent en compte des caractéristiques du signal de parole issues non pas du traitement du signal, mais surtout de contextes articulatoires ou prosodiques.

2.8 Conclusion

Dans ce chapitre, différentes manières d'extraire des paramètres acoustiques pertinents pour la reconnaissance de la parole ont été décrites. Ces techniques sont fondées sur des analyses du signal différentes comme l'analyse en ondelettes, l'analyse spectrale où la transformation des probabilités *a posteriori* issues d'un réseau de neurones. Le traitement RASTA peut être intégré à une analyse (le plus souvent PLP) pour augmenter la robustesse des paramètres au bruit. Ces paramètres peuvent être complétés par d'autres traits caractéristiques qui capturent une information différente (tel que le voisement).

Le fait que les paramètres acoustiques soient calculés de manières très différentes ne nous assure pas de la complémentarité des hypothèses que le décodeur permet de générer. Aussi, il est nécessaire d'analyser les forces et les faiblesses de chacun des jeux de paramètres et de comparer leurs performances en terme de reconnaissance. L'objectif de ces analyses est d'identifier les points forts de chacun pour les exploiter lorsqu'ils seront combinés (voir chapitre 3).

Deuxième partie

Combinaison de systèmes de RAP

Chapitre 3

Contexte d'étude et état de l'art

Sommaire

Les systèmes de reconnaissance automatique de la parole commettent des erreurs qui limitent le potentiel de leurs applications (Sarikaya et al., 2005). Les causes de ces erreurs sont multiples. D'une part, les modèles acoustiques sont construits de manière à minimiser globalement le WER sur tout l'espace acoustique. De ce fait, ils ne modélisent pas parfaitement la totalité de l'espace acoustique. En effet, certains paramètres peuvent souffrir du manque de données et être mal estimés. Ces imperfections présentes dans certaines zones mal modélisées peuvent aboutir à l'échec de la reconnaissance. D'autre part, les paramètres acoustiques obtenus par une analyse du signal de parole sont extraits de manière à mettre l'accent sur certaines caractéristiques du signal de parole. Il en résulte une perte d'information due au fait que les paramètres sont limités quant à l'extraction de l'information contenue dans le signal.

Dans le but d'accroître la robustesse des systèmes, il a été proposé de combiner plusieurs systèmes de reconnaissance différents afin de profiter de leur éventuelle complémentarité. En observant les différents constituants d'un système de reconnaissance, il apparaît 3 niveaux principaux dans lesquels on peut combiner différents systèmes.

En premier lieu, on peut travailler au niveau des paramètres acoustiques et effectuer une fusion précoce. L'hypothèse que certaines caractéristiques du signal de parole sont accentuées par certains jeux de paramètres et ignorées par d'autres motive l'idée de vouloir combiner ces flux d'observations acoustiques. Ces méthodes sont passées en revue dans la section 3.1.

Une autre manière de procéder consiste à combiner les probabilités obtenues avec des modèles acoustiques différents, comme décrit dans la section 3.2, on parlera alors de fusion intermédiaire.

L'ensemble des approches consistant à combiner plusieurs jeux de paramètres ou plusieurs distributions de probabilité issues de modèles utilisant des paramètres acoustiques différents est regroupé sous le nom d'approche multi-flux (*multi-stream*). Un exemple particulier de système multi-flux est le système multi-bandes (voir section 3.3).

La combinaison peut aussi se faire après le décodage (fusion tardive), on parlera alors de combinaison post-décodage. Des méthodes utilisant le graphe d'hypothèses entier ou les N-meilleures hypothèses issues de différents systèmes sont présentées en section 3.4.

Mais la combinaison de systèmes ne s'arrête pas là. À chaque niveau, de nombreuses stratégies peuvent être mises en œuvre lors de la combinaison. On peut notamment envisager l'introduction de mesures de confiance afin d'exploiter au mieux les performances locales de chaque système. Un aperçu des différents moyens pour estimer la confiance des hypothèses d'un reconnaisseur

est présenté dans la section 3.5.

3.1 Combinaison de paramètres acoustiques

L'extraction de paramètres acoustiques est la première phase du processus de reconnaissance. Elle consiste à transformer le signal de parole en vecteurs de paramètres acoustiques représentant l'information utile pour la reconnaissance (voir chapitre 2).

La combinaison de paramètres acoustiques avant le modèle acoustique comme présenté dans la figure 3.1 est une technique largement utilisée dans les systèmes de reconnaissance actuels. Un système combinant plusieurs jeux

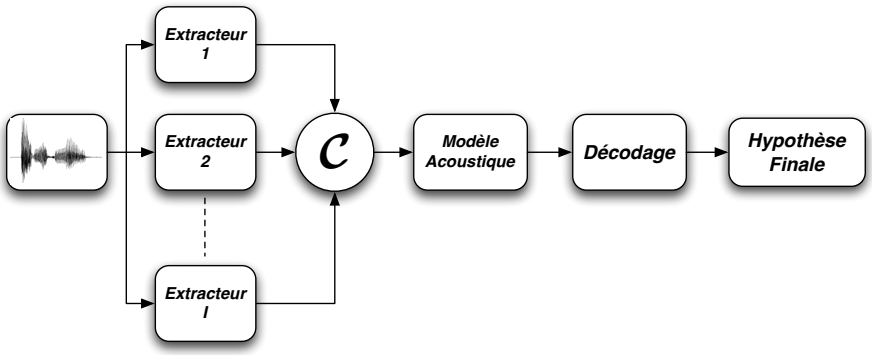

FIG. 3.1: *La combinaison de paramètres acoustiques.*

de paramètres se base sur le principe que certaines caractéristiques du signal de parole qui sont peu accentuées par un jeu de paramètres acoustiques particulier sont mises en relief par un autre, et que par conséquent le flux combiné résultant capture l'information complémentaire présente dans chaque type de traits acoustiques.

Ce genre de combinaison se fait directement au niveau des vecteurs d'observations qui sont intégrés dans un seul flux de paramètres, utilisé ensuite de manière classique pour l'apprentissage d'un modèle acoustique. La combinaison de paramètres avant le modèle acoustique a été proposé par Ellis (2000), où il est présenté des techniques de combinaison de paramètres ayant des caractéristiques différentes.

L'un des problèmes dont il faut tenir compte pour la combinaison de paramètres acoustiques concerne la trame d'analyse du signal de parole. En effet, la construction d'un long vecteur acoustique support de diverses informations

nécessite que ces informations concernent la même portion de signal afin de conserver une certaine cohérence.

3.1.1 Utilisation des dérivées premières et secondes

Dans les systèmes de reconnaissance actuels, il est très courant de compléter un jeu de paramètres par les dérivées premières (Δ) et secondes ($\Delta\Delta$) de ces paramètres. Les dérivées permettent d'inclure des caractéristiques dynamiques des paramètres acoustiques (vitesse et accélération). Le calcul des dérivées se fait sur des fenêtres centrées sur la trame analysée, ce qui assure la cohérence des informations présentes dans le vecteur.

L'utilisation de ces Δ et $\Delta\Delta$ est précisément un cas de concaténation de paramètres acoustiques. Une méthode de combinaison complète de modèles utilisant un jeu de paramètres (PLP), les Δ et les $\Delta\Delta$ de ces paramètres est présentée dans Misra et al. (2003). Chaque type de paramètres (statiques, Δ et $\Delta\Delta$) sont combinés de toutes les manières possibles pour former 7 jeux de paramètres acoustiques utilisés pour apprendre 7 modèles acoustiques différents, dont les probabilités sont ensuite combinées (voir section 3.2).

3.1.2 Augmentation du vecteur de paramètres

L'augmentation du vecteur de paramètres consiste à concaténer des paramètres additionnels dans un vecteur de paramètres acoustiques issu d'une méthode d'analyse classique. Les paramètres additionnels sont utilisés afin d'augmenter la quantité d'information présente dans le vecteur de paramètres acoustiques. Une particularité de ces paramètres est qu'ils ne pourraient pas être utilisés seuls dans un système de reconnaissance. Ils servent à représenter d'autres caractéristiques du signal de parole afin de compléter le vecteur de paramètres acoustiques.

Dans Eide (2001), des paramètres distinctifs basés sur des critères linguistiques sont ajoutés au cepstre afin d'obtenir un nouveau vecteur d'observations. Thomson et Chengalvarayan (1998) proposent l'utilisation de paramètres acoustiques basés sur des caractéristiques de voisement comme données supplémentaires. Les paramètres de périodicité et de *jitter* (fluctuations de la longueur des cycles d'ouverture/fermeture de la glotte[1]) indiquent si les cordes vocales vibrent ou non. Des paramètres acoustiques ou articulatoires additionnels comme le voisement (typiquement la périodicité) (Zolnay et al., 2002, 2005)

[1] À noter que la fréquence fondamentale de la voix dépend directement du nombre de cycles d'ouverture-fermeture de la glotte par seconde.

ou la phase (Schluter et Ney, 2001) ont été associés à des paramètres classiques comme MFCC ou PLP afin d'augmenter l'information contenue dans le vecteur de paramètres..

3.1.3 Concaténation de jeux de paramètres

Une généralisation de l'augmentation de paramètres consiste à concaténer différents jeux de paramètres acoustiques complets en un seul flux. Dans Hegde et al. (2005), un flux de paramètres unique construit à partir de différentes analyses du signal est proposé. Cela correspond à concaténer différents flux de paramètres en un seul et créer des modèles acoustiques capables de modéliser la distribution jointe de ces nouveaux longs vecteurs.

Dans le but de réduire la complexité de modélisation, des algorithmes ont été développés afin de sélectionner des sous-ensembles de paramètres dans un long flux en utilisant un critère qui optimise la classification automatique des données de parole en classes (phonèmes ou traits phonétiques). Malheureusement, les algorithmes pertinents ont une complexité trop élevée avec ce genre de classes comme décrit dans Kamal Omar et al. (2002) et Kamal Omar et Hasegawa-Jonhson (2002). Une solution sous optimale est cependant proposée. Elle consiste à sélectionner un ensemble de mesures acoustiques qui garantit une grande valeur de l'information mutuelle entre ces mesures acoustiques et des paramètres de discrimination phonétique.

L'utilisation de paramètres acoustiques hétérogènes pour caractériser différentes classes acoustico-phonétiques a été proposée dans Halberstadt et Glass (1998). Les différents paramètres utilisés sont les coefficients cepstraux MFCC et PLP, l'énergie à court terme, l'énergie dans les basses fréquences et le taux de changement de signe du signal (*Zero Crossing Rate* - ZCR). Les auteurs soulignent que, compte tenu de la limitation de la résolution temps/fréquence et de la non-réversibilité de la plupart des algorithmes d'analyse, il est nécessaire pour les systèmes de reconnaissance d'intégrer différents types de paramètres acoustiques afin de conserver un maximum d'information acoustico-phonétique. Différents jeux de paramètres sont créés en concaténant ceux présentés ci-dessus pour apprendre N modèles acoustiques. Ensuite, les résultats de ces systèmes sont combinés de trois manières différentes : par vote majoritaire, par combinaison linéaire des vraisemblances normalisées et par combinaison log-linéaire des vraisemblances normalisées. Les résultats obtenus sur le corpus petit vocabulaire TIMIT montrent que l'ajout d'information non présente dans les paramètres cepstraux améliore légèrement les performances des systèmes au prix d'une augmentation des coûts de calcul.

D'autres paramètres acoustiques ou articulatoires ont été utilisées pour

améliorer la reconnaissance. Dans Vaseghi et al. (1997), un décodage en deux passes est effectué. La première passe est classique et permet de générer les N-meilleures hypothèses avec les estimations de leurs bornes temporelles. Puis dans une seconde passe, les estimations des bornes sont utilisées pour extraire des paramètres phonétiques segmentaux qui seront utilisés, en conjonction de modèles segmentaux cette fois-ci, comme paramètres additionnels pour la ré-évaluation de chaque phonème.

Un ensemble de paramètres articulatoires proposés par Kirchhoff (1998) est présenté dans le tableau 3.1. Un réseau de neurones a été entraîné pour

Groupe	Paramètres
Voisement	+voisé, -voisé, silence
Manière	stop, voyelle, fricative, *approximant*, nasale, latérale, silence
Position	dentale, labiale, coronale, palatale, vélaire, glottale, haute, moyenne, basse, silence
Avant-Arrière	avant, arrière, nul, silence
Rondeur des lèvres	+ronde, -ronde, nul, slience

TAB. 3.1: *Paramètres articulatoires proposés dans Kirchhoff (1998).*

chaque groupe de paramètres. Les sorties de ces ANNs correspondent aux différents paramètres articulatoires présents dans ce groupe. Ces paramètres sont extraits par l'intermédiaire d'une segmentation forcée du corpus avec un système hybride utilisant des paramètres acoustiques classiques (ici RASTA-PLP). Ensuite, les caractéristiques articulatoires de chaque unités phonétiques permettent d'identifier quelles trames contient l'information sur tel ou tel paramètre articulatoire.

Ces paramètres sont utilisés tels quels dans un système hybride ANN/HMM et montrent de meilleures performances sur certaines classes phonétiques, surtout lorsqu'il y a beaucoup de bruit. Cela conforte l'idée que les traits acoustiques montrent des forces et faiblesses selon la zone de l'espace acoustique considérée et la nature de l'environnement d'enregistrement.

3.1.4 Réduction du nombre de paramètres

La taille des vecteurs de paramètres est un problème important qui se pose lors de l'ajout de paramètres. Pour remédier à cela, des techniques permettant de réduire le nombre de paramètres sont utilisées. On peut citer notamment l'analyse en composantes principales (*Principal Component Analysis* - PCA) et l'analyse linéaire discriminante (*Linear Discriminant Analysis* - LDA).

Dans Zolnay et al. (2005), la projection d'un long vecteur de paramètres acoustiques dans un espace plus petit (environ 30 paramètres) permet de générer un seul et unique flux de paramètres y_t que l'on utilise pour apprendre un modèle acoustique de manière classique. Le long vecteur correspond à la concaténation de $2L+1$ trames successives de paramètres MFCC, PLP ou des variantes de ceux-ci incluant un paramètre de voisement. La matrice de projection V^T est déterminée par LDA de manière à conserver l'information de classification la plus pertinente dans le petit vecteur y_t.

$$y_t = \left[V^T\right] \begin{bmatrix} \begin{bmatrix} x_{t-L}^1 \\ \dots \\ x_{t-L}^F \end{bmatrix} \\ \dots \\ \begin{bmatrix} x_t^1 \\ \dots \\ x_t^F \end{bmatrix} \\ \dots \\ \begin{bmatrix} x_{t+L}^1 \\ \dots \\ x_{t+L}^F \end{bmatrix} \end{bmatrix} \tag{3.1}$$

En dépit de la structure commune des jeux de paramètres utilisés (tous basés sur un modèle auditif), leur combinaison par LDA produit une réduction significative du WER.

Conclusion

Les techniques d'analyse du signal fournissent des paramètres cepstraux contenant de l'information spécifique à cette analyse. L'information présente est généralement incomplète. L'utilisation de paramètres acoustiques ou articulatoires supplémentaires a donc naturellement été envisagée pour les combiner avec des jeux de paramètres classiques.

Lorsque l'on augmente le nombre de paramètres acoustiques, la principale contrainte concerne la difficulté de modélisation. En effet, l'augmentation de la taille des vecteurs acoustiques complique l'estimation des fonctions de distribution de probabilité. Dans ce cas, la réduction de la taille des vecteurs en sélectionnant les plus pertinents (du point de vue support d'information) apparaît comme une solution adéquate. Une autre solution consiste à utiliser plusieurs jeux de paramètres « autonomes » et de les combiner à un stade ultérieur. Un exemple de ce type d'approche est la combinaison de probabilités, présentée dans la section suivante.

3.2 Combinaison de probabilités

La combinaison peut se faire au niveau des probabilités calculées avec des modèles acoustiques différents. Les différences entre modèles acoustiques se situent à deux niveaux. D'une part, ils peuvent être fondés sur des principes de modélisation différents, comme par exemple les HMM/GMM et les HMM/ANN. D'autre part, ils peuvent utiliser des jeux de paramètres acoustiques différents en entrée.

Ce genre de combinaison a été proposé, entre autre, par Kirchhoff (1998) et Zolnay et al. (2005). La combinaison de probabilités obtenues avec des modèles acoustiques différents (voir figure 3.2) est effectuée après l'obtention des vecteurs de paramètres et avant le décodage. Ce genre de combinaison néces-

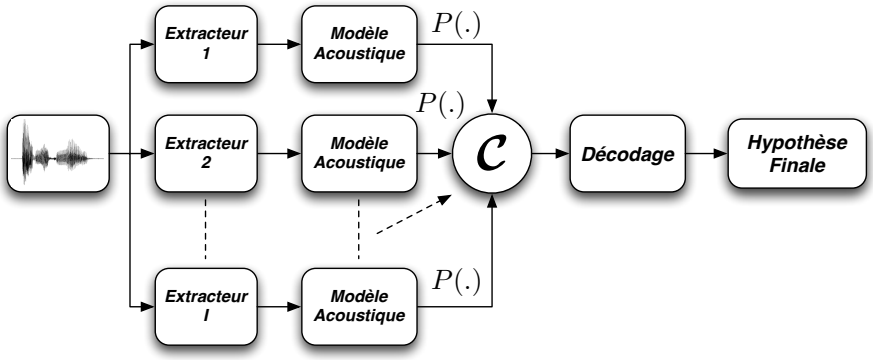

FIG. 3.2: *La combinaison de probabilités* a posteriori *ou de vraisemblances.*

site l'apprentissage de plusieurs modèles acoustiques différents, utilisant des paramètres acoustiques identiques ou non. Selon le type du modèle, les probabilités obtenues seront des probabilités *a posteriori* (pour les systèmes hybrides HMM/ANN) ou des vraisemblances (pour les HMM classiques à base de GMM). Pour que la combinaison de probabilités se fasse de manière cohérente, il faut prendre en compte plusieurs aspects dépendants des flux d'observations acoustiques et des modèles considérés.

3.2.1 Synchronisme des observations acoustiques

Le premier aspect important concerne le synchronisme des flux d'observations. Selon le niveau segmental auquel on souhaite combiner les probabilités, des contraintes plus ou moins fortes sur le synchronisme des flux d'observations sont nécessaires.

Combinaison de probabilités synchrone à la trame

La combinaison synchrone à la trame (ou à l'état d'un HMM) nécessite quelques hypothèses. Les flux d'observations acoustiques doivent être synchrones. Cela signifie que les trames des différents flux doivent être calculés sur les mêmes séquences d'échantillons (ou du moins sur des séquences ayant le maximum d'échantillons en commun). Ceci, afin de s'assurer que l'on combine des probabilités concernant la même portion de signal.

De plus, il est nécessaire d'avoir une topologie strictement identique pour tous les modèles. De manière générale, les systèmes à base de HMM ont des modèles acoustiques qui fournissent des vraisemblances $P(x|q)$ (dans le cas des GMM) ou des probabilités postérieures $P(q|x)$ (dans le cas des ANN) pour chaque trame x à reconnaître et pour chaque état q. Chaque modèle fournit sa propre estimation des vraisemblances ou probabilités. Par conséquent, il faut que les modèles contiennent des états correspondant à la même unité phonétique afin de combiner les probabilités de manière cohérente.

Combinaison synchrone à un niveau segmental supérieur à la trame

La combinaison peut être synchrone au niveau du phonème ou d'une unité plus grande (syllabe, mot). Dans ce cas, on peut envisager d'avoir une topologie identique pour les modèles et effectuer la combinaison de la même manière que pour la combinaison synchrone à la trame en ajoutant la contrainte que les transitions doivent être franchies au même moment pour tous les modèles.

Si les topologies des modèles sont différentes, alors il est nécessaire d'utiliser des méthodes de recombinaison afin de retrouver le synchronisme après l'unité considérée. Dans (Bourlard et al., 1996), il est présenté une architecture permettant la recombinaison de probabilités obtenues avec des modèles de topologies différentes, comme présenté dans la figure 3.3. Dans ce cas, l'utilisation de pseudo-états de recombinaison intégrés dans le HMM est proposé. Ces pseudo-états ne sont pas de réels états émetteurs du HMM et sont utilisés seulement pour la recombinaison des probabilités.

Combinaison asynchrone

Les flux ne sont pourtant pas toujours synchrones comme l'ont montré Mirghafori et Morgan. (1998) pour un système multi-bandes (voir 3.3). Certains événements ne se déclenchent pas en même temps pour différents flux. Relâcher la contrainte de synchronisme entre les flux consiste à permettre qu'une transition dans un modèle arrive plus tôt (ou plus tard) que la même transition

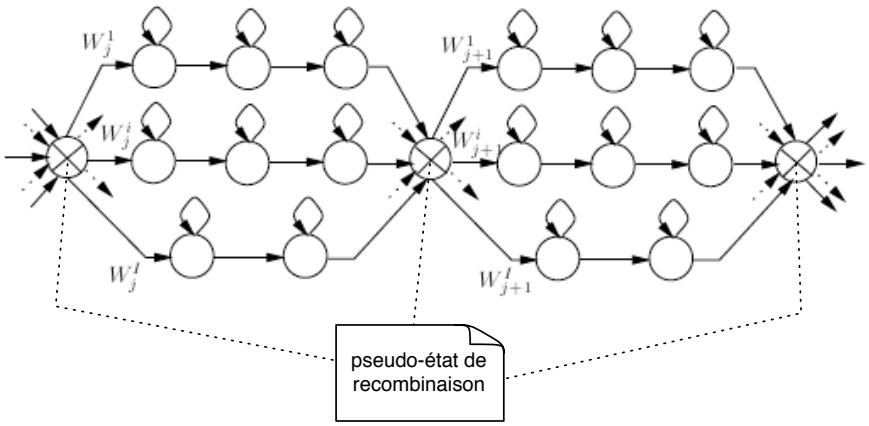

FIG. 3.3: *HMM comportant des pseudo-états de recombinaison.*

dans un autre modèle. Plus précisément, si une transition entre les phonèmes a et b a un maximum de probabilité à la trame t pour le flux f_i, et que cette même transition a un maximum de probabilité à la trame $t + \tau$ pour le flux f_j, alors les probabilités de la trame t pour le flux f_i seront combinées avec celles de la trame $t + \tau$ pour le flux f_j. Beaucoup de travaux, notamment en reconnaissance de la parole audio-visuelle (Gravier et al., 2002b) ont été conduit afin de modéliser l'asynchronie des flux de paramètres (Mirghafori et Morgan, 1999; Cerisara et al., 2000). Cependant, les résultats montrent que la prise en compte de l'asynchronie ne produit pas de réelle amélioration des performances.

3.2.2 Estimation des probabilités

Les modèles acoustiques permettent de calculer différents scores concernant la présence d'un symbole dans le signal de parole. Les vraisemblances ou les probabilités *a posteriori*, sont calculées en fonction du type de modèle considéré.

Vraisemblances : les modèles acoustiques utilisant des GMM pour modéliser les fonctions de densité de probabilité fournissent des scores $f(x|q)$ correspondant à la vraisemblance qu'une observation acoustique x ait été émise par un état q du HMM. Ces vraisemblances sont ensuite normalisées pour obtenir une estimation de la probabilité *a posteriori* $P(x|q)$ de la manière suivante :

$$P(x|q) = \frac{f(x|q)}{\sum_{v \in Q} f(x|v)} \tag{3.2}$$

avec Q l'ensemble des états du HMM. f est la fonction de vraisemblance pour

une mixture de gaussiennes définie comme suit :

$$f(x|q) = \sum_{l=1}^{L} \alpha_l N(x, \mu_l, \sigma_l) \tag{3.3}$$

avec α_l le poids de la composante l ($l = 1 \dots L$) et $N(x, \mu_l, \sigma_l)$ la fonction de densité de probabilité de la gaussienne l définie par :

$$N(x, \mu_l, \sigma_l) = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp^{-\frac{1}{2}\frac{(x_i-\mu_l)^2}{\sigma_l^2}} \tag{3.4}$$

La magnitude de ces vraisemblances est dépendante de la taille et de la nature des vecteurs de paramètres acoustiques utilisés pour les calculer. Aussi, on ne peut pas comparer (et donc combiner) directement des vraisemblances obtenues avec des jeux de paramètres de tailles différentes. Cependant, la combinaison de plusieurs modèles différents dont l'apprentissage a été réalisé avec le même jeu de paramètres montre une amélioration des performances de reconnaissance, comme par exemple dans Fischer et al. (2002) où les vraisemblances issues de modèles multi-linguaux sont combinées.

Probabilités *a posteriori* : dans un système hybride HMM/ANN, les probabilités *a posteriori* $P(q|x)$ sont directement disponibles à la sortie du réseau de neurones. Ces probabilités ne sont pas dépendantes de la taille des vecteurs de paramètres et peuvent donc être combinées plus naturellement.

Dans Zolnay et al. (2005), des résultats montrent que la combinaison log-linéaire de probabilités postérieures des mots connaissant la suite d'observations acoustiques calculée à l'aide de modèles différents augmente les performances globales du système. Plusieurs modèles acoustiques générant différentes probabilités $P^f(X^f|W)$ d'une suite de mots W sont considérés. Un seul modèle de langage est utilisé, ce qui permet de réécrire la règle de décision de Bayes comme suit :

$$W_{opt} = \underset{W}{\operatorname{argmax}}\, P(W)^{\lambda_{LM}} \prod_i P_i^f(X_i^f|W)^{\lambda_{f_i}} \tag{3.5}$$

avec $P(W)^{\lambda_{LM}}$, la probabilité de la suite de mots W donnée par le modèle de langage, et $P_i^f(X_i^f|W)^{\lambda_{f_i}}$, la probabilité de W donnée par le modèle acoustique utilisant les paramètres acoustiques f_i pondéré par le facteur d'échelle (*fudge*)λ_{f_i} correspondant à ce jeu de paramètres.

Dans Evermann et Woodland (2000), les probabilités *a posteriori* des mots sont obtenues à partir de plusieurs systèmes combinés par réseaux de confusion (*Confusion Network Combination* - CNC) décrit dans Mangu et al. (1999) et

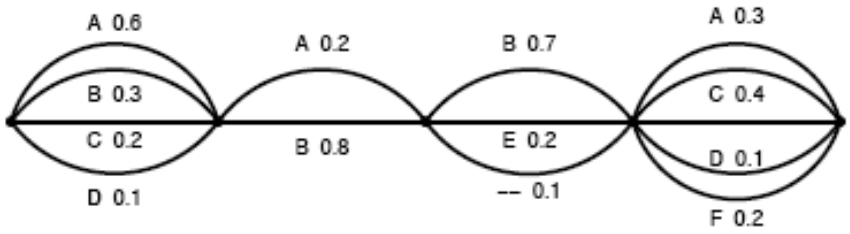

FIG. 3.4: *Exemple de réseau de confusion.*

dont un exemple est présenté dans la figure 3.4. Les CNC offrent une meilleure estimation de la probabilité *a posteriori* des hypothèses de mots menant à une amélioration du WER.

3.2.3 Génération de modèles différents

Dans le but de générer des systèmes différents, Siohan et al. (2005) proposent une méthode de partage d'états (*"tying"*) aléatoire. Le *tying* consiste à partager les paramètres entre les modèles ou parties de modèles. Cette technique est utilisée dans le but de réduire la quantité de données nécessaire pour l'apprentissage robuste. Il se base sur le principe qu'il vaut mieux apprendre un seul modèle en utilisant les données de deux unités très proches acoustiquement mais peu représentées dans le corpus d'apprentissage, plutôt que d'entaîner deux modèles séparés utilisant le peu de données à disposition. Le *tying* d'états se base sur un arbre de décision permettant de regrouper les unités acoustiques proches, sur un critère de MAP ou de maximisation de l'information mutuelle (*Maximum Mutual Information Estimation* - MMIE). Au lieu d'utiliser la séparation qui maximise le critère de MAP, Siohan et al. (2005) sélectionnent la séparation aléatoirement parmi les N meilleures séparations. Les modèles ainsi générés vont donc modéliser différents groupes d'unités acoustiques dépendantes du contexte.

La combinaison de systèmes de reconnaissance de la parole est bénéfique particulièrement lorsque les systèmes sont vraiment complémentaires. Une approche permettant de générer des modèles explicitement complémentaires est présentée dans Breslin et Gales (2006). Elle utilise un critère de risque bayésien minimum (*Minimum Bayesian Risk* - MBR). La fonction objective, composée d'une fonction de perte prenant en compte la dépendance des modèles précédents $\mathcal{M}^0 \dots \mathcal{M}^{s-1}$ se présente sous la forme suivante :

$$\mathcal{F}(\mathcal{M}) = \sum_{r=1}^{R} \sum_{H_w \in \mathcal{H}} P(H_w|\mathcal{O}_r;\mathcal{M}) \tilde{\mathcal{L}}(H_w, \tilde{H}|\mathcal{M}^0 \dots \mathcal{M}^{s-1}) \tag{3.6}$$

où $\tilde{H}$ est l'hypothèse correcte pour les données $\mathcal{O}_r$, $\mathcal{H}$ est l'ensemble des hypothèses possibles et $\mathcal{M}$ est le modèle courant.

Le but est de créer un modèle $\mathcal{M}^s$ se concentrant sur les erreurs commises par les modèles précédents $\mathcal{M}^0 \dots \mathcal{M}^{s-1}$. La fonction de perte $\tilde{\mathcal{L}}$ reflète le fait que les données d'entraînement ont été bien modélisées ou pas. Pour cela, elle utilise les probabilités *a posteriori* des mots calculées avec les modèles précédents de la manière suivante :

$$\tilde{\mathcal{L}}(W_m, \tilde{W}; S) = \begin{cases} 1 \; si \; P(W|\mathbf{O}; S) > \alpha, W_m \neq \tilde{W} \\ 0 \; sinon \end{cases} \tag{3.7}$$

Si un mot a été correctement modélisé, la fonction de perte sera minimale pour ce mot qui ne sera donc pas utilisé pour l'apprentissage des modèles suivants. Les résultats obtenus montrent que les modèles entraînés pour être complémentaires fournissent en général des performances moins bonnes lorsqu'ils sont utilisés seuls, mais leur combinaison fait baisser légèrement le WER.

3.2.4 Stratégies de combinaison

Un cadre théorique ainsi que de nombreux types de combinaison de probabilités sont proposés dans Kittler et al. (1998). En règle générale, le problème de reconnaissance de motif consiste à assigner une des m classes $\{\omega_1 \dots \omega_m\}$ à un motif Z. Considérons maintenant que nous avons R classifieurs (dans notre cas des modèles acoustiques) qui représentent chacun le motif Z par un vecteur d'observations x_i distinct. Dans chaque espace des observations, la classe ω_k est modélisée par la probabilité $P(x_i|\omega_k)$. D'après la théorie bayésienne, connaissant les observations $x_i, i = 1 \dots R$, la classe w_j devant être associée au motif Z est celle fournissant la probabilité *a posteriori* maximum, *i.e.* sélectionner la classe w_j vérifiant :

$$\omega_j = \underset{\omega_k}{\operatorname{argmax}} \, P(\omega_k|x_1 \dots x_R) \tag{3.8}$$

Cette équation n'est pas utilisable telle quelle puisque toutes les observations devraient être traitées simultanément, ce qui a pour conséquence d'augmenter considérablement la difficulté de modélisation et le temps de calcul.

Cependant, en appliquant le théorème de Bayes, la probabilité *a posteriori* devient :

$$P(\omega_k|x_1 \dots x_R) = \frac{P(x_1 \dots x_R|\omega_k)P(\omega_k)}{P(x_1 \dots x_R)} \tag{3.9}$$

où $P(\omega_k)$ est la probabilité *a priori* de la classe ω_k et $P(x_1 \dots x_R|\omega_k)$ est la fonction de densité de probabilités (*Probability Density Function* - p.d.f.) jointe. C'est cette p.d.f. que nous allons représenter de manière différente selon certaines hypothèses, plus ou moins réalistes.

La règle « produit »

L'indépendance statistique des observations x_i peut être vraie dans certaines applications, puisque ces observations sont distinctes. Dans ce cas, la p.d.f. jointe peut se ré-écrire de la manière suivante :

$$P(x_1 \dots x_R|\omega_k) = \prod_{i=1}^{R} P(x_i|\omega_k) \tag{3.10}$$

Il en résulte que l'équation 3.8 devient (en utilisant l'équation 3.9 et en omettant le dénominateur commun) :

$$\omega_j = \underset{\omega_k}{\text{argmax}} \prod_{i=1}^{R} P(x_i|\omega_k)P(\omega_k) \tag{3.11}$$

Ce genre de combinaison est relativement sévère car la probabilité d'une classe sera proche de 0 si un seul des modèles propose une probabilité proche de 0.

La règle « somme »

Dans le cas du traitement de parole totalement disparate (signal fortement altéré), il peut être recommandé de ne considérer que l'information *a priori* pour une trame de parole. On peut alors approximer les probabilités *a posteriori* comme suit :

$$P(\omega_k|x_i) = P(\omega_k)(1 + \delta_{ki}) \tag{3.12}$$

où $\delta_{ki} << 1$.

