Université du Maine

Université du Maine

MESURES DE CONFIANCE EN TRAITEMENT AUTOMATIQUE DE LA PAROLE ET APPLICATIONS

THÈSE

présentée et soutenue publiquement le 5 décembre 2006

pour l'obtention du

Doctorat de l'Université du Maine

(spécialité informatique)

par

JULIE MAUCLAIR

Composition du jury

Rapporteurs :	Renato De Mori	Professeur	LIA, Université d'Avignon et des Pays du Vaucluse
	Kamel Smaïli	Professeur	LORIA, Université de Nancy II
Examinateurs :	Claude Barras	Maître de Conférences	LIMSI-CNRS, Université de Paris XI
	Paul Deléglise	Professeur	LIUM, Université du Maine
	Yannick Estève	Maître de Conférences	LIUM, Université du Maine
Présidente :	Béatrice Daille	Professeure	LINA, Université de Nantes

Laboratoire d'Informatique de l'Université du Maine

lium

Il ne suffit pas de parler, il faut parler juste.
William Shakespeare (extrait de “Le songe d’une nuit d’été”)

Remerciements

Voilà, c'est fini. Trois ans de ma vie (presque quatre) résumés en quelques pages qui n'auraient pas été écrites sans le soutien des personnes à qui je rend hommage dans ces quelques lignes.

Ce doctorat doit beaucoup à MM. Paul Deléglise et Yannick Estève.
Tout d'abord, merci à Paul d'avoir dirigé ma thèse en ayant su me prodiguer soutien, conseils, éclaircissements, relecture intensive et aide en tous points.
Merci Yannick de m'avoir encadrée tout au long de ces trois années avec la chaleur méditerranéenne dont tu sais faire preuve. Je souhaite à tout thésard d'avoir un encadrant comme toi.
Merci à Bruno, Simon, Sylvain et Téva qui ont su répondre à mes questions quand je n'osais pas les (re)poser à Paul ou à Yannick.

Je remercie MM. Kamel Smaïli et Renato De Mori d'avoir accepté d'être les rapporteurs de cette thèse et de m'avoir offert des conseils et des encouragements très motivants et enthousiasmants.
Merci également à Mme Béatrice Daille et M. Claude Barras d'avoir accepté de faire partie de mon jury et de m'avoir prodigué des remarques très pertinentes.

En filigrane de cette thèse, il y a le laboratoire qui m'a accueilli dans cette ville du Mans.
Je remercie vraiment toutes les personnes que j'ai rencontrées pendant ces quatre ans au LIUM et au Mans. Merci à Bérangère, Bruno, Dorothée, Catherine, Christelle, Christophe, Etienne, Johan, Julien, Naïma, Mathilde, Mickaël, Neil, Marie-Laure, Pierre, Stéphane et tous ceux que j'oublie, notamment pour les pauses en salle café, les gourmandes et les autres, toujours dans la bonne humeur.
Merci également pour les soirées (en particulier les soirées couscous de Naïma) qui resteront longtemps dans les mémoires.
Un grand merci aussi à mes collègues de bureau qui m'ont apporté beaucoup de fous rires et de chocolat : Arnaud et mon prof de chant Mathieu.
À tous ceux de Toulouse qui ne m'ont pas oubliée, je tiens à dire que c'est grâce à eux que j'ai voulu faire une thèse, même si ça voulait dire quitter Toulouse. Un merci tout particulier à Julien qui a été mon tout premier encadrant et à Régine.

Je dédie également ce manuscrit à mes amies de toujours et aux membres de ma famille qui ont su endurer mes sautes d'humeur avec une infinie patience. Merci donc à Céline, Emeline, Stéphanie et à mes parents.

Et un merci rempli d'amour à Thomas, qui a toujours été là.

Table des matières

Table des figures

Liste des tableaux

Acronymes

AC	Acoustique
BE	Bande Étroite
BIC	Bayesian Information Criterion
CA	Confidence Acuracy
CER	Confidence Error Rate
CMLLR	Constrained Maximum Likelihood Linear Regression
CMU	Carnegie Mellon University
DET	Detection Error Tradeoff
EER	Equal Error Rate
EM	Expectation Maximisation
ESTER	Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophonique
FA	Fausses Acceptations
FE	Faux Rejets
HMM	Hidden Markov Model
LB	Large Bande
LMBB	Language Model Back-off Behavior
LPC	Linear Prediction Coding
LSA	Latent Semantic Analysis
MAP	Maximum *a posteriori*
MFCC	Mel-scale Frequency Cepstral Coefficient
MLLR	Maximum Likelihood Linear Regression
NCE	Normalised Cross Entropy
NIST	National Institute of Standards and Technologies
PAP	Probabilité *a posteriori*
PLP	Perceptually-based Linear Prediction
RAL	Reconnaissance Automatique du Locuteur
RAL	Reconnaissance Automatique de la Parole
SAT	Speaker Adaptative Training
SRAP	Système de Reconnaissance Automatique de la Parole
WER	Word Error Rate

Introduction

Au cours des dernières décennies, l'évolution technologique et scientifique a permis le développement considérable de nouveaux outils pour le traitement de la parole. De manière générale, des domaines tels que la reconnaissance de la parole, la reconnaissance du locuteur, les systèmes de dialogue, la traduction automatique, l'indexation de documents sonores connaissent un essor conséquent.

La quantité de documents sonores devenant très importante, il est devenu indispensable de développer des outils automatiques pour exploiter ces documents à des fins d'indexation ou d'annotation par exemple. Les systèmes de reconnaissance automatique de la parole (SRAP) sont généralement au coeur de ces outils. Le développement de ces outils nécessite encore des efforts de recherche.

Pour favoriser le développement du traitement de la parole et fédérer les efforts de recherche des acteurs de ce domaine, des campagnes d'évaluation sont proposées. Ces campagnes permettent de faire travailler ensemble des laboratoires et de dresser un état de l'art sur le domaine du traitement de la parole.

Aux États-Unis, ces campagnes d'évaluation sont généralement organisées chaque année par NIST (National Institute of Standards and Technologies)[1]. Les premières concernaient l'évaluation des systèmes de transcription d'émissions (HUB-4 Broadcast News). Des campagnes de détection de thèmes et d'entités nommées sont apparues ensuite avec Automatic Content Extraction et Topic Detection and Tracking. Plus récemment, la campagne Rich Transcription concerne la transcription enrichie avec des informations sur le locuteur.

Ces campagnes permettent de dynamiser les travaux dans le domaine du traitement de la parole mais se limitent pour la plupart à la langue anglaise et dans une moindre mesure à l'espagnol.

L'objectif du projet ESTER (Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophonique)[2] [Gravier 2004] est d'organiser une campagne d'évaluation autour de ces systèmes de reconnaissance automatique de la parole en français. La campagne est organisée dans le cadre du projet EVALDA sous l'égide scientifique de l'Association Francophone de la Communication Parlée avec le concours du Centre d'Expertise Parisien de la Délégation Générale de l'Armement et de ELDA (Evaluations and Language resources Distribution Agency). Elle est financée par le Ministère de la Recherche dans le cadre de l'appel à projet TECHNOLANGUE.

[1] `http://www.nist.gov/`
[2] `http://www.afcp-parole.org/ester/index.html`

La campagne d'évaluation ESTER constitue le contexte de mise en place du système de Traitement Automatique de la Parole (TAP) du LIUM sur lequel se basent les travaux développés dans ce mémoire. Cette campagne a permis d'accéder à des corpora conséquents ainsi qu'à un cadre d'évaluation reconnu par la communauté scientifique.

1 Description de la campagne d'évaluation ESTER

La campagne d'évaluation ESTER vise à l'évaluation des performances des systèmes de transcription d'émissions radiophoniques. Les transcriptions sont enrichies par un ensemble d'informations annexes comme le découpage automatique en tours de parole, le marquage des entités nommées, etc. La transcription enrichie vise à obtenir une transcription lisible et une représentation structurée du document à des fins d'extraction d'informations. Les émissions comportent des phases de parole lue et des phases de parole spontanée (interviews, débats, conversations téléphoniques).

Pour la phase 1, menant à un test blanc, l'évaluation s'articule autour de deux tâches : la transcription orthographique qui consiste à évaluer les systèmes de TAP, et la segmentation qui vise à évaluer les systèmes de suivi d'évènements sonores ou de locuteurs ainsi que les systèmes d'indexation selon le locuteur. La phase 2 (test final) inclue également la tâche de recherche d'information.

Les données mises à disposition des participants de la campagne (voir tableau 1.1) pour l'apprentissage et la mise au point des systèmes sont les enregistrements sonores d'émissions radiophoniques francophones (90 heures d'émissions provenant de France Inter, France Info, Radio France Internationale (RFI), Radio Télévision Marocaine (RTM)), les transcriptions enrichies de ces 90 heures d'enregistrements, ainsi que des dictionnaires de phonétisation de mots en français. Ces données ont été enregistrées sur trois périodes : 1998, 2000 et 2003. De plus, un corpus textuel correspondant aux années de 1987 à 2003 du journal "Le Monde" augmenté du corpus MLCC contenant des transcriptions des débats du Conseil Européen est fourni. Enfin, un corpus audio non transcrit d'environ 2000 heures datant du dernier trimestre 2003 à septembre 2004 est fourni. Ce corpus contient les mêmes radios que celles comprises dans les données transcrites plus des enregistrements de France Culture.

La phase de test, qui a eu lieu début 2005, porte sur de nouvelles données : 10 heures d'émissions provenant de RFI, RTM, France Info, France Inter, France Culture et une radio "surprise". Les données de test ont été enregistrées entre le 01/10/2004 et le 31/12/2004.

Les différents corpora ainsi que le package d'évaluation contenant aussi les protocoles d'évaluation et les outils de mesure de performance sont diffusés par ELDA.

2 Les différentes tâches d'ESTER

Dans la campagne ESTER, on retrouve trois catégories de tâches :

- **La transcription orthographique (TRS, TTR) :** cette catégorie propose une transcription orthographique de l'émission radiophonique en temps réel (TTR) ou non-contraint (TRS). Pour évaluer cette transcription, le taux d'erreur mot (Word Error Rate) est calculé. On retrouve dans cette catégorie diverses tâches visant à déterminer l'influence de la taille du vocabulaire, des données d'apprentissage ou encore du temps de calcul.
- **La segmentation en évènement sonore :** cette catégorie vise à la détection et au regroupement d'évènements sonores. Elle comporte :
  - SES : Suivi d'évènements sonores
  - SRL : Segmentation et regroupement de locuteurs
  - SVL : Suivi du locuteur
- **L'extraction d'informations de haut niveau :**
  - La détection d'entités nommées
  - La segmentation thématique du document
  - Le suivi thématique
  - Une tâche de Question-Réponse

Le LIUM participait aux tâches TRS, SES et SRL et son système de transcription basé sur le projet CMU Sphinx a terminé second dans la tâche TRS avec un score de **23.7%** en taux d'erreur mot [Deléglise 2005]. Pour la tâche SES, le LIUM a également terminé second avec un score de 16,9%.

3 Problématique

Dans le cadre de cette campagne, plusieurs facteurs ont déterminé le bon fonctionnement d'un système.

La plupart des erreurs rencontrées sont dûes à un parasitage du signal que le système essaie d'interpréter comme étant de la parole. Elles peuvent être de plusieurs types : soit ce sont des bruits dûs à l'environnement sonore tels que les claquements de portes et autres bruits de rue, soit elles proviennent de la qualité du canal d'enregistrement (microphone, enregistrement téléphonique, acoustique de la pièce...).

La variabilité du signal de parole peut également provenir du locuteur lui-même. Les systèmes doivent prendre en compte les variabilités intra-locuteurs telles que l'émotivité du locuteur, son débit de parole, son style de vocabulaire ainsi que les variabilités inter-locuteurs

telles que leur sexe, leur âge, leur langue ou encore leur accent. Une des sources d'erreur peut également provenir d'un tour de parole non respecté où deux locuteurs interviennent au même instant et ainsi, créer une perturbation du signal sonore.

Dans les tâches liées à la transcription, ces erreurs peuvent aller de la simple substitution d'un mot par un homophone à l'insertion d'un mot non pertinent pour la compréhension globale de la séquence de mots. Les erreurs commises par le système peuvent même se répercuter sur les mots voisins et créer toute une zone de mots erronés. Il serait alors intéressant de pouvoir détecter de telles erreurs pour décider comment les gérer. Pour les tâches liées à la segmentation en événements sonores, il serait aussi intéressant de pouvoir détecter quand le système de détection a commis une erreur en omettant un évènement ou en insérant un événement.

Pour identifier ces erreurs de reconnaissance, des indicateurs peuvent être estimés pour aider à leur détection, pour déterminer quelles parties de la transcription sont pertinentes ou encore sur quelles parties du signal le système ne peut pas décoder efficacement : ce sont les mesures de confiance.

Les mesures de confiance peuvent être utilisées à plusieurs niveaux : au niveau du phonème, au niveau du mot ou encore au niveau de la phrase. En appliquant ces mesures de confiance à un ou plusieurs niveaux, on peut prendre une décision sur la gestion de certains mots ou zones de parole en fonction de leur degré de confiance.

Objectif Un des objectifs de ce travail de thèse est de proposer des mesures de confiance, soit pour repérer les séquences de mots qui semblent sûres (et respectivement celles qui sont peu dignes de confiance) parmi les mots proposés comme hypothèse de reconnaissance fournie par un SRAP, soit pour quantifier la pertinence d'une décision locale dans l'objectif d'établir une décision globale.

Dans ces travaux, les applications présentées sont de deux types.

Premièrement, dans le cadre de la transcription automatique, les mesures de confiance permettent d'effectuer un filtrage de données. Ce filtrage vise à accroître les performances du SRAP utilisé par le LIUM en augmentant de manière non-supervisée les données d'apprentissage des modèles acoustiques. Ainsi, les zones de parole qui ont un degré de confiance élevé sont prélevées à partir d'un deuxième corpus transcrit automatiquemnt et ajoutées au corpus d'apprentissage initial transcrit manuellement.

Deuxièmement, dans le cadre de l'identification du locuteur au sein d'un document sonore, des scores de confiance sont utilisés pour déterminer le nom du locuteur en tentant d'extraire cette information de la transcription.

Ce document sera organisé comme suit : une première partie évoquera le contexte de travail avec le fonctionnement d'un SRAP et ses applications, les différentes mesures de confiance utilisées dans la littérature et le système de RAP utilisé au LIUM. La seconde partie portera sur les contributions de cette thèse et décrira trois mesures de confiance utilisées, dont une mesure acoustique connue dont nous proposerons une normalisation afin de pouvoir la combiner avec d'autres mesures, une nouvelle mesure exploitant le corportement du système de repli (*backoff*) d'un modèle de langage, ou l'appropriation d'une mesure basée sur la probabilité *a posteriori* d'un mot, mesure communément utilisée dans la littérature.

Nous aborderons également le problème de la fusion de ces mesures dans l'optique d'obtenir une mesure de confiance plus efficace.

Enfin, cette mesure finale sera appliquée à différents domaines de RAP tels que l'identification du locuteur, la transcription de réunions ou encore l'amélioration des performances d'un système de RAP grâce à l'apprentissage non supervisé de modèles acoustiques.

Première partie

Contexte de travail

Chapitre 1

Reconnaissance de la parole

Sommaire

L'explosion des technologies multimédia et des systèmes d'information a entrainé une création et une diffusion plus importantes des communications audio, notamment grâce aux technologies d'acquisition et de stockage qui deviennent bon marché et de plus en plus ergonomiques et grâce à la facilité des échanges via Internet. La manipulation automatique de ces documents audio repose sur l'utilisation d'un SRAP permettant le décodage automatique du signal de parole.

Le signal de parole est caractérisé par de nombreux paramètres qui rendent complexe son interprétation. En effet, ce signal possède une grande variabilité. Il est différent d'un locuteur à un autre et même un locuteur ne prononce jamais un mot deux fois de la même façon. Les différences d'âge, de sexe, d'accent, d'émotivité entre locuteurs rendent délicates l'extraction d'informations pertinentes concernant le signal, cette extraction se voulant être indépendante du locuteur. L'acoustique du milieu ambiant lors de la prise de son (bruits extérieurs, bruits de bouche, respirations, éternuements...) ainsi que la qualité de l'enregistrement génèrent encore des difficultés que le SRAP doit surmonter. La segmentation du signal en mots s'avère également un processus complexe à réaliser pour un système de RAP. En effet, pour un SRAP, le signal de parole est un flux continu et il n'a pas la capacité d'interpréter ce signal comme étant une suite de mots. La transcription fournie par le SRAP est la transcription *verbatim* du document sonore, une unité de cette transcription est appelée un mot.

L'objectif d'un système de RAP probabiliste est d'associer une séquence de mots $\hat{W} = w_1 w_2 ... w_k$ (avec w_i qui est un mot de cette séquence) à une séquence d'observations acoustiques X. Le système recherche la séquence de mots qui maximise la probabilité *a posteriori* $P(W|X)$, où $P(W|X)$ est la probabilité d'émission de W sachant X. On obtient, après application de la règle de Bayes :

$$\hat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} \frac{P(W)P(X|W)}{P(X)} \tag{1.1}$$

Comme la séquence d'observations acoustiques X est fixée, $P(X)$ peut être considérée comme une valeur constante inutile dans l'équation 1.1. On a donc :

$$\hat{W} = \arg\max_{W} P(W)P(X|W) \tag{1.2}$$

Deux types de modèles probabilistes sont utilisés pour la recherche de la séquence de mots la plus probable : des modèles acoustiques qui fournissent la valeur de $P(X|W)$, et un modèle de langage qui fournit la valeur de $P(W)$. $P(X|W)$ peut se concevoir comme la probabilité d'observer X lorsque W est prononcée, alors que $P(W)$ se réfère à la probabilité que W soit prononcée dans un langage donné. La difficulté pour obtenir un système de RAP performant est

de définir les modèles les plus pertinents possibles pour le calcul de $P(W)$ et $P(X|W)$ (voir figure 1.1).

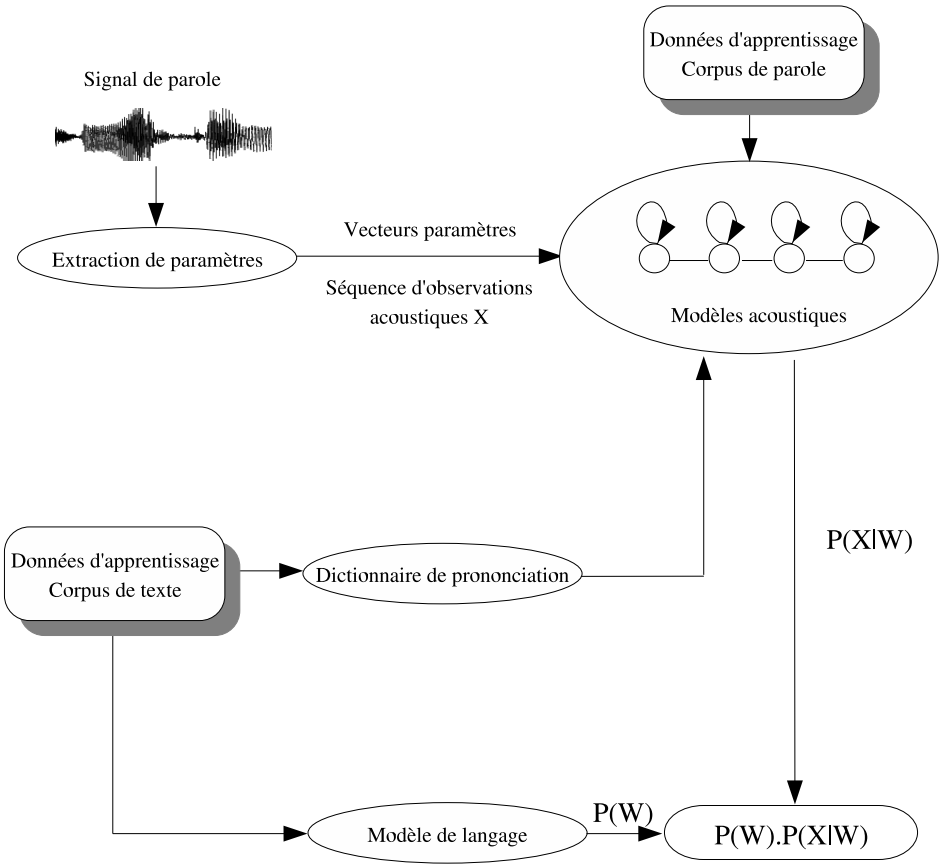


FIG. 1.1 – Système de RAP

Dans ce chapitre, nous évoquerons le fonctionnement d'un système de RAP en commençant par les caractéristiques du signal de parole et leur extraction. Ensuite, nous décrirons les différents modèles utilisés dans un système de RAP, les modèles acoustiques et le modèle de langage. Enfin, après avoir abordé la segmentation et l'adaptation des modèles à l'environnement et au locuteur, nous verrons comment évaluer un tel système.

1.1 Extraction de paramètres

Pour modéliser le signal de parole, un découpage du signal en trames de taille fixe (environ 25 ms) prises toutes les 10 ms est effectué. De chaque trame est extrait un vecteur de paramètres caractérisant celle-ci. Les deux méthodes les plus connues pour réaliser cette extraction sont l'analyse spectrale avec par exemple l'application de transformations non-linéaires de l'échelle

des fréquences (Mel-scale Frequency Cepstral) et l'analyse paramétrique avec par exemple l'utilisation d'une estimation spectrale par prédiction linéaire (Linear Prediction Coding). Ces analyses vont produire respectivement, des coefficients cepstraux (MFCC) et des coefficients PLP (Perceptually-based Linear Prediction). Par la suite, nous nous appuierons sur des exemples utilisant des MFCC (Mel-scale Frequency Cepstral Coefficients). [Calliope 1989], [Mariani 2002a] ou encore [Mariani 2002b] contiennent plus de détails sur les différentes méthodes utilisées en extraction de paramètres (et en reconnaissance de la parole en général).

Après l'extraction des différents paramètres d'une trame, on obtient donc une séquence d'observations acoustiques $X = x_1 x_2 ... x_n$ où x_i représente une observation acoustique.

1.2 Modèles acoustiques

1.2.1 Modèles de Markov cachés (HMM)

Les modèles acoustiques utilisés pour la reconnaissance de la parole sont depuis des années principalement basés sur les HMMs (Hidden Markov Models ou Modèles de Markovs Cachés) [Rabiner 1989, Calliope 1989]. Les HMMs sont des automates probabilistes à états finis qui permettent de calculer la probabilité d'émettre une séquence d'observations. Pour un système de RAP, les émissions sont donc les vecteurs de caractéristiques du signal de parole composés généralement de coefficients MFCC. Les HMMs respectent l'hypothèse markovienne d'ordre 1 : la connaissance du passé se résume à celle du dernier état occupé. Pour capter certains comportements et évolutions du signal dans le temps, on intègre dans les vecteurs de caractéristiques du signal les dérivées premières et secondes des coefficients MFCC.

Les systèmes de RAP à base de HMMs reposent ainsi sur les postulats suivants :

1. la parole est une suite d'états stationnaires, représentés par des vecteurs de caractéristiques (MFCC par exemple) et leur dérivées premières et secondes,
2. l'émission d'une séquence de ces vecteurs est générée par un HMM respectant l'hypothèse markovienne d'ordre 1.

La figure 1.2 présente un exemple de HMM, avec la topologie la plus courante (modèle gauche-droit, avec saut d'état possible).

À chaque intervalle de temps, un HMM transite d'un état i à un état j (avec $j \geq i$: un état peut boucler sur lui-même) avec une probabilité discrète a_{ij}. À chaque instant t un état j est donc atteint et une émission o_t est générée associée à une densité de probabilité $b_j(o_t)$.

L'apprentissage d'un modèle acoustique revient principalement à estimer les paramètres suivants :

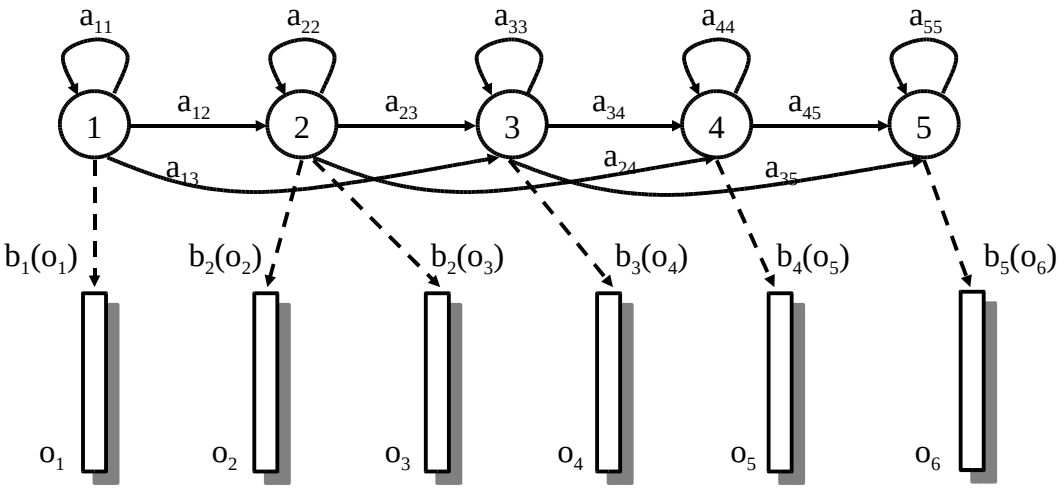


FIG. 1.2 – HMM à 5 états

1. Les probabilités d'émissions $b_j(o)$ des observations pour chaque état. Il s'agit généralement de mélanges de densités de probabilités gaussiennes, définies par leurs vecteurs de moyennes, leurs matrices de covariances (en pratique il s'agit de matrices diagonales), et une pondération associée à chaque densité de probabilité.
2. Les probabilités discrètes a_{ij} qui définissent la topologie du HMM en indiquant la probabilité de transition d'un état vers un autre.

Dans la pratique, l'unité de modélisation la plus courante est le phonème. Pour tenir compte de la variabilité de prononciation d'un phonème, un HMM est construit pour un phonème donné, associé à un contexte gauche et un contexte droit particuliers. Un contexte gauche (resp. droit) d'un phonème est un phonème qui précède (resp. succède à) ce phonème. Ce triplet (contexte gauche, phonème, contexte droit) est appelé triphone, ou phonème en contexte. Pour affiner la modélisation d'un phonème en contexte, la position de ce phonème dans un mot (début, milieu, fin ou phonème isolé) est parfois prise en compte. Une factorisation d'états similaires est effectuée afin de réduire la taille du modèle, on parle alors d'états partagés.

1.2.2 Apprentissage

L'apprentissage des modèles acoustiques (voir figure 1.3) consiste à estimer les paramètres des chaînes de Markov (probabilités de transitions) et des densités d'observation associées aux états, c'est à dire les vecteurs de moyennes et les matrices de covariances d'un ensemble de gaussiennes, ainsi que les pondérations permettant d'établir des mélanges à partir de ces

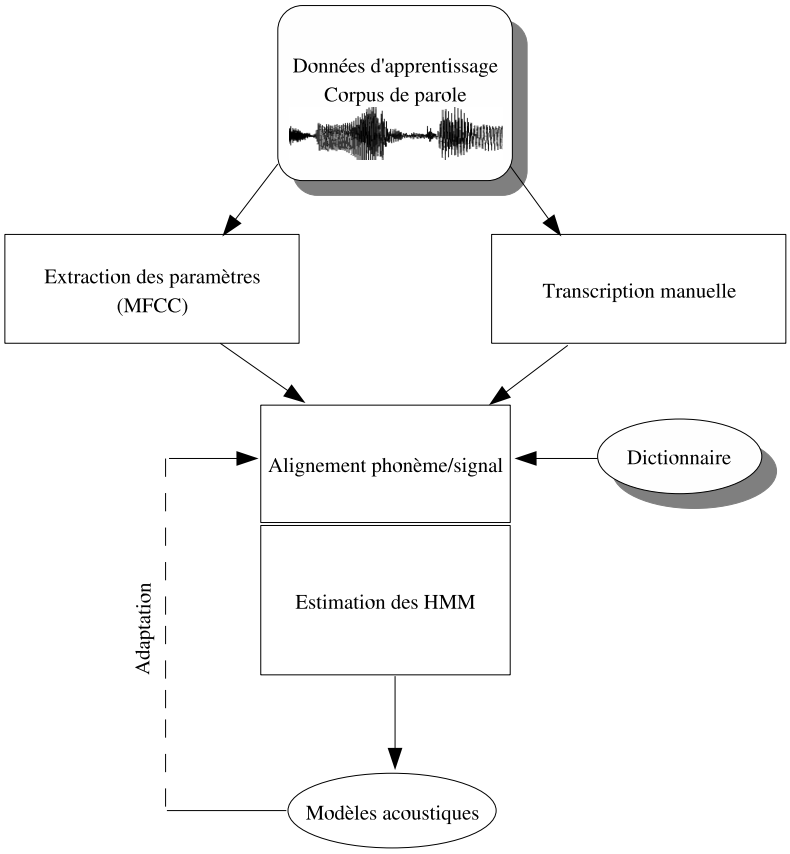


FIG. 1.3 – Apprentissage des modèles acoustiques

gaussiennes. L'algorithme généralement utilisé pour cette estimation est celui d'Expectation-Maximisation (EM) [Dempster 1977].

1.2.2.1 Lexique

L'apprentissage de modèles acoustiques ne peut être réalisé que si une transcription est disponible avec le signal de parole. La première étape du processus d'apprentissage consiste à obtenir une phonétisation de la transcription pour ensuite pouvoir aligner chaque phonème avec la partie du signal qui lui correspond.

Le lexique (ou vocabulaire, dictionnaire de phonétisation...) [Allauzen 2003, Allauzen 2004] regroupe alors tous les mots nécessaire au décodage ainsi que leur phonétisation. Sa composition repose sur le choix des entrées lexicales (graphèmes) et du jeu de phonèmes utilisé pour les décrire. Le jeu de phonèmes choisi dépend de la langue. En français, on répertorie environ 33 phonèmes, alors que l'anglais comprend 45 phonèmes et seulement 26 pour l'espagnol.

Dans le cas où plusieurs prononciations d'un même graphème sont possibles, les différentes séquences de phonèmes correspondantes doivent être inscrites dans le lexique. Dans le domaine de la transcripion graphèmes-phonèmes, deux approches sont principalement utilisées pour construire le lexique : la première repose sur l'utilisation d'un lexique phonétique de référence (exemple du projet BDLEX), la seconde propose l'utilisation d'une base de régles de phonétisation pour transcrire automatiquement les graphèmes en phonèmes (exemple du système LIA_PHON [Béchet 2001]). Les deux approches sont souvent utilisées conjointement. Lorsqu'un mot n'est pas inclut dans le dictionnaire BDLEX (qui ne contient pas de noms propres ni d'acronymes par exemple), on peut utiliser des systèmes automatiques de phonétisation. De même, dans les approches de transcription graphèmes-phonèmes automatiques, certaines règles donnent lieu à des exceptions qui nécessitent l'utilisation de listes.

Le lexique est une liste fermée de taille fixe. Pour les systèmes de transcriptions d'émissions radiophoniques en français, cette taille est généralement de 65000 mots environ[3]. Le choix des entrées lexicales contribue largement au bon fonctionnement d'un SRAP. En effet, tout mot absent du lexique ne peut pas être reconnu par le SRAP et peut engendrer des erreurs sur son voisinage.

Le lexique du système de RAP doit donc permettre de couvrir le maximum de données rencontrées au cours du processus de reconnaissance et ainsi d'éviter les problèmes dûs aux mots hors-vocabulaire.

1.2.2.2 Alignement phonème/signal

Les HMMs des modèles acoustiques nécessitent une phase d'estimation de leurs paramètres. Une fois la phonétisation des mots du lexique terminée, on procéde à une phase d'alignement des phonèmes sur le signal. Cette phase d'alignement phonème/signal va permettre l'association des vecteurs acoustiques aux états du HMM en utilisant soit l'algorithme forward-backward [Baum 1972] soit l'algorithme de Viterbi [Viterbi 1967]. Pour obtenir des modèles acoustiques performants, il est nécessaire que la phonétisation de chaque transcription soit la plus proche possible de la prononciation effective de la phrase correspondante. Le problème survient lorsqu'un mot possède plusieurs phonétisations alternatives : un choix doit être fait, et il est impossible de vérifier ce choix en écoutant le signal de parole lorsqu'il se mesure en dizaines d'heures de parole. Une solution consiste à estimer grossièrement des premiers modèles acoustiques en prenant la phonétisation la plus courte. Cette méthode permet de forcer un alignement phonème/signal en utilisant des informations acoustiques et textuelles. Dès que les premiers modèles acoustiques sont disponibles, on utilise un outil d'alignement phonème/signal. Ce dernier

[3]Cette taille dépend de la langue, certaine langue comme l'allemand nécessitant plus de 65000 mots.

s'appuie sur le dictionnaire de phonétisation créé précédemment pour envisager les différentes phonétisations possibles d'une transcription. Ensuite, il choisit la phonétisation la plus probable en fonction des scores de vraisemblance calculés par les premiers modèles acoustiques sur le signal de parole.

Pour avoir un grand nombre de paramètres et ainsi obtenir des modèles basés sur des HMMs robustes et des systèmes indépendants du locuteur, on utilise de grands corpora de données contenant les énoncés de plusieurs locuteurs. En s'appuyant sur des données provenant de plusieurs locuteurs, les variabilités inter-locuteurs sont mieux modélisées. Le système obtiendra ainsi un taux d'erreur sur les mots plus performant pour un locuteur test n'intervenant pas dans le corpus d'apprentissage que si le système était dépendant du locuteur.

1.3 Modèle de langage

Le modèle de langage a pour objectif de capturer les contraintes du langage naturel afin de guider le décodage acoustique. Il permet notamment de résoudre les ambiguités données par les nombreux homonymes que contient la langue française. Comme nous l'avons déjà noté, les modèles de langage probabilistes ont pour objet d'attribuer une probabilité à une séquence de mots. De manière générale, la probabilité de la séquence de mots W_1^k s'exprime :

$$P(W_1^k) = P(w_1) \prod_{i=2}^{k} P(w_i|w_1, ..., w_{i-1}) = P(w_1) \prod_{i=2}^{k} P(w_i|h_i) \tag{1.3}$$

Dans cette formule, h_i est l'*historique* du mot w_i. On a : $h_i = w_1, ..., w_{i-1}$

Le modèle de langage est estimé sur de grands corpora de textes pour avoir un maximum de couverture lexicale. Des données telles que des textes de journaux, de dépêches électroniques ou de transcriptions de documents audio sont utilisées. Les transcriptions enrichies comportant des annotations comme les hésitations ou encore les bruits de respirations sont favorisées, permettant ainsi une plus grande malléabilité du modèle de langage.

1.3.1 Modèles *n-grammes*

Le modèle de type *n-gramme* est le modèle probabiliste le plus généralement utilisé. Pour ce genre de modèle, l'historique d'un mot est représenté par les $n-1$ mots qui le précèdent.

Dans la pratique, la valeur de n dépasse rarement 3 : on parle de modèle *trigramme* (*unigramme* pour $n = 1$, *bigramme* pour $n = 2$).

Même si ce genre de modèle semble particulièrement réducteur en ne prenant en compte que des contraintes lexicales courtes, il contient suffisamment d'informations pour guider efficacement un système de RAP. Enfin, une qualité fondamentale des modèles *n-grammes* est la couverture totale des phrases pouvant être exprimées dans un langage. Ceci est intéressant pour le traitement de la parole spontanée : l'utilisation de modèles probabilistes de type *n-gramme* permet de modéliser certains aspects du langage oral spontané incorrects d'un point de vue grammatical : un modèle de langage à base de règles de grammaires formelles serait plus facilement mis en défaut dans ce type de situation. Bien entendu, il est évident que ces phénomènes typiques de la parole spontanée doivent être observés dans le corpus d'apprentissage pour être modélisés par le modèle *n-gramme*. En contrepartie, la précision des modèles *n-gramme* est limitée puisque ce type de modèle ne rejette aucune phrase, y compris celles n'appartenant pas au langage visé. Cependant, les scores affectés à ces phrases sont souvent pénalisés par rapport au score des phrases plus correctes car elles sont composées de séquences de mots peu fréquentes (voire inexistantes) dans le corpus d'apprentissage du modèle de langage, alors qu'il est plus probable de rencontrer les séquences de mots d'une phrase valide.

1.3.2 Estimation des probabilités

L'apprentissage d'un modèle de langage *n-gramme* consiste à estimer un ensemble de probabilités à partir d'un corpus d'apprentissage. Ce corpus d'apprentissage peut être composé de textes mais également de données orales transcrites (voir figure 1.4). Ces données permettront l'estimation des probabilités des *n-grammes* rencontrés. La probabilité d'un mot étant donné le passé dépend des $n-1$ étiquettes précédentes d'où la transformation de l'équation 1.3 en :

$$P(W_1^k) = \prod_{i=1}^{k} P(w_i|w_{i-1}, ..., w_{i-n+1}) \tag{1.4}$$

Il existe plusieurs méthodes pour procéder à l'estimation des paramètres du modèle de langage [Federico 1998]. La plus commune est l'estimation par *maximum de vraisemblance*, dont le nom indique que la distribution des probabilités du modèle de langage obtenue est celle qui maximise la vraisemblance du corpus d'apprentissage :

$$P_{MV}(w_i|h_i) = \frac{n(h_i, w_i)}{n(h_i} \tag{1.5}$$

où $n(x)$ indique la fréquence de x.