Après plusieurs étapes de simplification (Kittler et al., 1998), le critère de décision peut se ramener à sélectionner la classe ω_j vérifiant :

$$\omega_j = \underset{\omega_k}{\text{argmax}} \sum_{i=1}^{R} P(\omega_k|x_i) \tag{3.13}$$

Dans ce type de combinaison, la probabilité postérieure obtenue sera grande pour la classe considérée si au moins l'un des modèles fournit une grande probabilité pour cette classe.

La règle « maximum »

La règle « maximum » estime la probabilité postérieure d'une classe par la plus grande probabilité postérieure parmi celles proposées par les classifieurs

pour cette classe.

$$P(\omega_k|x) = \frac{\max_{i=1}^{R} P(\omega_k|x_i)}{\sum_{j=1}^{K} \max_{i=1}^{R} P(\omega_j|x_i)} \tag{3.14}$$

Ce type de combinaison réagit de la même manière que pour la règle « somme ».

La règle « minimum »

La règle « minimum », contrairement à la règle « maximum », estime la probabilité postérieure d'une classe par la plus petite probabilité parmi celles proposées par les classifieurs.

$$P(\omega_k|x) = \frac{\min_{i=1}^{R} P(\omega_k|x_i)}{\sum_{j=1}^{K} \min_{i=1}^{R} P(\omega_j|x_i)} \tag{3.15}$$

Ce type de combinaison réagit de la même manière que la règle « produit » dans le sens où la probabilité *a posteriori* d'une classe sera élevée seulement si tous les modèles fournissent une grande probabilité pour la classe considérée.

La règle « médiane »

Si on considère que les classes ont la même probabilité *a priori*, alors la règle « somme » devient une simple moyenne arithmétique des probabilités postérieures. Or, si l'un des classifieurs propose une grande probabilité pour une classe erronée, la probabilité moyenne s'en verra beaucoup affectée et pourra conduire à une mauvaise classification. La médiane (correspondant à sélectionner la probabilité qui sépare la distribution de probabilité en deux parties de même cardinal) est robuste à ce genre de problème.

$$P(\omega_k|x) = \frac{\operatorname{med}_{i=1}^{R} P(\omega_k|x_i)}{\sum_{j=1}^{K} \operatorname{med}_{i=1}^{R} P(\omega_j|x_i)} \tag{3.16}$$

Le vote majoritaire

Le vote majoritaire consiste à sélectionner la classe pour laquelle le plus grand nombre de classifieurs a donné le maximum de probabilité.

$$P(\omega_k|x) = \frac{\sum_{j=1}^{K} \Delta_{ij}}{K} \tag{3.17}$$

avec

$$\Delta_{ki} = \begin{cases} 1 : \text{si } P(\omega_k|x_i) = \max_{j=1}^{K} P(\omega_j|x_i) \\ 0 : \text{sinon} \end{cases}$$

L'avantage de cette méthode de combinaison est qu'il n'est pas nécessaire que les scores de chaque classifieurs soient comparables.

Combinaison par critère d'entropie

L'entropie d'une distribution de probabilité mesure le degré de chaos présent dans cette distribution.

$$E = -\sum_{j=1}^{K} P(\omega_k|x_i) \log P(\omega_k|x_i) \tag{3.18}$$

L'entropie d'une distribution de probabilité est minimale (égale à 0) lorsqu'une classe obtient la probabilité totale (égale à 1). Dans ce cas, le classifieur affiche une grande confiance pour la classe sélectionnée. Au contraire la décision du classifieur sera incertaine s'il propose des probabilités semblable pour un grand nombre de classes. Le cas extrême correspond à affecter la même probabilité à toutes les classes, ce qui maximise l'entropie. Le critère de décision consiste donc à favoriser la sortie du classifieur dont le vecteur de probabilités affiche la plus petite entropie.

3.3 Systèmes multi-bandes

Les travaux de Fletcher présentés dans Allen (1994) suggèrent que les messages acoustiques ne sont pas traités dans leur globalité par le système auditif humain. Au lieu de cela, de multiples décodages partiels du signal de parole, supposés indépendants, sont effectués. Les décodages exploitent des bandes de

fréquences plus étroites qui sont ensuite intégrées afin de recomposer le message.

Un cadre expérimental particulier tirant parti de ces résultats est l'analyse en sous-bandes du signal de parole. Le principe réside dans l'extraction de paramètres acoustiques dans des bandes de fréquences plus étroites. Le spectre est découpé en plusieurs bandes afin de séparer, le cas échéant, les fréquences contenant du bruit de celles qui n'en contiennent pas. Ensuite, des systèmes utilisant ces paramètres sont combinés (au niveau des probabilités ou des hypothèses de sortie).

Pour chaque sous-bande, les techniques de paramétrisation peuvent être les mêmes ou non (Cerisara et Fohr, 2001). Une fois les vecteurs de paramètres calculés, plusieurs méthodes pour les recombiner sont possibles. Dans un premier temps, on peut concaténer les vecteurs de paramètres acoustiques afin de reformer un vecteur contenant l'information de la bande de fréquence totale pour ensuite être utilisé de manière classique dans le système (Okawa et al., 1999). Une autre manière de procéder consiste à construire des modèles acoustiques spécifiques pour chaque bande de fréquence. Ces modèles permettent de calculer des probabilités qui sont combinées, avec l'une des techniques présentées dans la section 3.2, pour être utilisées dans le décodeur (voir figure 3.5).

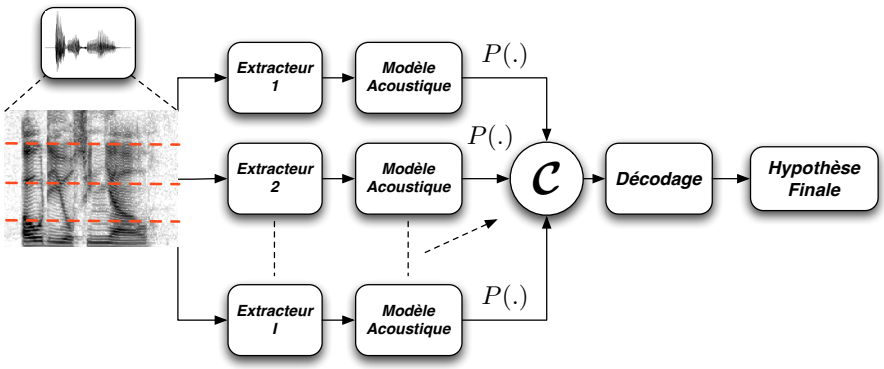

FIG. 3.5: *La combinaison de probabilités dans un système multi-bandes.*

Les avantages du système opérant en sous-bandes est que le bruit éventuellement présent dans une des sous-bandes ne corrompt pas les paramètres extraits dans les autres. Cependant, les systèmes multi-bandes ne proposent pas forcément de meilleures hypothèses qu'un système classique. En effet, Mirghafori et Morgan (1998) ont effectué une analyse des erreurs obtenues avec un système multi-bandes et *full-band*, montrant que les deux systèmes fournissent des performances différentes selon le contexte phonétique considéré. Cela motive l'idée de combiner les systèmes uniquement dans certains contextes, et pas systématiquement en chaque point de l'espace acoustique. Afin d'exploiter au

mieux le paradigme du multi-bandes, on peut utiliser des poids de combinaison proportionnels à la confiance que l'on peut attribuer à chacune des sous-bandes.

Berthommier et Glotin (1999) proposent l'utilisation du SNR pour détecter les sous-bandes non-bruitées et ensuite ajuster les poids lors de la combinaison trame à trame. Le seuil sur le rapport signal sur bruit calculé à partir d'une estimation de l'« harmonicité[2] » du signal permet d'attribuer une mesure de fiabilité à chaque bande. Les résultats montrent que l'ajout de paramètres identifiant des zones où la reconnaissance est supposée être plus ou moins facile permet d'améliorer les performances de reconnaissance (notamment dans un système multi-bandes).

Dans Daoudi et al. (2001), une généralisation du système multi-bandes est présenté. Au lieu d'utiliser les HMMs pour modéliser chaque sous-bande, des réseaux bayésiens (*Bayesian Networks* - BN) sont utilisés. Ceux-ci ont l'avantage d'intégrer la modélisation de la corrélation entre les sous-bandes et de permettre l'asynchronie entre les différentes sous-bandes sans nécessiter l'utilisation de méthodes de recombinaisons. Une amélioration de plus de 10% (précision en mots) a été observée sur un corpus de parole non bruitée petit vocabulaire.

3.4 Combinaison d'hypothèses de reconnaissance

La combinaison d'hypothèses de reconnaissance est faite sur le treillis de mots ou les N-meilleures hypothèses obtenues après le décodage du système de reconnaissance automatique de la parole. Le principe de la combinaison

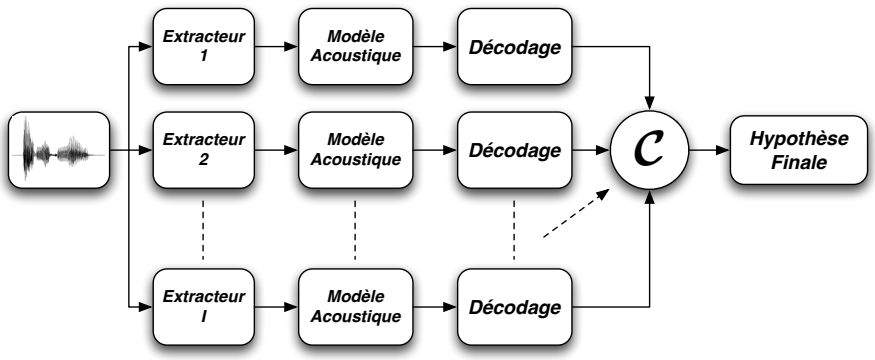

FIG. 3.6: *La combinaison d'hypothèses de reconnaissance.*

[2]L'harmonicité est une mesure indiquant que les harmoniques (fréquences de résonance) sont rigoureusement des multiples entiers de la fréquence fondamentale.

post-décodage est de considérer et comparer les différentes hypothèses générées par plusieurs systèmes de reconnaissance afin de constituer une nouvelle hypothèse. De nombreuses techniques peuvent être employées pour combiner, comme la comparaison des probabilités *a posteriori* des mots ou l'utilisation de mesures de confiance. Un ensemble non exhaustif est présenté dans les sous-sections suivantes.

3.4.1 Vote majoritaire pondéré : ROVER

Dans Fiscus (1997), un système de combinaison d'hypothèses de phrase est présenté : *Recognizer Output Voting Error Reduction* (ROVER). Les hypothèses

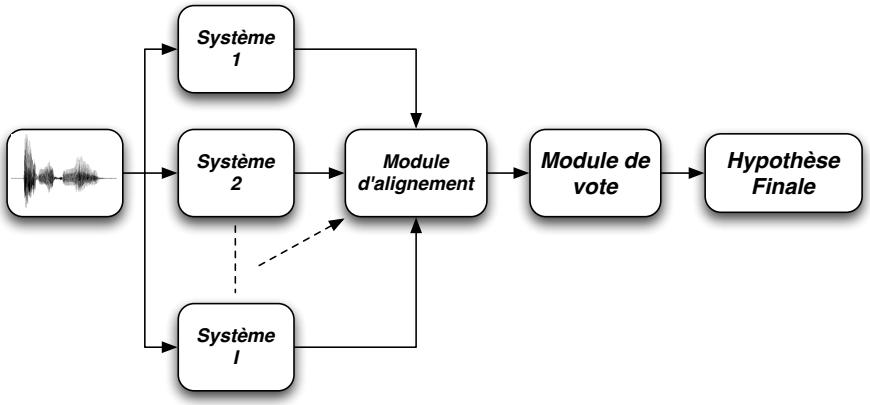

FIG. 3.7: *Combinaison par vote majoritaire pondéré : ROVER (Fiscus, 1997)*

sont d'abord alignées par programmation dynamique afin d'obtenir un réseau qui sera ensuite soumis au vote majoritaire pondéré. Pour ce faire, les sorties de plusieurs systèmes de reconnaissance sont combinées en un seul réseau de mots de coût minimal (voir equation 3.19) en appliquant itérativement des alignements par programmation dynamique. Le réseau obtenu est parcouru par un processus de vote majoritaire pondéré qui sélectionne la séquence de sortie de coût minimal.

Les scores d'évaluation sont divers mais tous basés sur une même formule générale :

$$Score(w) = \alpha(N(w,i)/Ns) + (1-\alpha)C(w,i) \tag{3.19}$$

avec $N(w,i)$ le nombre d'occurrences du mot w dans l'ensemble i, Ns le nombre de systèmes, C(w,i) le score de confiance associé au mot w, et α un poids permettant de faire le compromis entre la fréquence du mot et le score de confiance.

Fiscus (1997) propose trois schémas de combinaison :

- La fréquence d'apparition : c'est le vote majoritaire pur. Dans ce cas, le paramètre α est égal à 1.0.
- Le score de confiance moyen : le nombre d'occurrences et les scores associés à chaque hypothèse de mot par les différents systèmes permettent de sélectionner les meilleures hypothèses. α est estimé sur des données différentes du corpus de test.
- Le score de confiance maximum : l'hypothèse ayant le score maximum parmi les hypothèses proposées sera sélectionnée. α est également estimé sur un corpus.

Comme expliqué dans Hoffmeister et al. (2006), l'alignement est dépendant de l'ordre des permutations effectuées par le système. Le résultat est donc dépendant de l'ordre de combinaison des hypothèses de phrase de chaque système. Il a été montré que les meilleurs résultats sont obtenus lorsque les systèmes sont ordonnés par ordre croissant de taux d'erreur mot.

Cependant, les résultats de Schwenk et Gauvain (2000) montrent que la combinaison d'un grand nombre de systèmes peut affecter les performances globales du système, surtout à cause des systèmes ayant des performances moindres. Les auteurs montrent aussi l'apport bénéfique du modèle de langage, notamment pour départager les ex-æquo.

3.4.2 Les réseaux de confusion : CNC

Une autre manière de combiner les systèmes après le décodage est de modifier le graphe de mots de telle sorte que tous les arcs partant d'un noeud aient le même noeud de destination (Mangu et al., 1999). On obtient alors un réseau de confusion (voir figure 3.8) dont tous les arcs sont pondérés par une probabilité calculées à partir des probabilités des mots du graphe initial. Grâce à ce réseau de confusion, on peut maximiser les probabilités *a posteriori* locales des mots de la phrase, ce qui donne de meilleurs résultats que de maximiser la probabilité *a posteriori* globale de la phrase étant donné que la mesure utilisée pour calculer les performances d'un système est le taux d'erreur mot.

3.4.3 Combinaison bayésienne : BAYCOM

Dans Sankar (2005), une combinaison bayésienne des sorties de systèmes de RAP utilisant des modèles différents est proposée. La vraisemblance des phrases et un ensemble de scores de confiance permettent de garantir l'optimalité sous certaines hypothèses. L'indépendance entre les systèmes est supposée et les probabilités dépendent des performances du système global sans considérer les hypothèses de phonèmes locales, proposées par chaque système.

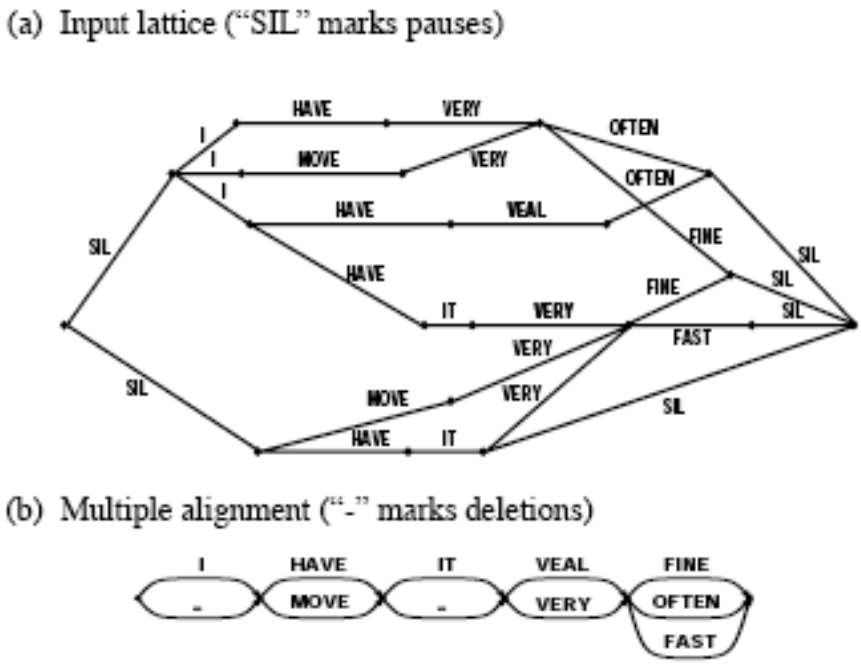

FIG. 3.8: *Du treillis au réseau de confusion (Mangu et al., 1999)*

BAYCOM permet, par une approche basée sur la théorie de la décision, de déterminer les poids optimaux pour la combinaison de plusieurs systèmes.

3.4.4 Autres méthodes

L'utilisation de réseaux de neurones, d'arbres de décision et d'autres techniques d'apprentissage automatique ont été utilisées pour combiner les résultats de plusieurs systèmes de RAP afin de réduire le WER.

Utsuro et al. (2003) utilisent des classifieurs à base de machines à support vectoriel (*Support Vector Machine* - SVM) pour sélectionner l'hypothèse la plus fiable parmi l'ensemble des hypothèses données par 26 systèmes de reconnaissance large vocabulaire. Les différences entre les systèmes résident dans leurs modèles acoustiques et linguistique. Les modèles acoustiques possèdent des topologies différentes, modélisent des unités différentes (phonèmes et syllabes) et sont dépendant du genre. Une diminution du taux d'erreur mot d'environ 36% par rapport à une combinaison par vote majoritaire pondéré ROVER a été observée.

Conclusion sur la combinaison post-décodage

Les techniques de combinaison de systèmes après le décodage produisent une amélioration conséquente des performances des systèmes de RAP. Cependant, elles possèdent certaines faiblesses qui nuisent à la pleine exploitation de ce genre de technique. La première d'entre elle est que la structure du graphe de mots est généralement perdue, tout comme les bornes de ces mots. Il y a donc perte de synchronisme entre les différents systèmes, ce qui peut provoquer la mise en confrontation de mots qui, au niveau temporel, ne seraient pas du tout en concurrence. De plus, les meilleures hypothèses proposées par les systèmes ne sont pas remises en cause. L'algorithme tente de sélectionner l'hypothèse correcte si elle a été produite par au moins l'un des systèmes.

Dans Hoffmeister et al. (2006), les résultats des combinaison par ROVER et CNC sont comparées à une approche basée sur la trame. Les résultats ne montrent pas d'amélioration significative des performances sur les corpus utilisés.

3.5 Mesures de confiance

Étant donné que les systèmes de reconnaissance sont enclins à commettre des erreurs, il est grandement souhaitable d'avoir une information sur la fiabilité de leurs hypothèses. La fiabilité d'un système de reconnaissance automatique de la parole est quantifiée par des mesures de confiance que le système apporte sur ses propres hypothèses. Ces mesures de confiance sont fondées sur différents types d'information.

Dans les systèmes de dialogue, l'aspect le plus important est de limiter le taux de fausse acceptation, c'est-à-dire le taux de mauvaise compréhension qui entraîne un traitement incorrect de la requête formulée par l'utilisateur. Il est donc très utile de connaître la confiance que possède le système dans les hypothèses qu'il propose, afin de modifier la prise de décision qui en résulte. On pourra, par exemple, demander une répétition de la part de l'utilisateur lorsque la confiance n'est pas suffisamment grande.

Dans le cadre de la combinaison de systèmes, les mesures de confiance calculées pour chaque système permettent par exemple de pondérer les probabilités de ce système afin de lui donner plus ou moins d'importance.

Un cadre général pour combiner des mesures de confiance au niveau de la phrase est proposé dans Souvignier et Wendemuth (1999). Étant donné une liste des N-meilleures hypothèses et le score de chaque phrase de cette liste, la

mesure de confiance suivante pour un motif générique A (qui est une séquence de I mots ou attributs) est proposée :

$$C(A) = \frac{\sum_{i \in I} \exp^{-\alpha sc_i}}{\sum_{i=1}^{N} \exp^{-\alpha sc_i}} \tag{3.20}$$

avec α un facteur d'échelle permettant de distribuer la probabilité dans la liste des N-meilleures hypothèses. Le numérateur est étendu à toutes les phrases contenant le motif A dans sa liste des N-meilleures phrases.

Un autre schéma de combinaison est donné par Garcia-Mateo et al. (1999) où l'utilisation de modèles et antimodèles permet le calcul d'un rapport de vraisemblance logarithmique (*Log Likelihood Ratio* - LLR).

$$LLR = f(llr_i(O_i)) \text{ avec } llr_i(O_i) = \log P(O_i|m_i) - \log P(O_i|a_i) \tag{3.21}$$

où m_i et a_i sont respectivement le modèle et l'antimodèle du $i^{\text{ème}}$ phonème. La fonction f peut être une simple somme ou une fonction sigmoïde.

Des mesures de confiance plus complexes sont proposées dans Siu et Gish (1999). Des régressions logistiques (*Logit*) effectuant une combinaison linéaire des scores sont considérées. Une extension de ces régressions logistiques (« *generalized additive models* ») consiste à appliquer des transformations non-linéaires sur les paramètres d'entrée.

De nombreux paramètres fondés sur les reconnaisseurs sont utilisés pour dériver des mesures de confiance (Zhang et Rudnicky, 2001). Ces paramètres sont utilisés pour la classification des mots comme corrects ou incorrects et peuvent se décomposer en quatre catégories : les paramètres acoustiques, les paramètres du modèle de langage, les probabilités postérieures logarithmiques des mots issues du treillis de mot et les paramètres calculés sur la liste des N-meilleures hypothèses. Ces paramètres sont utilisés dans différents types de classifieurs comme les réseaux de neurones, les arbres de décision, les classifieurs linéaires et les SVM.

De la même manière, Moreno et al. (2001) utilisent la technique du boosting pour évaluer la confiance des hypothèses de mot. Cette approche se compare favorablement à deux autres approches standard : les SVMs et les arbres de classification et de régression (*Classification And Regression Trees* - CART).

Dans chaque cas (Zhang et Rudnicky, 2001; Moreno et al., 2001), les mesures de confiance sont fondées sur le décodeur. On peut alors se demander si ces mesures de confiance peuvent être utilisées dans d'autres contextes.

Dans Cox et Dasmahapatra (2002), un vecteur de mesures de confiance, basées sur de l'information subsidiaire, est dérivé du système de reconnaissance

pour calculer la probabilité qu'un mot soit correct. Différentes approches haut niveau, se voulant moins spécifiques au reconnaisseur, permettent d'assurer la généralité et donc la ré-utilisabilité de ces mesures de confiance. Il est montré que la probabilité qu'un mot soit correct est dépendante du fait que le mot précédent soit correct. Une tentative pour découpler les modèles acoustique et linguistique a également été effectuée. Dans ce but, des reconnaisseurs de phonèmes et de mots sont utilisés sans modèle de langage. L'hypothèse la plus probable issue du reconnaisseur de mot est ensuite utilisée pour générer la mesure de confiance.

Un compte-rendu des techniques utilisées pour l'élaboration de mesures de confiance est présenté dans Lee (2001) et Jiang (2005). L'utilisation de mesures de confiance permet donc de faire un auto-diagnostic et ainsi évaluer le résultat de la reconnaissance, avant que celui-ci ne soit traité par la suite (dans le cadre d'un système de dialogue). La connaissance d'une telle information permet des traitements plus intelligents qui aboutissent en général à une amélioration des performances du système.

3.6 Conclusion

Dans ce chapitre, les différentes manières de combiner les systèmes de reconnaissance de la parole ont été présentées. Chaque composant du système peut être sujet à la combinaison. Les paramètres acoustiques peuvent être combinés afin qu'ils transportent une plus grande quantité d'information. Les modèles acoustiques quant à eux représentent et structurent l'espace acoustique. La modélisation de l'espace acoustique varie selon l'architecture du modèle acoustique et le type de paramètre acoustiques utilisés. Ainsi, deux modèles de même topologie fourniront des résultats différents s'ils utilisent des paramètres acoustiques différents. La combinaison de ces modèles espère donc profiter de leur éventuelle complémentarité. La combinaison d'hypothèses de reconnaissance consiste à générer une nouvelle hypothèse à partir de celles proposées par plusieurs systèmes. Des techniques permettant d'évaluer la confiance que l'on peut attribuer à un système permettent de sélectionner les mots qui constitueront l'hypothèse finale.

Chapitre 4

Combinaison acoustique au niveau phonétique

Sommaire

4.1 Introduction

Une trame de parole est décrite par un vecteur de paramètres acoustiques pouvant être représenté par un point de l'espace acoustique, comme présenté dans la figure 4.1.

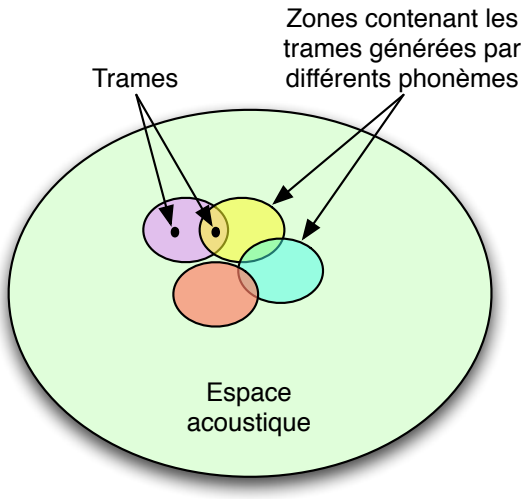

FIG. 4.1: *Schématisation d'un espace acoustique en deux dimensions, contenant des zones associées à des contextes phonétiques différents.*

Les paramètres acoustiques jouent un rôle essentiel dans un système de reconnaissance. Ils sont à la base de la construction des modèles acoustiques et sont porteurs de l'information nécessaire à la reconnaissance. Les paramètres acoustiques ont été développés afin de représenter globalement et le mieux possible l'ensemble de l'espace acoustique. On peut donc naturellement supposer qu'il existe certaines zones où ils extraient très précisément les caractéristiques des symboles phonétiques, affichant une grande fiabilité, et d'autres zones où ce n'est pas le cas, et sont donc peu fiables.

L'analyse de la variabilité des paramètres acoustiques permet de déterminer les régions d'un espace acoustique pour lesquelles un jeu de paramètres est approprié pour la reconnaissance. L'objectif de cette étude est d'identifier des conditions permettant d'isoler les situations pour lesquelles la fiabilité d'un jeu de paramètres acoustiques est grande.

Considérons une zone d'un espace acoustique peuplée d'un ensemble de points correspondant aux trames générées par la prononciation de symboles (phonème, syllabe, mot, ...). Afin de décrire la variabilité des paramètres acoustiques, nous allons étudier l'ambiguïté (ou confusion) des symboles phonétiques qui ont généré les trames appartenant à ces zones.

Le modèle acoustique a pour rôle de partitionner et de structurer l'espace acoustique en zones correspondant aux observations produites par la prononciation de chaque symbole. La variabilité des paramètres acoustique perturbe cette partition et crée de l'ambiguïté au sein du modèle acoustique. Plusieurs causes à cette ambiguïté peuvent être identifiées. Tout d'abord, le signal de parole est modifié par le bruit environnant, ce qui provoque des déformations du spectre pouvant aboutir à une mauvaise reconnaissance. Ensuite, les aspects morphologiques et environnementaux impliquent que deux locuteurs ne prononcent pas les phonèmes de la même manière. De plus, la discrimination de certains symboles phonétiques peut être complexe. Cette confusion se caractérise par le fait que deux vecteurs acoustiques correspondant à des phonèmes différents peuvent être très proches dans l'espace acoustique.

La reconnaissance de symboles phonétiques est fondée sur les probabilités estimées avec des modèles acoustiques. Par conséquent, il est difficile d'évaluer si l'ambiguïté d'une zone de l'espace acoustique est due à la variabilité des paramètres acoustiques ou à l'imperfection des modèles acoustiques.

Une architecture permettant d'une part de mesurer l'ambiguïté des symboles phonétiques, et d'autre part de séparer l'influence des modèles acoustiques de celle des paramètres acoustiques a été mise en place. Cette architecture nous permet également de faire une analyse comparative de la participation à l'ambiguïté de différents jeux de paramètres et ainsi d'évaluer leur complémentarité. Des stratégies sont dérivées de ces analyses pour exploiter différents jeux de paramètres. Une mesure de confiance dérivée de l'estimation de la variabilité attendue des paramètres acoustiques dans un segment de parole est introduite.

Afin de déterminer les forces et les faiblesses de différents jeux de paramètres acoustiques, plusieurs corpus avec des petits et grands vocabulaires, ainsi que différentes langues et différents types de modèles acoustiques ont été utilisés. Les analyses et expériences correspondantes sont présentées dans les sections suivantes.

La section 4.3 présente l'analyse comparative de deux jeux de paramètres permettant de mettre en évidence leurs différences. Une première stratégie de combinaison est alors proposée. La section 4.4 relate l'étude de l'ambiguïté des symboles phonétiques. L'influence du modèle acoustique est séparée de celle des paramètres acoustiques par l'intermédiaire d'une architecture spécifique basée sur le modèle de « canal de transmission ».

4.2 Matériel expérimental

4.2.1 Jeux de paramètres.

Deux jeux de paramètres acoustiques ont été exploités. Ils correspondent à des analyse du signal fondées sur des algorithmes différents. Le premier jeux de paramètres correspond à l'analyse à résolution multiple MRA (voir section 2.5) suivie d'une analyse en composantes principales. (*Principal Component Analysis* - PCA). Le deuxième jeu de paramètres correspond aux paramètres cepstraux PLP suivie par le filtrage J-RASTA (voir section 2.4), que nous appellerons RPLP.

La même technique de débruitage est appliquée pour chaque jeu de paramètres acoustiques (Gemello et al., 2004). Le débruitage est une composante essentielle pour le calcul des paramètres acoustiques. Il est basé sur la soustraction spectrale non-linéaire et peut, lorsque le rapport signal sur bruit (*Signal to Noise Ratio* - SNR) est faible, modifier substantiellement les paramètres acoustiques.

4.2.2 Modèles acoustiques.

Deux types de modélisation ont été utilisées lors des expériences. Le premier correspond au modèle hybride ANN/HMM (voir section 1.2.3) de Loquendo, repéré par l'indice A dans les expériences. Il a été entraîné sur un corpus de phrases phonétiquement équilibrées complètement indépendant des corpus d'Aurora3. Les données d'apprentissage sont issues d'un corpus de parole téléphonique classique. Le réseau de neurones possède 636 sorties, une pour chaque phonème et chaque transition entre deux phonèmes successifs. Une description de l'architecture globale est fournie dans Gemello et al. (1997, 1999).

Un modèle ANN a été entraîné pour chaque type de paramètres acoustiques. On appellera ANN.MRA le modèle utilisant les paramètres MRA et ANN.RPLP le modèle utilisant les paramètres RPLP.

Le second modèle est un ensemble de GMMs, chacun correspondant à un phonème. Il sera repéré par l'indice G dans les équations. Les GMMs ont été construits à partir d'un GMM "*du monde*" appris sur le corpus d'apprentissage d'Aurora3. Ensuite, les modèles de phonèmes ont été générés en adaptant le modèle du monde selon un critère de MAP (cf. section 1.5.1). Cet ensemble de modèles permet de générer un vecteur de probabilités des phonèmes étant donné un segment de parole. Les modèles appris avec les deux jeux de paramètres seront notés GMM.MRA et GMM.RPLP.

4.2.3 Description des corpus

Les corpus utilisés proviennent de l'ensemble Aurora3 (Pearce et Hirsch, 2000). Les corpus d'Aurora3 sont des corpus à petit vocabulaire (les dix chiffres) de différentes langues (ici italien et espagnol) enregistrés dans des automobiles en fonctionnement. Leurs caractéristiques sont présentées dans le tableau 4.1. Chacun de ces corpus est divisé en trois parties. Le corpus d'entraînement

	Langue	Corpus	Type	# phrases	# mots
AURORA	Italien (ITA)	TRAIN	CH0	1466	9082
			CH1	1485	9288
		DEV	CH0	664	4133
			CH1	645	3927
		TEST	CH1	626	3810
	Espagnol (SPA)	TRAIN	CH0	1696	9167
			CH1	1696	9167
		DEV	CH0	761	4028
			CH1	761	4028
		TEST	CH1	631	3325

TAB. 4.1: *Description des corpus d'Aurora3 (CH0 : partie non bruitée, CH1 : partie bruitée). Les phrases sont des suites de mots isolés.*

(TRAIN) sert à entraîner les GMMs. Il ne sert pas à l'apprentissage des réseaux de neurones, par conséquent il a été utilisé pour certaines analyses n'utilisant pas les GMMs. Le corpus de développement (DEV) est utilisé pour les différentes évaluations nécessaires. Le corpus de test (TEST) est utilisé pour les expériences de reconnaissance de la parole.