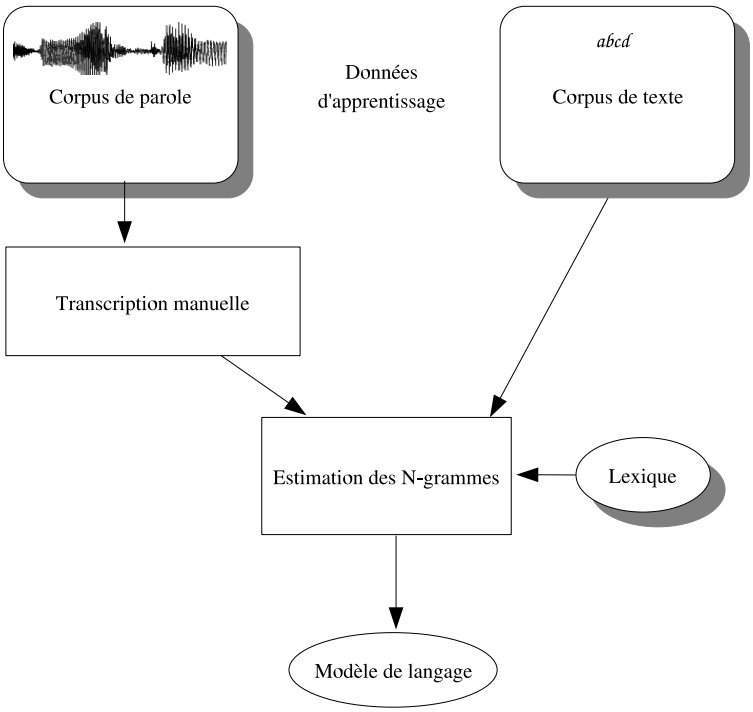


FIG. 1.4 – Apprentissage du modèle de langage

1.3.3 Lissage

C'est donc à partir de la valeur des fréquences d'apparition des *n-grammes* dans les données d'apprentissage que sont estimés les paramètres d'un modèle de langage. Malheureusement, la quantité de données est en général insuffisante et certains *n-grammes* n'apparaissent que peu de fois dans le corpus d'apprentissage. Il peut même arriver que certains mots du lexique soient absents du corpus d'apprentissage lorsque la construction de ce lexique n'impose pas cette présence.

Les techniques de lissage tentent de compenser cette carence : elles peuvent être vues comme une sorte de généralisation qui permet d'attribuer une probabilité non nulle à un événement non vu dans le corpus d'apprentissage. Les principales techniques de lissage sont décrites dans [Chen 1996] où est également présentée une discussion sur leurs performances respectives. Il existe deux grands types de lissage : le repli (ou *back-off*) et l' *interpolation linéaire*. Le repli est un mécanisme qui permet à un modèle de langage de type *n-gramme* d'utiliser une probabilité issue d'un ordre inférieur ($n-1$, $n-2$, ...) lorsqu'aucune probabilité n'est disponible à l'ordre n pour un mot et un historique donné. Pour chaque repli vers un ordre inférieur, la taille de l'historique est diminué et les chances d'obtenir une probabilité estimée sur le corpus d'apprentissage augmente. En contrepartie, un coefficient de repli est habituellement associé

à cette probabilité qui modifie la valeur finale de la probabilité proposée par le modèle pour le mot et l'historique donné. Pour l' *interpolation linéaire*, la probabilité du *n-gramme* est une combinaison linéaire des modèles d'ordre 0 à n. Chaque méthode de lissage peut être envisagée en version *back-off* ou en version *interpolation linéaire*.

1.4 Segmentation

De part la complexité du flux audio dans les documents sonores, il est nécessaire de segmenter le signal afin d'obtenir des zones homogènes (même conditions d'enregistrement, même locuteur...). Chaque segment est caractérisé par des conditions acoustiques spécifiques (parole téléphonique ou en studio, présence de parole, présence de musique, genre du locuteur, identité du locuteur...). Ces différentes indications sont autant de pistes qu'un système de reconnaissance utilise pour choisir les modèles acoustiques les plus appropriés pour décoder le segment considéré. Il s'agit alors de disposer de modèles acoustiques robustes vis-à-vis de toutes les conditions auxquelles le système de reconnaissance peut être confronté.

La première méthode employée pour obtenir ces segments s'appuie sur une recherche par programmation dynamique des points de discontinuité les plus probables dans le spectre comme le montre l'article [Cohen 1981]. La deuxième méthode repose sur une idée plus basique tout aussi efficace où on considère simplement des segments de 20 trames acoustiques (d'une durée d'environ 25 ms chacune) qui sont étiquetés par la suite.

1.5 Adaptation

L'adaptation des modèles acoustiques apporte une amélioration sensible des performances de reconnaissance. Elle peut permettre de compenser le manque de données d'apprentissage ou encore de compenser des différences entre conditions d'apprentissage et conditions d'utilisation, liées à l'environnement ou au locuteur. Les techniques d'adaptation de modèles préappris à de nouvelles données permettent de compenser ces écarts. L'apprentissage nécessitait la collecte de grands ensembles de données, l'adaptation quant à elle doit permettre d'obtenir rapidement un nouveau modèle proche des données de test à partir du modèle initial et d'un ensemble réduit de données d'adaptation. Par exemple, si un nouveau locuteur doit intervenir dans un flux audio, il sera nécessaire de prendre en compte sa voix et de prendre quelques phrases représentatives pour permettre au système de modifier progressivement ses paramètres jusqu'à ce qu'il redevienne stable. L'adaptation à la fatigue, la vitesse d'élocution, à la bande passante proposée ainsi qu'aux différents accents rencontrés sont autant de problèmes à prendre en considération.

Ainsi, les données d'adaptation appartenant aux données d'apprentissage permettront d'adapter les modèles pour obtenir des modèles spécifiques ou s'adapter à une nouvelle tâche. On peut ici faire de l'adaptation supervisée lorsque la transcription des données est disponible. Les données d'adaptation peuvent aussi être limitées aux données à décoder. On pourra alors envisager des techniques d'adaptation non supervisées.

Trois techniques d'adaptation sont communément utilisées : la composition de modèles, les méthodes de transformation linéaire et l'adaptation bayésienne.

1.5.1 Composition de modèles

La composition de modèles [Gales 1992] sert généralement à compenser un bruit additif en combinant un modèle de bruit avec un modèle de parole non bruitée.

1.5.2 Transformations linéaires

La régression linéaire (MLLR : Maximum Likelihood Linear Regression) [Leggetter 1995, Gales 1998] ou plus généralement, les méthodes de transformations linéaires sont efficaces pour l'adaptation non supervisée. Ces transformations linéaires sont utilisées dans la procédure d'adaptation de modèles indépendants du locuteur [Anastasakos 1996]. La méthode employée pour modéliser les variabilités inter-locuteurs consiste à estimer les paramètres des modèles acoustiques à partir de grands ensembles de données (modèle indépendant du locuteur). Mais ces modèles sont moins performants en termes de taux d'erreur que les modèles dépendants du locuteur. Pour résoudre ce problème, la technique consiste à adapter le modèle indépendant du locuteur à un locuteur spécifique afin d'obtenir un taux de reconnaissance aussi proche que possible de celui obtenu avec un modèle dépendant du locuteur. Pour l'adaptation au locuteur, les techniques utilisées appliquent les mêmes transformations linéaires que [Leggetter 1995] (techniques MLLR). Ces transformations sont utilisées dans la procédure d'apprentissage de modèles indépendants du locuteur et permettent d'estimer les transformations propres à chacun des locuteurs du corpus d'apprentissage ainsi que les paramètres des HMMs [Anastasakos 1996]. Les modèles ainsi calculés peuvent être plus efficacement adaptés à un nouveau locuteur. Il existe deux types de transformations linéaires : le cas non-contraint où les transformations sur les moyennes et variances des Gaussiennes sont décorrellées les unes des autres [Leggetter 1995] et le cas contraint (CMLLR : Constrained Maximum Likelihood Linear Regression) [Digalakis 1995]. La technique CMLLR utilisée dans la processus d'adaptation SAT (Speaker Adaptative Training)du LIUM est détaillée dans la suite de ce paragraphe.

1.5.2.1 SAT-CMLLR

Contrairement au cas non contraint, les transformations de la variance et de la moyenne de la technique CMLLR doivent être liées. Ces transformations sont de la forme :

$$\nu^{'} = A\nu - b \tag{1.6}$$

et

$$\sum^{'} = A \sum A^{'T} \tag{1.7}$$

où ν et ν' sont les moyennes avant et après transformation, $\sum$ et $\sum^{'}$ sont les variances, A est la matrice de régression et b, le facteur de décalage. Au moyen de l'algorithme EM, les paramètres A et b sont optimisés selon le maximum de vraisemblance sur les données d'adaptation.

Si les mêmes transformations linéaires sont utilisées lors de la phase d'apprentissage de modèles indépendants du locuteur, les transformations propres à chacun des locuteurs de ce corpus ainsi que les paramètres des modèles markoviens peuvent être estimées conjointement. Les modèles qui en résultent sont ensuite plus facilement adaptables à un nouveau locuteur.

1.5.3 Adaptation Bayésienne : méthode MAP

L'adaptation bayésienne (MAP : Maximum *a posteriori*) [Gauvain 1994] permet d'introduire dans l'apprentissage des contraintes probabilistes sur les paramètres des modèles. Le critère MAP est appliqué aux modèles ayant fait l'objet d'un apprentissage préalable et pour lesquels on dispose de données *a priori*. Les modèles markoviens sont toujours estimés avec l'algorithme EM mais en maximisant la vraisemblance *a posteriori* (MAP) au lieu de la vraisemblance des données. Elle permet d'obtenir de nouveaux modèles en réduisant la variance des modèles initiaux et ainsi d'obtenir des modèles plus spécifiques en utilisant un nombre restreint de données d'adaptation. Cette technique permet de créer à partir de modèles indépendant du locuteur, des modèles spécifiques au genre ou à un locuteur en particulier. De même, elle permet d'obtenir des modèles adaptés à des conditions acoustiques particulières. La méthode MAP [Gauvain 1994] est utilisée dans le SRAP du LIUM pour adapter les modèles à la bande passante.

1.6 Evaluation des systèmes de reconnaissance

Une fois l'hypothèse de reconnaissance fournie par le système de reconnaissance, on effectue une comparaison entre la transcription manuelle de référence et cette hypothèse. Pour cela,

les deux transcriptions sont alignées et on dénombre trois types d'erreurs. Les substitutions correspondent aux mots qui ont été reconnus à la place d'un mot de la transcription manuelle. Les insertions sont les mots reconnus qui se sont insérés par erreur entre deux mots corrects de la transcription de référence. Les suppressions correspondent aux mots de la référence qui ont été oubliés dans l'hypothèse de reconnaissance.

Pour évaluer et aussi pouvoir comparer les systèmes de RAP entre eux, on détermine leur taux d'erreur (ou Word Error Rate) sur les mots par la formule :

$$\text{Taux d'erreur} = \frac{nombre\ de\ substitutions + nombre\ d'\ insertions + nombre\ de\ suppressions}{nombre\ de\ mots\ dans\ la\ référence} \tag{1.8}$$

Ce taux d'erreur est calculé sur un corpus de test. Pour pouvoir comparer plusieurs systèmes, il faut qu'ils soient évalués sur les mêmes données. Pour ceci, il existe des campagnes d'évaluations comme ESTER ou encore les différentes évaluations proposées par NIST qui fournissent des corpora d'apprentissage de développement et de test ainsi qu'un protocole de test commun.

Après avoir étudié les différents modules qui composent un système de reconnaissance automatique de la parole et décrit leur méthode d'évaluation, nous allons décrire le système sur lequel s'appuient les expériences relatées dans ce manuscrit.

1.7 LIUM-RT : le système de transcription enrichie du LIUM

Ce paragraphe présente le système de TAP utilisé par le LIUM au moment de la campagne ESTER. Toutes les expériences réalisées lors de l'étude relatée dans ce manuscrit reposent sur ce système basé sur le décodeur CMU Sphinx 3.3. Nous présenterons donc ce décodeur mais également les différents éléments qui y ont été ajoutés par le LIUM tels que des outils de segmentation, d'adaptation des modèles acoustiques ou de rescoring de graphe de mots.

1.7.1 CMU Sphinx III

Le projet CMU Sphinx a été développé par DARPA pour réaliser un système robuste de reconnaissance automatique de la parole multi-locuteur et grand vocabulaire. Les différents packages développés par l'Université Carnegie Mellon tels que le décodeur CMU Sphinx II, Sphinx Train et le décodeur CMU Sphinx III sont disponibles en open-source. Le LIUM utilise le décodeur s3.3 [Ravishankar 1997, Chan 2004] qui est une version rapide du décodeur du projet CMU Sphinx III.

Le projet CMU Sphinx III auquel appartient le décodeur s3.3 a été développé pour améliorer les performances de CMU Sphinx III en vitesse (il fonctionne dix fois plus vite sans dégradation

notable des résultats de reconnaissance) avec l'implémentation de modèles acoustiques regroupés en sous-vecteurs ou encore l'utilisation améliorée d'arbres lexicaux. Ce décodeur utilise des modèles acoustiques continus multi-gaussiens à base de triphones (voir le paragraphe de l'état de l'art sur la reconnaissance de la parole) limités à 3 ou 5 états émetteurs par modèle de phonème, avec ou sans saut. Seuls des modèles de langage bigrammes ou trigrammes peuvent être utilisés et la taille du vocabulaire est limité à 65.000 mots.

1.7.2 Extraction de paramètres

Le décodeur utilise des paramètres cepstraux classiques : 13 MFCC (voir chapitre 1.1) sont extraits par trames d'une durée de 25 ms avec recouvrement de 10 ms. On complète ensuite le vecteur de paramètres de la trame avec les dérivées et les dérivées secondes des MFCC. Deux ensembles de paramètres sont ainsi calculés, correspondant aux analyses bande large (130Hz-6800Hz) et bande étroite (440Hz-3500Hz) qui donneront lieu à deux types de HMMs spécifiques à la bande passante. Les paramètres des HMMs sont appris avec l'algorithme Baum-Welch [Baum 1972]. La figure 1.5 montre le système utilisé, basé sur une segmentation par mot.

1.7.3 Segmentation en locuteur

Le processus de segmentation découpe le signal en parties homogènes en termes de locuteur, genre et largeur de bande. Pour une tâche de transcription, l'exactitude des frontières des segments en terme de largeur de bande et de genre est importante. En effet, les modèles acoustiques utilisés sont précalculés et spécialisés en fonction du genre du locuteur et de la largeur de bande. Le processus de segmentation acoustique en locuteur développé par le LIUM est basé sur le Critère d'Information Bayésien (BIC) [Gish 1991, Chen 1998] calculé en trois étapes :

- Le signal est décomposé en petits segments homogènes
- Les segments sont ensuite regroupés par locuteur sans changer les frontières
- Les frontières sont ajustées dans une phase finale

Les frontières initiales des segments sont déterminées grâce au calcul du Rapport de Vraisemblance Généralisé (GLR : Generalized Likelihood Ratio) [Willsky 1976] sur les paramètres acoustiques. Ces paramètres sont composés de 12 MFCC auxquels est rajouté l'énergie. Le signal est sur-segmenté pour minimiser les détections manquées lors de la détection des frontières. Une longueur de segment raisonnable est tout de même conservée pour permettre une estimation correcte et suffisante d'un modèle de locuteur associé à ce segment.

Le regroupement repose sur une classification hiérarchique ascendante. Au départ, chaque segment est placé dans un cluster. Les deux clusters les plus proches sont ensuite regroupés à

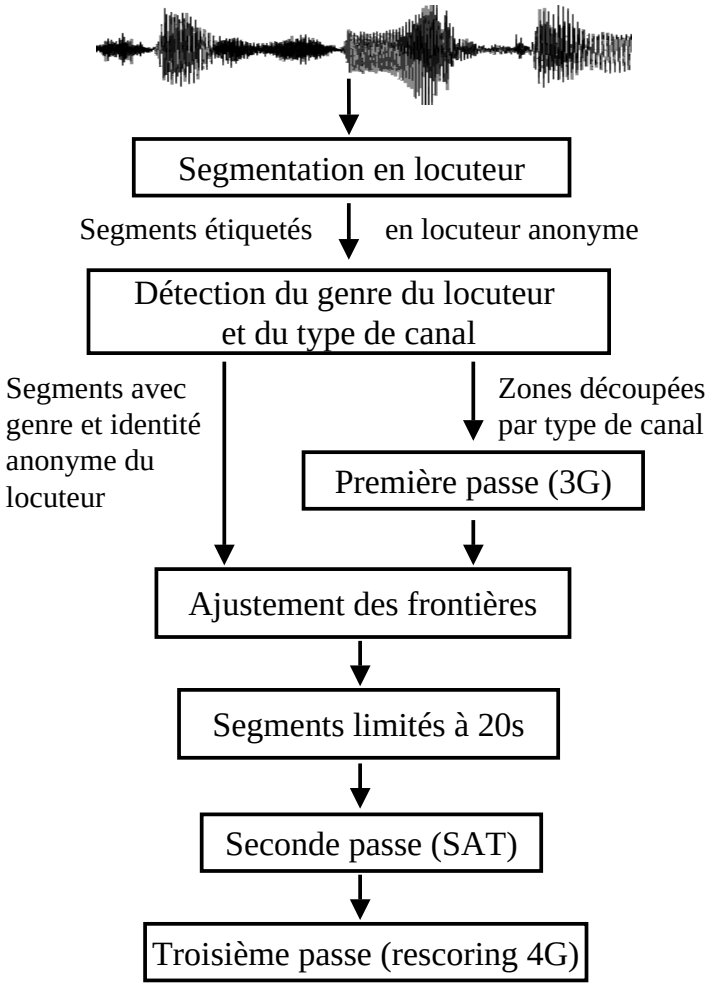


FIG. 1.5 – Architecture générale du système de reconnaissance de la parole utilisé par le LIUM

chaque itération jusqu'au critère d'arrêt. La mesure de similarité entre clusters et le critère d'arrêt pour le regroupement reposent sur la métrique BIC. Le locuteur, *ie* le cluster, est modélisé par une gaussienne à covariance pleine.

Un décodage avec l'algorithme de Viterbi [Viterbi 1967] est effectué en post-traitement pour affiner les frontières des segments. Chaque locuteur est modélisé par un HMM à un état. Les états sont modélisés par une mixture de gaussiennes (GMM) de 8 composantes à matrice de covariance diagonale appris avec l'algorithme EM-ML [Dempster 1977] sur l'ensemble des segments de ce locuteur. Le changement de modèle est pénalisé en soustrayant une constante à la vraisemblance du chemin.