L'ensemble CH0 correspond aux données non bruitées, enregistrées avec un microphone personnel (positionné très près du locuteur), ce qui permet de capter moins de bruit environnant. Les données CH1 ont été enregistrées avec un microphone captant beaucoup de bruits comme le bruit de la fermeture ou de l'ouverture des vitres électriques, le bruit du moteur à l'arrêt, ...

4.3 Comparaison de différents jeux de paramètres

L'analyse comparative de plusieurs jeux de paramètres nécessite de comparer des paramètres représentant l'information extraite sur les mêmes portions de signal. Un signal d'entrée échantillonné S est composé d'une suite d'échantillons $\{s(k\tau)\}$, où τ est la période d'échantillonnage. Comme le signal de pa-

role montre une grande variabilité pour la même phrase, les séquences d'échantillons sont transformées en vecteurs de paramètres acoustiques, plus stables.

Les paramètres acoustiques de types différents sont calculés à partir de séquences d'échantillons (aussi appelées fenêtres) qui peuvent être de tailles différentes. Considérons la séquence d'échantillons dans une fenêtre temporelle de durée T et représentons une telle séquence pour la $n^{\text{ième}}$ fenêtre par $Y_n = [s(k\tau)]_{nT}^{(n+1)T}$ avec $n = 0, ..., N$. Pour chaque valeur de n, la séquence d'échantillons Y_n du signal est transformée en un vecteur de paramètres Y_n^i représenté par un point dans un espace de paramètres acoustiques $\Im^i$.

Différents types de paramètres acoustiques sont considérés synchrones s'ils sont calculés sur des fenêtres centrées sur le même échantillon. Cela nous assure qu'ils partagent une grande proportion d'échantillons et établit ainsi une correspondance entre leurs espaces acoustiques respectifs.

4.3.1 Analyse comparative des performances

La figure 4.2 montre l'architecture de diagnostic permettant d'exploiter et de comparer les probabilités obtenues avec les différents jeux de paramètres acoustiques.

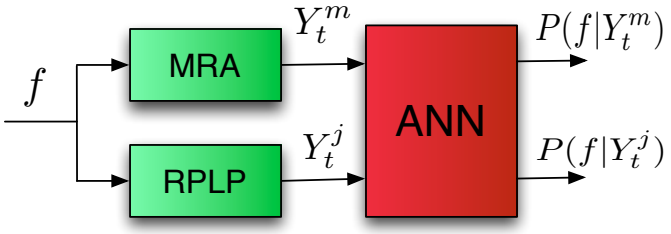

FIG. 4.2: *Architecture de diagnostic pour la comparaison de deux jeux de paramètres différents avec un type de modélisation (ANN).*

Les symboles f à l'entrée correspondent aux phonèmes de chaque langue (sans contexte ni transitions). $P(f|Y_t^m)$ est la probabilité *a posteriori* du phonème f étant donné la trame de paramètres MRA Y_t^m calculée au temps t. $P(f|Y_t^j)$ correspond à la probabilité calculée avec le modèle utilisant les paramètres RPLP.

Considérons l'espace dont les coordonnées sont définies par $P(f|Y_t^m)$ et $P(f|Y_t^j)$. Un tel espace peut être partitionné comme dans la figure 4.3. On peut alors comptabiliser l'occupation des zones de cet espace pour chaque paire de probabilités. Les coordonnées de chaque point de ces zones représentent les probabilités obtenues avec les deux modèles acoustiques.

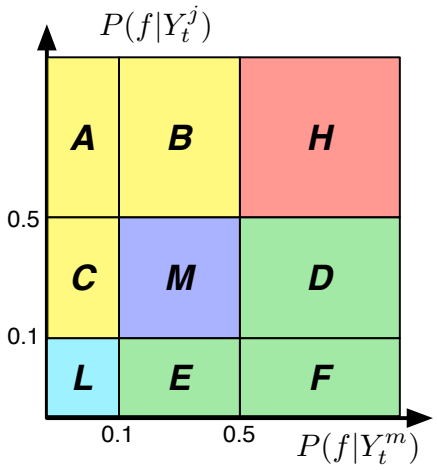

FIG. 4.3: *Partition de l'espace des probabilités a posteriori.*

La **zone H** représente les cas dans lequel le phonème f est l'hypothèse la plus probable pour les deux jeux de paramètres (probabilité supérieure à 0.5). Dans ce cas, le phonème devrait être correct puisque deux jeux de paramètres différents proposent le même phonème.
La **zone M** représente les cas dans lesquels les paramètres montrent une possibilité pour le phonème f mais que cette décision n'est pas vraiment très sûre. En effet, il se peut qu'un ou plusieurs autres phonèmes aient une probabilité relativement grande, ce qui diminue la confiance que l'on peut apporter à l'hypothèse.
La **zone L** correspond aux cas dans lesquels sélectionner le phonème f revient à faire un choix aléatoire parmi des candidats (donc très peu fiable).
Les **zones A**, **B**, et **C** correspondent aux zones où les paramètres RPLP sont plus confiants que les paramètres MRA, et les **zones D**, **E** et **F** correspondent au contraire.

Une première expérience a été menée avec les données bruitées (CH1) des corpus ITA.TRAIN et SPA.TRAIN. Les modèles ANN.MRA et ANN.RPLP ont été utilisés pour cette expérience. Il est important de souligner que le corpus Aurora3 n'a pas servi à entraîner les réseaux de neurones.

Des statistiques sur la répartition des points définis par les probabilités *a posteriori* dans les différentes zones décrites ci-dessus sont accumulées. L'objectif est de mettre en évidence les possibles confusions dues à l'inadéquation des paramètres ou à l'imperfection des modèles. Les résultats groupés par classes de phonèmes sont présentés dans le tableau 4.3.

Les classes sont constituées des phonèmes suivants :

Ces résultats sont obtenus en effectuant un **alignement forcé** des données

Langue	Classe	Phonèmes
Espagnol	Consonnes sourdes	t&, s, t, k, dz, d, v
	Consonnes sonores	R, n
	Voyelles	a, e, i, o, u
Italien	Consonnes sourdes	s, t, k, B, T, d, tS
	Consonnes sonores	r, n, j, w, N
	Voyelles	a, e, i, o, u

TAB. 4.2: *Composition des classes de phonèmes.*

du corpus d'entraînement pour chaque flux de paramètres acoustiques. L'alignement forcé consiste à trouver les bornes des phonèmes effectivement prononcés qui maximisent la probabilité *a posteriori* globale de la phrase étant donné le jeu de paramètres acoustiques et le modèle acoustique.
Seules les trames incluses dans les intersections des intervalles proposés par les deux systèmes pour le même phonème f sont utilisées.
La moyenne des probabilités postérieures pour chaque segment et pour chaque jeu de paramètres est calculée et la paire de valeurs est représentée par une étiquette de la figure 4.3.

Langue	Étiquette	Consonnes sourdes	Consonnes sonores	Voyelles
Espagnol	H	40,36%	63,94%	66,4%
	M	2,31%	1,79%	1,53%
	L	21,55%	7,08%	3,64%
	D∪E∪F	20,74%	13,39%	23,12%
	A∪B∪C	15,05%	13,80%	5,3%
Italien	H	37,54%	61,04%	58,02%
	M	8,29%	3,87%	7,01%
	L	17,67%	5,86%	7,54%
	D∪E∪F	20,49%	18,14%	12,59%
	A∪B∪C	16,01%	11,09%	14,85%

TAB. 4.3: *Distributions des zones de probabilité pour les données CH1 des parties espagnole (37933 pho.) et italienne (34043 pho.) du corpus Aurora3.*

Un premier résultat est que les deux jeux de paramètres fournissent des probabilités postérieures différentes en fonction des classes phonétiques. Il apparaît clairement que les phonèmes sourds sont plus difficiles à reconnaître que les phonèmes sonores. En effet, pour la colonne 3, une assez grande quantité de données se trouve dans la zone L. Une explication est que les occlusives ont une durée courte et leurs caractéristiques ne sont pas forcément bien captées par le module d'extraction de paramètres.

Les voyelles sont peu affectées par le débruitage puisque, de manière générale, leurs segments ont un SNR assez élevé. De plus, leur durée est assez longue, ce qui rend leur reconnaissance plus facile. Ceci peut être observé dans la cinquième colonne du tableau, où la majorité des données est classée dans la zone H.

Le résultat le plus étonnant est que, dans beaucoup de cas, les probabilités obtenues avec un jeu de paramètres sont hautes alors qu'elles sont basses avec l'autre jeu. En effet, une grande quantité de phonèmes est étiqueté D,E ou F. Cela signifie que les paramètres MRA ont de meilleures performances, et sont plus fiables pour représenter les phonèmes.

Il est peu probable que ce résultat soit juste dû aux limitations des techniques de modélisation. On observe de nombreux cas pour lesquels un modèle fournit une probabilité supérieure à 0.5, signifiant que le phonème a été bien reconnu, alors que l'autre fournit une probabilité inférieure à 0.1, signifiant le contraire (zones A et F).

Deux modélisations identiques utilisant deux jeux de paramètres différents fournissent des résultats différents pour un même phonème. On peut donc supposer qu'une partie des divergences (cas A, B, D et F) est probablement due aux variabilités intrinsèques des paramètres acoustiques qui rendent difficile la bonne estimation des distributions de probabilité.

Le nombre de cas dans les zones D,E et F est généralement plus élevé que dans les zones A,B et C, ce qui montre que les phonèmes proposés par MRA sont plus fréquemment corrects que ceux proposés par RPLP. Ce résultat se retrouve dans les performances de reconnaissance sur ce corpus, puisque MRA obtient des résultats légèrement meilleurs que RPLP.

4.3.2 Consensus entre les hypothèses de phrase

Des statistiques sur les cas de consensus ont été accumulées. Cette étude nous permet d'évaluer dans quelle mesure les systèmes sont différents en comparant les hypothèses qu'ils fournissent.

Différents ensembles (Q0, Q1, Q2 et Q3) correspondant à différents degré de consensus ont été définis comme suit : Q0 regroupe les phrases pour lesquelles les deux systèmes donnent la même hypothèse (consensus sur la phrase). Q1 regroupe les phrases pour lesquelles les systèmes proposent des hypothèses ayant un mot différent, Q2 celles ayant deux mots différents et Q3 celles ayant plus de deux mots différents[1].

[1]La couverture ne somme pas à 100% puisque 14 phrases (180 mots) ont été rejetées à cause

Les hypothèses générées par les systèmes sur la partie bruitée du corpus SPA.TEST ont été alignées. Pour chaque ensemble, la couverture et le nombre de mots sur lesquels les deux reconnaisseurs sont en accord ont été calculé. Les résultats sont résumés dans le tableau 4.4.

	Couv. (# mots)	Corrects	Couverture cumul.	Erreur cumul.
Phrases de Q0	1491	1479	1491 (44.84%)	16 (1.07%)
Phrases de Q1	452	440	1937 (58.25%)	28 (1.46%)
Phrases de Q2	265	261	2198 (66.10%)	32 (1.46%)
Phrases de Q3	715	662	2860 (86.01%)	85 (2.97%)

TAB. 4.4: *Taux d'erreur et consensus. Couverture en mots, nombre de mots corrects, couverture cumulée et taux d'erreur cumulé en %.*

Les phrases pour lesquelles il y a consensus couvrent 410 phrases sur les 631, ce qui correspond à une couverture en mots de 44.8% (1491 mots sur les 3325 au total). Parmi ces 410 phrases, seules 16 sont erronées et ne contiennent pas plus d'une erreur par phrase. Cet ensemble affiche un WER de 1.07%.

On observe que le taux d'erreur augmente progressivement en fonction de la couverture. Il en ressort que le degré d'accord des résultats de reconnaissance générés par des systèmes utilisant des paramètres acoustiques différents semble être un bon indice de confiance. De plus, lorsque les divergences entre les résultats sont peu nombreuses, il est possible d'appliquer des critères de décision permettant de sélectionner la bonne hypothèse si elle est fournit par au moins l'un des systèmes. Ce principe est présenté dans la section suivante, où l'utilisation des résultats obtenus est appliqué à une tâche de reconnaissance (également avec un corpus petit vocabulaire).

Divergences au niveau des mots

Lorsqu'il n'y a pas consensus entre les hypothèses de phrases générées par les deux reconnaisseurs, l'alignement de celles-ci est effectué. Les expériences décrites dans Gemello et al. (2004) montrent que le système utilisant les paramètres MRA a des meilleures performances que celui utilisant les paramètres RPLP. Les hypothèses qu'il propose sont donc prises pour référence lors de l'alignement des résultats.

Soit $w_m(b,e)$, un mot ou une séquence de mots proposé par le système MRA dans l'intervalle de temps commençant au temps b et finissant au temps e. Soit $w_j(b_j,e_j)$, la séquence de mots proposée en compétition par le système RPLP dans un segment de temps (b_j,e_j) chevauchant fortement l'intervalle (b,e).

de la trop mauvaise qualité de l'alignement des hypothèses.

Considérons maintenant les situations définies dans la table 4.5 décrivant les divergences possibles entre les reconnaisseurs.

MRA : $w_m(b,e)$	RPLP : $w_j(b_j,e_j)$	TYPE DE DIVERGENCE
w_i	w_k	substitution (sub)
w_i	$sil.$	délétion (del)
$sil.$	w_k	insertion (ins)
w_i	$w_i w_k$	j-insertion (i_j)
$w_i w_k$	w_i	m-insertion (i_m)
w_i	$w_q w_k$	substitution + j-insertion (si_j)
$w_q w_k$	w_i	substitution + m-insertion (si_m)
Tous les autres cas		plusieurs divergences (md_{mj})

TAB. 4.5: *Divergences possibles entre les sorties de reconnaisseurs différents*

Les différents cas possibles pour un intervalle (b,e) sont définis de la manière suivante :

1. substitution (sub) : un seul mot w_k chevauche fortement w_i,
2. délétion (del) : une hypothèse de silence (segment non parlé) est générée par RPLP.
3. insertion (ins) : une hypothèse de mot w_k est générée dans un intervalle de temps où MRA a généré une hypothèse de silence.
4. j-insertion : RPLP a inséré une hypothèse de mot w_k dans l'intervalle de temps (b,e).
5. m-insertion : même chose que précédemment, mais les rôles sont inversés.
6. substitution + j-insertion : RPLP propose deux mots différents dans (b,e).
7. substituion + m-insertion : MRA a inséré une hypothèse de mot dans l'intervalle correspondant à un mot.

4.3.3 Stratégie de décision

La stratégie de décision suivante se base sur les divergences entre les systèmes. L'alignement des meilleures séquences de mots générées par les deux systèmes de reconnaissance est effectué. Si les deux systèmes proposent la même hypothèse, alors elle est validée et conservée. Si ce n'est pas le cas, alors les hypothèses de mot $w_m(b,e)$ et $w_j(b,e)$ sont soumises à validation.

Soit f_w la séquence des N phonèmes composant le mot w (f_w : $[f_1, \ldots, f_n, \ldots, f_N]$). Pour chaque phonème f_n, son compétiteur[2] est identifié. Les probabilités *a posteriori* de chaque phonème sont utilisées pour désigner le symbole σ_n correspondant à une zone de l'espace décrit dans la figure 4.3 (p. 85). Notons σ :$[\sigma_1, \ldots, \sigma_n, \ldots, \sigma_N]$ la séquence d'étiquettes pour le mot w.

La règle de décision consiste à sélectionner le jeu de paramètres dont l'hypothèse de mot a la plus grande probabilité d'être correct étant donné la séquence d'étiquettes σ associée au mot w.

La formulation mathématique est la suivante :

$$\begin{aligned} w^* &= \underset{a \in m,j}{\operatorname{argmax}} P(w_{ref} = w_a | w_m, w_j, \sigma) \\ &= \underset{a \in m,j}{\operatorname{argmax}} P(\sigma | w_{ref} = w_a, w_m, w_j) P(w_{ref} = w_a | w_m, w_j) \end{aligned} \tag{4.1}$$

Les probabilités $P(\sigma | w_{ref} = w_a, w_m, w_j)$ et $P(w_{ref} = w_a | w_m, w_j)$ sont déterminées à partir d'un corpus développement.

Expériences de reconnaissance automatique de la parole

Des expériences de reconnaissance ont été menées sur les corpus ITA.TEST et SPA.TEST. Seules les données bruitées (CH1) ont été utilisées. Les systèmes de reconnaissance utilisent les modèles ANN.MRA et ANN.RPLP.

La probabilité $P(w_{ref} = w_a | w_m, w_j)$ a été calculée à partir du corpus de développement d'Aurora3 et utilisée pour sélectionner une hypothèse parmi celles proposées. La règle de décision appliquée correspond à l'équation 4.2.

Les résultats en terme de WER sont présentés dans le tableau 4.6. La sélec-

Performances	Italien	Espagnol
Système MRA	20.34%	15.19%
Nouvelle strat. - sub	6.2%	5.71%
Nouvelle strat. - del	9.79%	1.99%
Nouvelle strat. - ins	1.57%	4.72%
Nouvelle strat. - WER global	17.56%	12.42%

TAB. 4.6: *Performances, en terme de WER, de la nouvelle stratégie de décision comparée au meilleur système*

[2]Le compétiteur est le phonème proposé par l'autre système de reconnaissance qui a le plus grand nombre de trames en commun avec f_n).

tion d'hypothèse fondée sur la comparaison de deux jeux de paramètres a permis d'obtenir une réduction relative du WER de 13.67% pour l'italien et 18.24% pour l'espagnol.

Conclusion

L'analyse de la variabilité de deux jeux de paramètres différents a été effectuée. Certaines zones pour lesquelles un jeu de paramètres fournit des probabilités de phonèmes plus fiables que l'autre ont été mises en évidence.

L'analyse comparative des deux systèmes fournit une bonne mesure de confiance pouvant être utilisée directement dans un système de reconnaissance. Une stratégie utilisant ces résultats a été développée. Cette stratégie utilise une mesure de confiance combinant le consensus et un prédicat permettant de sélectionner l'hypothèse la plus probable. Cette mesure de confiance a permis une réduction significative du taux d'erreur mot sur un corpus bruité à petit vocabulaire.

D'un point de vue qualitatif, il est important de bien distinguer les situations dans lesquelles les systèmes de reconnaissance sont unanimes de celles dans lesquelles ils ne le sont pas. Considérons les segments de parole pour lesquels les systèmes proposent la même hypothèse de phonème. Si ce phonème est erroné, alors cela est probablement du à l'imperfection des modèles ou aux limites des stratégies de décodage. En effet, si deux systèmes de reconnaissance, dont la seule différence réside dans les jeux de paramètres, produisent la même hypothèse de phonème erronée, alors il est fort probable que l'erreur ne provienne pas de la paramétrisation.

4.4 Analyse de la confusion introduite par les paramètres acoustiques

On peut représenter schématiquement un système de reconnaissance comme un **canal de transmission** représenté dans la figure 4.4. Une source S génère des symboles d'entrée, f, représentés par des séquences de vecteurs de paramètres acoustiques. Ces vecteurs acoustiques sont ensuite traités par le système de reconnaissance qui génère des symboles g au niveau du récepteur. Les symboles d'entrée et de sortie appartiennent à un vocabulaire Q, *i.e.* $f \in Q$ et $g \in Q$. Par la suite, nous considérerons que Q est l'ensemble des phonèmes d'une langue.

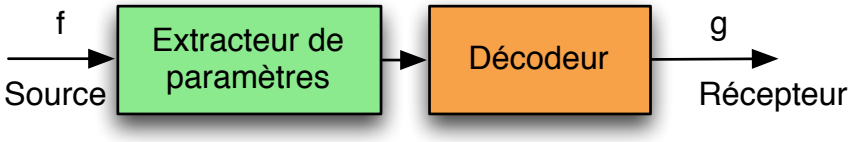

FIG. 4.4: *Canal de transmission schématisant un système de reconnaissance.*

Dans la théorie de l'information de Shannon (1948), le taux de transmission d'information est décrit comme la différence entre deux incertitudes.

$$R = H(S) - H_R(S) \tag{4.2}$$

avec $H(S)$ l'entropie de la source, et $H_R(S)$ l'entropie conditionnelle moyenne du récepteur étant donné la source. Shannon a nommé cette entropie conditionnelle moyenne « equivocation », terme que nous utiliserons par la suite.

Les erreurs introduites par le système sont donc la cause d'equivocation définie par l'entropie conditionnelle moyenne comme suit :

$$H_R(S) = -\sum_{f,g} P(f,g) \log_2 P(g|f) \tag{4.3}$$

Cette equivocation mesure le degré d'incertitude observable en sortie du canal de transmission. Chaque symbole g participe plus ou moins à l'augmentation de l'incertitude.

Il est important de noter que l'equivocation dépend du critère de décision pris pour reconnaître un phonème. En effet, si ce critère n'est pas purement le choix du phonème ayant la plus grande probabilité *a posteriori*, alors connaître la probabilité *a posteriori* avec laquelle un phonème transmis est reçu n'est pas suffisant pour déterminer le phonème g qui sera effectivement sélectionné par le système. En pratique, les systèmes de reconnaissance utilisent de l'information lexicale et linguistique et plus de l'information acoustique pour déterminer le symbole de sortie. Dans notre cas, l'hypothèse générée par le système correspond au symbole (phonème) pour lequel le modèle acoustique fournit le maximum de probabilité. Nous n'utilisons pas de modèle de langage dans cette architecture.

4.4.1 Equivocation en fonction de zones de l'espace acoustique

Une zone d'un espace acoustique $\Im^i$ est peuplée par des trames de paramètres acoustiques Y_n^i issues des transformations des séquences d'échantillons

du signal de parole. Soit $\{Y_n^i\}$ un ensemble de trames appartenant à la zone z^i de l'espace acoustique $\Im^i$. Ce même ensemble de trames peut être transformé en $\{Y_n^j\}$ dans l'autre espace acoustique $\Im^j$. Les trames appartiennent alors la zone z^j. Il y a donc correspondance entre les zones z^i et z^j, où un représentant de la première zone correspond à un et un seul représentant de l'autre zone.

L'objectif de cette analyse est d'identifier d'éventuelles zones affichant une confusion différente pour les jeux de paramètres acoustiques analysés.

Pour ce faire, la relation entre la confusion et certaines zones de l'espace acoustique doit être mise en évidence. Il y a deux difficultés majeures à faire cela.
La première difficulté vient du fait que la qualité de l'estimation des probabilités dépend du type et de la quantité des données utilisées pour le calcul. Il est difficile d'avoir suffisamment de données pour toutes les zones d'un espace acoustique. La fiabilité des mesures dans ces zones peut être faible. Aussi, un compromis doit être fait pour mesurer la confusion moyenne dans une zone suffisamment grande de l'espace acoustique. Par la suite, nous ne considérerons que les phonèmes non contextuels de manière à assurer un nombre suffisant de représentants pour chaque classe.

La seconde est que le calcul de la confusion requiert des probabilités qui sont calculées en utilisant des modèles acoustiques. Par conséquent, la qualité des modèles a un fort impact sur le résultat. La séparation des contributions à l'ambiguïté dues aux modèles de celles dues aux paramètres acoustiques est donc difficile. Une solution à ce problème consiste à utiliser plusieurs modèles acoustiques différents et faire des considérations spécifiques selon qu'ils produisent des distributions de probabilité similaires ou non.

Canal de transmission pour le calcul de l'equivocation

Afin de séparer l'influence des modèles acoustiques sur l'equivocation de celle des paramètres acoustiques, plusieurs modélisations acoustiques sont utilisées. Si, en utilisant plusieurs modélisations différentes avec le même jeu de paramètres acoustiques, on obtient des distributions de probabilité postérieure très similaires pour les phonèmes, on peut alors en déduire que cette confusion est principalement due à la variabilité des paramètres et non aux modèles acoustiques.

L'architecture pour mesurer l'equivocation dans ce cas est présentée dans la figure 4.5. Une source génère un signal audio codant le symbole f. Ce signal de parole est traité par le module d'extraction de paramètres acoustiques qui fournit une suite de vecteurs acoustiques correspondant aux modèles acoustiques.

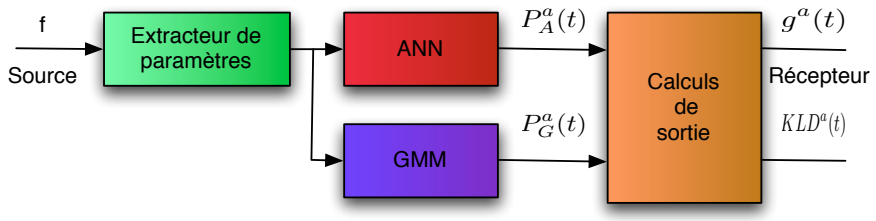

FIG. 4.5: *Architecture pour le calcul de l'entropie conditionnelle moyenne (equivocation) avec deux modèles différents utilisant un jeu de paramètres acoustiques unique.*

Dans notre cas, les modèles correspondent à des manières très différentes de modéliser l'espace acoustique. Les différentes modélisations acoustiques ANN et GMM ont été utilisées. Ces modélisations génèrent des distributions de probabilité à partir des trames qui leur sont données en entrée. Ces distributions sont ensuite traitées par le module de calcul de sortie. Les probabilités fournies par les deux modèles acoustiques sont fusionnées (produit des probabilités) pour déterminer l'hypothèse de phonème, $g^a(S_t)$, du système global.

4.4.2 Analyse comparative de la variabilité

Les deux flux de paramètres acoustiques sont indiqués par m et j et font respectivement référence à MRA et RPLP. Les vecteurs Y_n^m et Y_n^j représentent deux observations différentes de la $n^{\text{ième}}$ trame de parole Y_n. Ces vecteurs ont été calculés sur des segments centrés sur le même échantillon afin que les flux soient synchrones.

La variabilité des paramètres acoustiques est décrite par l'intermédiaire des probabilités *a posteriori*. Considérons un segment dans lequel un phonème a été reconnu par un système. Notons b la trame de début, e la trame de fin, et t la trame centrale. Les paramètres MRA extraits dans ce segment sont représentés par une séquence de vecteurs acoustiques : $Y_{b,e}^m = \{Y_b^m, \ldots, Y_t^m, \ldots, Y_e^m\}$. Nous indiquerons un tel segment par S_t.

Les distributions de probabilités *a posteriori* des phonèmes g connaissant S_t, $P_A^a(g|S_t)$ et $P_G^a(g|S_t)$, sont obtenues avec les modèles acoustiques utilisant le jeu de paramètres $a \in \{m, j\}$.

L'écart entre les deux distributions de probabilité est mesurée par leur divergence de Kullback-Leibler (*Kullback-Leibler Divergence* - KLD), définie comme suit :

$$KLD^a(S_t) = \sum_{g \in Q} P_A^a(g|S_t) \log \frac{P_A^a(g|S_t)}{P_G^a(g|S_t)} \tag{4.4}$$

Le symbole avec la plus grande probabilité *a posteriori* est considéré comme l'hypothèse générée par un modèle acoustique dans le segment S_t. Nous nommerons respectivement ces hypothèses $g_A^a(S_t)$ et $g_G^a(S_t)$.

Afin de réduire les effets des erreurs de segmentation et afin de se concentrer sur les effets de la variabilité des paramètres acoustiques, l'alignement forcé du corpus de développement a été effectué. L'alignement forcé consiste à trouver les bornes des phonèmes réellement prononcés, de manière à maximiser la probabilité *a posteriori* globale. La $KLD^a(S_t)$ est ensuite calculée et comparée à un seuil variable X. L'equivocation $E(X) = -\sum_{f,g} P(f,g) \log P(g|f)$ est mesurée sur tous les segments pour lesquels $KLD^a(S_t) < X$. En faisant varier X, on obtient une relation entre l'ambiguïté et le seuil de KLD X.

Résultats et observations

Une expérience initiale a été effectuée sur les phonèmes du corpus ITA.TEST. Les modèles ANN et GMM utilisant les paramètres MRA et RPLP ont été utilisés.

La figure 4.6 montre la relation entre l'ambiguïté et le seuil de KLD, X, pour les deux jeux de paramètres acoustiques et pour les deux conditions d'enregistrement (CH0 pour le signal propre et CH1 pour le signal bruité).

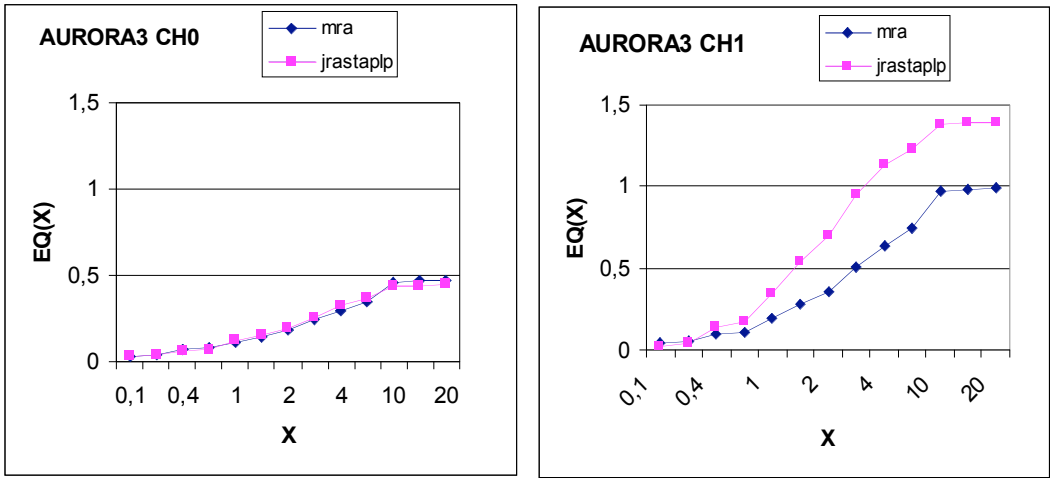

FIG. 4.6: *Comparaison des equivocations obtenues avec les paramètres MRA et RPLP pour le corpus ITA.TEST d'Aurora3 dans des conditions propres (CH0) et bruitées (CH1).*

On observe que les paramètres acoustiques se comportent différemment se-

lon les conditions d'enregistrement. Lorsque les données ont un SNR élevé (partie CH0), l'equivocation évolue de la même manière en fonction de la KLD, en dépit du fait que les paramètres acoustiques MRA et RPLP soient calculés avec des algorithmes très différents. Ceci tend a montrer que les deux jeux de paramètres sont globalement équivalent sur tout l'espace acoustique.

On peut remarquer des différences lorsque les données ont un SNR faible (partie CH1). Dans ce cas, il apparaît clairement que les paramètres MRA sont plus adéquats lorsque la KLD est grande (supérieure à 0.4). Cette différence peut s'expliquer par le fait que les paramètres acoustiques calculés sur les données bruitées sont considérablement modifiés par le débruitage. Le débruitage est appliqué sur toutes les sorties des filtres de l'arbre de l'analyse MRA. Or, le nombre de filtres présents dans l'arbre de la MRA est supérieur au nombre de filtres utilisés pour RPLP. Le débruitage plus fin opéré sur les paramètres MRA pourrait expliquer le fait qu'ils fournissent de meilleures performances dans ce cas.

Ce résultat nous permet de prédire qu'un jeu de paramètres aura de meilleures performances qu'un autre selon le niveau du SNR. Ces résultats sont en accord avec les résultats décrits dans Gemello et al. (2006), dans le sens où les performances obtenues avec MRA et RPLP sont très similaires sur les données propres alors que pour les données bruitées, MRA fournit des résultats significativement meilleurs que RPLP.

4.4.3 Effets du consensus entre les modélisations

Considérons le prédicat $CONS^a(S_t)$ défini par $\{g_A^a(S_t) = g_G^a(S_t)\}$, désignant le consensus entre les deux modélisations acoustiques utilisant les mêmes paramètres acoustiques sur le segment S_t. Si $CONS^a(S_t)$ est vrai, alors le phonème considéré comme reconnu au niveau du récepteur est $g^a(S_t) = g_A^a(S_t) = g_G^a(S_t)$. Si le prédicat n'est pas vérifié, alors c'est l'hypothèse ayant la plus grande probabilité obtenue par combinaison log-linéaire des probabilités *a posteriori* qui est choisie.

L'analyse de l'ambiguïté introduite par les paramètres acoustiques a été effectuée. Elle consiste à mettre en évidence le lien entre la divergence des distributions de probabilités obtenues avec 2 modèles acoustiques différents utilisant le même jeu de paramètres.

Les résultats obtenus sur la partie bruitée d'Aurora3 italien sont présentés dans la figure 4.7. Ils ont été regroupés en fonction du prédicat $CONS^a(S_t), a \in \{m, j\}$

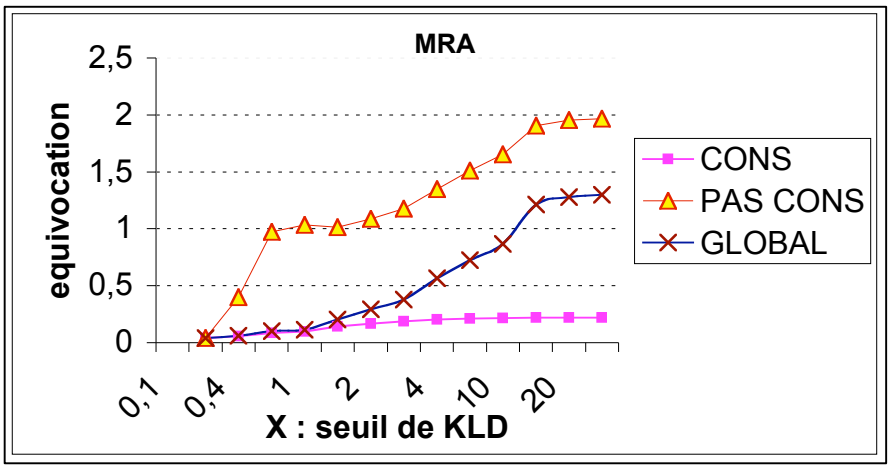

FIG. 4.7: *L'ambiguïté en tant que fonction d'intervalles de KLD quand les paramètres MRA sont utilisés. La courbe « CONS » correspond aux cas où le prédicat $CONS^m(S_t)$ est vrai, « PAS CONS » à ceux lorsqu'il est faux et « GLOBAL » à la totalité des cas.*

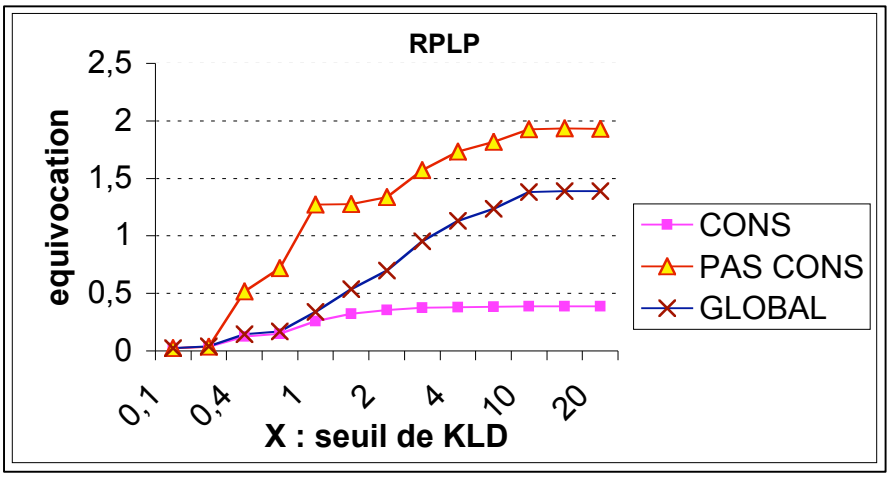

FIG. 4.8: *L'ambiguïté en tant que fonction d'intervalles de KLD quand les paramètres RPLP sont utilisés. La courbe « CONS » correspond aux cas où le prédicat $CONS^j(S_t)$ est vrai, « PAS CONS » à ceux lorsqu'il est faux et « GLOBAL » à la totalité des cas.*

On remarque que lorsque le prédicat $CONS^a(S_t)$ est vrai, l'ambiguïté est très faible et n'augmente pas beaucoup avec la KLD.

Il faut noter qu'une valeur faible de $KLD^a(S_t)$ ne signifie pas forcément que les deux distributions de probabilité donnent le maximum de probabilité au même symbole. En effet, deux distributions de probabilité très proches peuvent donner un maximum de probabilité pour des phonèmes différents, même si cela est relativement peu fréquent. Une explication de cela est que les paramètres peuvent afficher une certaine variabilité dans cette zone de l'espace acoustique, ce qui perturbe les modèles acoustiques.

Les principaux résultats des expériences effectuées sur les corpus de test d'Aurora3 sont résumés dans le tableau 4.7. Les abréviations suivantes sont utilisées :

- LV : $CONS^a(S_t) \wedge \{KLD^a(S_t) < 0.4\}$ est vrai
- H(S) : l'entropie de la source en bits
- Eq : l'equivocation en bits
- total Eq. : equivocation du corpus en entier
- Couv. : pourcentage de couverture

Corpus	Paramètres	H(S)	total Eq.	Eq. $\wedge$ LV	Couv.
ITA.TEST CH0	MRA	3.58	0.47	0.083	88.0
	RPLP	3.58	0.45	0.091	88.5
ITA.TEST CH1	MRA	3.58	1.03	0.14	55.2
	RPLP	3.58	1.39	0.20	45.3
SPA.TEST CH1	MRA	3.82	1.26	0.13	47.4
	RPLP	3.82	1.48	0.12	32.1

TAB. 4.7: *Entropie de la source, equivocation et couverture pour les corpus de test d'Aurora3.*

Lorsque les distributions de probabilité *a posteriori* de phonèmes, obtenues avec des modèles différents utilisant les mêmes paramètres acoustiques sont similaires ($KLD^a(S_t)$ faible), alors la confusion est principalement due à la variabilité des paramètres acoustiques. En effet, si deux modélisations acoustiques très différentes proposent des distributions de probabilité très proches, il est peu probable que les erreurs commises dans ce cas soient dues au fait que la zone soit mal modélisée.

Le cas échéant, cela voudrait dire que les deux modèles ont des lacunes à représenter la zone de l'espace acoustique en question.

Quand les distributions sont différentes ($KLD^a(S_t)$ élevée) et qu'il y a consensus entre les hypothèses de phonèmes, alors on peut en déduire que l'ambiguïté est principalement due à la variabilité des paramètres acoustiques. Dans ce cas, deux modélisations différentes utilisant le même jeu de paramètres

fournissent la même hypothèse erronée. Il est alors peu probable que ces erreurs soient dues au fait que la zone soit mal modélisée.

Lorsque les distributions de probabilité *a posteriori* des phonèmes conduisent à une grande valeur de $KLD^a(S_t)$, alors les causes peuvent être multiples.

Les deux modèles donnent la probabilité maximum pour le phonème prononcé. Dans ce cas, on peut dire que les traits acoustiques affichent une variabilité certaine qui se répercute de manière différente dans les deux modélisations. En effet, les deux modèles ont donné une bonne hypothèse malgré le fait que les probabilités aient été distribuées différemment entre les phonèmes, et ce en utilisant les mêmes vecteurs acoustiques. On peut donc dire que la variabilité contenue dans les vecteurs de paramètres acoustiques a provoqué une moins bonne discrimination des symboles phonétiques de la part des modèles acoustiques. Au final, les hypothèses générées sont correctes, mais avec une confiance moindre. On peut associer ce résultat à la robustesse des modèles.

L'un des deux modèles donne la probabilité maximum pour le bon phonème. Dans ce cas, on peut juste dire que l'un des modèles caractérise bien la zone de l'espace acoustique correspondante et l'autre non. La caractérisation de la variabilité ne peut se faire correctement étant donné que la séparation de l'influence du modèle et des jeux de paramètres n'est pas possible dans ce cas.

Aucun jeu ne donne une bonne hypothèse. Soit les deux modèles ne caractérisent pas bien la zone correspondant aux vecteurs acoustiques, soit les vecteurs contiennent une grande variabilité et ne correspondent pas à un phonème normalement présent dans cette zone (ou les deux).

Conclusion

Des analyses permettant de comparer plusieurs jeux de paramètres et plusieurs types de modélisation acoustiques ont été effectuées. D'une part, la variabilité des paramètres acoustiques a été mise en évidence par l'intermédiaire de l'equivocation mesurant la confusion introduite par les jeux de paramètres. D'autre part, l'accord entre différents modèles acoustiques permet d'isoler des zones dans lesquelles l'equivocation est faible.

Globalement, les deux jeux de paramètres acoustiques présentent le même comportement vis-à-vis de l'equivocation, ce qui tend à montrer que les jeux de paramètres sont peu différents. Cependant, les résultats ont été obtenus sur des corpus petit vocabulaire correspondant aux 10 chiffres italiens et espagnols. Par conséquent, tous les contextes phonétiques de ces langues ne sont pas rencontrés.

Dans le but de mieux comprendre les effets de la variabilité des paramètres acoustiques, une analyse des erreurs commises sur les phonèmes d'un corpus grand vocabulaire, où un plus grand nombre de contextes phonétiques peuvent être rencontrés, a été effectuée.

4.4.4 Analyse de la confusion sur un corpus grand vocabulaire

Présentation du corpus SpeechDat italien

Le corpus SpeechDat.IT, est un corpus téléphonique de parole continue grand vocabulaire en italien. Ce corpus est phonétiquement équilibré, ce qui signifie que les fréquences d'occurrence de chaque phonème respectent celles de la langue italienne. Il a été enregistré dans des conditions propres et contient donc peu ou pas de bruit. Le lexique correspondant contient 9400 mots différents.

Les caractéristiques du corpus SpeechDat sont présentées dans le tableau 4.8.

Corpus	Partie	Nombre de phrases	Nombre de mots
SpeechDat Italien (SpeechDat.IT)	Test	1074	7758
	Train	3222	22253

TAB. 4.8: *Description des différentes parties du corpus SpeechDat en italien.*

Analyse de la KLD et de l'equivocation

La figure 4.9 montre une comparaison entre les deux jeux de paramètres acoustiques obtenue avec ce corpus. On remarque que les deux jeux de paramètres ont des comportements similaires. Cela confirme les considérations faites pour le corpus Aurora3 italien partie CH0 (voir figure 4.6), ce qui correspond à l'environnement d'enregistrement de ce corpus.

Les résultats de l'analyse de l'equivocation sur SpeechDat.IT sont présentés dans le tableau 4.9. On constate que l'entropie de la source $H(S)$ est plus élevée que pour le corpus Aurora. Un plus grand nombre de contextes phonétiques sont fréquemment rencontrés dans ce corpus, ce qui implique que l'ambiguïté des symboles d'entrée est supérieure. On observe également une equivocation faible lorsque l'on se situe dans une zone à faible variabilité (colonne 5). Le consensus apparaît de nouveau comme un bon indicateur de confiance.

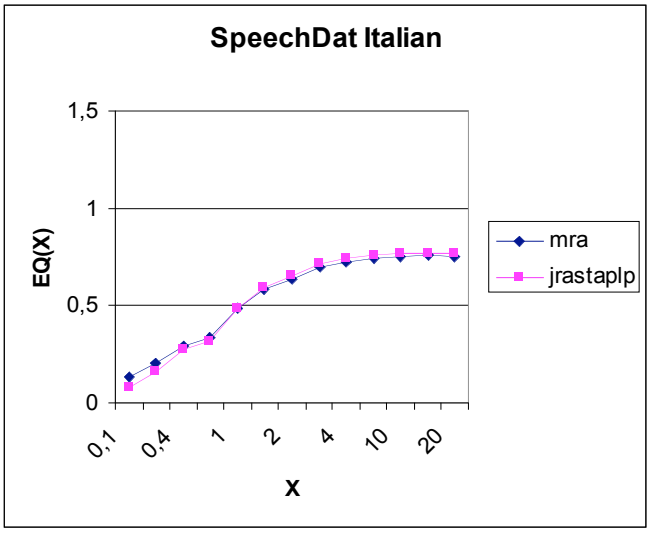

FIG. 4.9: *Comparaison de l'equivocation obtenue avec les paramètres acoustiques MRA et RPLP pour le corpus italien SpeechDat.IT*

Corpus	Traits acoustiques	H(S)	total Eq.	Eq. $\wedge$ LV	Couv.
SpeechDat Italien	MRA	4.1	0.75	0.42	69.9
SpeechDat Italien	RPLP	4.1	0.75	0.42	70.5

TAB. 4.9: *Entropie, equivocation et couverture pour le corpus SpeechDat.IT pour une KLD < 0.4.*

La figure 4.9 confirme que, même dans les zones où l'influence des modèles acoustiques est faible, les jeux de paramètres participent de manière très similaire à l'ambiguïté.

Cependant, ces résultats sont obtenus globalement sur tout le corpus (tous phonèmes confondus). On peut alors se demander si la distribution de l'equivocation est la même pour tous les phonèmes. L'analyse détaillée de la participation à l'equivocation des phonèmes ou classes de phonèmes est présentée dans la section suivante.

4.4.5 Distinction des classes de phonèmes

En dépit du fait que la relation globale entre la KLD et l'ambiguïté soit semblable pour les jeux de paramètres acoustiques considérés, il est intéressant d'étudier cette relation pour chaque phonème (ou classe de phonèmes) afin d'identifier d'éventuelles divergences locales.

Le corpus SpeechDat.IT est adéquat pour une telle étude puisqu'il a été conçu pour approcher une distribution de phonèmes proche de celle de la langue italienne. Il contient la totalité des contextes phonétiques de la langue.

Les phonèmes ont été également regroupés en classes selon certaines de leurs caractéristiques. Il y a cinq voyelles majeures en italien pouvant être décrites en terme de position et de manière d'articulation dans le tableau 4.10.

Voyelle	Position	Manière
a	centrale	haute
e	avant	milieu
i	avant	basse
o	arrière	milieu
u	arrière	basse

TAB. 4.10: *Position et manière d'articulation des voyelles italiennes.*

Les contributions à l'equivocation ont été calculées pour des seuils de KLD (X) petits, correspondant aux zones où l'influence des modèles est faible. La figure 4.10 montre les résultats pour les voyelles « avant » (en haut) et « arrière » (en bas). Les résultats pour la voyelle « centrale » /*a*/ ne sont pas présentés puisqu'ils ne montrent pas de différences significatives entre les deux jeux de paramètres.

Il s'avère que pour les deux types de voyelles (« avant » et « arrière »), l'equivocation la plus petite est observée lorsque la manière d'articulation est basse.

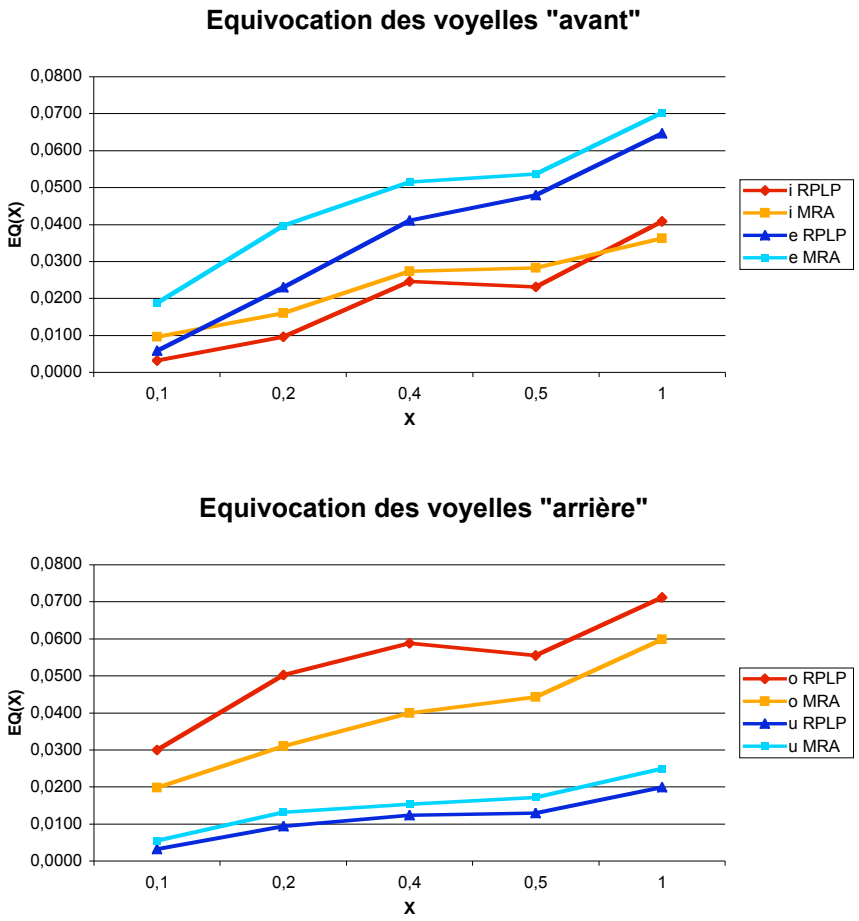

FIG. *4.10:* *Comparaison de l'ambiguïté des voyelles « avant » (en haut) et « arrière » (en bas) de l'italien.*

On peut observer des différences entre les participations à l'equivocation, surtout pour les voyelles /o/ et /e/. Ces observations suggèrent que les deux jeux de traits acoustiques possèdent une variabilité intrinsèque différente dans les zones de leurs espaces acoustiques correspondants à des voyelles dont la position d'articulation est moyenne.

Les contributions à l'equivocation ont également été calculées pour les plosives voisées et non voisées. Des différences ont été observées pour les plosives non voisées /p/ et /t/ (voir figure 4.11). Ceci est probablement dû au fait que le spectre de l'explosion («burst») a souvent une énergie faible et une durée courte pour les plosives voisées. Or, MRA nécessite un voisinage relativement grand pour analyser une trame, ce qui la rend plus faible à détecter les événements très courts.

Cela est aussi dû à la réduction de la largeur de bande de fréquences ayant court pour la parole téléphonique. En effet, le spectre de la parole téléphonique est compris entre 300 et 3800 Hz. Les hautes fréquences ne sont pas représentées et forment donc un manque d'information nécessaire à la bonne discrimination de ce genre de phonèmes.

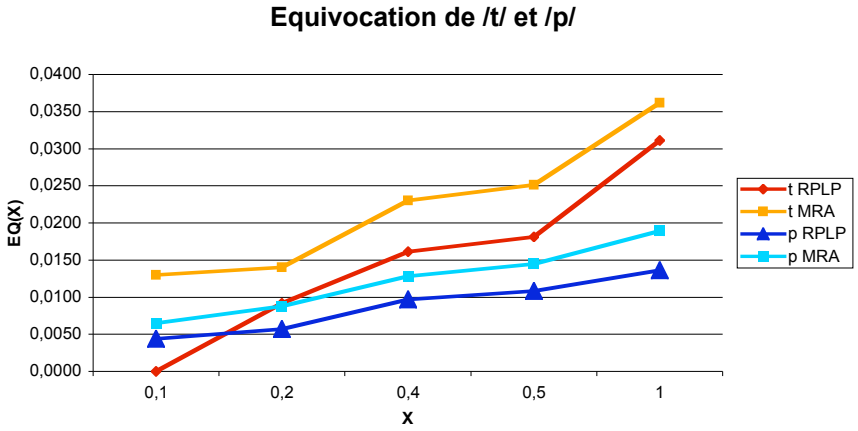

FIG. *4.11:* *Plosives non voisées.*

Parmi les fricatives, seuls /v/ et /s/ montrent des contributions différentes à l'ambiguïté pour les valeurs faibles de KLD, comme présenté dans la figure 4.12. Les paramètres RPLP montrent de meilleures performances pour les phonèmes /s/. Le contraire apparaît pour /v/. Une explication possible est que pour la parole téléphonique, la consonne /v/ est essentiellement caractérisée par ses traits voisés qui sont mieux analysés par l'arbre de filtres de la MRA. Les mêmes observations ont été faites avec le corpus non bruité d'Aurora3.

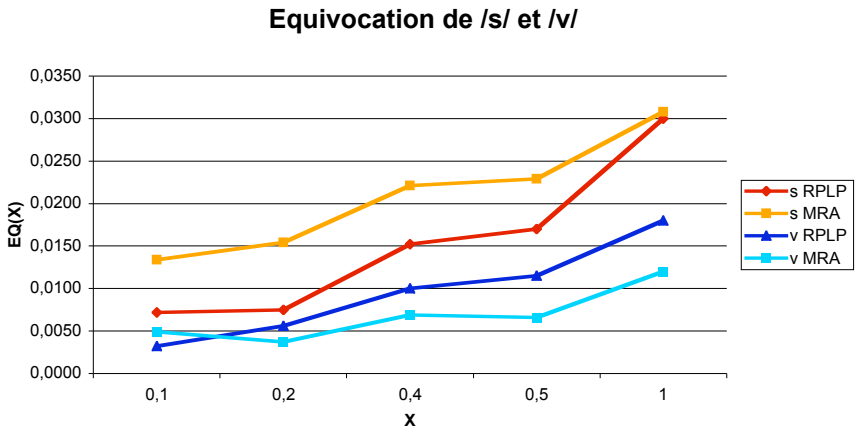

FIG. *4.12:* *Fricatives.*

Des comportements similaires ont été observés pour les consonnes sonores, dans lesquelles les nasales /*n*/ et les liquides /*r*/ affichent la plus grande participation à l'ambiguïté.

Les résultats présentés jusqu'à présent montrent que deux jeux de paramètres différents affichent des contributions locales à l'equivocation relativement différentes, même quand deux techniques de modélisation très différentes sont utilisées pour calculer les probabilités *a posteriori* des phonèmes. Cela indique qu'il y a des zones de leurs espaces acoustiques où ils affichent des degrés de variabilité intrinsèque différents.

Une telle variabilité est évidente même quand un ensemble limité de contextes co-articulatoires sont présents comme dans le corpus propre d'Aurora3.

Obtenir des modèles appropriés réduisant l'equivocation dans ces zones apparaît difficile. Cependant, certains symboles phonétiques affichent des participations à l'equivocation différentes selon le type de paramètres acoustiques utilisé. L'exploitation commune de ces jeux de paramètres dans un processus de combinaison apparaît donc comme une solution prometteuse afin de tirer parti des différences de chaque jeu de paramètres.

4.5 Sélection dynamique de paramètres acoustiques.

Dans les sections précédentes, nous avons fait le diagnostic et la comparaison de la fiabilité de différents jeux de paramètres acoustiques en fonction de certaines zones de l'espace acoustique.

Le principal objectif du travail présenté dans cette section est de prendre en compte le manque de fiabilité affiché par un jeu de paramètres. Le principe est d'identifier les zones peu fiables et d'utiliser un autre jeu de paramètres pour représenter le signal dans ces zones. En d'autres mots, il est raisonnable de penser que la variabilité intrinsèque des paramètres acoustiques mène à une ambiguïté qui ne peut pas être évitée. Cependant, comme l'ambiguïté n'est pas distribuée uniformément pour les deux jeux de paramètres, on peut envisager l'utilisation de paramètres différents lorsque les premiers mènent à une grande equivocation.

Les paramètres MRA et RPLP peuvent avoir un impact différent sur la confusion phonétique. Des critères de décision peuvent être étudier afin de guider leur utilisation.

On se propose d'introduire un critère de fiabilité C^i dépendant du type de paramètres acoustiques utilisés (exposant i) dans le calcul des probabilités. Ce critère permet de mesurer la qualité de représentation du signal Y_n fournie par Y_n^i.

Les modèles acoustiques permettent de calculer les probabilités *a posteriori* des mots w étant donné un segment de parole S_t (représenté par une séquence de vecteurs de paramètres acoustique $Y_n^i, \ldots, Y_{n+t}^i$). La probabilité *a posteriori* d'un mot w est alors définie par :

$$P(w|S_t) = \sum_{i=1}^{I} P^i(w|S_t)C^i \text{ avec } \sum_{i=1}^{I} C^i = 1 \tag{4.5}$$

En pratique, il est difficile d'évaluer le facteur de fiabilité C^i. C'est pourquoi, plutôt que d'estimer ce facteur, nous avons cherché à mettre en évidence des contextes indiquant le degré de fiabilité que l'on peut accorder à un jeu de paramètres. Le contexte permet ensuite de sélectionner le jeu de paramètres à utiliser

4.5.1 États de variabilité

On a vu dans les sections précédentes que le prédicat de consensus au niveau du phonème, $CONS^a(S_t)$, entre différents types de modèles acoustiques est un bon indicateur de confiance. De plus, la KLD des distributions de probabilité des phonèmes calculées avec des modèles acoustiques différents permet d'isoler des zones dans lesquelles la variabilité des paramètres acoustiques est élevée.

Considérons les états de variabilité σ_k présentés dans le tableau 4.11. Ces états de variabilité peuvent prendre différentes valeurs définies à partir des mesures de KLD et du prédicat de consensus.

	$KLD^a(S_t) < 0.4$	$KLD^a(S_t) > 0.4$
$CONS^a(S_t)$	$\sigma 1$	$\sigma 2$
$\neg CONS^a(S_t)$	$\sigma 3$	$\sigma 4$

TAB. *4.11:* *États de variabilité pour les paramètres acoustiques.*

Le prédicat $CONS^a(S_t)$ est vérifié lorsque les modèles ANN et GMM fournissent la plus grande probabilité pour le même phonème.

L'état $\sigma 1$ correspond aux zones où la variabilité des paramètres acoustiques est faible. Dans l'état $\sigma 4$, les paramètres sont peu fiables, et donc la probabilité d'erreur est grande. Les états $\sigma 2$ et $\sigma 3$ correspondent aux zones où la variabilité des paramètres acoustiques et/ou l'imperfection des modèles sont présentes.

Ces états permettent de séparer l'espace acoustique en zones pour lesquelles la fiabilité du système est identifiée.

4.5.2 Fiabilité des paramètres acoustiques

Pour caractériser la fiabilité des paramètres acoustiques, l'interprétation suivante est proposée :

$$C^i = 1 - P(\overline{w}_h|\Sigma_h) \tag{4.6}$$

où $P(\overline{w}|\Sigma)$ est la probabilité que le mot w soit incorrect connaissant un ensemble Σ d'états de variabilité.

Dans ce cas, on considère qu'un jeu de paramètres est fiable lorsque la probabilité conditionnelle de $\overline{w}$ connaissant un ensemble d'états de variabilité Σ est

petite. Cette probabilité est calculée à partir de la probabilité que les phonèmes qui composent w soient erronés.

Les états de variabilité sont utilisés pour calculer la probabilité d'erreur des hypothèses de phonèmes $P(\overline{h}|\sigma)$. La probabilité d'erreur du mot est donc définie par :

$$P(\overline{w}|\Sigma) = \frac{1}{K}\sum_{k=1}^{K} P(\overline{h}_k|\sigma_k) \tag{4.7}$$

où $\Sigma = \sigma_1 ... \sigma_K$ est une séquence de symboles représentant la confiance accordée aux phonèmes de w (dépendant de la fiabilité des paramètres acoustiques).

En pratique, afin de réduire le temps de calcul pour la génération et la confirmation d'hypothèses de mots, un seul jeu de paramètres dans un espace acoustique de référence $\Im^r$ est calculé et utilisé pour une phrase entière. Si la fiabilité d'une hypothèse de mot est trop petite, alors une nouvelle hypothèse générée avec un autre jeu de paramètres est considérée pour le segment correspondant à ce mot.

4.5.3 Expériences et résultats

Des expériences ont été effectuées avec les corpus de test d'Aurora3 (ITA.TEST, SPA.TEST) et de SpeechDat.IT. Les résultats sont présentés dans le tableau 4.12.

Les modèles acoustiques ANN et GMM utilisant les paramètres acoustiques MRA et RPLP ont été utilisés.

Pour Aurora3, les résultats de base ont été obtenus avec les paramètres MRA qui montrent de meilleures performances que RPLP pour cette tâche (Gemello et al., 1999).

Avec la procédure proposée, 339 (54%) phrases pour l'italien et 309 (50.4%) pour l'espagnol ont été validées directement avec l'équation 4.7 (le second jeu de paramètres n'a pas été utilisé dans ce cas). Le WER pour ces phrases est de 4.8% pour l'italien et 0.5% pour l'espagnol.

Lorsque la phrase n'est pas validée directement, les hypothèses peu fiables sont remises en cause, et l'autre jeu de paramètres est invoqué. Dans environ 90% des cas pour l'italien et 80% pour l'espagnol, le second jeu de paramètres propose une hypothèse différente de celle générée par le jeu de référence. Pour les mots où le second jeu de paramètres propose la même hypothèse, on observe un WER de 33.7% pour l'italien et 12.8% pour l'espagnol. Ces taux d'erreur sont relativement élevés si on les compare à ceux obtenus lorsque l'on considère le

consensus mot sur tout le corpus (11.2% WER pour l'italien et 2.05% WER pour l'espagnol). Cela montre que cette zone est peu fiable.

Corpus	Param. de référence	Baseline (%)	Stratégie (%)	Oracle (%)
ITA.TEST CH1	MRA	21.13	17.66	15.03
SPA.TEST CH1	MRA	12.3	8.68	6.8
SpeechDat.IT	RPLP	35.5	31.0	29.6

TAB. *4.12:* *Résultat de la sélection de paramètres acoustiques en terme de WER.*

Quand elle est applicable et quand les systèmes ne donnent pas les mêmes hypothèses, la stratégie apporte une réduction significative du WER. Le gain relatif est de 16.42% pour l'italien et 29.4% pour l'espagnol.

Pour le corpus SpeechDat.IT, le décodage a été effectué sans l'utilisation d'un modèle de langage. La raison est que l'étude se concentre sur l'impact des différents jeux de paramètres sur les performances du système.

Les paramètres RPLP donnent les meilleurs résultats de base pour ce corpus, il est donc choisi comme jeu de paramètres de référence. L'utilisation dynamique des paramètres MRA réduit le WER d'environ 13% relatifs.

Le score Oracle consiste à sélectionner la bonne hypothèse de mot si elle est proposée par au moins l'un des deux systèmes. On observe que le gain relatif maximal que l'on peut espérer avec ce type de méthode est de 16.62% pour SpeechDat.IT.

Conclusion

Dans ce chapitre, différents jeux de paramètres sont comparés. La variabilité des paramètres acoustiques a été mise en évidence grâce à l'exploitation de différentes architectures de diagnostic. Le diagnostic révèle que des jeux de paramètres calculés avec des algorithmes très différents affichent globalement la même ambiguïté pour l'ensemble de l'espace acoustique. Des différences apparaissent lorsque les phonèmes sont analysés séparément.

Ces différences ont été exploitées dans des techniques de combinaison au niveau des mots et des phonèmes aboutissant à une amélioration des performances du système de reconnaissance.

Même si une évaluation précise de l'impact sur le WER de la variabilité intrinsèque des paramètres acoustique est difficile à réaliser, il est probable qu'elle

soit responsable de la plupart des confusions observées. Le nombre considérable d'erreurs communes aux différents jeux de paramètres acoustiques indique qu'il y a des limitations dans la manière dont les paramètres sont calculés.

Chapitre 5

Combinaison acoustique à très bas niveau segmental

Sommaire

Jusqu'à présent, la combinaison de systèmes post décodage à un niveau segmental élevé (phonème, mot) a montré de bons résultats sur des applications à petit vocabulaire. Nous allons maintenant nous intéresser aux applications grand vocabulaire.

Les méthodes de combinaison travaillant *a posteriori* sur les résultats de décodage sont appliquées sur une partie réduite des hypothèses : les N-meilleures hypothèses ou le treillis que l'on peut exploiter en sortie du système contiennent les hypothèses de mots les plus probables étant donné le signal d'entrée et les modèles acoustique et linguistique. La réduction de l'ensemble d'hypothèses limite le domaine de recherche qui est susceptible de ne pas contenir la solution.

Pour remédier à ce genre de problème, nous avons considéré la combinaison de systèmes avant le décodage. Dans ce cadre, une des possibilités est la combinaison de probabilités *a posteriori* au niveau de la trame.

L'objectif de ce type de combinaison est d'obtenir une meilleure estimation des probabilités *a posteriori* d'un symbole étant donné une trame. Il en découle deux contraintes majeures.

La première est que les jeux de paramètres doivent être synchrones. Cela est nécessaire afin de combiner l'information capturée par les différentes analyses sur une même portion de signal. Il faut s'assurer que chaque trame de chaque jeu de paramètres est calculée avec un nombre maximum d'échantillons en communs. Dans la section 2, nous avons vu que les fenêtres d'analyses peuvent avoir des tailles différentes, mais dès lors qu'elles sont centrées sur les mêmes échantillons, on peut alors considérer que les trames contiennent la même information. Ceci nous permet raisonnablement de relâcher cette contrainte et de considérer que nos flux de paramètres sont synchrones.

La seconde contrainte concerne l'estimation des probabilités *a posteriori*. Pour les calculer, il est nécessaire de se situer dans un espace probabiliste. Un espace probabiliste peut être défini par un ensemble fini d'événements (dans notre cas, l'apparition d'un symbole q appartenant à un vocabulaire Q) se partageant la probabilité totale, égale à 1. Ici, l'ensemble des symboles correspond aux états des modèles acoustiques. Cet ensemble doit être le même pour tous les jeux de paramètres acoustiques. De ce fait, il est nécessaire que les modèles acoustiques possèdent le même ensemble d'états, ou en d'autres termes, la même topologie. Pour ce faire, nous avons développé une technique d'apprentissage permettant de générer des modèles acoustiques fondés sur des paramètres acoustiques différents et ayant une topologie strictement identique.

Plan du chapitre

La section 5.2 présente une analyse de l'equivocation présente dans les états du HMM. Cette analyse tente de caractériser les différences qualitatives de deux jeux de paramètres acoustiques en comparant leur contribution à l'ambiguïté. La section 5.3 décrit la procédure permettant d'obtenir des modèles acoustiques ayant la même topologie, condition nécessaire à la combinaison cohérente des probabilités postérieures. La section 5.4 présente les différentes possibilités pour combiner plusieurs distributions de probabilités postérieures. La section 5.6 relate les différentes stratégies employées pour l'estimation des facteurs de pondération pour la combinaison des probabilités. Dans la section 5.7, différents types d'adaptation des modèles acoustiques sont présentés dans le but de générer des modèles différents en vue de les combiner.