Pour le processus de reconnaissance qui fonctionne en continu sur des segments de bande homogène, la largeur de bande puis le genre sont détectés directement sur chaque segment

du locuteur avec des GMMs. Les frontières des segments sont finalement affinées avec les frontières des phrases fournies par la transcription de la première passe de décodage.

1.7.4 Modélisation acoustique

Les modèles acoustiques ont été appris sur environ 80h d'émissions radiophoniques provenant de la campagne ESTER (voir tableau 1.1) accompagnées de leurs transcriptions manuelles. Ceci correspond à 72h de données large bande et 8h de données bande étroite. Ce corpus permet de modéliser 35 phonèmes et 5 sortes de fillers (bruits de bouches, hésitations, etc...) différents. Pour cet apprentissage, le toolkit SphinxTrain associé aux décodeurs CMU Sphinx a été utilisé. Les modèles large bande (LB) ont été appris avec les données large bandes seulement (72h) alors que les modèles bande étroite (BE) ont été appris avec l'ensemble des données, puis adaptés avec la méthode MAP [Gauvain 1994] (voir chapitre 1.5.3 sur l'adaptation bayésienne) aux 8h de données bande étroite. De même, un modèle par genre a été appris avec une méthode MAP ce qui résulte en quatre modèles au total : LB-homme, LB-femme, BE-homme, BE-femme. Pour compléter les données d'apprentissage, les transcriptions automatiques provenant de 75h de différentes radios ont été ajoutées. Les modèles résultants avant injection de données sont composés de 5500 états partagés, chaque état étant modélisé avec un mélange de 22 Gaussiennes.

source	Apprentissage/Développement		Test
	transcrit	non transcrit	
France Inter	33h/2h	337h	2h
France Info	8h/2h	643h	2h
RFI	23h/2h	445h	2h
RTM	18h/2h	-	2h
France Culture	-	252h	1h
"suprise"	-	-	1h
Total	**82h/8h**	**1677h**	**10h**

TAB. 1.1 – Ressources fournies par ESTER

Un module d'apprentissage adaptatif du locuteur (Speaker Adaptative Training) basé sur une transformation CMLLR [Gales 1997] (voir chapitre 1.5.2.1 sur l'adaptation CMLLR) a également été ajouté pour améliorer les performances de CMU Sphinx III [Deléglise 2005]. Les modèles spécifiques genre-bande (LB-homme, LB-femme, BE-homme, BE-femme) sont utilisés pour calculer la transformation CMLLR pour chaque segment (ou chaque locuteur). Quatre modèles en découle : SAT-LB-homme, SAT-LB-femme, SAT-BE-homme et SAT-BE-femme.

1.7.5 Modélisation du langage

Comme pour les modèles acoustiques, le corpus textuel utilisé pour l'apprentissage des modèles de langage doit être le plus proche possible des données qu'aura à traiter le système de reconnaissance. Du fait d'un coût élevé, les transcriptions manuelles sont difficiles à obtenir et des ressources linguistiques relativement similaires peuvent être utilisées. Aisni, les données d'apprentissage fournies par ESTER contiennent des transcriptions manuelles de broadcast news mais aussi en grande partie des données provenant d'articles du journal "Le Monde". En fait, dans les émissions radiophoniques, les intervenants ont la plupart du temps une intervention travaillée où la parole spontanée n'apparaît pas. À défaut de transcriptions d'émissions de radio, des articles de journaux sont utilisés. Les données d'apprentissage ont été découpées en trois ensembles homogènes :

1. Les transcriptions manuelles de 89h (sur 90h d'apprentissage + développement) d'émissions radiophoniques fournies par ESTER. La dernière heure est conservée pour tester les modèles de langage obtenus. Ces transcriptions sont constituées de 1,35 million de mots dont 34.000 différents.
2. Des articles du journal "La Monde" datant de l'année 2003. Ces articles sont constitués de 19 millions de mots dont 220.000 différents. Cet ensemble comprend les articles les plus récents par rapport à la période des données de test de la campagne d'évaluation ESTER (2004).
3. Des articles du journal "Le Monde" entre l'année 1987 et l'année 2002 comprenant environ 300 millions de mots.

1.7.5.1 Vocabulaire

Les mots du vocabulaire proviennent de chacun de ces trois ensembles. Premièrement, l'ensemble des 34.000 mots provenant des transcriptions manuelles y ont été incorporés. Comme cet ensemble est très proche des données de test, il semble intéressant de les garder tous dans le vocabulaire. Ensuite, les mots apparaissant plus de 10 fois dans le deuxième ensemble (soit environ 19.000 mots) sont conservés. Au final, les mots les plus fréquents du troisième ensemble sont incorporés jusqu'à obtenir la taille maximum autorisée par le décodeur s3.3 de 65.000 mots.

1.7.5.2 Estimation des modèles n-grammes

Grâce à ce vocabulaire et à chacun des trois ensembles, trois modèles de langage trigramme sont appris. Le toolkit SRILM est alors utilisé pour estimer et interpoler ces modèles en un unique modèle de langage trigramme. Tous les unigrammes et bigrammes sont conservés, tandis

que les trigrammes n'apparaissant qu'une seule fois sont retirés. Les poids d'interpolation ont été appris sur 80h du corpus d'ESTER et sur les deux autres corpora en utilisant l'implémentation de l'algorithme EM fourni par le toolkit CMU SLM. Ces poids ont ensuite été optimisés sur les 9h restantes du corpus ESTER. Le modèle trigramme résultant a été utilisé lors des deux premières passes du processus de reconnaissance de la parole.

1.7.6 Processus de transcription de la parole

Le processus de transcription est composé de trois passes (voir figure 1.5) :

1. La première passe utilise les modèles acoustiques correspondant au genre et à la largeur de bande détectée lors du processus de segmentation ainsi que le modèle de langage trigramme.
2. La seconde passe applique une transformation CMLLR par segment ou par locuteur et utilise le même modèle de langage trigramme que lors de la première passe. Un graphe de mots est généré, contenant les mots ainsi que leurs scores acoustiques.
3. Le graphe de mots obtenu est rescoré dans une dernière passe avec un modèle de langage quadrigramme. Ce modèle quadrigramme est estimé de la même façon que le précédent modèle trigramme, en rejetant tous les quadrigrammes et trigrammes n'apparaissant qu'une seule fois dans les données d'apprentissage. Nous avons modifié l'outil de rescoring de graphes distribué par Sphinx à partir de la version s3.5 afin qu'il soit capable d'utiliser des modèles de langage quadrigrammes. Les modifications apportées impliquent un élagage du graphe pour éviter l'explosion combinatoire. Ensuite, une exploration combinatoire du graphe est menée en utilisant le modèle de langage quadrigramme.

Model	1-grammes	2-grammes	3-grammes	4-grammes
trigramme	65.5K	18.4M	25.4M	-
quadrigramme	65.5K	18.4M	22.2M	19.7M

TAB. 1.2 – Nombre de n-grammes dans les modèles de langages trigramme et quadrigramme utilisés lors du processus de reconnaissance de la parole

Le tableau 1.2 montre le nombre de n-grammes dans les modèles de langages trigramme et quadrigramme résultants.

1.8 Conclusion

Dans ce chapitre, nous avons présenté les systèmes de reconnaissance automatique de la parole et plus particulièrement celui qui est utilisé par le LIUM. Ce dernier est basé sur le

projet CMU Sphinx III, auxquels quelques éléments ont été ajoutés telles que l'adaptation de modèles acoustiques et une méthode de rescoring de graphe de mots [Deléglise 2005]. Les études menées dans la suite de ce manuscrit s'appuient sur ce système. Ce système a atteint un score de 23,7% WER pendant la campagne d'évaluation d'ESTER.

Le chapitre suivant présentera les différentes applications possibles pour un système de reconnaissance de la parole dont certaines seront utilisées pour mettre en oeuvre les capacités des mesures de confiance.

Chapitre 2

Applications

Sommaire

Un SRAP peut être vu comme un module appartenant à un système plus important. Les hypothèses fournies par le SRAP sont généralement utilisées dans diverses applications comme la commande vocale, les systèmes de dialogue (demande d'informations), la dictée vocale, la transcription automatique, la traduction, ou encore pour l'indexation de données audio et audiovisuelles. Certains systèmes comme les systèmes de demande de mot de passe utilisent en plus un module de reconnaissance du locuteur. Ce module est également présent lors de l'indexation en locuteur de documents sonores, domaine exploré dans la suite de ce manuscrit au chapitre portant sur l'identification automatique des segments par nom de locuteur de la partie II.

Le chapitre suivant évoquera après un historique rapide quelques unes des applications utilisant un SRAP. Pour mieux appréhender l'indexation en locuteurs de documents sonores, quelques principes de la reconnaissance du locuteur et ses applications seront décrits. Enfin, l'indexation de documents sonores ainsi que la transcription de documents sonores, aspect applicatif de ce manuscrit, seront détaillées.

2.1 Historique

Le tableau 2.1 propose un historique succinct de l'évolution des systèmes de reconnaissance de la parole. Les premiers systèmes de reconnaissance de la parole ne reconnaissent que les mots isolés et nécessitent une phase d'apprentissage longue et fastidieuse [Néel 2002]. L'apparition dans les années 60 des méthodes numériques et l'utilisation généralisée des ordinateurs apportent un renouveau à ce domaine dont les résultats restent toutefois modestes, quelques 500 mots isolés sont reconnus avec des systèmes dépendants du locuteur [Vicens 1969]. Les difficultés liées en particulier à la parole continue avaient été sous-estimées, telles que la variabilité du signal due au locuteur (état émotionnel ou physique, genre du locuteur, accent), la variabilité due aux conditions acoustiques (type de microphone, bruits...), la variabilité du canal de transmission (téléphone, radio...) et la variabilité due à la langue (discours spontané, hésitations, silences, reprises...). Vers 1970, on s'intéresse à l'apport de contraintes linguistiques dans le processus de décodage automatique de la parole [Vicens 1969, Tubach 1970]et les connaissances en micro-électronique conduisent à une augmentation de la puissance des ordinateurs. À partir de ce moment, deux voies de recherche ont été inspectées : la reconnaissance globale et la reconnaissance analytique [Calliope 1989]. Les systèmes à démarche globale ont été conçus à l'origine pour reconnaître directement les mots dans une tâche de reconnaissance limitée. Le contexte est ici dépendant du locuteur en mots isolés dans une ambiance peu bruitée. Les systèmes à démarche analytique [Lesser 1975, Lowerre 1976, Simon 1983, Stern 1986] ont été développés pour la plupart pour reconnaître de la parole continue, multilocuteur, à grands

Année	Événement
1952	Reconnaissance de chiffres par dispositif électronique câblé Système dépendant du locuteur utilisant les densités de passage par zéro [Davis 1952]
Années 60	Méthodes centi-secondes où une liste d'étiquettes phonétiques est attribuée à chaque 10ms de signal Systèmes dépendants du locuteur
1965	Reconnaissance de phonèmes en parole continue pour le japonais [Doshita 1965]
1968	Reconnaissance de mots isolés (500 mots) [Vicens 1969]
1969	Utilisation d'informations sémantiques et syntaxiques [Vicens 1969, Tubach 1970]
Années 70	Méthodes basées sur la programmation dynamique (DTW : Dynamic Time Warping [Higgins 1991]) Efficace pour des vocabulaire de petites tailles et des systèmes dépendants du locuteur
1971	Lancement du projet ARPA aux USA visant à tester la faisabilité de la compréhension automatique de la parole avec des contraintes raisonnables
1972	Premier appareil commercialisé de reconnaissance de mots isolés (24 mots) : le VIP100 [Hersher 1972]
1976	Fin du projet ARPA : les systèmes opérationnels sont HARPY [Lowerre 1976], HEARSAY II [Lesser 1975] et HWIM [Woods 1976]
Années 80	Méthodes statistiques utilisant les HMMs [Jelinek 1976] Amélioration du taux de reconnaissance SRAP indépendants du locuteur à grands vocabulaires en parole continue
1983	Première utilisation mondiale d'un système à commande vocale à bord d'un avion de chasse
1985	Les systèmes de reconnaissance dépassent le millier de mots reconnus
1986	Lancement du projet japonais ATR utilisant la traduction automatique en temps réel par le téléphone [Fujisaki 1987]
1988	Premières machines de dictée vocale par mots isolés
1989	Premier système de reconnaissance CMU Sphinx [Lee 1989]
Années 90	Méthodes hybrides utilisant les HMMs et les réseaux de neurones [Bourlard 1994, Franco 1992] Systèmes plus robustes au bruit, plus rapides et plus performants
1993	Premier SRAP de parole continue (langue allemande) fonctionnant en quasi temps réel présenté par Phillips à la conférence Eurospeech [Steinbiss 1993]
1993	IBM lance son premier système de reconnaissance vocale sur PC : *Speech Server Series* [Derouault 1993]
1997	IBM lance une machine à dictée vocale en parole continue : *IBM Voice Type-Dictée Personnelle* [Crépy 1997]
Années 2000	Compétition de plus en plus vive des différents laboratoires de recherche et des industriels notamment avec les campagnes d'évaluation NIST[4] Chutes des prix des produits proposés et amélioration notable des performance Démocratisation des produits notamment avec l'arrivée des serveurs vocaux par téléphone Dynamisation de la communication parlée grâce aux nouvelles technologies telles qu'Internet

TAB. 2.1 – Historique de la reconnaissance de la parole et de ses applications

vocabulaire et langage peu contraint. Les systèmes Summit [Zue 1989] et Arial [Caelen 1981] sont des exemples de systèmes à démarche analytique. Depuis, les systèmes sont capables de s'adapter à n'importe quel locuteur, ils gèrent la parole continue avec un vocabulaire de plusieurs centaines de milliers de mots, voire illimité en milieu calme [Mariani 2002b].

2.2 Commande vocale

Il s'agit ici de systèmes le plus souvent dépendants du locuteur pour la reconnaissance de mots isolés. De nombreux systèmes à commande vocale sont utilisés dans des avions de chasse, des automobiles, pour manoeuvrer des objets à distance ou encore pour l'aide aux personnes handicapées. En effet, dans des endroits exigus comme la cabine de pilotage d'un avion, la parole permet au pilote ou au conducteur de disposer un nouveau moyen d'interaction avec sa machine sans pour autant gêner son attention visuelle. Le projet VODIS (1995-1999) intègre un prototype de reconnaissance de mots clés (AudioNav) pour l'aide à la navigation embarquée à bord d'une automobile. Le projet CARIN [Cardeilhac 1995] permet le positionnement du véhicule, la planification de l'itinéraire et notamment le guidage du conducteur par des messages vocaux grâce aux réseaux GTTS (Global Transport Telematic System) intégrant GSM, GPS et Internet. Dans le domaine de l'avionique, les études menées par [Pastor 1993] (système TOP-VOICE) et par [Gerlach 1993] (système CASSY) s'intéressent à la robustesse du système vis-à-vis du bruit. Pour améliorer la productivité humaine, les systèmes à commande vocale permettent par exemple d'effectuer l'inventaire du stock d'une entreprise commerciale (mise en place dans les entrepôts du groupe Super U dans l'ouest de la France) de manière plus efficace que de manière écrite.

2.3 Systèmes de dialogue

Ce sont des systèmes multilocuteurs qui, en plus d'un module de reconnaissance de la parole, incluent des modules de compréhension, de synthèse de la parole, de génération de phrases et d'interrogation de bases de données. La figure 2.1 montre le fonctionnement d'un tel système. La plupart de ces systèmes fonctionnent à partir du téléphone et permettent d'orienter l'utilisateur à travers une base de données comme les réservations de billets de train (Railtel [Billi 1997] et ARISE [Lamel 1999]), des informations touristiques (système GEORAL[Siroux 1995] donnant des informations sur la région du Trégor en Bretagne)... Un autre exemple de système de dialogue est le système CMUCommutator, provenant de l'université de Carnegie Mellon. C'est un système permettant à l'utilisateur de planifier un voyage en avion, de louer une voiture, de réserver une chambre via un serveur vocal et le réseau téléphonique [Rudnicky 1999]. Le

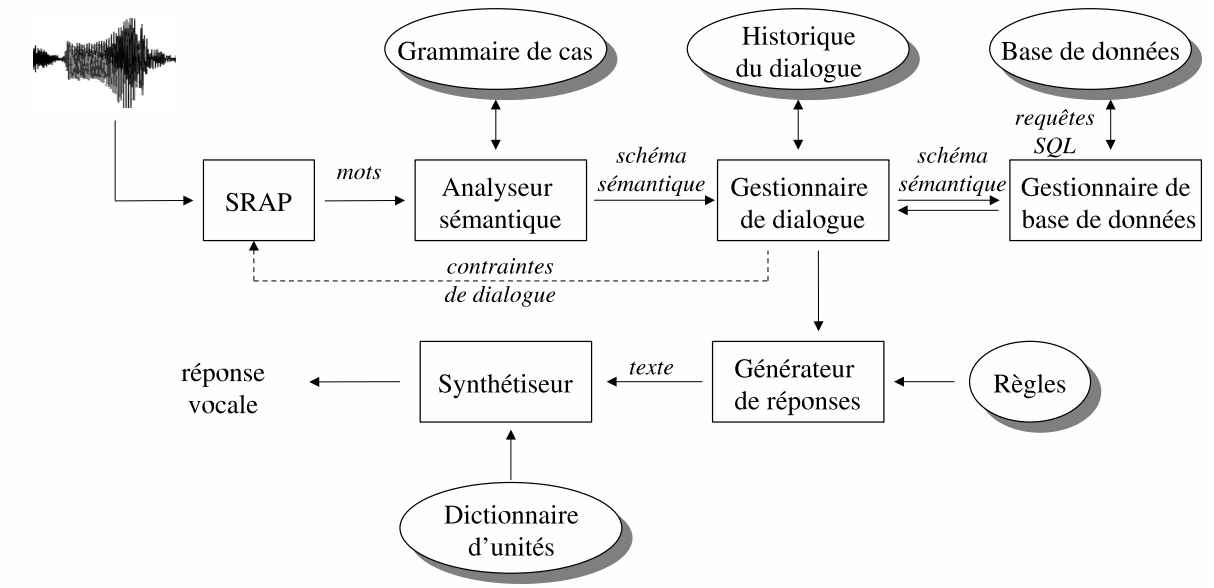


FIG. 2.1 – Système de dialogue

système Artimis de France Télécom permet en langue naturelle une réelle activité conjointe dans plusieurs domaines applicatifs telles que la transaction boursière, le tourisme ou les voyages basée sur une hiérarchie d'objectifs clairement établis [Sadek 1997, Panaget 1998]. Les exigences de tels systèmes sont par exemple le fonctionnement en temps réel, l'indépendance du locuteur et le traitement de la parole spontanée avec des hésitations, des retours en arrière et des reformulations. Si la taille du vocabulaire de ces serveurs vocaux est limitée, un module de gestion de mots hors vocabulaire est nécessaire. Les performances de tels systèmes varient en fonction de la complexité de la tâche. Elles sont de moins de 5% d'erreurs pour un vocabulaire limité à une dizaine de mots et un microphone de proximité et atteignent 25% pour un système téléphonique avec un vocabulaire d'un millier de mots [Néel 2002].

2.4 Dictée vocale

Pour le grand public et les professionnels, on trouve bon nombre de logiciels facilitant la prise de notes grâce à la transcription de la parole de l'utilisateur. Pour cette application, le locuteur est connu (système dépendant du locuteur) et les conditions de prise de son sont généralement optimales du fait de la proximité du micro dans un environnement assez peu bruité. Elle nécessite également un long temps d'apprentissage afin que le système s'adapte à la voix, notamment aux accents régionaux, aux défauts d'élocution et de prononciation... Pour la parole spontanée, les performances de tels systèmes se situent aux alentours de 14%. Pour

des textes lus, le taux d'erreur est d'environ 7% sur de l'anglais américain [Pallett 1996]. Des résultats similaires ont été observés pour le français (campagne AUPELF [Dolmazon 1997]) et l'allemand [Young 1997].

2.5 Traduction automatique

Ces systèmes tentent de pallier la barrière de la langue. Des applications comme la traduction de mails oraux, de cours magistraux en direct ou tout simplement la traduction instantanée d'un locuteur étranger sont envisagées par ce domaine. La figure 2.2 montre le diagramme

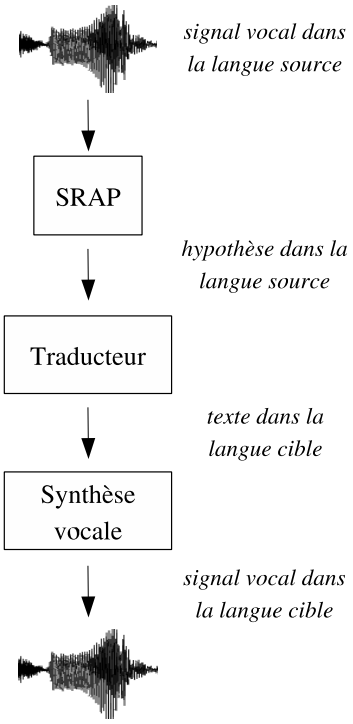


FIG. 2.2 – Diagramme d'un traducteur parole-parole

d'un traducteur automatique parole-parole. Le signal vocal en langue source est transcrit au moyen d'un SRAP. Ce texte est traduit dans la langue cible grâce au composant de traduction. Ensuite, la synthèse vocale permet de transformer ce texte en signal vocal dans la langue cible. La traduction automatique de la parole nécessite de résoudre plusieurs problèmes : d'une part reconnaître la parole continue prononcée par un locuteur quelconque, puis en comprendre le sens pour générer un énoncé dans la langue cible et enfin le synthétiser avec une formulation et une voix les plus naturelles possibles.

Par exemple, les travaux de [Fürgen 2006] offrent une application de la traduction automatique parole-texte pour des séminaires ou encore des cours magistraux. Dans le cadre de

la traduction parole-parole, le consortium C-STAR [5] (Consortium for Speech Translation Advanced Research) fait coopérer plusieurs laboratoires (Japonais, Coréen , Anglais, Français, Allemand et Italien) et permet le regroupement de systèmes de traduction d'une langue à un autre. Ce consortium permet de dynamiser les recherches de chacun des laboratoires et permet également les interactions entre chacun des partenaires. Pour le moment, ce consortium est centré autour des domaines touristiques tels que la demande d'informations, de réservations et de planning. Une campagne d'évaluation européenne dans le domaine de la traduction automatique parole-parole a été également mise en place à partir de mars 2005. Il s'agit de la campagne TC-STAR [6] (Technology and Corpora for Speech to Speech Translation) [Mostefa 2006] qui permet d'encourager les avancées dans ce domaine sur trois langues : l'anglais européen (le meilleur système est à 10,6% WER), l'espagnol (le meilleur système est à 11,5% WER) européen et le mandarin (le meilleur système est à 10,7% WER). Les tâches de traduction comprennent la traduction vers l'anglais d'émissions radiophoniques chinoises, des sessions plénières du parlement européen traduites de l'espagnol à l'anglais et vice et versa. Il y a également dans cette campagne des tâches de transcription automatique et de synthèse vocale.

2.6 La reconnaissance du locuteur

La reconnaissance automatique du locuteur (RAL) vise à déterminer si un échantillon de voix a été prononcé par une personne donnée. Elle peut être scindée en deux catégories :

- **Identification du locuteur**
  Parmi un ensemble de locuteurs connus, il s'agit de rechercher l'identité du locuteur possédant la référence la plus proche du signal vocal donné.
- **Vérification du locuteur**
  Étant donné un signal vocal et une identité proposée par un locuteur, il s'agit d'accepter ou de rejeter l'hypothèse que le locuteur considéré l'ait prononcé.

Les articles [Bimbot 2002, Merlin 2004, Bimbot 2004]offrent plus de détails sur la reconnaissance du locuteur.

Le diagramme 2.3 représente une application d'authentification du locuteur comme dans les applications de serrure vocale. Le système connaît les locuteurs autorisés à ouvrir la serrure alors que des imposteurs peuvent tenter de la franchir. Le système d'authentification du locuteur peut par exemple être obtenu comme sur cette figure en mettant en cascade un système d'identification du locuteur avec un système de vérification.

[5] http://www.c-star.org/
[6] http://www.tc-star.org/

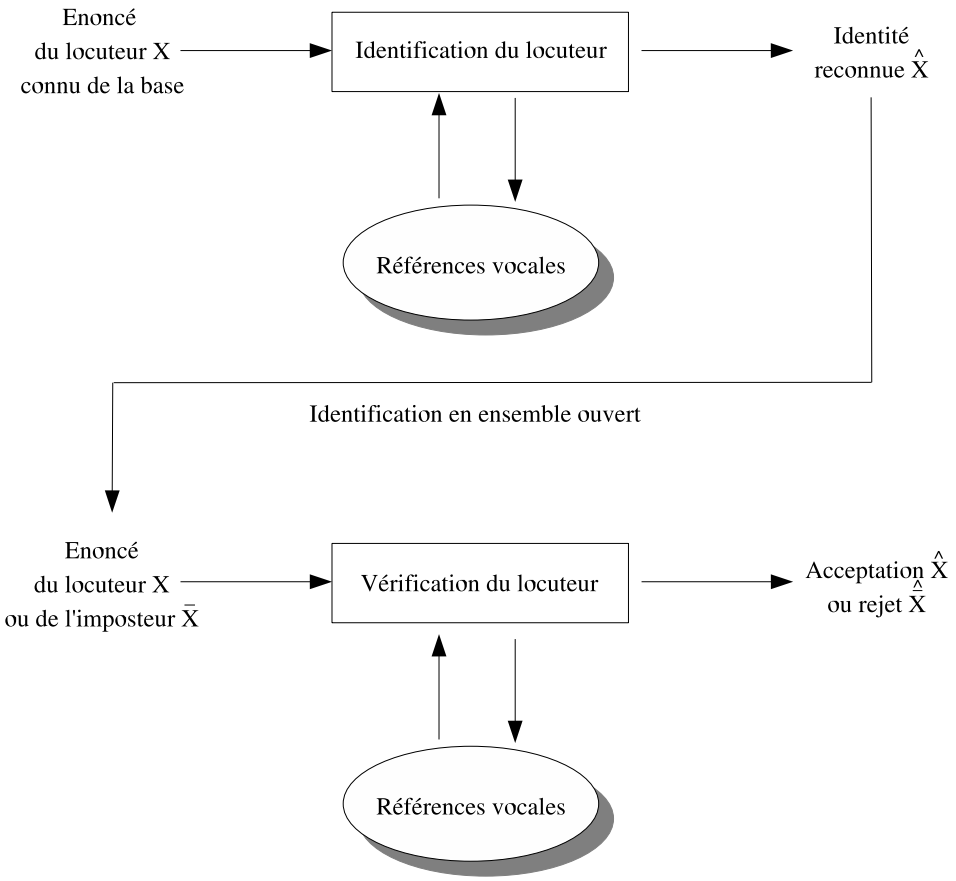


FIG. 2.3 – Diagramme représentant un système d'identification du locuteur (en haut) couplé avec un système de vérification du locuteur (en bas)

Les applications liées à la reconnaissance du locuteur sont de l'ordre du contrôle d'accès, de la vérification de présence, de la protection d'équipements, de l'authentification, de la personnalisation d'informations...

Certaines applications comme le contrôle d'accès font appel à des mots de passe où la vérification du locuteur est dépendante de la reconnaissance d'un mot isolé. Ici, l'environnement de prise de son est généralement isolé du bruit. L'énoncé de chaque utilisateur a été appris et est stocké dans une base de données de références vocales. Des méthodes de comparaison sont enclenchées lors de la vérification du mot de passe telles que les méthodes DTW (Dynamic Time Warping) [Higgins 1991], des méthodes basées sur les HMMs [Jelinek 1997, Rabiner 1993] ou encore des méthodes de modélisation statistique combinée à des informations temporelles (GDW : Gaussian Dynamic Warping) [Bonastre 2003].

Les systèmes de reconnaissance du locuteur sont comparés grâce au taux d'égale erreur (EER : Equal Error Rate), c'est à dire quand le taux de fausses acceptations (locuteur incorrectement accepté) est égal au taux de faux rejets (locuteur incorrectement rejeté). Ce taux correspond au point de fonctionnement du système. Si ce sont des systèmes dépendants du texte (contraintes sur le contenu linguistique de l'énoncé du locuteur), il est de l'ordre de 0,5% pour de bonnes conditions de prises de son. Pour des systèmes indépendants du texte où il n' y a pas de contraintes sur l'énoncé que doit prononcer l'utilisateur, il est de plus de 2%. Pour la parole téléphonique, les taux se dégradent de 5 à 10% [Bimbot 2002]. Ce point d'EER sert uniquement à comparer les différents systèmes car dans des applications telles que le contrôle d'accès à des transactions bancaires, il est nécessaire de sélectionner un seuil que le taux de fausses alarmes ne doit pas dépasser [Bengio 2004]. Ainsi, pour les systèmes de vérification du locuteur, une fonction de coût de détection DCF est calculée. Cette fonction est définie par NIST Speaker Recognition Evaluation [NIST04]. Il s'agit de chercher à minimiser la formule :

$$DCF = C(FA).P(imposteurs).TFA + C(FR).P(client).TFR \tag{2.1}$$

où $C(FA)$ et $C(FR)$ sont respectivement le coût d'une fausse acceptation et le coût d'un faux rejet, $P(imposteurs)$ et $P(client)$ sont respectivement la probabilité *a priori* qu'un imposteur fasse une tentative de vérification et la probabilité *a priori* qu'un client fasse une tentative de vérification. À chaque valeur du seuil utilisé pour prendre la décision correspond une valeur de la fonction de coût de détection.

En termes de performances, la reconnaissance du locuteur, pour des systèmes de vérification nécessitant généralement une fiabilité accrue, ne permet pas une confiance absolue. Plusieurs études montrent néanmoins que la modalité de la parole s'avère très conviviale pour les utilisateurs [Bimbot 2002] et le matériel nécessaire à la prise de son et au système d'authentification est à un coût plus accessible que d'autres techniques d'authentification. Le projet SuperSid[7] [Reynolds 2003] montre qu'en exploitant des données de plus haut niveau que les données acoustiques telles que la prosodie, la prononciation ou encore les interactions conversationnelles, on peut atteindre un taux de 0,2% soit 71% de gain relatif par rapport à l'état de l'art.

2.7 Transcription enrichie de documents sonores

Les documents sonores contiennent des données non homogènes acoustiquement et sont donc difficiles à transcrire. Par exemple, une émission de radio contient des publicités, de la musique, qui ne devront pas être mis au même plan qu'une interview suivant les informations

[7] `http://www.clsp.jhu.edu/ws2002/groups/supersid/`

qu'on veut mettre en exergue pour caractériser le document sonore. En moyenne, pour un annotateur humain, il faut une dizaine d'heures pour transcrire seulement une heure de parole d'un document sonore en l'enrichissant d'informations telles que le locuteur, ses hésitations ou encore la transcription de ce qui est dit. Grâce à un SRAP, la transcription est automatisée, évitant aux usagers de longues heures de transcription. Dans cette tâche, les tours de parole non respectés, le jargon souvent spécifique employé par les protagonistes, les diverses sources microphoniques difficiles à séparer et la captation du bruit environnant sont autant de difficultés à gérer par le système de transcription.

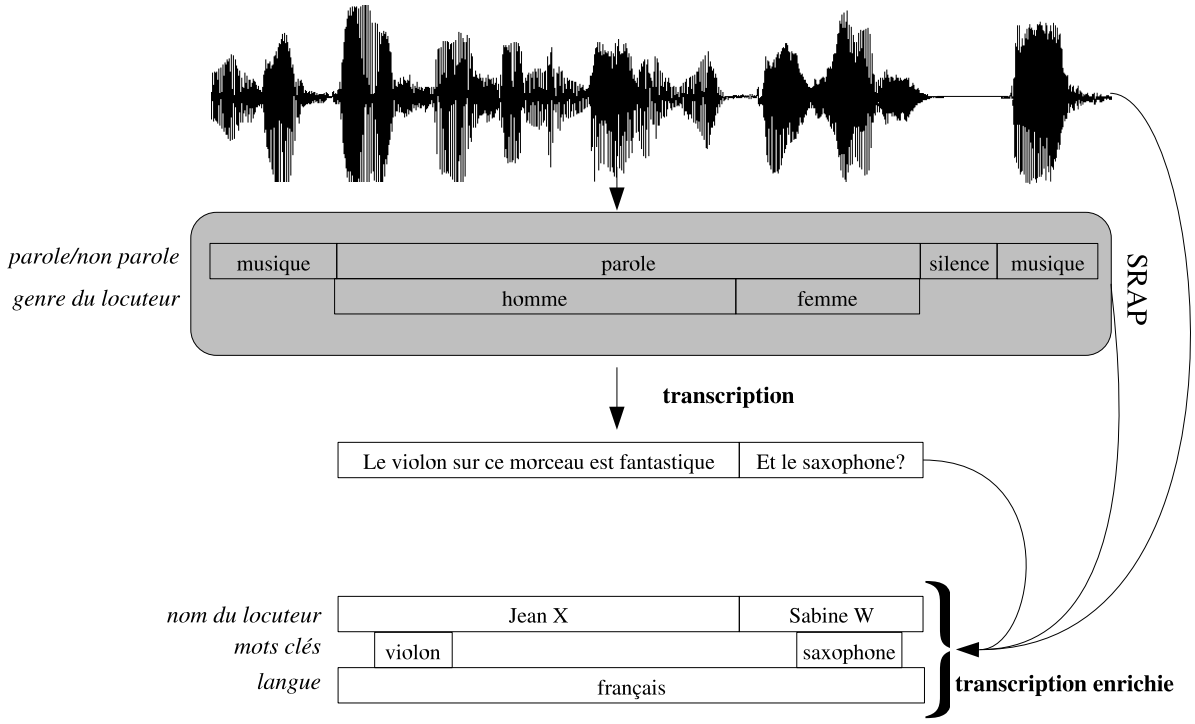


FIG. 2.4 – Système de transcription enrichie

La première étape d'un système de transcription (voir figure 2.4) consiste à segmenter le flux audio en fonction de la présence ou non de parole, en fonction des conditions d'enregistrement (parole téléphonique, studio, rue...), pour ensuite permettre au système de transcription de choisir le modèle le plus adapté pour retranscrire telle ou telle partie du signal. Ensuite, pour enrichir la transcription, un annotateur humain peut ajouter des informations sur le locuteur, la langue employée, les mots-clés caractérisants un passage du document, les thèmes abordés dans ce document... Cette annotation peut également être automatisée au moyen de systèmes de détection d'entités nommées, d'identification des langues ou de reconnaissance du locuteur.

2.7.1 Indexation

Une des perspectives de la transcription enrichie est l'indexation de documents. En effet, l'annotation mettra en correspondance les différents segments du document avec ses caractéristiques (langue du locuteur, son nom, le thème de l'émission, s'il s'agit ou non de musique...). Un index de ces caractéristiques pourra être créé, regroupant les informations de plusieurs documents. Grâce à un moteur de recherche ayant accès à cette base de donnés, un utilisateur pourra retrouver et consulter un document ou seulement certaines de ses parties correspondant à sa requête.

Un des projets actuels dans le domaine de l'indexation est le projet Audiosurf [8]. Le but de ce projet est de créer une plate-forme d'indexation de l'audio afin de permettre à des utilisateurs de retrouver une information dans un corpus d'émissions radiophoniques ou télévisuelles comme Radio France, France Télévision... comme s'il s'agissait de documents texte. Le projet THISL [9] [Abberley 1999] a pour but de permettre une démonstration de recherche d'émissions radiophonique pour la BBC. Dans une première étape, le système de RAP ABBOT [10] fourni la transcription des émissions. Ensuite le système procède à l'indexation de celle-ci via le système de recherche d'informations thisIIR.

Un document sonore peut par exemple être indexé en locuteurs. Les mesures de confiance seront utilisées dans le chapitre 7 pour déterminer le nom du locuteur directement à partir de la transcription. Le but est ici d'annoter le document sonore en fonction de qui parle pour avoir une indication sur le locuteur [Delacourt 2000a, Meignier 2002]. Pour cela, on procède en quatre étapes :

1. Le document est annoté suivant les locuteurs :
   - découpage du flux audio en segments homogènes (même caractéristiques acoustiques),
   - regroupement des segments par locuteur et détermination du nombre de locuteurs intervenant dans le document.
2. À partir de cette segmentation en locuteurs, un index du document est crée. Cet index associe ici à un locuteur, les segments correspondants.
3. Dans le cas où il y aurait plusieurs documents à indexer, une troisième phase permet de rechercher les locuteurs intervenant dans ces différents documents. Le but est de répartir en classes les locuteurs intervenant dans plusieurs documents. C'est l'appariement en locuteurs [Meignier 2002].

[8] `http://www.sinequa.com/html-fr/fr-recherche.audiosurf.html` et `http://www.audiosurf.org`

[9] `http://www.dcs.shef.ac.uk/spandh/projects/thisl`

[10] `sheffield`

4. Enfin, on organise les documents grâce à un index qui met en correspondance les locuteurs avec les documents. L'utilisateur peut ainsi faire une requète au système qui grâce à cet index peut retrouver les documents et les segments d'un locuteur.

Dans la terminologie NIST, cette étape s'appelle la *diarization* dont les principales méthodes reposent uniquement sur des paramètres acoustiques [Chen 1998, Delacourt 2000b, Ajmera 2003, Ben 2004, Barras 2006, Meignier 2006]. Cependant, à la sortie du processus de *diarization*, les segments sont étiquetés anonymement, le locuteur étant représenté soit par un mot défini arbitrairement, soit par des exemples de sa voix. Il semble pourtant pertinent, pour enrichir la transcription de documents sonores, de connaître l'identité du locuteur. Ce problème sera traité dans le chapitre 7 où sera décrite une méthode d'identification nommée du locuteur.