5.1 Matériel expérimental

Trois jeux de paramètres correspondant à des manières plutôt différentes de transformer le signal de parole ont été utilisés. Le premier jeu de paramètres est un vecteur de coefficients cepstraux PLP présentés dans la section 2.3, le second est un vecteur de paramètres RPLP (voir section 2.4) et le troisième jeu est calculé avec l'analyse MRA (voir section 2.5). Le système utilisé est Speeral, développé au Laboratoire Informatique d'Avignon (LIA) et décrit dans Nocera et al. (2002). Un modèle tri-gramme de 64k mots est utilisé pour modéliser le langage. Les modèles acoustiques sont des HMMs utilisant des GMMs pour modéliser les états. Ils sont composés de 10040 modèles de phonèmes dépendants du contexte, 3600 états émetteurs (qui peuvent être partagés parmi des modèles ayant le même phonème central) et 232716 gaussiennes.

L'ensemble des résultats en terme de WER présentés dans le tableau 5.2 ont été obtenus avec le corpus de test de MEDIA, présentés dans le tableau 5.1.

Langue	Corpus	Nombre de phrases	Nombre de mots
MEDIA	TRAIN	13641	82639
	DEV	1377	10434
	TEST	3771	26092
ESTER	TRAIN	80217	$\sim$ 1 million
	TEST_TEL	512	4813

TAB. 5.1: *Description des corpus Media et Ester.*

MEDIA est un corpus de parole téléphonique constitué de dialogues de réservation d'hôtels obtenus par une méthode de magicien d'Oz. La méthode de

magicien d'Oz consiste à faire croire à l'utilisateur qu'il s'adresse à un serveur téléphonique alors qu'en fait c'est une personne qui lui répond.

Les trois modèles acoustiques ont été entraînés séparément avec l'approche jumeau décrite dans la section 5.3, en utilisant les 82 heures de parole du corpus français ESTER (Galliano et al., 2005). Ce corpus large bande a préalablement été sous échantillonnées à 8 kHz puisque les modules d'extraction de paramètres acoustiques ont été développés pour de la parole téléphonique.

Les modèles ont ensuite été adaptés par MAP (voit section 1.5.1 p. 36) avec le corpus d'entraînement de MEDIA.

Certains résultats ont été obtenus sur la partie téléphonique du corpus de test d'ESTER (noté ESTER.TEL). Ce corpus correspond aux appels téléphoniques passés durant les émissions de radio suivantes :

20041006_0700_0800_CLASSIQUE	20041006_0800_0900_CULTURE
20041025_1930_2000_RFI_ELDA	20041026_1930_2000_RFI_ELDA
20041027_1230_1300_RFI_ELDA	20041124_1230_1300_RFI_ELDA

5.2 Analyse de la confusion au niveau de l'état

Considérons un espace acoustique $\Im^i$ peuplé par des vecteurs de paramètres acoustiques Y_n^i issus de la paramétrisation de séquences d'échantillons de parole. Considérons également des modèles acoustiques dépendant du contexte composés de HMMs dans lesquels des GMMs modélisent les densités de probabilités pour chaque état. Chaque état est représenté par un symbole q appartenant à un vocabulaire Q représentant l'ensemble des états du modèle.

Les résultats de deux systèmes dont la seule différence réside dans les paramètres acoustiques d'entrée sont exploités pour comparer l'ambiguïté présente dans chaque état . Le premier système utilise les paramètres obtenus par analyse à résolution multiple MRA (voir section 2.5) et le second utilise les paramètres cepstraux RPLP (voir section 2.4).

Les modèles acoustiques structurent l'espace $\Im^i$ en zones dans lesquelles un symbole a la plus grande probabilité *a posteriori*. Soit g_z le symbole qui a la plus grande probabilité dans la zone $z \in Z$, ou Z est l'ensemble des zones. Considérons maintenant un canal de transmission identique à celui représenté dans la figure 4.4 (p. 92). Soit f le symbole transmis et g_z le symbole reçu dans la zone z. Notons que dans la zone z le symbole reçu est toujours le même, même si le symbole transmis peut être différent. La participation à l'equivocation, définie par l'équation 4.3 (p. 92), varie selon les zones considérées.

L'objectif de cette étude est de vérifier s'il existe des zones pour lesquelles la participation à l'equivocation est haute pour un type de paramètres et faible pour l'autre. C'est un moyen de mettre en évidence le degré de complémentarité des jeux de paramètres.

5.2.1 Equivocation globale et equivocation locale

À partir de ce cadre de travail, on peut exprimer l'equivocation globale à partir de la participation locale de chaque zone. En supposant que les symboles sont générés indépendamment dans une zone de l'espace acoustique, l'equivocation globale s'exprime comme une somme pondérée des participations locales de chaque zones :

$$H_R(S) = -\sum_{z \in Z} P(z) H_R(z) \tag{5.1}$$

où

$$H_R(z) = -\sum_{f} P(f, g_z) \log P(g_z|f) \tag{5.2}$$

est la participation de la zone z à l'equivocation et $P(z)$ est la probabilité qu'un symbole transmis soit reconnu comme un symbole g_z appartenant à la zone z.

Les probabilités de l'équation 5.2 peuvent être exprimées de la manière suivante :

$$\begin{aligned} P(f, g_z) &= P(f|g_z)P(g_z) \\ P(g_z|f) &= \frac{P(f|g_z)P(g_z)}{P(f)} \end{aligned} \tag{5.3}$$

$P(f|g_z)$ est la probabilité des symboles transmis f dont les vecteurs acoustiques sont dans la zone z. On peut donc réécrire cette probabilité comme suit : $P(f|g_z) = P_z(f)$. L'équation 5.2 devient alors :

$$H_R(z) = -\sum_{f} P_z(f) P(g_z) \log \frac{P_z(f) P(g_z)}{P(f)} \tag{5.4}$$

Si tous les phonèmes transmis dont les vecteurs acoustiques sont dans la zone z sont des instances de g_z, alors $P_z(g_z) = 1$ et $H_R(z) = 0$. Lorsque le système ne commet aucune erreur dans une zone alors la participation de cette zone à l'equivocation est nulle, et donc le système est très fiable dans cette zone.

Dans cette étude, nous allons considérer les zones correspondant aux états d'un HMM. Soit $E^a(q)$ l'equivocation des paramètres acoustiques a dans l'état

q. Les contributions à l'equivocation sont calculées pour tous les états et séparément pour les deux jeux de paramètres acoustiques considérés.

Afin d'évaluer les différences des degrés de confusion présents dans chaque état q entre les deux jeux de paramètres, considérons $DE(q) = E^j(q) - E^m(q)$, la différence entre les mesures d'equivocation calculées respectivement avec les paramètres RPLP (indice j) et MRA (indice m). Cette mesure représente l'écart d'equivocation entre les deux jeux de paramètres dans chaque trame générée par l'état q.

La comparaison entre plusieurs jeux de paramètres est possible seulement pour les zones où il y a un nombre suffisant de trames de parole. Pour cette raison, nous avons sélectionné les 898 états générant le plus de trames sur les 3435 disponibles.

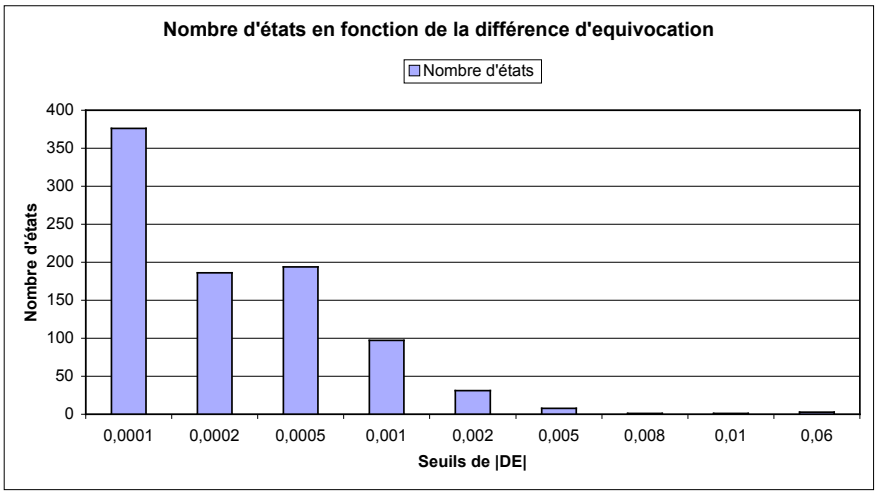

FIG. 5.1: *Distribution du nombre d'états en fonction de la valeur de $|DE|$.*

La figure 5.1 montre le nombre d'états ayant une valeur moyenne de $|DE|$ comprise entre certains seuils (axe des abscisses). On peut remarquer que la plupart des états ont une valeur de $|DE(q)|$ très faible. Cela signifie que les états contribuent de façon très similaire à l'equivocation. Il apparaît peu de contextes dans lesquels la participation à l'equivocation de l'un des jeux de paramètres soit vraiment différente de l'autre.

5.2.2 Equivocation et KLD

Soit $KLD(n, q)$ la divergence de Kullback-Leibler moyenne mesurée entre les distributions de probabilités *a posteriori* de chaque trame n étant donné l'état q. Ces distributions de probabilités sont calculées avec des modèles utilisant deux

jeux de paramètres différents. Cette mesure évalue le degré de désaccord entre les jeux de paramètres conditionnellement à un état.

Si pour un état donné, les distributions de probabilités sont souvent éloignées, alors la KLD moyenne pour cet état sera élevée. On peut alors supposer que si les deux jeux de paramètres fournissent des distributions de probabilités divergentes pour des trames émises par un état donné, alors la différence d'equivocation pour cet état devrait être grande. Les distributions de $KLD(n,q)$ ont été analysées afin d'observer si les états ayant une haute densité de trames pour lesquelles la valeur de KLD est élevée ont aussi une haute valeur de $DE(q)$. Le seuil de 6.0 pour la KLD a été déterminé empiriquement en analysant la distribution des trames en fonction de leur KLD. Nous n'avons conservé que les états ayant une valeur de $DE(q)$ moyenne supérieure à 0.001 bit. Ceci restreint à 45 le nombre d'états analysés.

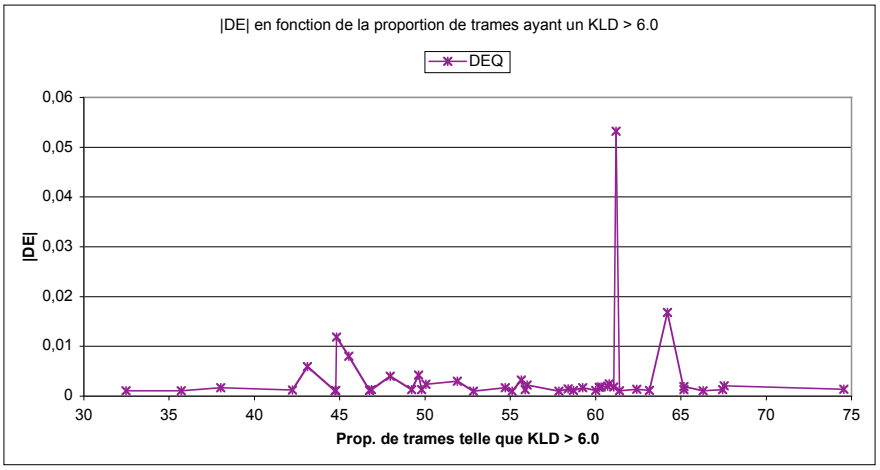

FIG. 5.2: *$|DE(q)|$ en fonction de la proportion de trames ayant une KLD > 6.0.*

La figure 5.2 montre, les valeurs de $DE(q)$ pour les états ayant une $|DE(q)| > 0.001$ dans le corpus MEDIA.DEV, ordonnés en fonction de la proportion de trames ayant une KLD élevée.

Il apparaît que les différences d'equivocation semblent assez décorrélées de la quantité de trames ayant une KLD élevée.

Parmi ces 45 états, il y a seulement 3 états pour lesquels la mesure de confusion de MRA est plus grande que celle de RPLP, *i.e.* $DE(q) < 0$. Les états pour lesquels cela arrive appartiennent aux classes de phonèmes suivantes :

- burst ou plosives non voisées (résultat identique à celui obtenu avec Aurora3 italien comme présenté dans la figure 4.11, p. 104)
- voyelles nasalisées dans le contexte de sons non voisés

– liquides et glissements dans certains contextes intervocaliques
– voyelles avant-milieu dans un grand nombre de contextes, comme pour le corpus italien présenté dans la figure 4.10, p. 103.

Les résultats montrent que RPLP a tendance à mieux modéliser les bursts et certains effets coarticulatoires que MRA. Ceci est probablement dû au fait que les réponses de certains filtres utilisés pour le calcul des paramètres MRA sont longues, et qu'il est donc difficile de capter des événements courts.

5.2.3 Conclusion

L'equivocation est un indice de confusion. Les analyses présentées dans cette section montrent que, pour un symbole donné, l'equivocation est globalement la même pour les différents jeux de paramètres. L'analyse de la confusion des paramètres acoustiques au niveau de l'état révèle que deux jeux de paramètres calculés de manières très différentes ont des zones de faiblesses identiques.

Cependant, il y a également une grande quantité de trames pour lesquelles la valeur de KLD est grande. En effet, l'analyse de la KLD trame à trame montre que les distributions de probabilités des trames fournies par les différents modèles peuvent être éloignées pour certains états. Un grand nombre d'états génèrent plus de 50% de trames ayant une KLD élevée. Cela signifie que les distributions de probabilités *a posteriori* des états sont souvent éloignées pour les deux jeux de paramètres et qu'au final, il serait profitable de combiner les jeux de paramètres au niveau de la trame.

5.3 Modèles acoustiques *jumeaux*.

La production d'hypothèses de reconnaissance est effectuée par l'intermédiaire d'une stratégie de décodage qui évalue des séquences de symboles. Au niveau acoustique, le décodage est essentiellement fondé sur des probabilités comme les probabilités *a posteriori* $P(q|Y_n)$ (q étant un symbole d'un vocabulaire Q et Y_n est une trame). Ce processus exploite un vecteur de probabilités par modèle acoustique et par trame, défini comme suit :

$$\begin{pmatrix} P(q_0|Y_n^i) \\ \vdots \\ P(q_k|Y_n^i) \\ \vdots \\ P(q_K|Y_n^i) \end{pmatrix}$$

où Y_n^i est la $n^{\text{ième}}$ trame de paramètres acoustiques i et q_k, $k = 0 \dots K$ sont les symboles à évaluer.

Considérons $\{\Im^i\}, i = 1 \dots I$, l'ensemble des espaces acoustiques correspondants aux différents jeux de paramètres acoustiques $\{Y^i\}$ disponibles, et $\{Y_n^i\}, i = 1 \dots I$, les instances de la trame Y_n dans ces espaces acoustiques.

La combinaison de probabilités *a posteriori* issues de jeux de paramètres différents nécessite l'apprentissage de plusieurs modèles (au moins un par jeu de paramètres). De plus, étant donné que les probabilités *a posteriori* des états sont combinées au niveau de la trame, il est nécessaire que les modèles possèdent la même topologie (même ensemble d'états).

Nous avons donc développé un protocole, fondé sur le *single-pass retraining* (Woodland et al., 1996) permettant de créer rapidement des modèles *"jumeaux"*.

5.3.1 Protocole d'apprentissage

L'apprentissage en mode *jumeau* nécessite d'avoir un premier modèle appris de manière classique et qui fournit de bonnes performances sur les tâches pour lesquelles il a été entraîné. Appelons ce modèle source M^0, entraîné avec le jeu de paramètres Y^0. Ce modèle a été appris de manière classique et sert dans un premier temps à effectuer un alignement forcé des données du corpus d'apprentissage utilisé pour estimer les paramètres des nouveaux modèles.

Notre objectif est de créer de nouveaux modèles *jumeaux* M^i qui utilisent des jeux de paramètres différents Y^i mais ayant le même ensemble de modèles de phonèmes et le même ensemble d'états que le modèle source M^0.

La figure 5.3 décrit l'approche d'apprentissage des modèles *"jumeaux"*.

Deux vecteurs de paramètres «source» et «cible», respectivement Y_n^0 et Y_n^i, sont calculés à partir de la trame de parole Y_n. Ces trames sont considérées comme synchrones. Le modèle M^0 est utilisé pour calculer la probabilité de chaque gaussienne g connaissant le vecteur de paramètres Y_n^0 (Étape **E**). Les statistiques concernant ces probabilités sont accumulées, en utilisant le vecteur de paramètres Y_n^i, pour chaque gaussienne correspondante dans le modèle M^i. L'étape **M** de l'algorithme **EM** est ensuite effectuée pour estimer les paramètres des gaussiennes de M^i. Le modèle M^i est ré-estimé en utilisant plusieurs itérations de l'adaptation maximum *a posteriori* (MAP). La segmentation du corpus d'entraînement est mise à jour (réalignement) en utilisant le modèle M^i obtenu à chaque itération.

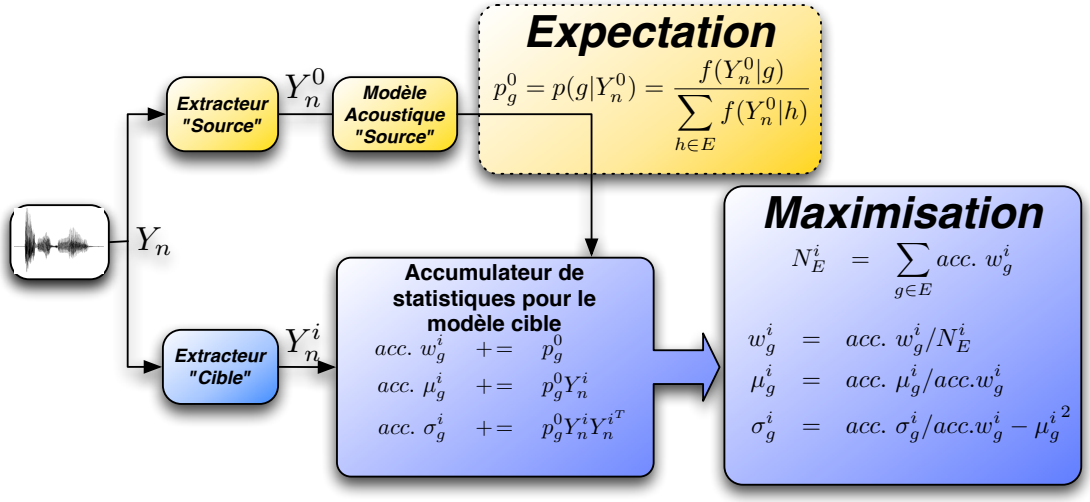

FIG. 5.3: *Apprentissage des modèles* "jumeaux".

5.3.2 Propriétés du modèle *jumeau*

Le modèle source permet de générer une première version d'un modèle jumeau utilisant des paramètres acoustiques différents. Ce premier modèle est mal adapté à ce nouveau jeu de paramètres. En effet, la segmentation forcée a été générée avec le modèle source et peut différer de celle obtenue avec un autre modèle utilisant des paramètres acoustique différents. De plus, la première étape *E* est effectuée en utilisant le modèle et le jeu de paramètres source, ce qui contraint fortement l'apprentissage de la première version du modèle jumeau. Aussi, après avoir construit le premier modèle jumeau, on réaligne le corpus avec celui-ci et on adapte ses paramètres afin de prendre en compte les différences introduites par le changement de paramètres acoustiques.

Il est important de remarquer que cette procédure assure que les états correspondants dans chaque modèle font référence au même contexte acoustico-phonétique. De plus, pour chaque trame Y_n^i d'un espace acoustique donné, la vraisemblance $L(Y_n^i|q)$ est calculée en considérant le voisinage de trames propre à cet espace, ce qui assure que les spécificités de chaque jeu de paramètres sont conservées et non contraintes par le jeu de départ.

En conclusion, le modèle source et les modèles jumeaux font une partition équivalente de leurs espaces acoustiques respectifs (même nombre de modèles phonétiques, d'états et de gaussiennes), mais les distributions des symboles à l'intérieur de ces zones sont différentes parce qu'elles sont ré-estimées en utilisant une segmentation forcée produite avec des jeux de paramètres différents.

5.4 Combinaison des probabilités *a posteriori*

Dans le cadre de la combinaison de probabilités, nous proposons de combiner les probabilités *a posteriori* des états de HMMs au niveau de la trame, comme présenté dans la figure 5.4.

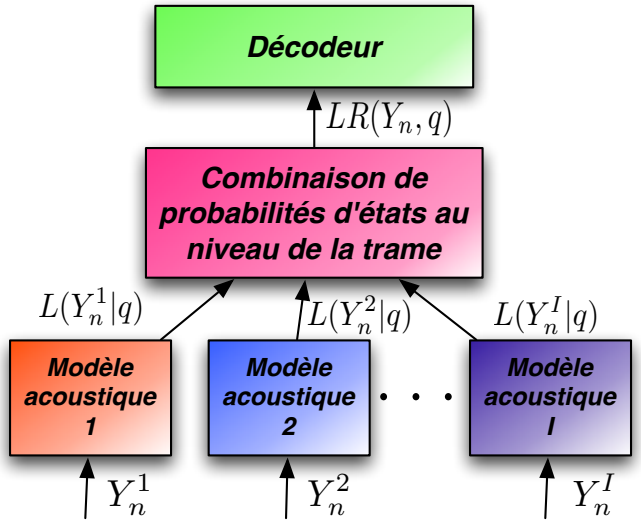

FIG. 5.4: *Architecture pour la combinaison trame à trame.*

Deux manières principales de combiner les probabilités *a posteriori* sont envisagées : la combinaison linéaire et la combinaison log-linéaire.

5.4.1 Combinaison linéaire des probabilités *a posteriori*

La combinaison linéaire (notée LC) des probabilités *a posteriori* est effectuée en supposant que chaque espace acoustique fournit une quantité d'information proportionnelle aux probabilités *a posteriori*.

La relation suivante peut donc être utilisée pour la combinaison linéaire trame à trame des probabilités postérieures des états :

$$\hat{P}(q|Y_n) = \sum_{i=1}^{I} \alpha_i P(q|Y_n^i) \text{ avec } \sum_{i=1}^{I} \alpha_i = 1 \tag{5.5}$$

où α_i est le poids de combinaison que l'on attribue à chaque jeu de paramètres en fonction de la confiance qu'on lui accorde et $P(q|Y_n^i)$ est la probabilité *a posteriori* de l'état q connaissant l'observation acoustique Y_n^i.

Les modèles acoustiques ont des GMMs qui modélisent des fonctions de densités de probabilités associées à chaque état. Les mixtures de gaussiennes permettent de calculer les vraisemblances $L(Y_n^i|q)$ qu'un état q ait générée une trame Y_n^i (voir l'équation 3.4 pour la définition). La probabilité $P(q|Y_n^i)$ est calculée à partir des vraisemblances de la manière suivante :

$$P(q|Y_n^i) = \frac{L(Y_n^i|q)}{\sum_{g \in Q} L(Y_n^i|g)} \tag{5.6}$$

La probabilité $P(q|Y_n)$ est donc proportionnelle à la combinaison linéaire des rapports de vraisemblances (LCLR). Dans un premier temps nous attribuerons la même fiabilité à chaque jeu de paramètres, ce qui signifie que l'on peut omettre α_i :

$$\hat{P}(q|Y_n) = LCLR(n,q) = \frac{1}{I} \sum_{i=1}^{I} \left[\frac{L(Y_n^i|q)}{\sum_{g \in Q} L(Y_n^i|g)} \right] \tag{5.7}$$

La normalisation $\sum_{g \in Q} L(Y_n^i|g)$ permet d'estimer les probabilités *a posteriori* à partir des vraisemblances. Cette normalisation est indispensable pour la combinaison car les vraisemblances ne sont pas comparables entre les différents jeux de paramètres.

La combinaison linéaire correspond à la « règle somme » définie dans Kittler et al. (1998).

5.4.2 Combinaison log-linéaire des probabilités *a posteriori*

Une autre possibilité pour combiner les probabilités *a posteriori* des états au niveau de la trame est de supposer que le système est dans l'état q si tous les jeux de paramètres acoustiques sont d'accord avec cela. En supposant l'indépendance statistique des jeux de paramètres, on obtient :

$$\hat{P}(q|Y_n) = P(q|Y_n^1, \ldots, Y_n^I) = \prod_{i=1}^{I} P(q|Y_n^i)^{\alpha_i} \tag{5.8}$$

En fusionnant les équations 5.6 et 5.8 et en prenant le logarithme afin de sim-

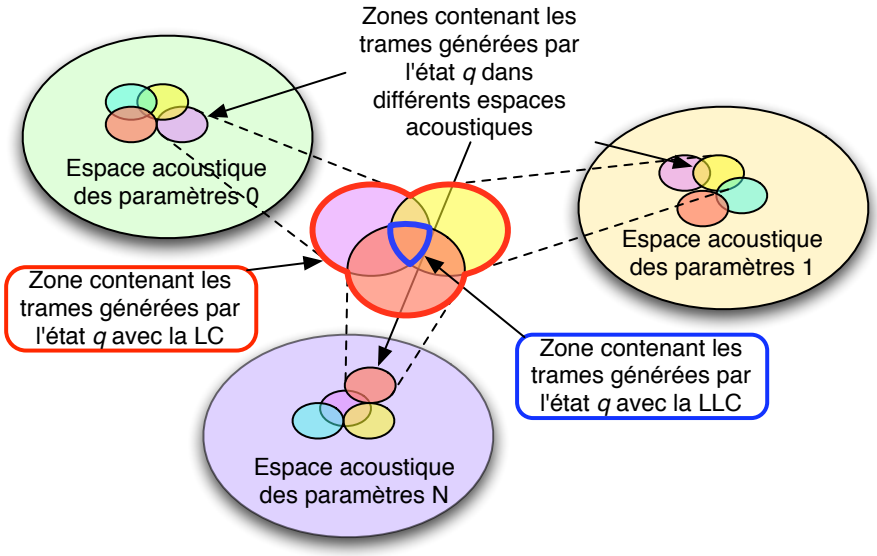

FIG. 5.5: *Vue ensembliste de la combinaison linéaire et log-linéaire de probabilités a posteriori.*

plifier[1], on obtient la combinaison log-linéaire des rapports de vraisemblance (LLCLR) suivante :

$$\log LLCLR(q, Y_n) = \sum_{i=1}^{I} \alpha_i \log \left[\frac{L(Y_n^i|q)}{\sum_{g \in Q} L(Y_n^i|g)} \right] \tag{5.9}$$

Les scores de la LLCLR sont ensuite normalisés afin d'obtenir une distributions de probabilité considérée comme un estimateur de la distribution de probabilité *a posteriori*.

$$\log \hat{P}(q|Y_n) = \frac{LLCLR(q, Y_n)}{\sum_{g \in Q} LLCLR(g, Y_n)} \tag{5.10}$$

Il est important de noter que ces équations considèrent l'indépendance statistique des différents jeux de paramètres.

La répartition des probabilités et des logarithmes des probabilités, présentées dans la figure 5.6, montre que la distribution est gaussienne pour les log-probabilités d'un état. Ce n'est pas le cas pour la distribution de probabilités. Cela conforte l'idée que les probabilités peuvent être fusionnées par combinaison log-linéaire.

[1] À noter que les systèmes de reconnaissance utilisent les logarithmes des probabilités afin de ne pas atteindre les limites techniques de représentation des nombres flottants des ordinateurs.

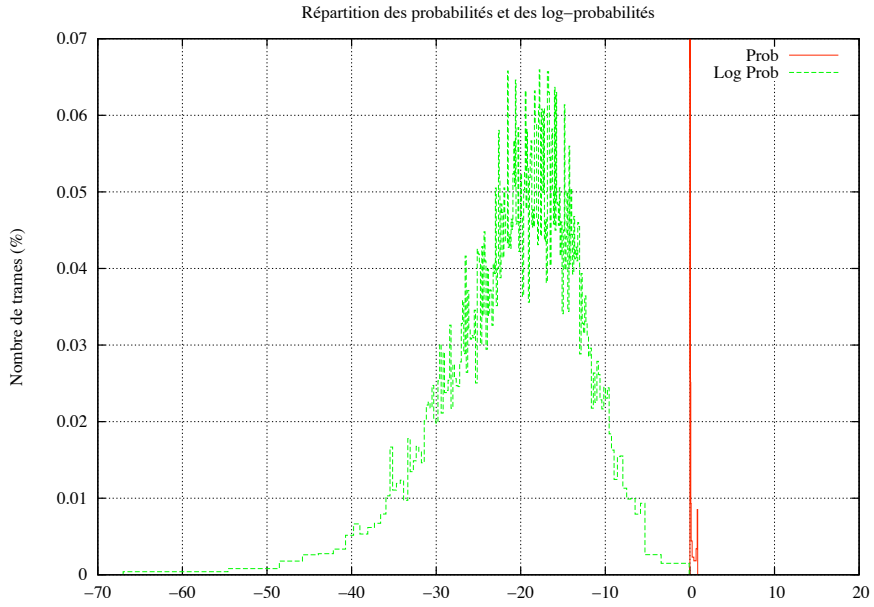

FIG. 5.6: *Répartition des probabilités et des log-probabilités pour un état étant donné une trame de parole.*

Une vue ensembliste représentant les différents espaces de paramètres acoustiques $\Im^i, i = \{1, \ldots, I\}$ est représenté dans la figure 5.5. Considérons dans chaque espace, l'ensemble $\{Y_q^i\}$ des trames générées par un état q Lorsque l'on effectue une combinaison linéaire, on considère que la trame Y_n appartient à l'union des ensembles $\{Y_q^i\}$. Pour la combinaison log-linéaire, on considère que la trame appartient à l'intersection de ces ensembles.

À partir de ces deux types de combinaison principaux, plusieurs techniques permettant d'estimer les coefficients α_i peuvent être utilisées. Elles se basent sur des critères visant à estimer la confiance que l'on peut donner à chacun des modèles. Plusieurs types de pondération sont considérés dans la section 5.6.

5.5 Expériences de reconnaissance

5.5.1 Résultats et analyses

Les résultats avec la combinaison trame à trame des paramètres MRA, RPLP et PLP sont présentés accompagnés par leurs intervalles de confiance[2] ainsi que les gains obtenus relativement au meilleur système utilisant un seul jeu de paramètres (dans le cas de MEDIA.TEST, PLP montre les meilleures performances). À noter que les poids α_i des équations 5.5 et 5.8 sont égaux pour les différents modèles utilisés (aucune différence n'est faite entre les jeux de paramètres).

Jeu(x) de paramètres	WER (%)	Gain relatif (%)	Int. de conf. (%)
MRA	33.2	-	0.57
RPLP	32.2	-	0.57
PLP	32.1	-	0.57
En utilisant la combinaison linéaire			
MRA+RPLP	29.5	8.4	0.55
MRA+PLP	28.2	12.1	0.55
RPLP+PLP	28.0	13.0	0.54
MRA+RPLP+PLP	28.1	12.7	0.55
En utilisant la combinaison log linéaire			
MRA+RPLP	29.2	9.3	0.55
MRA+PLP	28.2	12.1	0.55
RPLP+PLP	28.2	12.1	0.55
MRA+RPLP+PLP	**27.6**	**14.0**	0.54
ROVER	29.3	8.7	0.55
ORACLE	25.4	20.8	0.52

TAB. 5.2: *Résultats de la combinaison trame à trame sur le corpus de test de MEDIA (3771 phrases et 26092 mots).*

On constate que les combinaisons linéaire et log-linéaire fournissent une réduction très significative du WER sur ce corpus[3].

[2] Les intervalles de confiance en % sont calculés de la manière suivante :

$$\frac{1.96 \cdot \sqrt{WER \cdot (1 - WER)}}{\sqrt{N}} \cdot 100$$

avec N, la taille du lexique du corpus.

[3] Les résultats sur le corpus de développement MEDIA.DEV sont disponibles en annexe p. 175.

Les résultats obtenus sur le corpus ESTER avec les mêmes méthodes de combinaison sont présentés dans le tableau 5.3. Les meilleurs résultats ont également été obtenus avec la combinaison log-linéaire des trois jeux de paramètres.

Jeu(x) de paramètres	WER (%)	Gain relatif (%)	Int. de conf. (%)
MRA	41.1	-	1.39
RPLP	37.9	-	1.37
PLP	46.6	-	1.41
En utilisant la combinaison linéaire			
MRA+RPLP	35.2	7.1	1.35
MRA+PLP	33.0	12.9	1.33
RPLP+PLP	33.7	11.1	1.34
MRA+RPLP+PLP	35.1	7.4	1.35
En utilisant la combinaison log-linéaire			
MRA+RPLP	35.5	6.3	1.35
MRA+PLP	34.8	8.2	1.35
RPLP+PLP	35.9	5.3	1.36
MRA+RPLP+PLP	32.2	15.0	1.32

TAB. 5.3: *Résultats de la combinaison trame à trame sur ESTER.TEST.TEL (512 phrases et 4813 mots).*

La première remarque est que les meilleurs résultats ont été obtenus en effectuant une combinaison log-linéaire des probabilités issues des trois modèles acoustiques disponibles. Une réduction du WER de plus de 14% pour MEDIA et d'environ 15% pour ESTER relativement au meilleur système utilisant un seul jeu de paramètres a été observée. L'apport de l'approche proposée pour la combinaison trame à trame est évident même dans le cadre de tâches à très grands vocabulaires.