2.7.2 Transcription de réunions

Une autre perspective de la transcription est la transcription de réunions. Lors de réunions dans une entreprise, il est intéressant pour les protagonistes de bénéficier d'un compte-rendu exact des échanges qui ont eu lieu. Le projet européen CHIL (Computers in the Human Interaction Loop) [11] [Waibel 2004] s'intéresse au problème de la transcription de réunions en anglais. Aucun système n'est à ce jour développé pour le français.

2.7.3 Campagnes d'évaluation

Les campagnes d'évaluation dans les domaines de la transcription et de l'indexation sont centrées sur une tâche en particulier et permettent aux applications et aux systèmes qui les composent d'être toujours plus performants. Par exemple, la campagne Spoken Data Retrieval (SDR) qui étend la campagne d'évaluation TREC (Text Retrieval Conference) de NIST à la parole permet de rechercher des documents sonores par l'indexation de mots-clés et de thèmes [12]. La campagne TDT [13](Topic Detection and Tracking) de NIST a pour ambition de détecter des thèmes, des sections, de les suivre à travers des émissions radiophoniques en anglais et en mandarin. L'évaluation RT-04 (Rich Transcription 2004 Spring Meeting Recognition Evaluation) [Garofolo 2004] fait partie des évaluations NIST sur les transcriptions enrichies et inclut les tâches de segmentation en locuteur et de transcriptions de réunions. En France, la campagne ESTER [Galliano 2005] portait uniquement sur la transcription d'émissions radiophoniques (voir l'introduction de ce manuscrit). Pour la transcription automatique, les performances se situent entre 10 et 30% d'erreur.

[11] `http://chil.server.de`
[12] `http://www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.htm`
[13] `http://www.nist.gov/speech/tests/tdt/index.htm`

2.8 Conclusion

Différentes applications du traitement de la parole ont été décrites dans ce chapitre, permettant de survoler les différents domaines qui sont susceptibles d'utiliser les sorties d'un SRAP.

Les SRAP possèdent des corpora d'apprentissage toujours plus importants et des modèles linguistique d'ordres plus élevés garantissant une amélioration de leurs performances. En effet, ils sont plus robustes aux changements de conditions acoustiques et aux changements de locuteur. Malgré ces nouvelles performances, le taux d'erreur d'un système de reconnaissance reste insatisfaisant (environ 20% d'erreurs) comparé au taux d'erreur observé pour un annotateur humain. Pour réduire cette différence, et ainsi pouvoir réduire considérablement l'intervention humaine, il est nécessaire de trouver des techniques d'amélioration des performances en faisant par exemple appel à des mesures de confiance. Ces mesures de confiance permettront de détecter les erreurs de reconnaissances du système et, grâce à leur gestion, ce système proposera une meilleure assise à toutes les applications qui en sont tributaires.

Chapitre 3

Mesures de confiance

Sommaire

Une mesure de confiance associée à une hypothèse de reconnaissance est une estimation de la fiabilité de cette hypothèse. La mesure de confiance $CM(h)$ associée à une hypothèse h appartient à l'intervalle $[0, 1]$ et peut être interprétée comme étant la probabilité que l'hypothèse soit correcte ou non. Idéalement, $CM(h)$ vaut 0 si l'hypothèse h est incorrecte, 1 si elle est correcte.

Les mesures de confiance sont utilisées dans de nombreux domaines du traitement de la parole [Lee 2001] comme la reconnaissance de la parole [Wessel 2005, Cox 2002], les systèmes de dialogue [San-Segundo 2001] ou encore l'identification des langues [Metze 2000].

Dans les différents domaines du traitement de la parole, les mesures de confiance sont applicables à plusieurs niveaux : au niveau du phonème (principalement pour la reconnaissance de la parole), sur le mot, sur une phrase entière ou encore sur un concept (notamment utile en dialogue, voir [San-Segundo 2001, Raymond 2004, Raymond 2005]). Dans la suite de ce chapitre, on se placera généralement dans le cas où w est un mot mais la plupart des mesures suivantes sont applicables au niveau du phonème, d'une phrase ou d'un concept.

L'article [Jiang 2005] propose de classer les mesures de confiance en trois catégories :

1. la majeure partie des travaux dans ce domaine tendent à calculer des mesures de confiance en combinant des paramètres prédictifs. Ces paramètres sont collectés durant le décodage et incluent aussi bien des paramètres acoustiques que des paramètres issus du modèle de langage ou du comportement de l'algorithme de recherche. Ensuite, tous les paramètres choisis sont combinés en une seule mesure attestant du degré de véracité d'un mot.
2. la mesure de confiance dérivée de la probabilité *a posteriori* d'un mot est souvent utilisée dans la littérature. En pratique, cette mesure peut être estimée de plusieurs façons comme nous le verrons par la suite.
3. à la fin du processus de reconnaissance d'une observation X, nous obtenons une hypothèse W. La fiabilité de cette hypothèse est évaluée au moyen d'une mesure de confiance. L'estimation des mesures de confiance est ici formulée comme un test statistique pour vérifier si l'hypothèse est correcte ou non et la ranger dans l'une des deux classes *Correct/Incorrect*.

3.1 Propriétés d'une mesure de confiance

Dans toutes les applications du traitement de la parole, les hypothèses pour lesquelles sont estimées les mesures de confiance sont fournies par un SRAP. Soit la séquence de K mots reconnus par ce système que l'on note $\{w_1, \ldots, w_K\}$. Chaque mot w peut être associé à une mesure de confiance $CM(w)$. Une mesure de confiance idéale doit posséder deux propriétés :

1. Elle doit être égale à 0 si w est incorrect et égale à 1 si w est correct.
2. Soit la moyenne des mesures de confiance des K mots reconnus :

$$\mu(CM) = \frac{1}{K}\sum_{i=1}^{K} CM(w_i)$$

On définit alors la dernière propriété d'une mesure de confiance idéale par le fait que $\mu(CM)$ doit être une approximation du taux de mots émis bien reconnus. Par rapport au taux WER (voir l'équation 1.8), le taux de mots émis bien reconnus ne prend pas en compte les suppressions car il s'applique uniquement aux mots émis par le SRAP. Une mesure de confiance doit donc appartenir à l'intervalle $[0, 1]$.

3.2 Rapports de vraisemblance

Dans le cadre de la reconnaissance de la parole, le principal objectif d'utilisation d'une mesure de confiance est de pouvoir ranger l'observation X dans l'une des classes suivantes [Moreau 2000, Moreau 2001, Lee 2001] :

- $H_{cor}(W)$: hypothèse selon laquelle le résultat W du processus de reconnaissance est correct.
- $H_{inc}(W)$: hypothèse selon laquelle ce même résultat est incorrect.

Dans le cadre là, le rapport de vraisemblance $LR(X|W)$ constitue généralement une mesure de confiance sur l'hypothèse de reconnaissance W :

$$LR(X|W) = \frac{P(X|H_{cor}(W))}{P(X|H_{inc}(W))} \tag{3.1}$$

Pour l'estimation de ce rapport de vraisemblance, il s'agit de trouver un moyen de calculer $P(X|H_{cor}(W))$ et $P(X|H_{inc}(W))$ et donc, de modéliser $H_{cor}(W)$ et $H_{inc}(W)$. Le problème principal réside dans la modélisation de $H_{inc}(W)$ qui regroupe toutes les hypothèses alternatives erronées.

Le paragraphe suivant montre comment des HMMs et des distributions de probabilité sont utilisés pour modéliser $H_{cor}(W)$ et $H_{inc}(W)$.

3.2.1 Modélisation par HMMs

Si $H_{cor}(W)$ peut être estimé avec le score acoustique de W (fourni lors du décodage de X par le système), l'estimation de $H_{inc}(W)$ n'est pas immédiate. La méthode constitue à

calculer des scores “alternatifs” en construisant des cohortes, des anti-modèles ou des modèles de rejet [Charlet 2001, Sukkar 1996, Rahim 1997, Rose 2001, Henández-Ábrego 2000, Falavigna 2002].

La plupart de ces modèles fournissant des scores “alternatifs” s’appuyaient au départ sur des modèles de mot [Sukkar 1996, Rahim 1997]. Par la suite, des modèles de phonèmes ont été estimés pour obtenir ces scores [Falavigna 2002].

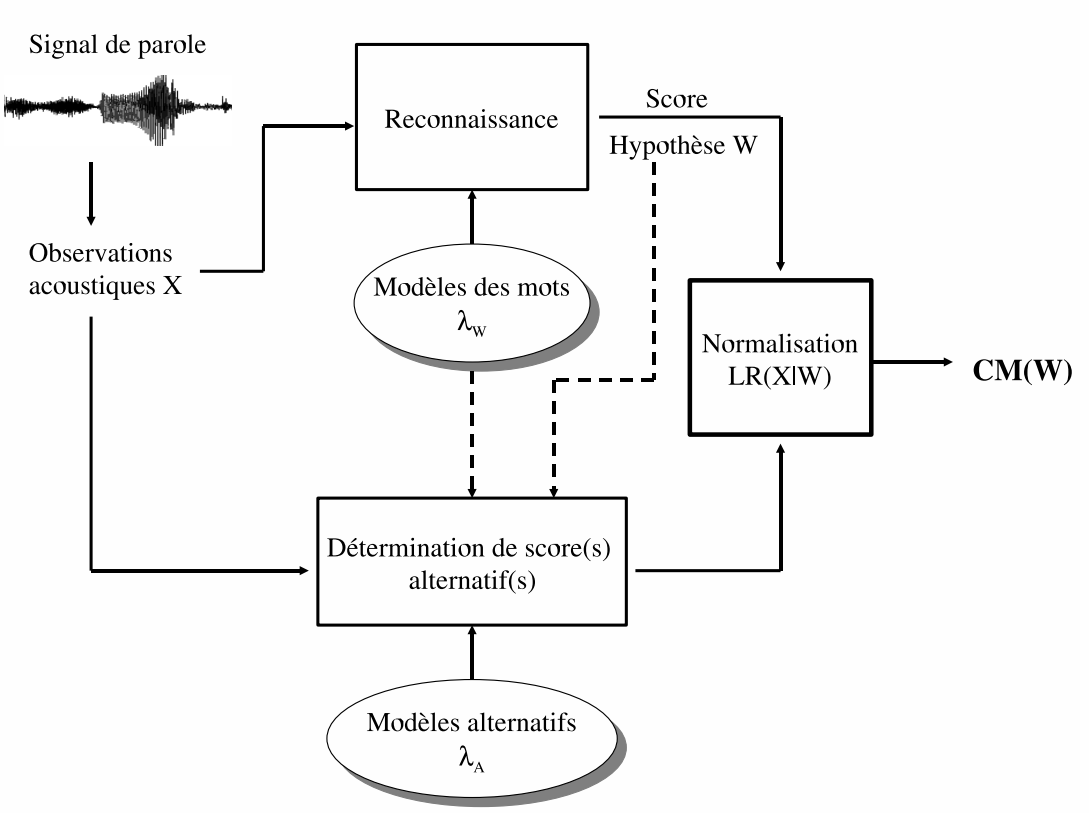


FIG. 3.1 – Estimation de la vraisemblance LR(X|W) par modélisation de HMMs alternatifs

Cohortes

Dans une application portant sur un vocabulaire de K mots [Boite 2000, Rahim 1997], le calcul de $LR(X|W)$ s’effectue en normalisant $P(X|W)$ par les vraisemblances de tous les autres mots possibles. Les auteurs de [Boite 2000] procèdent par simple somme des vraisemblances :

$$LR(X|W) = \frac{P(X|W)}{\sum_{i=1, W_i \neq W}^{K} P(X|W_i)} \tag{3.2}$$

Ceci n'est possible que pour une valeur de K réduite. Ainsi, pour un vocabulaire limité à 11 mots lors de la reconnaissance de chiffres, [Rahim 1997] normalise le score d'un mot par les scores obtenus par les 10 autres.

Cette méthode de résolution demande un coût trop élevé dès que K devient important. Si la liste des N meilleures hypothèses est disponible, une solution est de normaliser le score de la meilleure hypothèse par le score des $N-1$ hypothèses suivantes. Par exemple, [Charlet 2001] utilise le score de la deuxième meilleure hypothèse.

Anti-modèles

En général, la méthode consiste à définir pour chaque modèle λ_W du système un anti-modèle $\widetilde{\lambda_W}$ associé [Sukkar 1996, Rahim 1997]. Le HMM $\widetilde{\lambda_W}$ est entraîné sur les données du corpus d'apprentissage qui ne génèrent pas W. L'équation 3.1 devient :

$$LR(X|W) = \frac{P(X|\lambda_W)}{P(X|\widetilde{\lambda_W})} \tag{3.3}$$

Dans [Falavigna 2002], $H_{inc}(W)$ est modélisé grâce à un anti-modèle de phonèmes. Pour chaque HMM phonémique ph_i, un anti-HMM $\widetilde{ph_i}$ est estimé, il prend en compte toutes les observations acoustiques qui ne génèrent pas ph_i.

Modèles de rejet

Il s'agit ici de disposer d'un ou de plusieurs modèles d'entrées incorrectes. Ce sont des modèles poubelles ou modèles de rejet. Ces HMMs sont appris à partir de parole hors-vocabulaire, de bruits ... L'équation 3.1 devient dans ce cas :

$$LR(X|W) = \frac{P(X|\lambda_W)}{P(X|\lambda_{rejet})} \tag{3.4}$$

Pour [Sukkar 1996, Rahim 1997, Tsiporkova 2000], λ_{rejet} est un modèle de rejet unique modélisant l'ensemble des mots hors-vocabulaire.

3.2.2 Modélisation par distribution de probabilité

La modélisation de $H_{cor}(W)$ et $H_{inc}(W)$ peut se faire grâce à des distributions de probabilités (voir figure 3.2). Cette modélisation permet d'utiliser d'autres sources d'information que les scores acoustiques issus d'alignement sur des HMMs mais demande un corpus d'apprentissage spécifique. Un classificateur Bayésien permet de modéliser les deux classes du problème (*Correct* et *Incorrect*), par deux distributions de probabilités définies dans un espace de paramètres

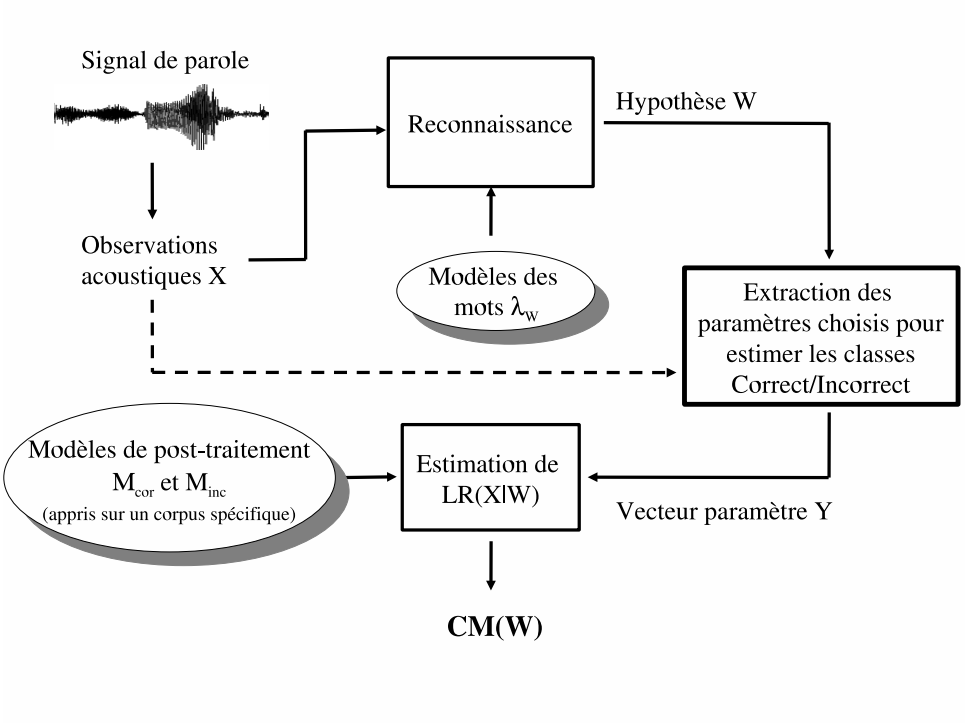


FIG. 3.2 – Estimation de la vraisemblance LR(X|W) par modélisation d'espaces paramétriques M_{cor} et M_{inc}

Y choisi auparavant. L'espace des mots corrects est appelé M_{Cor} et celui des mots incorrects M_{Inc}. Les paramètres de modélisation peuvent être :

- Les scores acoustiques des phonèmes composant W [Kamppari 2000]
- Le nombre d'hypothèses avec lesquelles W est en concurrence dans la liste N-Best[Kamppari 2000]
- La durée du mot W[Kamppari 2000]
- La distance entre le résultat d'une boucle de phonème non-contrainte et la transcription phonétique du système[Cox 2002]
- ...

Ces différents paramètres servent d'espace de représentation pour modéliser l'espace des mots corrects M_{Cor} et l'espace des mots incorrects M_{Inc}. Les scores de chacun de ces deux espaces sont des estimations des vraisemblances $P(X|H_{cor}(W))$ et $P(X|H_{inc}(W))$. On apprend alors les distributions de manière identique sur deux corpora, M_{Cor} qui est composé de données correctement reconnues et M_{Inc} qui est composé de données donnant lieu à des erreurs, indépendamment de l'hypothèse W considérée. Ces distributions peuvent être des densités gaussiennes comme c'est le cas dans [Kamppari 2000] ou encore deux histogrammes[Cox 2002].

3.3 Probabilités *a posteriori*

Pour un mot reconnu, les SRAP peuvent calculer sa probabilité *a posteriori* $P(w|X)$. C'est une estimation du fait que w correspond bien à l'information lexicale portée par les observations acoustiques X [Moreau 2000]. Dans la section 1.1 du chapitre traitant de la reconnaissance de la parole, l'équation 1.1 montre que, lors de l'attribution du score au mot reconnu w, les systèmes de reconnaissance omettent la normalisation par la probabilité $P(X)$, $P(X)$ étant constante sur l'ensemble des mots w. Le score délivré par le système de reconnaissance n'est plus $P(w|X)$ mais simplement $P(w).P(X|w)$. Avec ce score, l'hypothèse la plus vraisemblable du point de vue des connaissances du système est connue mais on ne sait pas à quel point elle est correcte. Le score délivré par le SRAP est donc inadéquat en tant que mesure de confiance.

Pour avoir une véritable mesure de confiance, il s'agit de déterminer $P(X)$. En théorie, on a :

$$P(X) = \sum_{hyp} P(hyp).P(X|hyp) \tag{3.5}$$

où toutes les hypothèses hyp possibles sont sommées pour l'ensemble des observations acoustiques X. Pour hyp, toutes les combinaisons de mots possibles doivent être prises en compte, phonèmes, bruits, etc. Dans la pratique, énumérer toutes les hypothèses possibles n'est pas réalisable. Des techniques d'approximation sont alors employées, les plus courantes se basant sur l'utilisation de graphes de mots ou encore sur les listes des N meilleures hypothèses.

3.3.1 Approximation par graphe de mots

Les auteurs de [Kemp 1997, Wessel 1998, Wessel 1999, Wessel 2000, Wessel 2001] et ceux de [Metze 2000] proposent de calculer directement la probabilité *a posteriori* $P(w|X)$ grâce au graphe de mots généré par un SRAP lors du processus de décodage : un SRAP fournit l'hypothèse la plus probable, mais peut également fournir un ensemble d'hypothèses sous la forme d'un graphe. Pour la majorité des systèmes, la taille de l'espace de recherche des hypothèses est une donnée critique en terme d'occupation mémoire, et surtout de vitesse de traitement. Pour restreindre cet espace de recherche, une technique d'élagage est pratiquée durant le décodage, supprimant les hypothèses qui semblent localement peu compétitives. Dans le graphe de mots, seules les hypothèses les plus prometteuses sont conservées. Même s'il ne s'agit que d'une approximation et que toutes les hypothèses ne sont pas prises en compte, ce sont les plus plausibles et elles dominent dans le calcul de $P(X)$. La somme des hypothèses du graphe est donc une bonne estimation de $P(X)$.

L'algorithme de calcul est le suivant :

Étant donné un arc a auquel est associé un mot w commençant au noeud s et finissant au noeud

e avec un poids $b(w)$, la probabilité *a posteriori* de l'arc pour un graphe $\mathcal{G}$ est le ratio entre la somme des probabilités de tous les chemins passant par l'arc a et la somme des probabilités de tous les chemins composant $\mathcal{G}$:

$$P(a|\mathcal{G}) = \frac{\sum_{C \epsilon \mathcal{G}, a \supset C} P(C|\mathcal{G})}{\sum_{C \epsilon \mathcal{G}} P(C|\mathcal{G})} \tag{3.6}$$

où $C \in \mathcal{G}$ signifie que C est un chemin complet du graphe $\mathcal{G}$ et $a \supset C$ signifie que le chemin complet C passe par a. La probabilité $P(a|\mathcal{G})$ peut être calculée grâce à l'algorithme forward-backward [Baum 1972] comme c'est le cas dans les articles [Wessel 2001, Jiang 2005, Metze 2000]). Cette probabilité peut être directement utilisée comme mesure de confiance du mot w supporté par l'arc a mais ses performances en tant que mesure de confiance ne sont pas optimales. En effet, l'arc a n'est pas le seul arc dont le mot associé est w. Il existe d'autres arcs de $\mathcal{G}$ qui sont associés au même mot w avec des instants début et fin légèrement différents et la chronologie du graphe de mot a alors un impact très fort sur les probabilités *a posteriori* des mots dont il est composé. Ne compter qu'un seul arc dans le calcul de la probabilité *a posteriori* de w va sous-estimer son score de confiance car la probabilité totale pour un mot est répartie entre tous les arcs associés à ce mot. [Wessel 2001] propose trois méthodes de calcul de mesures de confiance contournant ce problème.

La première méthode consiste à prendre en compte tous les arcs associés au mot w et qui ont un chevauchement temporel avec l'arc courant a. La seconde cumule les probabilités de tous les arcs associés à w et qui ont un chevauchement temporel avec la trame médiane de a La troisième quant à elle trouve les arcs associés à w et calcule la somme des probabilités de ceux qui ont un chevauchement temporel avec une des trames de a. On calcule la probabilité pour chacune des trames de a et on prend le résultat maximum. Cette méthode donne les meilleurs résultats en termes de mesure de confiance pour le mot w.

3.3.2 Approximation par la liste des N meilleures hypothèses

Une autre possibilité de calcul de $P(w|X)$ peut se faire grâce à la liste des N meilleures hypothèses [Wessel 1999, Wessel 2001, Stolcke 1997]. En effet, l'intérêt d'utiliser des listes N-best plutôt que le graphe de mots est que ces listes sont composées de phrases basées sur la position des mots. Le problème de chronologie rencontré lors de l'utilisation d'un graphe de mot est donc éliminé. Cependant, même pour des valeurs de N élevées, les performances de la probabilité *a posteriori* prise comme mesure de confiance sont moindres que lors de l'utilisation directe du graphe de mot. Le graphe de mot est plus dense et les informations chronologiques

qu'il contient et qui sont utilisées dans l'algorithme forward-backward (algorithme permettant de calculer la probabilité *a posteriori*) semblent influer sur la mesure de confiance résultante.

Les SRAP fournissent l'hypothèse la plus probable et non pas la liste des mots les plus probables pour chaque instant de la reconnaissance. Pour les approches par graphe de mots et par liste des N meilleures hypothèses, les calculs de la probabilité *a posteriori* se basent donc sur un mot appartenant à l'hypothèse minimisant les erreurs. Pourtant, les métriques d'évaluation de ces systèmes comme le WER s'appuient sur la minimisation des erreurs sur les mots pris séparément. Une technique minimisant explicitement le taux d'erreur sur les mots a été présentée dans [Mangu 2000], technique décrite au paragraphe suivant.

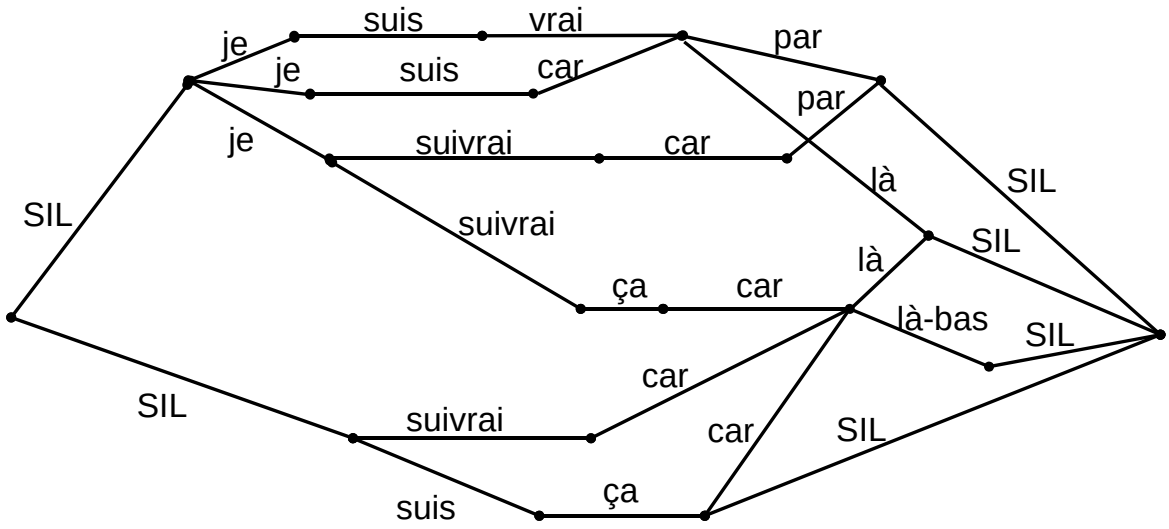
3.3.3 Approximation par réseau de confusion

Cette approche conserve les hypothèses du graphe de mots les plus probables (une technique d'élagage est employée pour enlever les hypothèses ayant une faible probabilité) en un seul alignement appelé réseau de confusion (voir figure 3.3). Cette technique est utilisée dans [Falavigna 2002, Evermann 2000] pour définir la probabilité *a posteriori* sur les mots comme mesure de confiance.

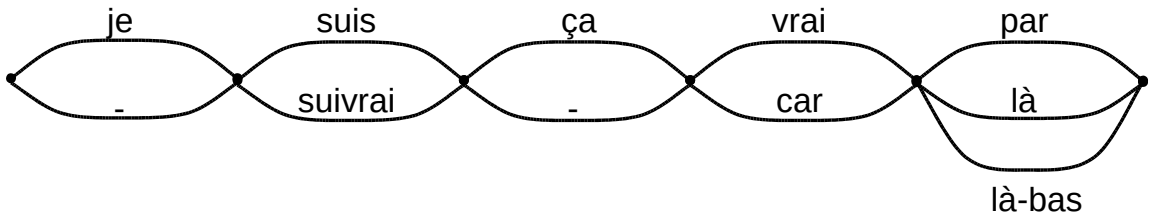
L'algorithme de création du réseau de confusion proposé par [Mangu 2000] comporte trois phases :

1. **Initialisation**
   Calcul de la probabilité *a posteriori* de chaque arc du graphe
   Création de classes d'équivalence où chaque classe est formée d'un même mot ayant des temps de début et fin égaux.
2. **Regroupement intra-mot**
   Regroupement des classes contenant les mêmes mots et qui préservent la consistance du graphe *i.e.* qui préservent les relations temporelles entre les mots
   La probabilité *a posteriori* est sommée sur la classe
3. **Regroupement inter-mot**
   Regroupement des classes par similarité phonétique des mots et préservant la consistance du graphe.

Un algorithme d'élagage est au préalable effectué, permettant d'enlever les arcs du graphes qui ont une probabilité faible et qui auraient compromis l'alignement final du réseau de confusion. Leur présence alourdit le traitement et les probabilités des mots associés sont négligeables dans le calcul de la probabilité *a posteriori* finale. Grâce au réseau de confusion, la probabilité



(a) Graphe de mots en entrée (“SIL” correspond à un silence)



(b) Réseau de confusion (“-” correspond à une suppression)

FIG. 3.3 – Graphe de mots en sortie du SRAP et le réseau de confusion lui correspondant

a posteriori d’un mot est approximée par la somme des probabilités *a posteriori* de toutes les transitions passant par ce mot et qui sont en concurrence avec lui.

3.3.4 Discussion

En comparant les performances d’un rapport de vraisemblance basé sur le calcul d’anti-modèles et des mesures de probabilités *a posteriori* calculées à partir de graphes de mots et d’un réseau de confusion, les auteurs de [Falavigna 2002] montrent que les mesures de probabilité *a posteriori* sont plus efficaces en termes de taux d’égale erreur et de courbe DET (métriques décrites au paragraphe 3.7).

Ces mesures s’appuient sur des modèles de mots isolés. Même en se mettant au niveau phonémique comme dans [Falavigna 2002], ces mesures sont moins robustes que les mesures basées sur la probabilité *a posteriori* des mots. Les deux estimations sont très proches, on cherche

dans les deux cas une modélisation du “monde”. Cependant, les rapports de vraisemblance ne tiennent pas compte de la probabilité *a priori* $P(W)$ et sont trop dépendants de l'ensemble de paramètres choisis pour estimer les HMMs.

D'autres critères peuvent également servir à l'estimation d'une mesure de confiance. Pour répondre aux propriétés énoncées au début du chapitre, les paramètres suivants sont adaptés et classifiés. Les méthodes de classification seront décrites dans la suite du chapitre. Certains de ces critères répondent à ces propriétés sans les classifieurs et sont assimilés directement à des mesures de confiance.

3.4 Critères de décision

3.4.1 Critères acoustiques

Les modèles acoustiques produisent une estimation de la vraisemblance des observations acoustiques par rapport à l'hypothèse de reconnaissance, ce qui peut être considéré comme un paramètre prédictif.

Un paramètre simple proposé dans [Jiang 2005] consiste alors à prendre le score de vraisemblance acoustique du mot et à le normaliser par le nombre de trames du mot. Cette simple normalisation n'offre pas une précision assez forte pour attester du niveau de pertinence d'un mot reconnu.

Un des critères acoustiques utilisé comme mesure de confiance est la stabilité acoustique du mot [Jiang 2005, Wessel 2001]. Les auteurs partent du principe que si on décode plusieurs phrases avec des poids différents des scores des modèles acoustiques et du modèle de langage, un mot se retrouvant toujours à la même place dans la majorité des phrases sera vraisemblablement correct. La stabilité acoustique du mot est alors son nombre d'apparitions dans les hypothèses alternatives sur le nombre total d'hypothèses.

Afin de s'affranchir des contraintes linguistiques et lexicales provenant du modèle de langage et du dictionnaire de phonétisation, les auteurs de [Young 1994, Raymond 2004] se concentrent sur les résultats des modèles acoustiques. Le mot reconnu a un score acoustique qui est la somme des scores acoustiques des phonèmes qui le compose. Ils proposent comme critère de confiance de calculer le ratio entre le score acoustique du mot reconnu et le score des phonèmes obtenus dans le même temps par une boucle de phonèmes non contrainte par un arbre lexical. Cette boucle permet à n'importe quel phonème d'en suivre un autre avec une probabilité égale. Soient X une séquence d'observations acoustiques et w le mot reconnu résultant de ces observations acoustiques. La différence de la vraisemblance logarithmique entre le score acoustique $\log P(X|\lambda_C)$ donné par le modèle contraint λ_C et le score acoustique $\log P(X|\lambda_L)$ donné par

la boucle de phonèmes non-contraints λ_L s'écrit :

$$AC(w) = \frac{1}{N_f(w)}[\log P(X|\lambda_C) - \log P(X|\lambda_L)] \tag{3.7}$$

où N_f est le nombre de trames du mot. Cette normalisation par le nombre de trames est nécessaire afin que la longueur du mot n'influe pas dans la mesure résultante.

Ce paramètre n'est pas compris entre 0 et 1, ce problème sera résolu dans le paragraphe 5.1 du chapitre intitulé "Mesures de confiance proposées" afin d'assimiler ce paramètre à une réelle mesure de confiance.

3.4.2 Critères linguistiques

Le modèle de langage d'un système de reconnaissance peut également contribuer à la constitution d'un paramètre prédictif.

Ainsi, [Uhrik 1997] propose d'utiliser le comportement du repli (voir la section 1.3 traitant du modèle de langage) directement comme mesure de confiance. En fonction du degré de repli du modèle de langage pour les deux mots précédents ou encore pour le mot précédent et le suivant, les auteurs associent un score de confiance arbitraire au mot considéré. Ce score étant attribué arbitrairement, il ne prend pas en compte le véritable taux d'erreur sur le mot. Ce taux d'erreur en fonction du comportement du repli est pris en compte lors du calcul de la mesure linguistique que nous proposons (voir paragraphe 5.2 du chapitre "Mesures de confiance proposées").

Le modèle de langage permet également d'obtenir un score de confiance sur un mot w. Le score fourni $P(w)$ est la probabilité de trouver ce mot en connaissant les $n-1$ mots qui le précèdent (type forward) [Duchateau 2002a]. n étant l'ordre avec lequel a été appris le modèle de langage (voir section 1.3.1). Ce score ne prend en compte que les mots précédents le mot suivant. Les auteurs proposent alors d'utiliser comme source supplémentaire d'informations la probabilité de trouver ce mot connaissant non plus les $n-1$ mots qui le précèdent mais les $n-1$ mots qui le suivent. Un modèle de langage de type backward est estimé sur le même corpus d'apprentissage que le modèle de langage forward mais les mots contenus dans les phrases sont considérés dans l'ordre inverse de la lecture (du dernier au premier mot). Une combinaison de scores (voir section 3.6.3) acoustiques, forward et backward est proposée, mettant en évidence l'apport de la mesure backward sur la pertinence de la mesure de confiance finale. Cette mesure prend en compte tout le contexte d'un mot et non plus uniquement sur les mots précédents. Grâce à la mesure backward, l'impact d'une erreur sur les mots suivants est estimé. Toutefois,

la mesure backward est contraignante car elle demande l'estimation d'un nouveau modèle de langage où l'ordre des mots est inversé.

Dans [Estève 2003], les auteurs proposent une mesure linguistique au niveau de la phrase. Elle est basée sur l'idée que les événements non vus dans le corpus d'apprentissage lors de l'estimation du modèle de langage sont mal modélisés. Cette mauvaise modélisation est la source d'un nombre potentiellement important d'erreurs de reconnaissance. Pour une phrase et un modèle de langage n-gramme donnés, cette mesure est obtenue simplement en calculant le rapport entre le nombre de n-grammes présents dans cette phrase et également présents dans le corpus d'apprentissage du modèle de langage sur le nombre de n-grammes présents dans cette phrase. Cette mesure est très facile à calculer et nécessite un temps de calcul négligeable. Cependant, cette mesure s'applique globalement sur une phrase et perd de la précision lorsqu'elle est utilisée au niveau des mots.

3.4.3 Autres critères de décision

Il existe également d'autres critères de décision permettant de déterminer des mesures de confiance calculées à partir de différents éléments du système de reconnaissance.

Grâce à la liste des N meilleures hypothèses [Gillick 1997, Guo 2004, Hazen 2002], on peut calculer :

- Le nombre de fois où un mot apparaît dans les différentes hypothèses
- La différence entre la meilleure hypothèse et les suivantes
- Le score des premières hypothèses
- ...

La durée du mot, du phonème mais également de l'état du HMM dans lequel on se trouve peuvent apporter une information [Vergyri 2000].

L'utilisation du graphe de mots peut également être une source d'informations. On y récupère par exemple la densité de l'hypothèse, c'est à dire le nombre d'arcs en concurrence aux limites de temps du mot W ou encore la profondeur du graphe après élagage (le nombre total d'hypothèses en concurrence) [Duchateau 2002a].

Dans [Henández-Ábrego 2000], les auteurs utilisent également le débit du locuteur comme mesure de confiance. En effet, ils estiment que si le débit change brutalement, il s'agira sûrement d'une erreur d'insertion ou de suppression.

3.5 Mesures intégrant des notions sémantiques et syntaxiques

Les mesures précédentes relèvent des sorties directes du système de reconnaissance. Nous aborderons ici quelques techniques utilisées afin d'apporter des connaissances de plus haut niveau aux mesures de confiance comme des connaissances sémantiques ou syntaxiques.

3.5.1 Analyse sémantique latente (LSA : Latent Semantic Analysis)

La technique LSA est une technique utilisée en recherche d'information où l'idée est d'associer les mots qui sont en co-occurrence dans les documents textuels qui sont eux-mêmes sémantiquement similaires [Cox 2002]. L'hypothèse formulée par cette technique est que ces mots seront également sémantiquement proches. Grâce à cette technique, une matrice mots/documents pourra être formée pour ensuite être réprésentée dans un sous-espace réduit à l'aide de la décomposition en valeurs singulières (SVD) [Golub 1989]. Les vecteurs des mots qui sont proches dans ce sous-espace correspondent à des mots sémantiquement proches. En projetant les mots du vocabulaire dans ce sous-espace, on peut estimer la similarité sémantique entre chaque paire de mots. La similarité entre deux mots w_i et w_j est alors utilisée pour estimer la vraisemblance de leur apparition simultanée dans une même phrase, en formulant l'hypothèse que les corpora d'apprentissage et de test comportent des données homogènes. Si w_i et w_j sont des mots représentés par les vecteurs ω_i et ω_j dans le sous-espace calculé par SVD, alors cette mesure s'écrit comme le cosinus de l'angle entre ces vecteurs :

$$S(w_i, w_j) = \frac{\omega_i.\omega_j}{|\omega_i||\omega_j|} \tag{3.8}$$

La similarité $S(w_i, w_j)$ est comprise dans l'intervalle $[-1; 1]$ et une valeur positive élevée de $S(w_i, w_j)$ signifie que les mots w_i et w_j sont fortement correllés sémantiquement.