La combinaison de seulement deux jeux de paramètres montrent des résultats différents. En effet, il apparaît que la combinaison linéaire fournit des résultats équivalents voire meilleurs (non significativement) que la combinaison log-linéaire. Ceci contraste avec la combinaison des trois jeux de paramètres où la LLC surpasse la LC.

ROVER

Afin de comparer nos résultats à une technique de combinaison reconnue, les hypothèses de reconnaissance des différents systèmes utilisant un seul jeu de paramètres ont été combinées avec ROVER, une technique de vote majoritaire pondéré. Ces résultats ont été produit en utilisant le vote majoritaire non

pondéré puisque le système actuel ne fournit pas les probabilités *a posteriori* des hypothèses de mot en sortie.

On observe que le gain obtenu est moindre qu'avec la combinaison au niveau de la trame. Cela peut s'expliquer par le fait que la combinaison après le décodage ne remet pas en cause les hypothèses de mots générées par chaque système et essaye de sélectionner la meilleure hypothèse avec des critères de confiance. Ici, le critère invoqué est la fréquence d'apparition d'un mot parmi les hypothèses proposées par les différents systèmes. On observe une différence de WER de 1.7% absolus (près de 6% relatifs) entre la combinaison trame à trame LLC et ROVER.

En terme de temps de traitement, la combinaison par ROVER est beaucoup plus coûteuse car elle nécessite un décodage complet pour chaque jeu de paramètres utilisé alors que la combinaison LLC n'en nécessite qu'un seul. Il est important de prendre en compte ce critère pour l'utilisabilité des systèmes qui doivent fournir des temps de réponse raisonnables.

Les résultats plus discrets obtenus avec ROVER nous confortent dans l'idée que la combinaison de systèmes après le décodage ne profite pas de toutes les forces de chaque jeux de paramètres ni de chaque modèles acoustiques.

Score Oracle

Le score oracle a été obtenu en sélectionnant le jeu de paramètres qui donne le vecteur de probabilité attribuant la plus haute probabilité pour l'état qui a émis chaque trame analysée. Pour ce faire, la segmentation forcée du corpus a été effectuée avec chaque jeu de paramètres. Ensuite, lorsque les segmentations forcées étaient en accord sur l'état ayant généré la trame considérée, le vecteur de paramètres proposé par le jeu de paramètres fournissant la plus grande probabilité pour l'état est sélectionné. La figure 5.7 présente le calcul du décodage oracle. Lorsque les segmentations forcées ne proposent pas le même états pour une trame, la combinaison log-linéaire des probabilités est effectuée. Ceci a pour but de ne pas privilégier un jeu de paramètres plutôt qu'un autre.

L'oracle fournit une borne maximale que l'on peut atteindre en combinant les modèles acoustiques et les jeux de paramètres disponibles. Elles sont par conséquent très difficiles voire impossibles à atteindre. Elles montrent cependant le gain potentiel important que l'on pourrait obtenir.

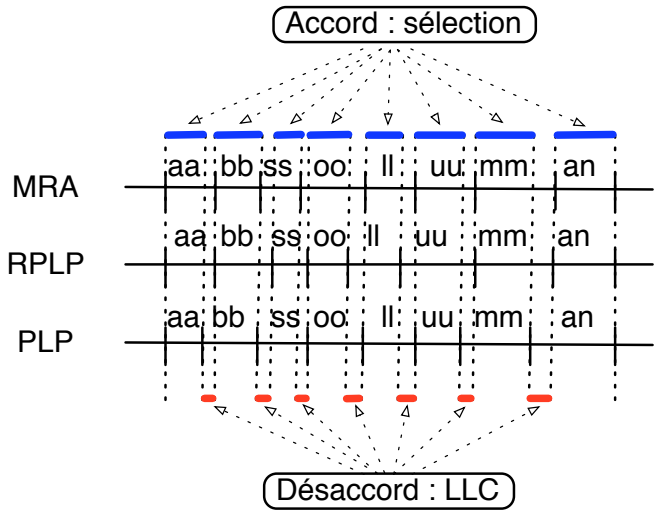

FIG. 5.7: *Décodage en mode oracle : lorsque les segmentations forcées sont en accord, le jeu de paramètres fournissant la plus grande probabilité pour le phonème prononcé (décomposé en états) est sélectionné, sinon la combinaison log-linéaire des probabilités fournies par les trois systèmes est effectuée.*

Perspectives qu'offrent ces résultats

Ces premiers résultats encourageants nous amènent à faire quelques commentaires. Les poids utilisés lors de la combinaison sont les mêmes pour tous les modèles. On pourrait envisager une pondération variable se basant sur la qualité des modèles ou des paramètres acoustiques relativement à un contexte connu. Plusieurs méthodes pour introduire de l'information sur la confiance que l'on peut accorder à un jeu de paramètres sont présentées dans la section 5.6.

En outre, les modèles ont été entraînés séparément avec l'approche jumeau puis adaptés avec plusieurs itérations d'adaptation MAP. Chaque itération comprend le réalignement forcé des données avec le modèle M^i obtenu à l'itération précédente, suivi d'une adaptation par maximum *a posteriori* de M^i pour obtenir M^{i+1}. On peut se poser la question de savoir si l'adaptation des modèles doit se faire librement pour chaque modèle ou si elle doit être contrainte par la combinaison. Plus précisément, faut-il forcer les modèles utilisant un seul jeu de paramètres acoustiques à générer une segmentation proche de celle obtenue avec la combinaison LLC ou alors faut-il les laisser libres d'affirmer leur forces et faiblesses ? Plusieurs expériences permettant de répondre à cette question sont présentées dans la section 5.7.

5.6 Calcul des poids de combinaison

Les probabilités des états issues de différents modèles acoustiques sont combinées au niveau de la trame selon les équation 5.10 ou 5.7. Ces équations comportent un facteur α_i correspondant au poids de combinaison de chaque jeux de paramètres. Jusqu'à présent, ces poids de combinaison étaient égaux pour tous les modèles acoustiques utilisés. Des méthodes visant à estimer des facteurs de pondération ont été explorées afin de mieux combiner les probabilités. Ces poids sont proportionnels à la confiance que l'on peut attribuer à un jeu de paramètres.

On distingue deux types de mesures de confiance. D'une part les mesures *a priori*, généralement entraînées sur un corpus de développement puis utilisées telles quelles dans le système. Parmi ce genre de mesure, nous avons utilisé des matrices de confusion et des régressions logistiques. D'autre part les mesures dynamiques qui dépendent d'un certain contexte. Nous allons considérer l'entropie du vecteur de probabilités et la divergence de Kullback-Leibler.

5.6.1 Matrices de confusion

Une matrice de confusion correspond à une matrice dont les lignes et les colonnes correspondent aux symboles devant être reconnus. Des statistiques correspondant à la probabilité *a priori* de reconnaître un symbole g lorsque le symbole f a été émis sont accumulées. La diagonale de cette matrice correspond aux cas $f = g$, et donc au taux de bonne reconnaissance du symbole f. Ces statistiques ont été accumulées pour chaque trame du corpus de développement. Les symboles considérés sont les états des HMMs. Chaque état correspond à la même unité phonétique puisque les modèles ont été appris avec la méthode *jumeau* (5.3).

Le protocole suivi est le suivant : le vecteur de probabilités pour chaque trame est d'abord calculé pour tous les modèles acoustiques. Cela nous permet d'identifier l'état ayant le maximum de probabilités pour chaque jeu de paramètres acoustiques et ainsi de déterminer le taux de bonne reconnaissance de ces états à l'aide de la matrice de confusion. Le poids attribué au vecteur de probabilités correspond à ce taux de bonne reconnaissance normalisé pour chaque jeu de paramètres.

Soit $\mathcal{C}$ la matrice de confusion de taille ExE, E étant le nombre d'états des modèles acoustiques. Considérons $q^n_{i_{best}}$, l'état le plus probable pour la trame n et le jeu de paramètre i. Le poids α_i des équations 5.9 et 5.7 peut s'exprimer

ainsi :

$$\alpha_i(n) = \frac{\mathcal{C}(q^n_{i_{best}}, q^n_{i_{best}})}{\sum_{i=1}^{I} \mathcal{C}(q^n_i, q^n_i)} \tag{5.11}$$

avec I le nombre de modèles acoustiques combinés.

Les résultats obtenus sont présentés dans le tableau 5.4. Ils ne montrent pas

	WER (%)	Int. de conf. (%)
Baseline (LLC)	27.6	0.54
Matrice de confusion	27.4	0.54

TAB. 5.4: *Utilisation de matrices de confusion sur MEDIA.TEST.*

de gain significatif (résultat confirmé par ceux obtenus sur MEDIA.DEV, voir sec. B.2 p. 175).

Ce résultat peut s'expliquer par le fait que l'approche ne considère pas le processus de décodage qui suit l'estimation des probabilités, notamment l'apport du modèle de langage et du lexique. De ce fait, lorsqu'un modèle acoustique fournit, pour une trame d'observation donnée, la plus grande probabilité pour un état, cela ne signifie pas forcément que cet état sera sélectionné en sortie du système.

5.6.2 Régressions logistiques

Motivations

Les différents pseudo-modèles acoustiques[4] issus de la combinaison linéaire et log-linéaire des probabilités fournissent des résultats différents sur le corpus MEDIA.TEST. Nous avons voulu savoir le degré de complémentarité de ces deux pseudo-modèles. Nous avons considéré un oracle choisissant l'hypothèse aboutissant au plus petit WER parmi les hypothèses données par les pseudo-modèles LC et LLC. Un taux d'erreur de 25.8% a été trouvé, signifiant que les hypothèses proposées sont différentes

Nous avons considéré l'utilisation de régressions logistiques afin de sélectionner le type de combinaison à utiliser pour chaque trame.

[4]Je nomme « pseudo-modèle » acoustique un modèle acoustique issu de la combinaison de plusieurs autres modèles. Ceci dans le but de le différencier des modèles classiques et autonomes que l'on retrouve généralement dans la littérature.

Principe

Le principe de la régression logistique est de trouver une relation entre des variables explicatives $X_1, ..., X_{n-1}$ et une variable expliquée Y (qui est une probabilité). Cette relation est définie comme suit :

$$\log \frac{P(y_i)}{1 - P(y_i)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{n-1} x_{in-1} + \epsilon_i \tag{5.12}$$

Le but est donc d'estimer les n paramètres inconnus $\beta_0, \beta_1, \beta_2, \ldots, \beta_{n-1}$. La forme matricielle de la régression logistique se présente comme suit :

$$\log \frac{P(y)}{1 - P(y)} = X\beta + \epsilon \tag{5.13}$$

La méthode des moindres carrés permet de résoudre une telle équation. On obtient l'estimateur suivant :

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{5.14}$$

Les coefficients de régression sont déterminés en utilisant un corpus d'apprentissage. La régression logistique permet de calculer la probabilité qu'une observation X multidimensionnelle appartienne à une certaine classe Y. On peut alors optimiser un seuil de probabilité permettant d'obtenir les meilleures performances en terme de classification.

Application à la sélection de la méthode de combinaison

Dans notre application, nous nous sommes attachés à vérifier si les trames de paramètres acoustiques contenaient d'emblée l'information caractéristique du type de combinaison à effectuer afin d'assurer un bon résultat de reconnaissance.

Pour ce faire, nous avons entraîné des régressions logistiques sur des trames de paramètres acoustiques (matrice X). Chaque trame est étiquetée en fonction des probabilités obtenues par combinaison de modèles acoustiques. Si la combinaison log-linéaire des probabilités *a posteriori* obtenues avec les trois modèles propose une probabilité supérieure à celle obtenue par combinaison linéaire pour l'état qui a émis cette trame, alors le label « 1 » est attribué à la trame. Dans le cas contraire, on lui associe le label « 0 ».

Des régressions logistiques sont estimées pour chaque état et pour chaque type de paramètres acoustiques. Ensuite, lors du décodage, les matrices de régression correspondant aux états ayant le maximum de probabilité avec les

deux techniques de combinaison sont utilisées afin d'obtenir la probabilité que la combinaison LLC ou LC soit la plus performante pour la trame considérée.

Les résultats obtenus sont présentés dans le tableau 5.5 On remarque que

WER (%)	Int. de conf. (%)
28.2	0.55

TAB. 5.5: *Résultats de la sélection par régression logistique sur MEDIA.TEST (3771 phrases et 26092 mots).*

le gain obtenu n'est pas très élevé. Les vecteurs d'observations acoustiques ne contiennent pas d'information évidente concernant le type de combinaison à utiliser pour calculer les probabilités.

Perspectives

Nous n'avons pas exploité toute la puissance des régressions logistiques dans cette étude. En effet, de nombreux autres paramètres auraient pu être étudiés. Cependant, les premières expériences n'ont pas montré de résultats incitant l'exploration de cette voie.

5.6.3 Entropie des vecteurs de probabilités

L'entropie en tant que mesure de confiance a de nombreuses fois été utilisée dans les système de reconnaissance (Chen et al., 2006; Gravier et al., 2002a; Misra et al., 2003; Schaaf et Kemp, 1997). Dans le but d'utiliser cette mesure pour l'amélioration de la combinaison, le comportement de l'entropie des vecteurs de probabilités obtenus avec les modèles simples (qui n'utilisent qu'un seul jeu de paramètres) et les pseudo-modèles combinés[5] a été effectuée. Un système utilisant les pseudo-modèles a été construit. Ce système introduit des poids dépendant de l'entropie des vecteurs de probabilités pour pondérer la participation des différents modèles acoustiques lors de la combinaison.

[5]On nommera

pseudo-modèle combiné, le modèle « virtuel » résultant de la combinaison de plusieurs modèles « simples ».

Analyse de l'entropie des trames provenant de différentes techniques de paramétrisation

L'analyse de l'entropie des vecteurs de probabilités des trames de trois jeux de paramètres acoustiques a été effectuée. Les trois techniques d'analyse acoustique considérées sont MRA, RPLP et PLP. Chaque modèle a été entraîné de la même manière, avec la procédure d'apprentissage des modèles acoustiques jumeau décrite dans la section 5.3. L'entropie des vecteurs de probabilités postérieures est un critère de confiance utilisé pour pondérer l'impact que l'on donne à un certain modèle.

L'entropie du vecteur de probabilités obtenu avec un modèle acoustique donne une information sur le pouvoir discriminant de ce modèle. Si l'entropie est faible, c'est que le modèle a fournit une grande probabilité pour un faible nombre d'état. Si elle est haute, alors la probabilité est distribuée entre un nombre plus élevé d'état.

Les résultats de l'analyse de l'entropie calculée sur les vecteurs de probabilités *a posteriori* des trames obtenus avec les trois modèles simples et les deux pseudo-modèles combinés sont présentés dans la figure 5.8.

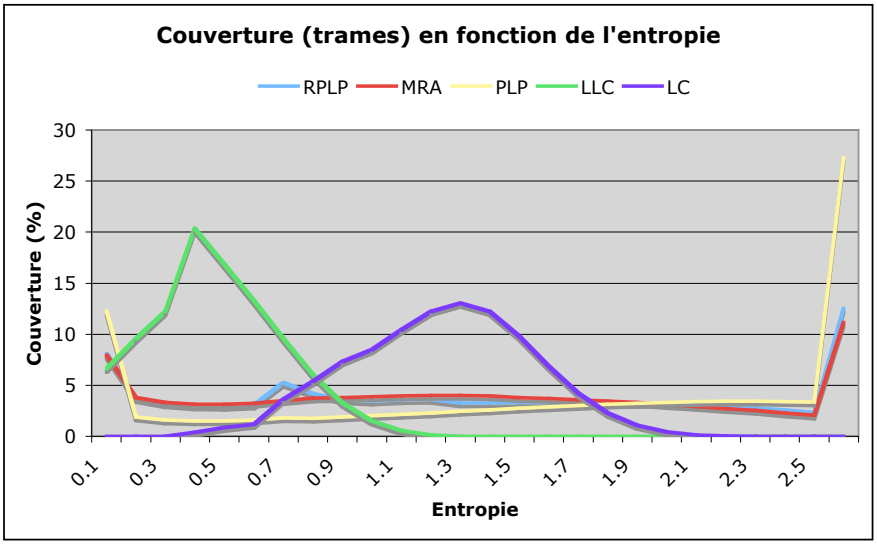

FIG. 5.8: *Distribution des trames du corpus MEDIA.TRAIN (3810560 trames) selon l'entropie de leur vecteur de probabilités.*

D'une part, on peut observer que la répartition des trames en fonction de l'entropie est plutôt uniforme pour les modèles simples. Cela signifie que le pouvoir de discrimination des modèles simples est relativement faible. Les distributions de probabilités ont une entropie variable pour ces modèles.

Les deux combinaisons considérées, LC et LLC, modifient considérablement la distribution des données relativement à l'entropie. En effet, les trames ne sont plus réparties uniformément en fonction de l'entropie. L'entropie moyenne est très réduite, surtout pour la combinaison log-linéaire. On peut donc en conclure que les pseudo-modèles sont plus discriminants que les modèles simples.

Lorsque l'on observe les résultats obtenus en fonction de la classification de la trame (une trame est considérée bien classifiée si l'état ayant le maximum de probabilité correspond à celui de la segmentation forcée), présentés dans l'annexe A.1 p. 167, on remarque que dans l'ensemble, les trames classifiées correctement sont plus nombreuses lorsque l'entropie est faible. Cependant, il en va de même pour les trames mal classifiées. En fait, une entropie faible signifie que le système a donné beaucoup de probabilité pour un symbole, cela ne signifie pas pour autant que ce symbole est le bon. De même, le fait qu'un système distribue plus uniformément les probabilités parmi les symboles ne signifie pas que la reconnaissance va échouer.

La combinaison linéaire fait la somme des probabilités données par les différents systèmes, alors que la combinaison log-linéaire effectue un produit de ces probabilités. Ainsi, en combinaison linéaire, si l'un des systèmes attribue une grande probabilité à un symbole, la probabilité finale de ce symbole sera grande. Pour la combinaison log-linéaire, c'est le contraire. Si l'un des systèmes donne une probabilité faible pour un symbole, alors la probabilité résultante sera faible. Ainsi pour qu'un symbole ait une grande probabilité en combinaison log-linéaire, il faut que les trois systèmes fournissent un probabilité relativement élevée pour ce symbole. Malgré cela, il apparaît que la combinaison log-linéaire diminue énormément l'entropie du vecteur de probabilités.

Expériences de reconnaissance

Plusieurs travaux, notament (Misra et al., 2003), ont introduit une mesure de confiance basée sur l'entropie permettant de pondérer la participation d'un système au score acoustique final. Cette mesure de confiance est inversement proportionnelle à l'entropie du vecteur de probabilités *a posteriori* obtenu avec le système en question.

Nous avons testé ce type de pondération avec notre système en combinant les 3 jeux de paramètres acoustiques différents à notre disposition.

Les résultats obtenus ne montrent pas d'amélioration des performances par rapport à la combinaison log-linéaire à poids égaux pour chaque système. Dans ce cadre de travail, l'entropie ne semble pas être un bon critère permettant de mesurer la confiance que l'on peut attribuer à un jeu de paramètres.

WER (%)	Int. de conf. (%)
28.8	0.55

TAB. 5.6: *Résultats de la pondération par l'entropie sur MEDIA.TEST (3771 phrases et 26092 mots).*

Cela peut s'expliquer de deux manières différentes. D'une part l'entropie du vecteur de probabilités *a posteriori* n'est peut-être pas un bon critère de qualité. Le fait qu'un système propose une grande probabilité pour un symbole et une probabilité faible pour tous les autres ne signifie pas pour autant que ce système est fiable. En effet, si la grande probabilité n'est pas affectée au bon symbole, l'entropie sera faible, mais le système se trompe.

Divergence de Kullback-Leibler

Comme nous l'avons vu dans la seconde partie, la divergence de Kullback-Liebler (KLD) calculée entre les vecteurs de probabilités des phonèmes obtenu avec des jeux de paramètres différents est un bon indicateur de confiance pour la reconnaissance. Nous avons donc tenté de transposer l'idée au niveau de la trame. L'idée est de sélectionner les deux systèmes ayant le minimum de KLD parmi les trois à notre disposition. Le propos est que si deux modèles de même type utilisant des traits acoustiques différents proposent une distribution similaires de probabilités postérieures pour les symboles acoustico-phonétiques, alors il est fort probable que leur meilleure hypothèse soit la bonne.

Une autre manière de faire est de considérer que la combinaison veut tirer parti des différences intrinsèques des paramètres acoustiques. Dans ce cas, il faut conserver au maximum les différences entre les jeux de paramètres utilisés pour la combinaison. Cela nous amène à utiliser les deux systèmes proposant des distributions de probabilités les plus éloignées, et donc ayant une KLD maximale.

Les résultats de reconnaissance par sélection de modèles au niveau de la trame et en fonction de la KLD sont présentés dans le tableau 5.7.

	WER (%)	Int. de conf. (%)
MIN KLD	29.8	0.56
MAX KLD	28.0	0.54

TAB. 5.7: *Résultats de la sélection par KLD sur MEDIA.TEST (3771 phrases et 26092 mots).*

On remarque que le WER obtenu en sélectionnant à chaque trame les deux systèmes proposant des distributions de probabilités les plus éloignées est rela-

tivement proche de celui obtenu en combinant les 3 modèles par interpolation log-linéaire. De plus, ce résultat est meilleur que ceux obtenus avec la meilleure combinaison de deux modèles (voir tableau 5.2).

Un point important est que la combinaison de différents modèles proposant des distributions de probabilités similaires ne donne pas de réel gain au final. L'idée que les paramètres acoustiques extraient des caractéristiques différentes du signal de parole et qu'ils possèdent des forces et des faiblesses pouvant être complémentaires est clairement mis en évidence avec ce genre d'expérience.

5.7 Adaptation des modèles acoustiques en vue de leur combinaison

À partir du cadre d'apprentissage en mode « jumeau » et en vue de combiner différents modèles acoustiques, plusieurs approches peuvent être envisagées pour leur adaptation. En effet, la combinaison de modèles acoustiques utilisant des jeux de paramètres différents repose sur le fait que l'information capturée par chaque jeu de paramètres est différente et parfois complémentaire. Aussi est-il intéressant d'étudier le comportement de chaque modèle en fonction de la manière dont il est adapté. Nous avons envisagé deux options principales. La première consiste à optimiser les paramètres de chaque modèle acoustique indépendamment afin que les jeux de paramètres apportent leur propre information et modélisent l'espace acoustique à leur manière. L'autre méthode se base sur le constat que la combinaison LLC fournit de meilleurs résultats que chaque système utilisant un seul jeu de paramètres. On peut donc naturellement supposer que la segmentation forcée du corpus avec ce système sera plus proche de la segmentation idéale.

Dans le premier cas, la procédure d'adaptation se déroule comme suit. On obtient un modèle initial utilisant un certain jeu de paramètres acoustique avec la procédure jumeau, comme décrit plus haut. Ensuite, avec ce modèle initial, on effectue l'alignement forcé du corpus d'entraînement, puis on réadapte avec MAP, le modèle afin d'obtenir un meilleur modèle. Dans ce cas, chaque modèle utilisant un jeu de paramètres est adapté séparément et ne profite pas de l'apport de la combinaison avec les autres modèles. En contre-partie, il affirme les caractéristiques propres du type de paramétrisation et influe directement sur la représentation de l'espace acoustique.

Dans le second cas, les modèles initiaux entraînés avec chaque jeux de paramètres acoustiques sont combinés par LLC afin d'effectuer un alignement forcé du corpus d'apprentissage. Cet alignement forcé est ensuite utilisé pour adapter chaque modèle initial. Ces modèles seront ensuite combinés à leur tour afin

de fournir une nouvelle segmentation forcée pouvant être utilisée pour l'itération suivante. On peut évidemment réitérer ces étapes plusieurs fois afin de converger vers les modèles optimaux.

5.7.1 Impact du taux de concordance des modèles

En plus de l'apport de l'adaptation des modèles, nous avons voulu mettre en évidence l'impact de la différence entre les modèles sur le résultat de la combinaison. En partant du principe que la complémentarité et la diversité des modèles acoustiques ont un impact direct sur le résultat de la reconnaissance obtenu par combinaison, il est intéressant de quantifier cette relation.

Nous avons pris comme mesure de différence entre les modèles acoustiques, le taux de trames affecté au même phonème contextuel par deux modèles différents lors d'une segmentation forcée du corpus d'adaptation MEDIA.TRAIN. Nous appelons cette mesure, la concordance des systèmes (voir tableau 5.8). Plus la concordance est grande, plus les systèmes donnent des résultats identiques, et donc moins la combinaison devrait être intéressante.

Le choix des paramètres acoustiques est primordial lorsque l'on veut combiner plusieurs systèmes. Étant donné que le but de la combinaison est de tirer profit des qualités de plusieurs jeux de paramètres, mieux vaut utiliser des paramètres les plus différents possible.

En utilisant les modèles issus de l'apprentissage par méthode jumeau, nous avons obtenus les résultats présentés dans le tableau 5.8. Les modèles *M*-0 correspondent aux modèles initiaux appris avec l'approche jumeau sur le corpus d'entraînement d'ESTER puis adaptés par une itération de MAP avec le corpus MEDIA.TRAIN. Les modèles *M*-1 correspondent aux modèles *M*-0 qui ont été adaptés en utilisant la segmentation forcée du corpus d'apprentissage calculée avec le modèle LLC-0. Chaque modèle *M*-0 a été adapté par maximum *a posteriori* en utilisant le corpus d'apprentissage dont la segmentation forcée a été recalculée avec ce même modèle *M*-0. Le résultat de cette adaptation est le modèle *M*-2 correspondant.

Chacun de ces modèles a été utilisé (seul ou en combinaison) afin de produire des résultats de reconnaissance sur le corpus MEDIA.TEST présentés dans la colonne *WER*. Ils ont aussi été utilisés pour générer une segmentation forcée du corpus d'apprentissage afin de mesurer leur degré de concordance avec les autres modèles (colonnes *Concordance*).

Jeu de paramètres	WER (%)	Concordance (%)		
		RPLP-*X*	PLP-*X*	LLC-*X*
MRA-0	33.2	61.25	75.48	78.80
RPLP-0	32.2	-	53.70	66.66
PLP-0	32.0	-	-	76.56
LLC-0	27.6	-	-	-
Adaptation avec une segmentation forcée identique (LLC)				
MRA-1	31.9	78.61	78.74	83.75
RPLP-1	31.8	-	74.23	79.45
PLP-1	29.1	-	-	81.95
LLC-1	26.9	-	-	-
Adaptation avec une segmentation forcée différente				
MRA-2	32.3	76.10	74.93	65.98
RPLP-2	32.1	-	70.71	63.18
PLP-2	29.2	-	-	60.80
LLC-2	27.1	-	-	-

TAB. 5.8: *Comparaison entre le taux de concordance mesuré sur le corpus MEDIA.TRAIN et le gain en WER obtenu avec le corpus MEDIA.TEST. La concordance mesure le pourcentage de trames pour lesquelles un même état a reçu la plus haute probabilité pour les jeux de paramètres considérés.*

5.7.2 Résultats et observations

Les meilleurs résultats sont obtenus en utilisant la segmentation forcée donnée par la combinaison LLC pour adapter les modèles.

On peut observer que les modèles acoustiques ne réagissent pas de la même manière par rapport à l'adaptation. En effet, le modèle PLP progresse d'environ 3% absolus en terme de WER lorsqu'il est adapté avec l'une ou l'autre des segmentations, alors que le gain n'est que de 1.2% maximum pour MRA et seulement 0.4% pour RPLP. Ce résultat est difficile à expliquer. Les modèles acoustiques ayant une topologie strictement identique, la seule différence qui puisse expliquer ces résultats est le type de paramètres acoustiques utilisé.

La différence entre les modèles *M*-1 et *M*-2 réside dans le fait que la même segmentation est utilisée pour adapter les 3 modèles *M*-1, ce qui n'est pas le cas pour les modèles *M*-2. On remarque que le taux de concordance entre les modèles *M*-1 et leur combinaison LLC est très grand comparée aux autres. Ceci s'explique par le fait que le système LLC semble converger vers une segmentation optimale pour ce type de combinaison.

On remarque aussi qu'adapter les modèles avec une segmentation obtenue

avec un modèle produisant de meilleurs résultats apporte plus d'information qu'une adaptation faite avec un modèle moins performant. En effet, les modèles *M*-2 (adaptés séparément) fournissent de moins bons résultats que les modèles *M*-1 (adaptés avec la même segmentation forcée). L'apport des spécificités de chaque jeu de paramètres ne surpasse donc pas la qualité de la segmentation utilisée pour l'adaptation.

La concordance varie entre 54 et 75% pour les modèles de base. On peut remarquer que les processus d'adaptation semblent rapprocher les modèles acoustiques. En effet, les modèles *M*-1 produisent des segmentations concordantes à environ 75% et les modèles *M*-2, quant-à eux, génèrent des segmentations forcées dont la concordance un peu plus faible et varie entre 70 et 76%.

Dans le second cas, la convergence s'explique par le fait que les modèles ont été adaptés avec la même segmentation. De ce fait, on peut supposer que les paramètres des modèles ont été modifiés de la même manière respectivement à chaque espace acoustique. Les segmentations obtenues se rapprochent d'une segmentation optimale que l'on peut atteindre avec la combinaison des probabilités calculées avec ce type de modélisation utilisant ces trois jeux de paramètres. D'ailleurs, le taux de concordance avec la segmentation fournie par la combinaison des trois modèles *M*-1 est bien supérieures à celui produite par les modèles de base *M*-0.

Les résultats obtenus avec les modèles *M*-2 sont plus étonnants. Même en laissant les spécificités de chaque jeu de paramètres acoustiques s'exprimer, les segmentations forcées obtenues se rapprochent fortement après adaptation. Cependant, ces segmentations sont beaucoup plus éloignées de celle obtenue avec leur combinaison, ce qui contraste avec les résultats obtenus avec les modèles *M*-1. Les modèles se rapprochent les uns des autres tout en s'éloignant du pseudo-modèle correspondant à leur combinaison. En terme de performances, les résultats obtenus avec la combinaison de ces modèles ne sont pas significativement plus mauvais puisque seulement 0.2% les séparent des résultats obtenus avec la combinaison des modèles *M*-1.

La concordance entre les segmentations forcées que proposent les pseudo-modèles LLC-1 et LLC-2 est relativement basse comparée à celle des modèles LLC-0 et LLC-1 (voir tableau 5.9). Ceci suggère que bien que les modèles LLC-1 et LLC-2 produisent des taux d'erreurs similaires, les résultats qu'ils proposent ne sont pas si proches.

Pour vérifier cela, nous avons produit les résultats de combinaison avec ROVER et LLC en mode oracle. L'oracle avec ROVER consiste à générer toutes les sorties possibles en combinant les hypothèses des systèmes mot à mot et à ne conserver que celle fournissant le plus petit taux d'erreur. L'oracle avec la LLC consiste à choisir le vecteur de probabilités donné par le pseudo-modèle qui

fournit la plus grande probabilité pour l'état effectivement prononcé (obtenu par alignement forcé). Les résultats obtenus sont présentés dans le tableau 5.9, accompagnés de la concordance de ces trois pseudo-modèles mesurée sur le corpus d'entraînement.

Pseudo-modèles utilisés	Type d'oracle	WER (%)	Concordance (%)
LLC-0 et LLC-1	ROVER	24.6	97.42
LLC-0 et LLC2	ROVER	24.0	79.68
LLC-1 et LLC2	ROVER	25.0	79.05
LLC-0 et LLC1 et LLC2	ROVER	23.2	-
LLC-0 et LLC1 et LLC2	LLC	25.1	-

TAB. *5.9:* *ROVER mode oracle sur MEDIA.TEST et concordance des pseudo-modèles LLC-X sur MEDIA.TRAIN.*

Le premier résultat important est que les modèles génèrent des hypothèses différentes qui peuvent potentiellement être exploitées par des techniques de combinaison. Nous avons testé la combinaison de ces pseudo-modèles avec ROVER et la combinaison de probabilités LLC. Les résultats obtenus sont présentés dans le tableau 5.10.

Type de combinaison	WER (%)	Int. de conf. (%)
LLC	27.0	0.54
ROVER LLC0-1-2	26.9	0.54

TAB. *5.10:* *Combinaison par ROVER et par LLC des trois pseudo-modèles LLC-0, LLC-1 et LLC-2 sur MEDIA.TEST.*

Aucune amélioration n'est observée, que ce soit avec ROVER ou la combinaison LLC des probabilités des états. Il semble que les techniques de combinaison ne parviennent pas à capturer l'information complémentaire nécessaire à la bonne reconnaissance.

La seconde observation est que malgré la grande concordance des modèles LLC-0 et LLC-1, leur combinaison produit des résultats relativement différents (score Oracle à 24.6). La différence entre les deux systèmes réside dans l'adaptation des modèles utilisés pour la combinaison. Le fait de modifier les paramètres des modèles en fonction du jeux de paramètres acoustiques qu'ils utilisent permet de générer des résultats de reconnaissance différents et complémentaires.

Il est important de noter que ces résultats ne sont que des bornes maximales que l'on pourrait potentiellement atteindre avec une méthode de combinaison. Le résultat de la sélection de mots telle que le fait l'Oracle ROVER est très difficile à réaliser en pratique.

5.7.3 Conclusion

Nous avons développé une méthode pour générer des modèles différents en vue de leur combinaison. L'adaptation de modèles acoustiques utilisant des paramètres acoustiques différents avec des segmentations forcées issues de systèmes différents a permis de créer des modèles complémentaires dont la combinaison conduit à une amélioration des résultats. Une réduction relative du WER de plus de 15.9% par rapport au meilleur système utilisant un seul jeu de paramètres a été observée. En parallèle de cela, nous avons fait le lien entre la concordance des modèles et le résultats de leur combinaison. Il en ressort que plusieurs modèles relativement proches du point de vue de la segmentation forcée peuvent être combinés afin de produire des hypothèses qualitativement différentes de celles obtenues avec des modèles plus éloignés.

Les meilleurs résultats ont été obtenus en adaptant les modèles avec la segmentation forcée générée avec le meilleur système disponible. Cependant, l'écart avec le pseudo-modèle combinant les modèles adaptés séparément n'est pas significatif.

Ce phénomène pourrait s'expliquer comme suit. Le pseudo-modèle LLC-0 est le meilleur disponible initialement. Il en résulte que l'adaptation de chaque modèle de base *M*-0 avec la segmentation donnée par LLC-0 sera *a priori* la plus performante. En effet, on peut raisonnablement penser que l'augmentation de la qualité de la modélisation de l'espace acoustique permet de générer des segmentations plus proches de la segmentation *idéale* et donc les modèles adaptés seront plus représentatif de cet espace. Dans le second cas de figure, on ne profite pas de cette meilleure segmentation. Cependant, on peut penser que l'information apportée par chaque jeux de paramètres est mise en emphase lors de l'étape d'adaptation. Ainsi, ce sont les spécificités de chaque jeux de paramètres qui apportent l'information supplémentaire aboutissant à de meilleurs résultats. En conclusion, d'une part on a amélioré la partition de l'espace acoustique et d'autre part, on a mis en exergue la complémentarité de chaque jeu de paramètres. Chacune de ces manières de procéder conduit à une sensible amélioration des résultats.

En outre, les différentes adaptations ont permis de générer des pseudo-modèles acoustiques fournissant des résultats qualitativement différents, comme le montre les résultats du tableau 5.9. Cependant, aucune des stratégies de combinaison considérée n'a réussi à extraire l'information complémentaire permettant de profiter de cette complémentarité.

5.8 Discussion et conclusions

La combinaison de probabilités *a posteriori* des états offre une solution à plusieurs problèmes pouvant être rencontrés avec d'autres approches.

Les méthodes de combinaison actuelles, comme les réseaux de confusion (CNC) ou ROVER sont limitées par le fait qu'elles opèrent *a posteriori* sur des sorties asynchrones qui ne sont plus reliées au signal et qui sont obtenues en effectuant une sélection préliminaire des hypothèses les plus probables reposant sur de l'information ou un savoir partiel. De plus, ce genre de combinaison ne reconsidère pas les hypothèses de mots trouvées par les systèmes et ne produit pas de nouvelles hypothèses. Elles s'attendent à ce que un ou plusieurs systèmes donne une hypothèse correcte et tente de l'extraire en utilisant des mesures de confiance. Dans notre protocole, le processus de décodage est influencé par la combinaison qui est effectuée avant, ainsi des hypothèses qui aurait pu être élaguées puisqu'elles ont une faible probabilité peuvent être réévaluées à la hausse.

Cette combinaison à bas niveau ne repose pas sur les hypothèses faites par les systèmes et essaie d'extraire l'information complémentaire au niveau de la trame avant qu'une hypothèse de mot ne soit formulée.

Les résultats expérimentaux montrent que la combinaison trame à trame des probabilités *a posteriori* calculées avec des modèles utilisant des jeux de paramètres différents produit une réduction substantielle du WER.

Bilan

Des approches diverses peuvent être mises en œuvre pour améliorer les systèmes de reconnaissance de la parole. Dans cette thèse, des jeux de paramètres et des modèles acoustiques différents sont combinés à plusieurs niveaux de granularité afin d'augmenter les performances de ces systèmes. Les différents types de combinaison utilisés font suite au diagnostic dont l'objectif est d'identifier les points forts et les points faibles de chacun des jeux de paramètres utilisés.

Nous nous sommes attachés à mettre en évidence l'equivocation (mesure d'ambiguïté) introduite par les différents jeux de paramètres dans différentes zones de l'espace acoustique. L'equivocation peut être due aux deux types de composant qui interviennent : les paramètres acoustiques et les modèles acoustiques. Nous avons donc proposé une architecture où le système de reconnaissance est considéré comme un « canal de transmission ». Cette architecture permet de séparer la participation des paramètres acoustiques à l'equivocation de celle des modèles acoustiques. La divergence de Kullback-Leibler (KLD) mesurée entre deux distributions de probabilités permet d'identifier les zones dans lesquelles l'ambiguïté ne dépend pas des modèles acoustiques. Il est alors possible d'évaluer la participation des paramètres acoustiques à l'equivocation. Les résultats des analyses sont différents selon le niveau segmental considéré.

Au niveau du phonème ou du mot, le diagnostic révèle que les différents jeux de paramètres affichent globalement la même equivocation sur l'ensemble de l'espace acoustique. Cependant, des disparités se révèlent lorsque l'on considère les phonèmes séparément. Nous avons proposé une technique de sélection de paramètres acoustiques qui découle de ce résultat. Elle est fondée sur l'estimation de la probabilité qu'une hypothèse de mot est incorrecte. Celle-ci est calculée à partir de la probabilité que les phonèmes de ce mot soient incorrects étant donné des « états de variabilité » associés à chaque phonème. Ces états de variabilité ont été déterminés par le diagnostic préalablement effectué. Cette technique a permis d'obtenir des gains significatifs sur des tâches petit et moyen vocabulaire.

Au niveau de la trame, les résultats de diagnostic sont différents. Malgré l'utilisation de paramètres acoustiques calculés avec des analyses très diffé-

rentes (d'une part une analyse cepstrale et d'autre part une analyse en ondelette), l'equivocation au niveau de la trame est globalement la même (comme au niveau phonétique). L'equivocation est la même lorsque les états du HMM sont étudiés séparément (ce qui diffère du niveau phonétique). Néanmoins, l'analyse de la KLD montre qu'il existe un grand nombre de trames pour lesquelles les distributions de probabilités des états divergent fréquemment. Cela signifie qu'il peut être profitable de combiner les différents jeux de paramètres au niveau de la trame. La combinaison trame à trame des probabilités *a posteriori* a montré de très bonnes performances sur des applications grand voire très grand vocabulaire. L'estimation des probabilités *a posteriori* des états étant donnée une trame de parole est fortement améliorée par ce type de combinaison.

Il est toutefois difficile d'exploiter directement l'ensemble des résultats de diagnostic dans un système de reconnaissance. Dans cette thèse, nous proposons des solutions permettant de résoudre partiellement les problèmes dûs à la variabilité des paramètres acoustiques. Un grand nombre d'erreurs découle de l'application d'une règle qui ne convient pas pour le contexte du segment de parole analysé (critère de décision inapproprié, combinaison non fiable, etc.). Par conséquent, les méthodes s'appliquant systématiquement doivent être évitées au profit de méthodes appliquées ponctuellement en fonction d'un contexte bien identifié. Il en résulte que l'identification de ces contextes est nécessaire. C'est dans ce sens que nous proposons une méthode permettant d'identifier des zones dans lesquelles un jeu de paramètres acoustiques n'est pas fiable. De nouveaux paramètres spécialement développés pour ces zones doivent être utilisés afin d'augmenter la robustesse des systèmes de reconnaissance.

Un des points essentiels de cette thèse concerne la combinaison des probabilités *a posteriori* des états d'un HMM étant donnée une trame de parole. Afin d'effectuer une combinaison cohérente, les modèles acoustiques doivent avoir la même topologie, c'est-à-dire le même ensemble d'états. Une technique d'apprentissage rapide et efficace est proposée pour générer des modèles acoustiques respectant cette contrainte. Elle consiste à exploiter un modèle entraîné de manière classique afin d'estimer les paramètres d'un modèle «jumeau» utilisant un jeu de paramètres acoustiques différent. Cette technique a permis de produire rapidement des modèles acoustiques performants afin de les combiner.

Lorsque l'on combine plusieurs systèmes, la pondération de la participation de chaque système en fonction d'une estimation de leur qualité est *a priori* une solution permettant d'améliorer les résultats. Les résultats observés dans cette thèse ne vont pas dans ce sens pour les méthodes explorées. En effet, la pondération des vecteurs de probabilités n'a pas donné de résultats significativement meilleurs. Cela rejoint les résultats obtenus par Kantor et Hasegawa-Johnson (2008) où le réglage des poids de combinaison ne produit pas d'amélioration

voire même, dans certains cas, dégrade les résultats malgré le respect du critère d'optimalité bayésien. À l'avenir, la diversification des types de paramètres et de modélisations acoustiques sera probablement privilégiée à l'estimation de poids de combinaison optimaux.

La génération de modèles différents est une problématique intéressante pour la combinaison de systèmes. Dans cette thèse, j'ai analysé le comportement des modèles acoustiques selon la manière dont ils ont été adaptés. Il en découle que la segmentation de référence et la diversité des paramètres acoustiques ont tous deux une influence positive sur le résultat obtenu avec le modèle adapté. Ces résultats doivent être complétés en analysant l'impact de l'adaptation croisée des modèles. Cela consiste à adapter un modèle acoustique avec une segmentation produite par un autre modèle et inversement. Les modèles obtenus sont ensuite utilisés par une technique de combinaison.

Perspectives

À partir de l'architecture de combinaison présentée dans cette thèse, plusieurs perspectives sont envisageables.

Tout d'abord, de nouveaux jeux de paramètres peuvent être exploités afin de diversifier d'autant plus les sources d'information. Par exemple, l'utilisation de nouveaux jeux de paramètres cepstraux ou coarticulatoires pourra être considérée. Dans un avenir proche, il est prévu d'introduire des paramètres de type TRAPS (*temporal patterns*) extrayant une information sur la structure temporelle du signal de parole (mesure d'énergie dans les bandes critiques). On peut espérer que plus les paramètres acoustiques extraient des caractéristiques différentes du signal de parole, plus les gains obtenus seront importants lors de la combinaison.

L'utilisation de différents types de modélisation pour la combinaison trame à trame des probabilités peut également être envisagée. Ce type de combinaison a montré de bons résultats lorsqu'elle est utilisée au niveau du mot ou du phonème (voir chapitre 4, combinaison de probabilités calculées avec des GMMs et des ANNs). Porter cette technique au niveau de la trame nécessiterait la relaxation de certaines contraintes, notamment sur la topologie des modèles.

Enfin, une combinaison hiérarchique à résolution multiple peut être explorée. Dans un premier temps, les probabilités des états combinées au niveau de la trame améliorent l'estimation des probabilités *a posteriori* de ces états. Dans un second temps, les hypothèses de reconnaissance proposées par chaque système (combiné ou non) sont exploitées par une méthode de combinaison (CNC, ROVER, ou sélection de jeux de paramètres acoustiques comme présenté dans

la section 4.5). Une architecture intégrant ces deux façons de combiner les jeux de paramètres et les modélisations acoustiques, comme présentée dans la figure 5.9, est une idée intéressante à approfondir.

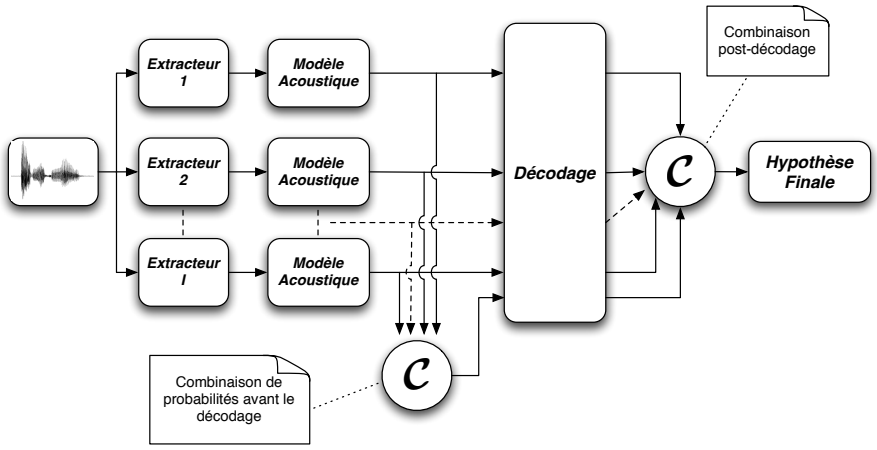

FIG. 5.9: *Architecture d'un système combinant les probabilités* a posteriori *avant le décodage et les hypothèses de reconnaissance (1-Best, N-Best, treillis de mots) après le décodage.*

Glossaire

KLD : Kullback-Leibler Divergence.

PER : Phoneme Error Rate, taux d'erreur de phonème.

HMM : Hidden Markov Model, modèle de Markov caché.

GMM : Gaussian Mixture Model, modèle de mixtures de gaussiennes.

ANN : Artificial Neural Network, réseau de neurones artificiel.

WER : Word Error Rate, taux d'erreur de mot.

MRA : *Multi Resolution Analysis*, analyse à résolution multiple.

PLP : *Perceptual Linear Prediction*, prédiction linéaire perceptive.

RPLP : RelAtive SpecTrAl - Perceptual Linear Prediction, PLP avec traitement RASTA.

JRASTA-PLP : voir RPLP .

MAP : Maximum *a posteriori*

ML : *Maximum Likelihood*, maximum de vraisemblance.

MLLR : *Maximum Likelihood Linear Regression*, régression linéaire avec critère de maximum de vraisemblance.

Liste des illustrations

Liste des tableaux

Bibliographie

(Acero, 1990) A. Acero, 1990. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Thèse de Doctorat, Carnegie Mellon University, Pittsburgh, Pennsylvania.

(Allen, 1994) J. Allen, 1994. How do humans process and recognize speech ? *IEEE Transactions on Speech and Audio Processing 2*, 567–577.

(Bahl et al., 1990) L. R. Bahl, F. Jelinek, et R. L. Mercer, 1990. *A maximum likelihood approach to continuous speech recognition*, 308–319. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

(Baum et al., 1966) L. Baum, T. Petrie, G. Soules, et N. Weiss, 1966. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics 37*, 1554–1563.

(Baum et al., 1970) L. E. Baum, T. Petrie, G. Soules, et N. Weiss, 1970. A maximisation technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics 41*, 164–171.

(Bellot, 2006) O. Bellot, 2006. *Adaptation au locuteur des modèles acoustiques dans le cadre de la reconnaissance automatique de la parolee*. Thèse de Doctorat, Laboratoire Informatique d'Avignon (LIA).

(Berthommier et Glotin, 1999) F. Berthommier et H. Glotin, 1999. A new snr-feature mapping for robust multistream speech recognition. Dans B. University Of California (Ed.), *Internationnal Congress on Phonetic Sciences (ICPhS)*, Volume 1 de *XIV*, San Francisco, 711–715.

(Boll, 1979) S. Boll, 1979. Suppression of acoustic noise in speech using spectral substraction. *IEEE Transactions on Acoustics, Speech and Signal Processing 27*(2), 113–120.

(Bourlard et al., 1996) H. Bourlard, S. Dupont, et C. Ris, 1996. Multi-stream speech recognition. Technical Report IDIAP-RR 07, IDIAP.

(Breslin et Gales, 2006) C. Breslin et M. Gales, 2006. Explicitly generating complementary systems for speech recognition. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Pittsburgh, PA, 525–528.

(Bristow, 1986) G. Bristow, 1986. *Electronic Speech Recognition : Techniques, Technology & Applications*. USA : McGraw-Hill Book Company.

(Cerisara et Fohr, 2001) C. Cerisara et D. Fohr, 2001. Multi-band automatic speech recognition. *Computer Speech and Language 15*, 151–174.

(Cerisara et al., 2000) C. Cerisara, D. Fohr, et J.-P. Haton, 2000. Asynchrony in multi-band speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, 1121–1124.

(Chen et al., 2006) Y. Chen, C. yu Wan, et L. shan Lee, 2006. Entropy-based feature parameter weighting for robust speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, 41–44.

(Cohen et al., 1995) J. Cohen, T. Kamm, et A. Andreou, 1995. Vocal tract normalization in speech recognition : compensating for systematic speaker variability. *The Journal of the Acoustical Society of America 97*(5), 3246–3247.

(Cox et Dasmahapatra, 2002) S. Cox et S. Dasmahapatra, 2002. High-level approaches to confidence estimation in speech recognition. *IEEE Transactions on Speech and Audio Processing 10*(7), 460–471.

(Daoudi et al., 2001) K. Daoudi, D. Fohr, et C. Antoine, 2001. Continuous multi-band speech recognition using bayesian networks. Dans les actes de *IEEE Automatic Speech Recognition and Understanding Workshop*, 41–44.

(Donoho, 1995) D. Donoho, 1995. De-noising by soft-thresholding. *IEEE Transactions on Information Theory 41*(3), 613–627.

(Eide, 2001) E. Eide, 2001. Distinctive features for use in automatic speech recognition. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, Aalborg, Denmark, 1613–1616.

(Ellis, 2000) D. Ellis, 2000. Feature stream combination before and/or after the acoustic model. Technical Report TR-00-007, ICSI.

(Evermann et Woodland, 2000) G. Evermann et P. Woodland, 2000. Posterior probability decoding, confidence estimation and system combination. Dans les actes de *NIST Speech Transcription Workshop*.

(Fischer et al., 2002) V. Fischer, E. Janke, et S. Kunzmann, 2002. Likelihood combination and recognition output voting for the decoding of non-native speech with multilingual hmms. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, 489–492.

(Fiscus, 1997) J. Fiscus, 1997. A post-processing system to yield reduced word error rates :recognizer output voting error reduction (rover). Dans les actes de *IEEE Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, CA, USA, 347–354.

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

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

(Galliano et al., 2005) S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, et G. Gravier, 2005. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, Lisbon, Portugal, 1149–1152.

(Garcia-Mateo et al., 1999) C. Garcia-Mateo, W. Reichl, et S. Ortmanns, 1999. On combining confidence measures in hmm-based speech recognizers. Dans les actes de *IEEE Automatic Speech Recognition and Understanding Workshop*, Keystone, 201–204.

(Gauvain, 2000) J. Gauvain, 2000. Systèmes de reconnaissance à grand vocabulaire : progrès et défis. Dans les actes de *Journées d'Études de la Parole*, Aussois, France, 31–38.

(Gauvain et Lee, 1994) J.-L. Gauvain et C.-h. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing 2*, 291–298.

(Gemello et al., 1997) R. Gemello, D. Albesano, et F. Mana, 1997. Continuous speech recognition with neural networks and stationary-transitional acoustic units. Dans les actes de *IEEE International Conference on Neural Networks*, Houston, USA, 2107–2111.

(Gemello et al., 1999) R. Gemello, D. Albesano, et F. Mana, 1999. Multi-source neural network for speech recognition. Dans les actes de *International Joint Conference on Neural Networks*, Volume 5, Washington, DC, USA, 2946–2949.

(Gemello et al., 2006) R. Gemello, F. Mana, D. Albesano, et R. De Mori, 2006. Multiple resolution analysis for robust automatic speech recognition. *Computer Speech and Language 20*(1), 2–21.

(Gemello et al., 2004) R. Gemello, F. Mana, et R. De Mori, 2004. A modified ephraim-malah noise suppression rule for automatic speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, 957–960.

(Gemello et al., 2002) R. Gemello, F. Mana, P. Perogaro, et R. De Mori, 2002. Robust multiple resolution analysis for automatic speech recognition. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Denver, Colorado.

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

(Gravier et al., 2002a) G. Gravier, S. Axelrod, G. Potamianos, et C. Neti, 2002a. Maximum entropy and mce based hmm stream weight estimation for audio-visual asr. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, 853–856.

(Gravier et al., 2002b) G. Gravier, G. Potamianos, et C. Neti, 2002b. Asynchrony modeling for audio-visual speech recognition. Dans les actes de *International Conference on Human Language Technology Research*, San Diego, California, 1–6.

(Halberstadt et Glass, 1998) A. K. Halberstadt et J. R. Glass, 1998. Heterogeneous measurements and multiple classifiers for speech recognition. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Sydney, Australia, 1379–1382.

(Hariharan et al., 2001) R. Hariharan, I. Kiss, et O. Viikki, 2001. Noise robust speech parameterization using multiresolution feature extraction. *IEEE Transactions on Speech and Audio Processing SAP-9*(8), 856–865.

(Haton et al., 2006) J. Haton, C. Cerisara, D. Fohr, Y. Laprie, et K. Smaïli, 2006. *Reconnaissance automatique de la parole : Du Signal à son Interprétation*. Dunod.

(Hegde et al., 2005) R. Hegde, H. Murthy, et G. Rao, 2005. Speech processing using joint features derived from the modified group delay function. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, Philadelphia, PA, 541–544.

(Hermansky, 1990) H. Hermansky, 1990. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America 87*, 1738–1752.

(Hermansky et al., 2000) H. Hermansky, D. Ellis, et S. Sharma, 2000. Tandem connectionist feature extraction for conventional hmm systems. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Istanbul, Turkey, 1635–1638.

(Hermansky et Morgan, 1994) H. Hermansky et N. Morgan, 1994. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing 2*(4), 578–589.

(Hoffmeister et al., 2006) B. Hoffmeister, T. Klein, R. Schluter, et H. Ney, 2006. Frame based system combination and a comparison with weighted rover and cnc. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, 537–540.

(Jelinek, 1998) F. Jelinek, 1998. *Statistical Methods for Speech Recognition*. Bradford Books.

(Jiang, 2005) H. Jiang, 2005. Confidence measures for speech recognition : A survey. *Speech Communication 45*, 455–470.

(Kamal Omar et al., 2002) M. Kamal Omar, K. Chen, M. Hasegawa-Jonhson, et Y. Brandman, 2002. An evaluation of using mutual information for selection of acoutic-features representations of phonemes for speech recognition. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Volume 1, Denver, CO, 2129–2132.

(Kamal Omar et Hasegawa-Jonhson, 2002) M. Kamal Omar et M. Hasegawa-Jonhson, 2002. Maximum mutual information based acoustic-features representation of phonological features for speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume 1, Orlando, FL, 81–84.

(Kantor et Hasegawa-Johnson, 2008) A. Kantor et M. Hasegawa-Johnson, 2008. Stream weight tuning in dynamic bayesian networks. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Las Vegas, Nevada, 4525–4528.

(Kim et Rahim, 2004) H. Kim et M. Rahim, 2004. Why speech recognizers make errors ? a robustness view. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Jeju, Korea, ThA1703o1.

(Kirchhoff, 1998) K. Kirchhoff, 1998. Combining articulatory and acoustic information for speech recognition in noise and reverberant environments.

Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Sydney, Australia, 891–894.

(Kittler et al., 1998) J. Kittler, M. Hatef, R. Duin, et J. Matas, 1998. On combining classifiers. Dans les actes de *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 20(3), 226–239.

(Lee, 2001) C.-H. Lee, 2001. Statistical confidence measures and their applications. Dans les actes de *International Conference on Speech Processing*, Daejon, CorÈe du Sud, 1021–1028.

(Leggetter et Woodland, 1995) C. J. Leggetter et P. C. Woodland, 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language 9*, 171–185.

(Mangu et al., 1999) L. Mangu, E. Brill, et A. Stolcke, 1999. Finding consensus among words : Lattice-based word error minimization. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, Volume I, 495–498.

(Mari et al., 1996) J.-F. Mari, D. Fohr, et J.-C. Junqua, 1996. A second-order hmm for high performance word and phoneme-based continuous speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Washington, DC, USA, 435–438.

(Mari et Haton, 1994) J.-F. Mari et J.-P. Haton, 1994. Automatic word recognition based on second-order hidden markov models. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Yokohama, Japan, 247–250.

(Markel et Gray Jr., 1976) J. D. Markel et A. H. Gray Jr., 1976. *Linear Prediction of Speech*. Communication and Cybernetics. Berlin Heidelberg New York : Springer-Verlag.

(Martin, 1994) R. Martin, 1994. Spectral subtraction based on minimum statistics. Dans les actes de *European Signal Processing Conference*, 1182–1185.

(Matrouf, 1997) D. Matrouf, 1997. *Adaptation des Modèles Acoustiques pour la Reconnaissance de la Parole Bruitée*. Thèse de Doctorat, LIMSI.

(Matrouf et al., 2001) D. Matrouf, O. Bellot, P. Nocera, G. Linares, et J.-F. Bonastre, 2001. A posteriori and a priori transformations for speaker adaptation in large vocabulary speech recognition systems. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, Aalborg, Denmark, 1245–1248.

(Mirghafori et Morgan, 1998) N. Mirghafori et N. Morgan, 1998. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Sydney, Australia, 743–746.

(Mirghafori et Morgan., 1998) N. Mirghafori et N. Morgan., 1998. Transmissions and transitions : A study of two common assumptions in multi-band asr. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume 2, Seattle, WA, 713–716.

(Mirghafori et Morgan, 1999) N. Mirghafori et N. Morgan, 1999. Sooner or later : exploring asynchrony in multi-band speech recognition. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, 595–598.

(Misra et al., 2003) H. Misra, H. Bourlard, et V. Tyagi, 2003. New entropy based combination rules in hmm/ann multi-stream asr. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume II, Hong-Kong, China, 741–744.

(Moreno et al., 2001) P. Moreno, B. Logan, et B. Raj, 2001. A boosting approach for confidence scoring. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, Aalborg, Denmark, 2109–2112.

(Morgan et al., 2004) N. Morgan, B. Chen, Q. Zhu, et A. Stolcke, 2004. Trapping conversational speech : Extending trap/tandem approaches to conversational telephone speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, Montreal, Canada, 537–540.

(Nocera et al., 2002) P. Nocera, G. Linarès, D. Massonié, et L. Lefort, 2002. Phoneme lattice based a* search algorithm for speech recognition. Dans les actes de *TSD*, Volume 5, Brno, République Tchèque, 301–308.

(Okawa et al., 1999) S. Okawa, T. Nakajima, et K. Shirai, 1999. A recombination strategy for multi-band speech recognition based on mutual information criterion. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, Budapest, Hungary, 603–606.

(Pearce et Hirsch, 2000) D. Pearce et H. Hirsch, 2000. The aurora experimental framework for the performance evaluation of speech recognition for mobile application. Dans les actes de *Proceedings of International Conference on Spoken Language Processing*, Beijin, China, 29–32.

(Perogaro, 2000) P. A. Perogaro, 2000. *Sviluppo di un Front-End Wavelet per il riconoscimento vocale*. Thèse de Doctorat, Università degli studi di Padova.

(Personnaz et Rivals, 2003) L. Personnaz et I. Rivals, 2003. *Réseaux de neurones formels pour la mod'elisation, la commande et la classification*. Lavoisier.

(Pujol et al., 2005) P. Pujol, S. Pol, C. Nadeu, A. Hagen, et H. Bourlard, 2005. Comparison and combination of features in a hybrid hmm/mlp and a hmm/gmm speech recognition system. *IEEE Transactions on Speech and Audio Processing SAP-13*(1), 14–22.

(Rabiner, 1989) L. Rabiner, 1989. A tutorial on hidden markov models and selected applications inspeech recognition. *Proceedings of the IEEE 77*(2), 257–286.

(Rabiner et Juang, 1993) L. Rabiner et B.-H. Juang, 1993. *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

(Raymond et al., 2006) C. Raymond, F. Béchet, R. De Mori, et G. Damnati, 2006. On the use of finite state transducers for semantic interpretation. *Speech Communication 48*(3-4), 288–304.

(Russell, 1993) M. Russell, 1993. A segmental hmm for speech pattern modelling. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume II, Minneapolis, MN, USA, 499–502.

(Sankar, 2005) A. Sankar, 2005. Bayesian model combination (baycom) for improved recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, Philadelphia, PA, 845–848.

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

(Schaaf et Kemp, 1997) T. Schaaf et T. Kemp, 1997. Confidence measures for spontaneous speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume II, Munich, Germany, 875–878.

(Schluter et Ney, 2001) R. Schluter et H. Ney, 2001. Using phase spectrum information for improved speech recognition performance. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, 133–136.

(Schwenk et Gauvain, 2000) H. Schwenk et J.-L. Gauvain, 2000. Combining multiple speech recognizers using voting and language model information. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Volume 2, Beijing, China, 915–918.

(Servan et al., 2006) C. Servan, C. Raymond, F. Béchet, et P. Nocéra, 2006. Conceptual decoding from word lattices : application to the spoken corpus media. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Pittsburgh, Pennsylvania, 1614–1617.

(Shannon, 1948) C. Shannon, 1948. A mathematical theory of communication. Dans les actes de *The Bell System Technical Journal*, Volume 27, 379–423, 623–656.

(Siohan et al., 2005) O. Siohan, B. Ramabhadran, et B. Kingsbury, 2005. Constructing ensembles of asr systems using randomized decision trees. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, Philadelphia, PA, 197–200.

(Siu et Gish, 1999) M. Siu et H. Gish, 1999. Evaluation of word confidence for speech recognition systems. *Computer Speech and Language 13*, 299–319.

(Souvignier et Wendemuth, 1999) B. Souvignier et A. Wendemuth, 1999. Combination of confidence measures for phrases. Dans les actes de *IEEE Automatic Speech Recognition and Understanding Workshop*, Budapest, Hungary, 217–220.

(Stevens, 1998) K. N. Stevens, 1998. *Acoustic Phonetics*. The MIT press.

(Stevens, 1957) S. S. Stevens, 1957. On the psychophysical law. *Psychological Review 64*, 153–181.

(Thomson et Chengalvarayan, 1998) D. Thomson et R. Chengalvarayan, 1998. Use of periodicity and jitter as speech recognition feature. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, Seattle, WA, 21–24.

(Utsuro et al., 2003) T. Utsuro, Y. Kodama, T. Watanabe, H. Nishizaki, et S. Nakagawa, 2003. Confidence of agreement among multiple lvcsr models and model combination by svm. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, Hong Kong, China, 16–19.

(Vaseghi et al., 1997) S. Vaseghi, N. Harte, et B. Miller, 1997. Multi resolution phonetic/segmental features and models for hmm-based speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Munich, Germany, 1263–1266.

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

(Woodland et al., 1996) P. Woodland, M. Gales, et D. Pye, 1996. Improving environmental robustness in large vocabulary speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume 1, Los Alamitos, CA, USA, 65–68.

(Zhang et Rudnicky, 2001) R. Zhang et A. Rudnicky, 2001. Word level confidence annotation using combinations of features. Dans les actes de *European Conference on Speech Communication and Technology, Interspeech*, Aalborg, Denmark, 2105–2108.

(Zolnay et al., 2002) A. Zolnay, R. Schluter, et H. Ney, 2002. Robust speech recognition using a voiced-unvoiced feature. Dans les actes de *International Conference on Spoken Language Processing, Interspeech*, Volume II, Denver, CO, 1065–1068.

(Zolnay et al., 2005) A. Zolnay, R. Schluter, et H. Ney, 2005. Acoustic feature combination for robust speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech and Language Processing*, Volume I, Philadelphia, PA, 457–460.

Publications personnelles

Conférences internationales

L. BARRAULT, C. SERVAN, D. MATROUF, G. LINARÈS ET R. DE MORI. « Frame-based acoustic feature integration for speech understanding » *dans les actes de IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'08)*, Las Vegas (Nevada), 2008.

L. BARRAULT, D. MATROUF, R. DE MORI, R. GEMELLO AND F. MANA. « Dynamic selection of acoustic features in an automatic speech recognition system » *dans les actes de EUropean SIgnal Processing COnference (EUSIPCO'06)*, Florence (Italy), 2006.

L. BARRAULT, D. MATROUF, R. DE MORI, R. GEMELLO AND F. MANA. « Characterizing feature variability in automatic speech recogntinion systems » *dans les actes de IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'06)*, Toulouse (France), 2006.

L. BARRAULT, R. DE MORI, R. GEMELLO AND F. MANA. « Variability of automatic speech recognition systems using different features » *dans les actes de European Conference on Speech Communication and Technology (Eurospeech'05)*, Lisbon (Portugal), pp. 221-224, Sept. 2005.

Conférences nationales

L. BARRAULT, D. MATROUF, G. LINARÈS ET R. DE MORI. « Combinaison de différents jeux de paramètres acoustiques pour la reconnaissance de la parole » *dans les actes de Journées d'Études de la Parole (JEP'08)*, Avignon (France),

2008.

L. BARRAULT. « Etude des variabilités des systèmes de reconnaissance automatique de la parole utilisant des paramètres acoustiques différents » *Rencontres des Jeunes Chercheurs en Parole (RJC'05)*, Toulouse (France), 2005.

Workshops, Ateliers

D. MATROUF, L. BARRAULT ET R. DE MORI. « A general method for combining acoustic features in an automatic speech recognition system » *Speech Recognition and Intrinsic Variation (SRIV) Workshop*, Toulouse (France), 2006.

Annexes

Annexe A

Résultats de diagnostic

A.1 Analyse de l'entropie des vecteurs de probabilités *a posteriori*

Avec pour objectif d'analyser l'information que peut apporter l'entropie des vecteurs de probabilités en terme de classification, nous avons calculé les distributions des trames en fonction de l'entropie pour chaque type de paramètres acoustiques et leur combinaison. Nous avons considéré deux classes : les trames bien classifiées et celles mal classifiées. Ces deux classes correspondent aux trames dont l'état ayant le maximum de probabilité correspond ou non à celui donné par l'alignement forcé.

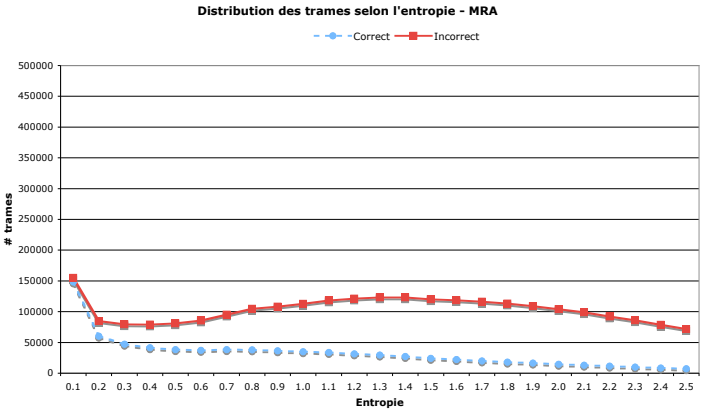

FIG. ***A.1:*** *Couverture des trames bien (« Correct ») et mal (« Incorrect ») classifiées en fonction de l'entropie avec le modèle utilisant les paramètres acoustiques MRA sur le corpus MEDIA.TRAIN.*

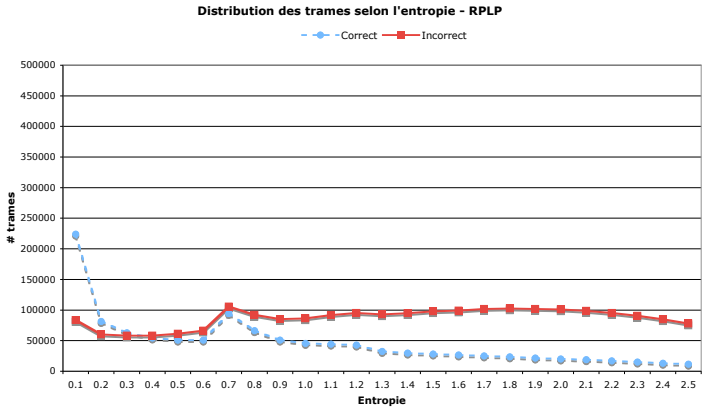

FIG. *A.2:* *Couverture des trames bien (« Correct ») et mal (« Incorrect ») classifiées en fonction de l'entropie avec le modèle utilisant les paramètres acoustiques RPLP sur le corpus MEDIA.TRAIN.*

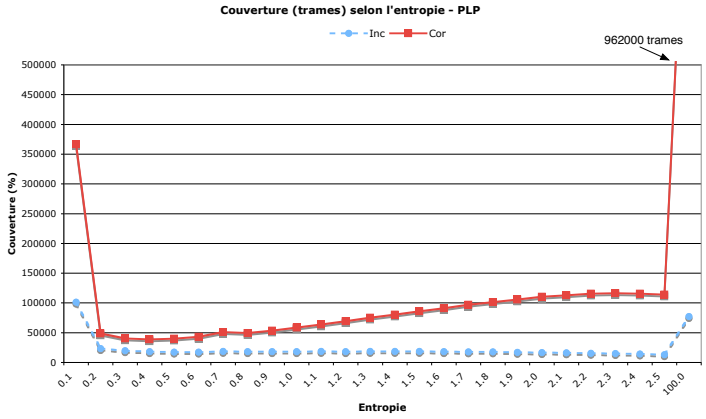

FIG. *A.3:* *Couverture des trames bien (« Correct ») et mal (« Incorrect ») classifiées en fonction de l'entropie avec le modèle utilisant les paramètres acoustiques PLP sur le corpus MEDIA.TRAIN.*

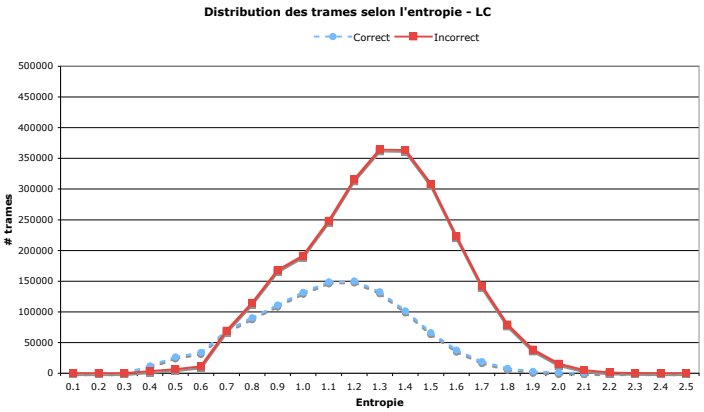

FIG. *A.4:* *Couverture des trames bien (« Correct ») et mal (« Incorrect ») classifiées en fonction de l'entropie avec la combinaison LC sur le corpus MEDIA.TRAIN.*

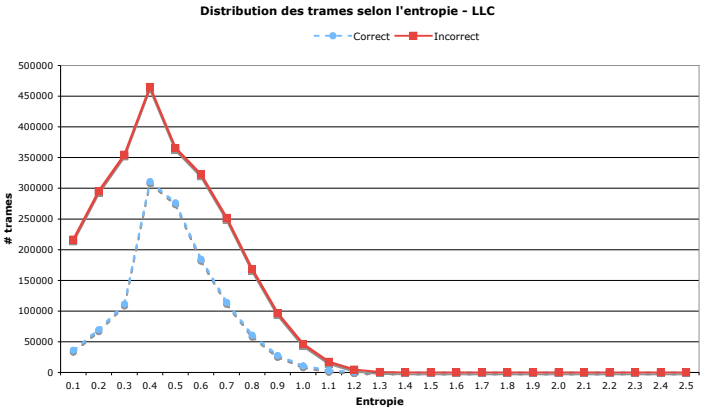

FIG. *A.5:* *Couverture des trames bien (« Correct ») et mal (« Incorrect ») classifiées en fonction de l'entropie avec la combinaison LLC sur le corpus MEDIA.TRAIN.*

On observe tout d'abord les mêmes tendances que sur la courbe 5.8 (p. 133), à savoir que l'entropie des vecteurs de probabilités obtenus par combinaison (linéaire ou log-linéaire) réduit fortement l'entropie moyenne. On remarque cependant que les distributions dans les deux classes considérées évoluent de manière très similaire. De ce fait, le pouvoir discriminant de l'entropie des vecteurs de probabilités ne semble pas être très prononcé. Ceci pourrait expliquer le fait que l'entropie des vecteurs de probabilités *a posteriori* ne fournit pas d'amélioration des résultats lorsqu'ils sont utilisés pour la pondération de la participation de chaque système.

A.1.1 Distribution normale de l'entropie

J'ai voulu mettre en évidence le pouvoir de discrimination des vecteurs de probabilités en modélisant la distribution de leur entropie par une gaussienne en séparant les trames bien classifiées de celles mal classifiées.

De manière générale, les trames mal classifiées ont une variance plus élevée que celles correctes. Un fait marquant est que les distributions pour les trames incorrectes ont toujours une moyenne plus faible que les distributions pour les trames correctes excepté pour la combinaison LLC.

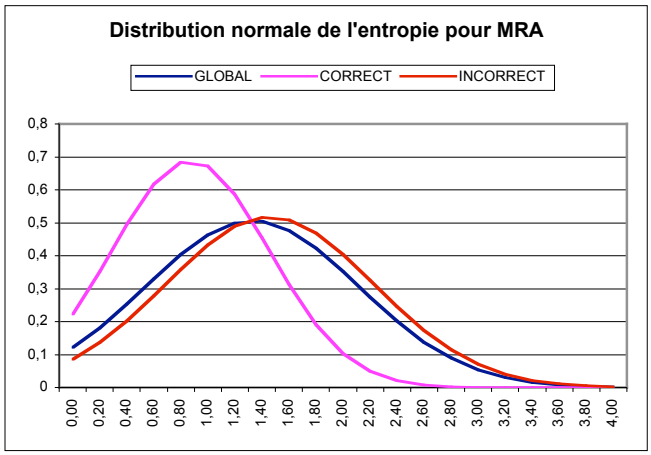

FIG. *A.6:* *Distribution gaussienne de l'entropie des vecteur de probabilités calculées avec le modèle utilisant les paramètres acoustiques MRA sur le corpus MEDIA.TRAIN.*

Pour LLC, on observe que les distributions sont très proches. De ce fait, l'entropie ne permet pas de distinguer une trame correcte d'une trame incorrecte (par mesure de vraisemblance). Par contre la quantité de trames incorrectes est beaucoup plus faible que la quantité de trames correctes. Ce qui peut expliquer les meilleurs résultats obtenus avec ce type de combinaison. Pour LC,

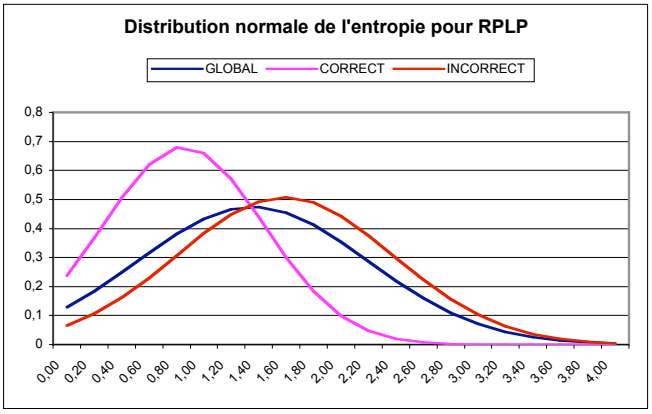

FIG. *A.7:* *Distribution gaussienne de l'entropie des vecteur de probabilités calculées avec le modèle utilisant les paramètres acoustiques RPLP sur le corpus MEDIA.TRAIN.*

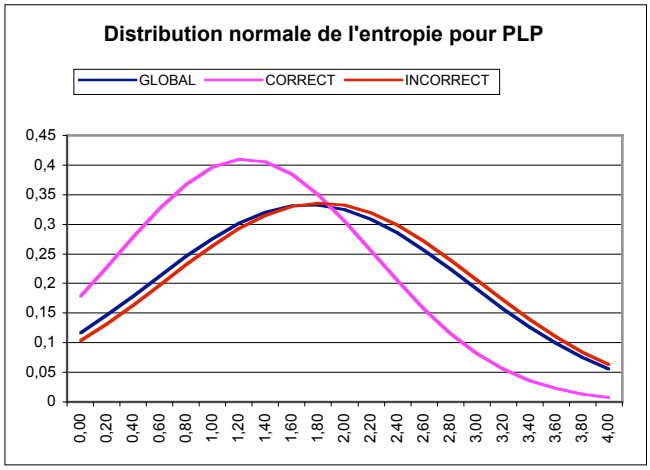

FIG. *A.8:* *Distribution gaussienne de l'entropie des vecteur de probabilités calculées avec le modèle utilisant les paramètres acoustiques PLP sur le corpus MEDIA.TRAIN.*

les conclusions sont différentes. En effet, les distributions ont la même forme (nombre de trames dans les deux classes relativement proche), mais la moyenne de la distribution de l'entropie des trames correctes est plus faible que celle de la distribution des trames incorrectes.

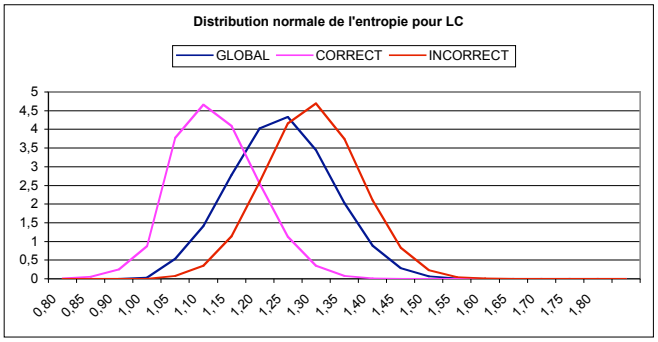

FIG. *A.9:* *Distribution gaussienne de l'entropie des vecteurs de probabilités calculées avec la combinaison LC sur le corpus MEDIA.TRAIN.*

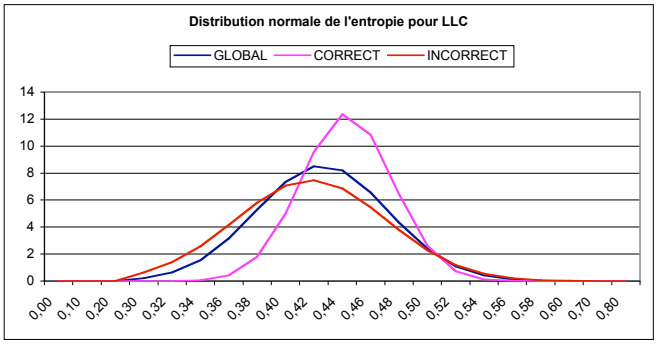

FIG. *A.10:* *Distribution gaussienne de l'entropie des vecteurs de probabilités avec la combinaison LLC sur le corpus MEDIA.TRAIN.*

Une perspective à cette analyse est d'exploiter l'information obtenue avec l'entropie des distributions de probabilités générées par la LC et par la LLC. On utilise la distribution de probabilité LLC sauf quand la vraisemblance est grande pour la classe « correct » avec la LC.

Annexe B

Résultats de reconnaissance complémentaires

Afin de tester les différentes approches sur un corpus plus petit, j'ai utilisé les diverses techniques de combinaison présentées dans cette thèse sur le corpus de développement de MEDIA (MEDIA.DEV). Ce corpus comporte 1377 phrases pour un total de 10434 mots.

Il est relativement proche du corpus de test, ce qui explique que les taux de reconnaissance sont proches de ceux obtenus sur le corpus de test.

Exemple de dialogue issu de MEDIA

Voici un exemple de dialogue complet.

WOz	Bienvenue sur le serveur MEDIA système d'informations touristiques et de réservation d'hôtel. Quelle information souhaitez-vous ?
Spk	**allô. Je souhaite une réservation pour euh une pour une une chambre pour Rennes début septembre deux chambres pour Rennes début septembre mini au maximum cent cinquante euros dans un quartier touristique**
WOz	Vous souhaitez faire une réservation à Rennes ?
Spk	**oui**
WOz	pour
Spk	**oui**
WOz	le mois de septembre
Spk	**oui**
WOz	Merci de patienter, je recherche vos informations à Rennes. Je vous propose deux hôtels : l'hôtel Anne de Bretagne, la chambre est à cent cinquante euros, l'hôtel duc
Spk	**oui**
WOz	duché la chambre est à cent trente euros. Voulez-vous réserver dans l'un de ces deux hôtels
Spk	**la première chambre à cent cinquante euros**
WOz	Vous souhaitez réserver à l'hôtel Anne de Bretagne ?
Spk	**allô**
WOz	Vous souhaitez réserver à l'hôtel Anne de Bretagne ?
Spk	**euh la première chambre**
WOz	Veuillez patienter, il reste des chambres libres dans cet hôtel, j'effectue votre réservation. Le montant de votre séjour s'élève à mille huit cents euros. Souhaitez-vous confirmer la réservation ?
Spk	**euh s() mille huit cents euros attendez hum mille huit cents euros c' est non c' est trop cher**
WOz	La réservation s'effectue pour six jours deux chambres
Spk	**oui tout à fait ah oui oui tout à fait s() oui oui oui je confirme**
WOz	Le numéro de dossier est le cent soixante treize deux cent quarante huit. Désirez-vous une
Spk	**d'acco(rd)**
WOz	autre information
Spk	**non merci par contre si par contre je voudrais savoi(r) je je je suis bien dans le quartier touristique**
WOz	Cet hôtel se situe dans un quartier touristique à Rennes
Spk	**d'accord ben c' est parfait**
WOz	Nous vous remercions d'avoir utilisé le serveur MEDIA. Au revoir
Spk	**Au revoir**

B.1 Résultats de reconnaissance sans pondération

Jeu(x) de paramètres	WER (%)	Gain relatif (%)	Int. de conf. (%)
MRA	33.1	-	0.90
RPLP	33.1	-	0.90
PLP	32.0	-	0.90
En utilisant la combinaison linéaire			
MRA+RPLP	28.2	11.5	0.86
MRA+PLP	27.2	15.0	0.85
RPLP+PLP	27.3	14.7	0.86
MRA+RPLP+PLP	27.9	12.2	0.86
En utilisant la combinaison log-linéaire			
MRA+RPLP	28.7	13.8	0.87
MRA+PLP	27.2	15.0	0.85
RPLP+PLP	27.4	14.4	0.86
MRA+RPLP+PLP	**26.7**	**16.5**	0.85
ROVER	28.4	11.3	0.87

TAB. *B.1:* *Résultats de la combinaison trame à trame sur le corpus de développement de MEDIA (1377 phrases et 10434 mots).*

On observe la même tendance que pour le corpus MEDIA.TEST. Les meilleurs résultats sont obtenus avec la combinaison log-linéaire des probabilités *a posteriori*. Ceux-ci dépassent la combinaison post-décodage par vote majoritaire ROVER.

B.2 Utilisation des différentes techniques de pondération

Type de pondération	WER (%)	Int. de conf. (%)
Baseline (LLC)	26.7	0.85
Matrice de confusion	26.5	0.85
Régressions logistiques	26.5	0.85
Entropie inverse	27.5	0.86
MAX KLD	27.3	0.85
MIN KLD	27.3	0.85

TAB. *B.2:* *Utilisation de différentes techniques de pondération sur MEDIA.DEV.*

En ce qui concerne la pondération par matrice de confusion et par régressions logistiques, les améliorations obtenues ne sont pas significatives. Tout comme pour le corpus MEDIA.TEST, on observe un très léger gain lorsque l'on utilise de telles fonctions de pondération.

L'utilisation d'un facteur de pondération inversement proportionnel à l'entropie du vecteur de probabilités a dégradé les résultats. Les analyses présentées dans l'annexe A.1 révèlent que l'entropie des vecteurs de probabilités montre une grande variabilité. En effet, il est très difficile de distinguer un vecteur correct d'un vecteur incorrect par l'intermédiaire de ce type de mesure.

La divergence de Kullback-Leibler permet de quantifier l'écart entre deux distributions de probabilités. À partir des trois systèmes considérés, j'ai combinés les deux systèmes selon plusieurs critères reliés à la KLD. Le premier critère correspond au fait que deux systèmes proposant des distributions différentes vont permettre de corriger certaines erreurs que l'autre système peut commettre. Ce critère revient à sélectionner et à combiner les distributions de probabilités dont la KLD est la plus grande. Le second critère considère que si deux systèmes fournissent une distribution de probabilités proches, alors ils ont plus de chance d'être correct. Ce critère revient à sélectionner et à combiner les vecteurs de probabilités dont la KLD est la plus petite.

B.3 Utilisation d'un quatrième jeu de paramètres

Nous avons considéré l'utilisation des paramètres MFCC afin d'augmenter la quantité d'information en entrée du système de reconnaissance. La procédure d'apprentissage de ce modèle est strictement la même que pour les autres modèles. La procédure jumeau a été utilisée pour apprendre un premier modèle avec le corpus d'apprentissage d'ESTER. Ensuite, ce premier modèle a été adapté en utilisant le corpus d'entraînement de MEDIA.

Jeu(x) de paramètres	WER (%)	Gain relatif (%)	Int. de conf. (%)
MFCC	30.9	-	0.89
LLC MRA+RPLP+PLP	27.6	10.7	0.86
LLC MRA+RPLP+PLP+MFCC	27.3	11.7	0.85

TAB. *B.3:* *Résultats de reconnaissance avec les paramètres MFCC sur le corpus MEDIA.TEST.*

Les résultats obtenus avec le système utilisant le jeu de paramètres MFCC seul sont meilleurs que ceux obtenus avec les autres paramétrisations. En effet, le meilleur système (celui utilisant les PLP) a un taux d'erreur de 32.0%, alors

que les MFCC fournissent un taux d'erreur de de 30.9%. On aurait pu espérer obtenir une amélioration conséquente des performances du système combinant les 4 jeux de paramètres. Cela n'a pas été le cas, puisqu'une diminution du WER de seulement 0.3% a été observée.

Cela peut vouloir dire plusieurs choses. D'une part l'information acoustique apportée par les MFCC n'est peut-être pas assez différente de celle déjà présente. Les paramètres MFCC sont également des paramètres cepstraux et ne sont pas fondamentalement différents des paramètres PLP. De ce fait, les caractéristiques extraites du signal ne sont probablement pas très différentes. D'autre part, la quantité d'information acoustique nécessaire à la bonne reconnaissance (si jamais elle a une limite) a été atteinte et par conséquent l'ajout d'une nouvelle source d'information acoustique n'apporte pas d'amélioration. Dans ce cas, c'est le modèle de langage et/ou le processus de décodage qui doivent être perfectionnés pour améliorer le système.

Annexe C

De la transcription au décodage conceptuel

Nous avons développé une technique de combinaison fournissant une amélioration significative des performances du système de reconnaissance de la parole. La plupart des applications actuelles nécessitent plus que la transcription du signal de parole, elles veulent comprendre le message contenu dans la suite de mots prononcée. Les résultats encourageants obtenus ne seront utiles que si les mots porteurs de sens sont mieux reconnus, et pas seulement les mots utilitaires.

Afin de vérifier l'impact de l'amélioration de la transcription sur la compréhension de la parole, nous avons comparé les résultats de décodage conceptuel obtenus avec les modèles utilisant un seul jeu de paramètres (MRA, RPLP et PLP) à ceux obtenus avec le pseudo-modèle combiné (LLC).

Pour ce faire, nous avons utilisé le corpus MEDIA.TEST. Ce corpus est composé de 2992 tours de parole correspondant à 3771 phrases et 26092 mots. Il a été transcrit manuellement et annoté conceptuellement selon une représentation sémantique décrite dans (Servan et al., 2006).

C.1 Description du décodeur conceptuel

Un système à base de transducteurs à nombre d'états fini (Raymond et al., 2006) est utilisé pour extraire les séquences de concepts associées aux phrases issues du module de transcription. Ce processus de décodage menant à la production d'une liste structurée des N-meilleures hypothèses conceptuelles est présenté dans le schéma C.1.

Les sorties du module ASR ont été générées avec les systèmes utilisant un seul jeu de paramètres puis avec le système effectuant la combinaison des probabilités au niveau de la trame.

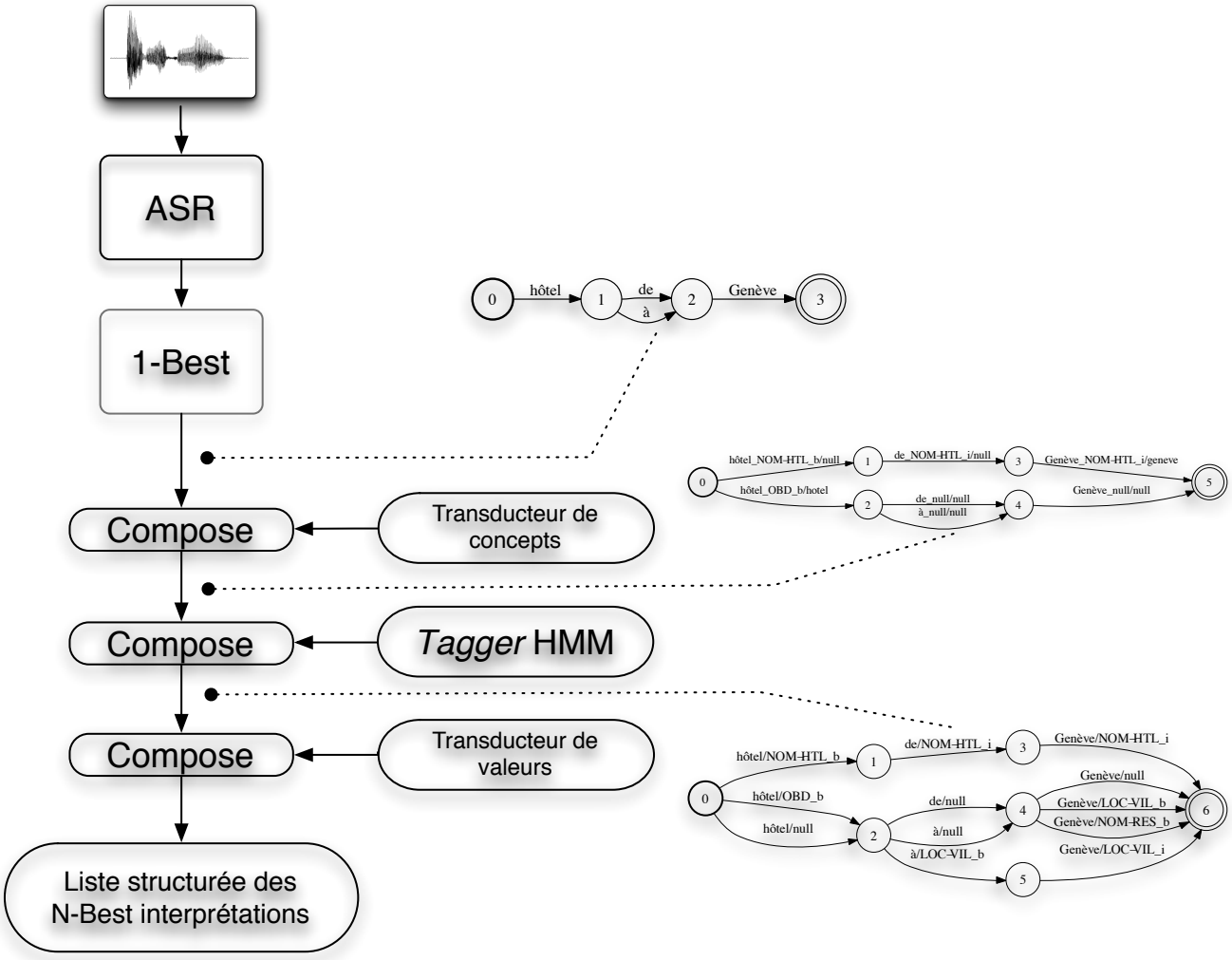

FIG. *C.1:* *Architecture du décodeur conceptuel.*

C.2 Résultats et observations

Les résultats en terme de taux d'erreur concept (CER) sont présentés dans le tableau C.1.

On remarque que l'amélioration des performances du systèmes de transcription conduit à une diminution du CER.

	MRA	RPLP	PLP	LLC
WER (%)	33.0	33.0	32.0	27.6
CER (%)	37.0	37.1	35.1	32.4

TAB. *C.1:* *CER (%) obtenus avec la meilleure hypothèse conceptuelle sur MEDIA.TEST.*

Le tableau C.2 rend compte des valeurs de CER et WER pour différentes situations. La colonne «LLC Gagnant» correspond aux cas où LLC fournit un taux d'erreur concept plus faible que le système utilisant un jeu de paramètres acoustique seul. La colonne «LLC Perdant» correspond aux cas opposés. La colonne «Consensus» correspond aux situations où les systèmes proposent la même hypothèse. Il y a consensus entre LLC et PLP dans plus de 71% des tours de parole. Dans ce cas, on observe un taux d'erreur très faible. Par conséquent, le consensus apparaît comme un indicateur de confiance valable.

Comparaison entre LLC et PLP				
	LLC Gagnant	Consensus	LLC Perdant	Total
% tours	13.1	71.5	8.7	100
CER LLC	35.5	**24.7**	57.8	32.4
CER PLP	61.6	**24.7**	30.9	35.1
WER LLC	32.6	**22.7**	36.1	28.1
WER PLP	42.8	**26.1**	35.2	32.8
Comparaison entre LLC et RPLP				
	LLC Gagnant	Consensus	LLC Perdant	Total
% tours	15.1	70.7	7.4	100
CER LLC	30,9	**27,3**	56,4	32,4
CER RPLP	60,2	**27,3**	34	37,1
WER LLC	30,7	**23,2**	36,5	28,1
WER RPLP	41,7	**25,8**	36	32,8
Comparaison entre LLC et MRA				
	LLC Gagnant	Consensus	LLC Perdant	Total
% tours	15.9	69.5	7.7	100
CER LLC	32,1	**26,7**	56,8	32,4
CER MRA	60	**26,7**	32,1	37.0
WER LLC	30,4	**23,8**	35,3	28,1
WER MRA	39,6	**28.0**	38,9	33,9

TAB. *C.2:* *Relation entre les performances de reconnaissance conceptuelle et de reconnaissance de la parole (2992 tours de parole au total).*

L'utilisation de plusieurs jeux de paramètres acoustiques fournit une réduction significative du CER. Lorsque les systèmes utilisant différents jeux de paramètres proposent les mêmes hypothèses conceptuelles, alors il est fortement

probable que les hypothèses sont correctes. Un faible WER est obtenu pour les phrases dont les différents systèmes fournissent la même interprétation sémantique. Le consensus parmi les hypothèses conceptuelles obtenues avec différents jeux de paramètres acoustiques est un bon indicateur de confiance autant pour la reconnaissance de la parole que pour le décodage conceptuel.

En outre, des comportements différents peuvent être observés pour les phrases où un jeu de paramètres obtient de meilleurs résultats que la combinaison. En effet, on observe un WER beaucoup plus faible pour les phrases pour lesquelles un jeu de paramètres utilisé seul fournit de meilleures hypothèses conceptuelles que la combinaison. Notamment pour PLP et RPLP, le taux d'erreur dans ces cas là est même plus faible que pour LLC.

Il y a un lien certain entre le résultat de reconnaissance conceptuelle et celui de la reconnaissance de la parole. L'amélioration de la transcription fournie par la combinaison LLC concerne effectivement des mots porteurs de sens et a donc un impact direct sur les interprétations sémantiques.

Score Oracle et perspectives

Nous avons considéré un Oracle sélectionnant la bonne hypothèse conceptuelle si elle est proposée parmi la liste des N-meilleures.

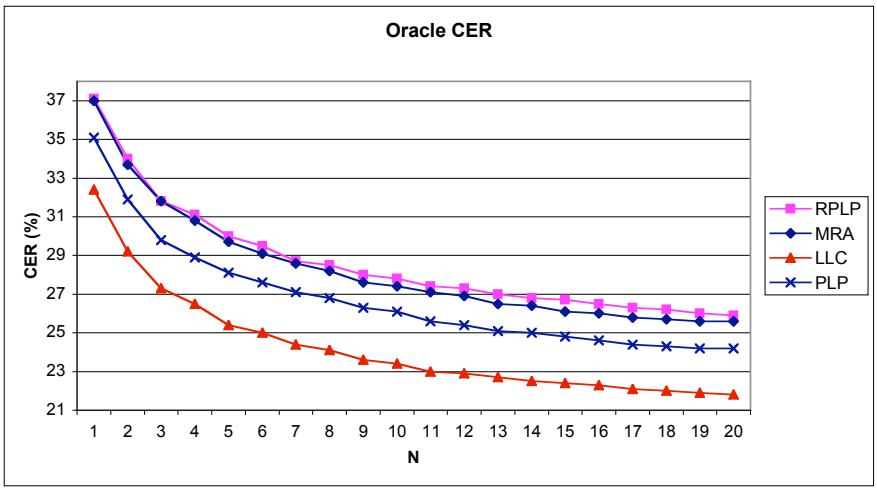

FIG. C.2: *Évolution du CER oracle en fonction de la taille N de la liste des N-meilleures hypothèses.*

On remarque qu'avec une méthode adéquate permettant de sélectionner la bonne hypothèse dans la liste des N-meilleures hypothèses conceptuelles, on pourrait diminuer le CER de plus de 30% relativement au résultat n'utilisant que la meilleure hypothèse.

Ce résultat encourage l'exploitation d'un plus grand nombre d'hypothèses de phrases pour le décodage conceptuel. Une extension de ce principe est l'utilisation du treillis de mots complet. Cette stratégie à déjà été mise en œuvre avec le jeu de paramètres PLP. Les résultats obtenus sont proches de ceux que l'on obtient en utilisant la meilleure hypothèse conceptuelle avec le système de reconnaissance de la parole combinant les différents jeux de paramètres. On peut alors espérer que l'utilisation du treillis de mots issu de la combinaison des jeux de paramètres apportera une amélioration supplémentaire par rapport à l'utilisation de seulement la meilleure hypothèse.