Supposons une phrase composée de N mots $w_1, w_2, ..., w_N$. La mesure de confiance m_{LSA} associée au mot w_i peut s'écrire :

$$CM_{LSA}(w_i) = \frac{1}{N} \sum_{j=1}^{N} S(w_i, w_j) \tag{3.9}$$

Cette mesure devrait être faible pour des mots qui ne sont pas sémantiquement pertinents dans la phrase et élevée pour des mots corrélés avec le reste de la phrase. Dans la pratique, cette mesure ne rend pas compte des mots non pertinent avec les autres mots de la phrase. Pour cela, il faut enlever le bruit engendré par les mots "fonctionnels" de la phrase qui sont corrélés avec tous les autres. Une solution est d'enlever du calcul les mots qui sont toujours en forte

corrélation avec la plupart des mots du vocabulaire. Quand un mot apparaît dans une phrase et que sa mesure moyenne de similarité avec les autres mots $\bar{L}_i$ dépasse un seuil, on ne le compte pas dans le calcul de la mesure de confiance des autres mots de la phrase. D'autres mesures basée sur l'information de similarité entre mots sont décrites dans [Cox 2002].

3.5.2 Information mutuelle inter-mots

Une mesure similaire est utilisée dans [Guo 2004]. Les auteurs proposent d'utiliser la notion de l'information mutuelle inter-mots pour donner une mesure de confiance au mot. À partir des diverses transcriptions des données d'apprentissage, l'information mutuelle donnée par chaque paire de mots du vocabulaire est calculée. Pour [Guo 2004], les données d'apprentissage sont composées de 50000 phrases provenant de données similaires à celles du corpus d'apprentissage des modèles de leur SRAP. Soient w_i et w_j deux mots et $N(w_i, w_j)$ leur nombre de co-occurence dans l'ensemble des données d'apprentissage. La probabilité d'avoir les mots w_i et w_j dans le même document s'écrit :

$$P(w_i, w_j) = \frac{N(w_i, w_j)}{\sum_{w', w''} N(w', w'')}$$

où le dénominateur est le nombre de co-occurrences des mots w_i et w_j dans le corpus d'apprentissage. Les probabilités marginales des co-occurrences des mots w_i et w_j dans un document sont :

$$P(w_i) = \sum_{w_j} P(w_i, w_j) \text{ et } P(w_j) = \sum_{w_i} P(w_i, w_j)$$

L'information mutuelle entre toute paire de mots w_i et w_j s'écrit alors :

$$MI = \log(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}) \tag{3.10}$$

La mesure de confiance associée au mot reconnu W est la moyenne de l'information mutuelle entre ce mot et tous les autres mots du même document (ou phrase).

$$CM_{MI}(W) = \frac{\sum_{y} \log(\frac{P(W, y)}{P(W)P(y)})}{\text{Nombre de mots} \neq \text{dans la phrase}} \tag{3.11}$$

Les auteurs de [Guo 2004] utilisent une technique de lissage afin de pouvoir calculer l'information mutuelle entre des paires de mots pour lesquelles ce calcul n'est pas possible. Cette technique permet de résoudre le problème de l'éparpillement des données d'apprentissage. $N(w_i, w_j)$ peut être estimée d'une autre façon dans laquelle on prend en compte une constante

C qui est la fréquence du mot dans le corpus :

$$N'(w_i, w_j) = N(w_i, w_j) + C$$

Ensuite, la probabilité $P(w_i, w_j)$ devient :

$$P_{lissage}(w_i, w_j) = \frac{P(w_i, w_j) + \alpha.P(w_i).P(w_j)}{1 + \alpha} \tag{3.12}$$

où α est un autre paramètre permettant avec C d'obtenir les meilleurs résultats sur un corpus de développement.

3.5.3 Discussion

Les mesures LSA et d'information mutuelle sont intéressantes car elles ne prennent pas en compte les sorties directes du système de RAP. Cependant les scores de confiance dérivés de ces mesures ne permettent pas de discriminer les mots incorrects des mots corrects, les deux histogrammes des classes ne permettant pas cette séparation en fonction de la valeur du score des mots. Ces mesures doivent être combinées avec des mesures dérivant des sorties du SRAP. Ainsi, la mesure LSA dans [Cox 2002] est combinée avec une mesure provenant de la liste des N-meilleures hypothèses, permettant d'obtenir une plus grande précision avec un plus faible rappel (voir paragraphe 3.7.2 pour les définitions de précision et rappel) que la mesure provenant de la liste des N-meilleures hypothèses prise seule. La mesure combinée est donc utile dans des applications où on cherche à obtenir une faible quantité de mots corrects. Pour [Guo 2004], CM_{MI} est combinée avec la probabilité *a posteriori* d'un mot à partir d'un graphe de mots comme dans [Wessel 2001] pour améliorer leur taux d'égale erreur de 2 % en relatif (les métriques d'évaluation sont décrites dans le paragraphe 3.7).

3.6 Combinaison de plusieurs mesures de confiance

Supposons que pour un mot w, nous disposions de K mesures de confiance $CM_j(w), j = 1, \ldots K$. Combiner ces K mesures de confiance peut permettre d'allier les qualités de chacune d'entre elles et de meilleures performances seraient atteintes. Cette combinaison de plusieurs paramètres pour obtenir une mesure unique peut être faite de plusieurs façons. Elle peut être obtenue par des opérateurs mathématiques ou encore à l'aide de classificateurs.

3.6.1 Sélection des mesures

Une fois les paramètres prédictifs et les mesures de confiance déterminés, il s'agit de calculer une mesure de confiance unique grâce à un classificateur. En effet, pour faire une bonne mesure de confiance, ces paramètres doivent permettre de séparer efficacement les hypothèses correctes des hypothèses incorrectes. Cependant, aucun de ceux décrits dans la section 3.4 n'est dédié à cette tâche c'est pourquoi il est nécessaire de recourir à un classificateur tel qu'un arbre de décision, un réseau de neurones ou encore une simple interpolation linéaire. De plus, la combinaison des paramètres et des mesures de confiance entre eux semblent être un moyen efficace d'accroître leur performance en termes de mesures de confiance.

Pour sélectionner les paramètres pertinents à combiner, [Schaaf 1997] détermine une matrice de corrélation entre ces paramètres. Les paramètres ayant un pouvoir discriminant *Correct/Incorrect* supérieur à un seuil par rapport à leur coefficient de corrélation sont conservés. Les auteurs de [Siu 1997] proposent quant à eux un algorithme incrémental de sélection. Pour chaque paramètre, les auteurs évaluent sa capacité à améliorer la mesure finale sur un corpus de développement quand il est ajouté aux paramètres déjà choisis. S'il améliore le tout, il est conservé dans la combinaison finale.

3.6.2 Opérations mathématiques

Les opérateurs les plus utilisés dans cette combinaison sont : minimum, maximum, moyenne arithmétique, moyenne géométrique, produit ou encore la moyenne quadratique. Ceci s'exprime par une formule de type :

$$CM(w) = \mathcal{O}\left(CM_1(w), \ldots, CM_K(w)\right)$$

où $\mathcal{O}$ est l'opération choisie.

Comme nous l'avons vu précédemment au paragraphe 3.1, la règle de combinaison choisie doit respecter certaines contraintes. En particulier, la moyenne de la mesure de confiance finale doit être une approximation du taux d'erreur global sur les mots émis. Les opérateurs de minimum, de maximum et de produit ne respectent pas cette contrainte mais peuvent en pratique être utilisés si le biais résultant n'est pas trop élevé.

3.6.3 Classificateurs

Les classificateurs permettent de regrouper les différents critères prédictifs ou mesures de confiance afin d'obtenir une mesure de confiance comprise entre zéro et un (voir les propriétés d'une mesure de confiance au paragraphe 3.1). Il s'agit de définir une méthode de combinaison

de ces paramètres pour obtenir une mesure de confiance sur le mot W. Ces classificateurs peuvent aussi permettre de combiner des critères prédictifs avec des mesures de confiance, comme les mesures intégrant des notions sématiques et syntaxiques.

Interpolation linéaire

Pour prendre en compte les qualités de chacune des mesures, une simple interpolation linéaire peut être utilisée :

$$CM(w) = \sum_{i=1}^{K} q_k CM_k(w), \text{ avec } \sum_{k=1}^{K} q_k = 1. \tag{3.13}$$

Les poids q_k peuvent par exemple être appris de manière à minimiser le taux d'erreur empiriquement sur un corpus de développement [Guo 2004]. Les auteurs de [Guo 2004] combinent une mesure de confiance de haut niveau (CM_{MI}) avec une mesure de confiance plus couramment utilisée, la probabilité *a posteriori*. Ces deux mesures ne sont pas dépendantes l'une de l'autre et les combiner devrait permettre d'allier leurs qualités. La combinaison des deux mesures par simple interpolation linéaire entraîne une amélioration des résultats de 10% en relatif sur le taux d'égale erreur (pour les métriques d'évaluation, voir paragraphe 3.7) par rapport aux deux mesures prises séparément.

Modèles linéaires

Utilisés dans [Duchateau 2002b, Duchateau 2002a, Gillick 1997]et [Siu 1997, Siu 1999], les modèles linéaires généralisés (Generalized Linear Model GLM) permettent également de combiner des paramètres prédictifs ($y_1 ... y_n$) par transformation linéaire afin d'obtenir une mesure de confiance. La relation entre la mesure de confiance CM(W) et Y, le vecteur des paramètres s'écrit :

$$CM(W) = f^{-1}B.Y = \sum_{i=1}^{n} \beta_i y_i \tag{3.14}$$

où f est une fonction monotone sur $]0, 1[$ et $B = (\beta_1, ...\beta_n)$ est un vecteur. Dans [Duchateau 2002b, Duchateau 2002a, Gillick 1997, Siu 1997, Siu 1999], la fonction f est la fonction *Logit* :

$$f(CM(W)) = Logit(CM(W)) = \log \frac{CM(W)}{1 - CM(W)} \tag{3.15}$$

La matrice B est estimée pour maximiser la probabilité de classification correcte moyenne sur un corpus de développement.

Les auteurs de [Siu 1997, Siu 1999] utilisent également une variante du modèle *Logit* appelé le modèle additif généralisé (Generalized Additive Model GAM). La différence avec le modèle précédent réside dans le fait que les paramètres sélectionnés peuvent subir des transformations au préalable, indépendament les uns des autres. L'équation 3.15 devient :

$$f(CM(W))) = \log \frac{CM(W)}{1 - CM(W)} = \sum_{i=1}^{n} f_i(y_i) \tag{3.16}$$

Le fait de transformer non linéairement certains paramètres permet de rendre redondants ceux qui sont corrélés et ainsi espérer réduire le nombre de paramètres nécessaires. L'utilisation de telles transformations nécessite l'estimation de plus de paramètres que le modèle *Logit* conventionnel et ajoute à la complexité du système.

Dans l'approche de [Schaaf 1997, Kamppari 2000] et [Sun 2003], la matrice B est estimée au moyen d'une analyse discriminante linéaire (Linear Discriminant Analysis LDA). L'analyse discriminante linéaire part de la connaissance de la partition en classes *Correct/Incorrect* des hypothèses et cherche les combinaisons linéaires des variables décrivant ces hypothèses qui conduisent à la meilleure discrimination entre les classes. Les auteurs de [Schaaf 1997] donnent plus de détails sur l'approche LDA.

Support Vector Machine SVM

L'idée d'une classification par SVM [Burges 1998] est de calculer l'hyperplan permettant de séparer au mieux deux nuages de points (représentant nos deux classes *Correct/Incorrect*). Il s'agit de reconsidérer le problème dans un espace de dimension supérieure. Dans ce nouvel espace, il existe un séparateur linéaire permettant de classer au mieux les deux nuages de points. Cette technique de combinaison de paramètres améliore les résultats trouvés avec un arbre de décision ou un réseau de neurones, bien qu'elle ne soit pas très robuste aux changements de paramètres choisis pour représenter les données de départs [Zhang 2001].

Arbre de décision

Les arbres de décision [Cornuéjols 2002, Breiman 1984] sont souvent utilisés en reconnaissance de la parole afin d'effectuer une séparation binaire *Correct/Incorrect* des hypothèses de reconnaissance [Fu 2005, Kemp 1997, Zhang 2001, Stemmer 2002]. L'arbre est appris à partir d'un corpus de mots étiquetés en Correct/Incorrect. Une mesure de confiance est attribuée à chacune des feuille de l'arbre qui doit permettre une classification efficace quand un nouvel exemple est soumis. Pour apprendre un arbre de décision, il s'agit de déterminer un critère de test sur les paramètres prédictifs représentant le degré de confiance d'un mot. À chaque

noeud, l'algorithme de construction de l'arbre détermine grâce à ce critère quel est le meilleur paramètre à tester pour progresser dans l'arbre et atteindre la meilleure répartition possible entre les classes *Correct* et *Incorrect*.

Une fois l'apprentissage terminé, on effectue une extraction d'un vecteur de paramètres prédictifs Y pour calculer la mesure de confiance d'un mot. L'arbre de décision s'appuie sur les valeurs de ces paramètres pour progresser à travers chacun de ses noeuds au moyen d'un test sur ces paramètres. Chaque noeud correspond à une décision binaire dont dépend la suite du parcours dans l'arbre. À chaque feuille de l'arbre correspond une mesure de confiance $CM(W)$. La figure 3.4 donne un exemple de combinaison de paramètres grâce à un arbre de décision afin d'obtenir une mesure de confiance.

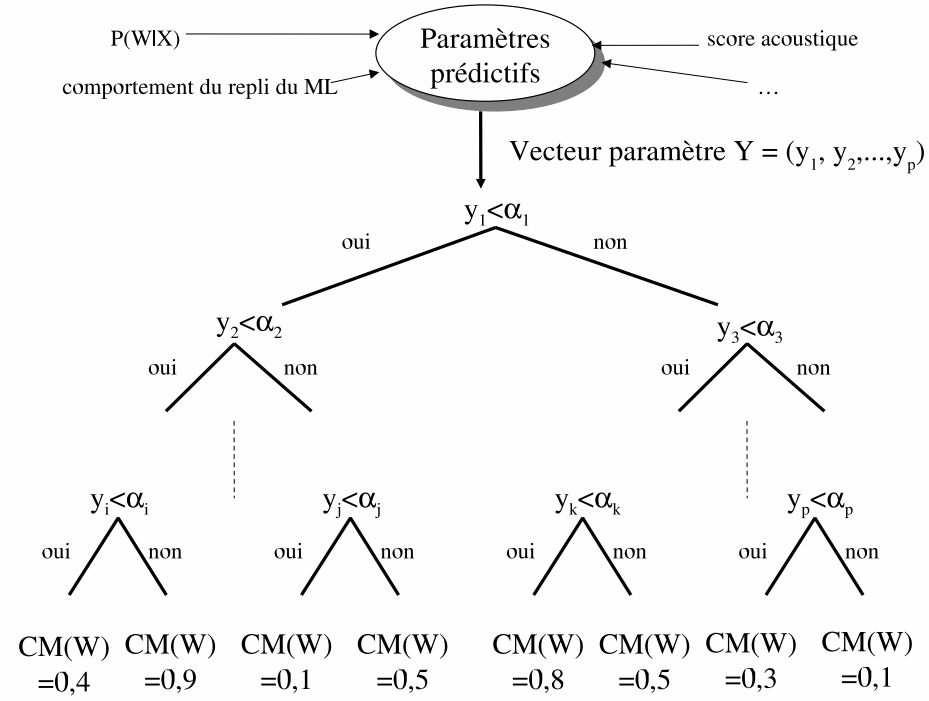


FIG. 3.4 – Exemple d'arbre de décision permettant la combinaison de paramètres en une mesure de confiance

Dans [Zhang 2001], les auteurs montrent que la métrique d'évaluation des performances d'un arbre doit être corrélée avec le critère choisi pour le test. En effet, si un arbre est appris avec un test portant sur un critère de minimisation du taux d'erreur sur les mots, il aura de meilleures performances en termes de taux d'erreur sur les mots qu'un arbre construit à partir d'un critère basé sur l'entropie. Il s'agit donc d'évaluer les performances d'un classificateur avec la métrique utilisée pour le construire.

À la différence des autres classificateurs, les arbres permettent de ne pas avoir une seule mesure finale. Ils permettent de visualiser les différentes interactions entre paramètres ainsi que l'importance de chacun d'entre eux.

Boosting

Les méthodes de boosting constituent une famille d'algorithmes d'apprentissage automatique qui construisent des modèles de classification en combinant des règles de classification. Les algorithmes de bases ont été introduits dans [Freund 1995]. [Moreno 2001] utilise une technique de boosting pour combiner leurs paramètres prédictifs. Cette technique applique à un vecteur de paramètres pondérés une procédure de classification itérative.

Le classificateur final est formé d'une somme pondérée de chaque classificateur intermédiaire appris à chaque étape de l'algorithme. Les auteurs de [Freund 1995] montre que l'erreur de classification à l'apprentissage engendrée par le classificateur final est proche de zéro de manière exponentielle par rapport au nombre d'itérations de l'algorithme. [Moreno 2001] a également testé des classificateurs tels que les SVM ou les arbres de décision, constatant que la technique du boosting améliorait les résultats en classification *Correct/Incorrect*.

Réseaux de neurones

Un autre classificateur utilisé pour la combinaison de paramètres est le réseau de neurones [Kemp 1997, Schaaf 1997, San-Segundo 2001, Charlet 2001, Stemmer 2002]. Le réseau prend en entrée un certain nombre de paramètres réunis en un vecteur Y. En sortie du réseau de neurones, on obtient la mesure de confiance du mot hypothèse $CM(W)$ comme le montre la figure 3.5.

Dans [Kemp 1997, Schaaf 1997, Weintraub 1997], l'apprentissage du réseau de neurone se fait par rétropropagation du gradient qui minimise l'erreur de classification des mots en *Correct/Incorrect* sur un corpus d'apprentissage. En comparaison avec les méthodes de combinaison par interpolation linéaire et arbres de décision [Kemp 1997, Schaaf 1997, Zhang 2001, Stemmer 2002], les réseaux de neurones sont les classificateurs qui donnent la meilleure combinaison de paramètres.

3.6.4 Théorie des probabilités

La fusion de données a depuis peu suscité un intérêt certain dans la communauté scientifique [Dubois 1992, Dubois 1994, Janez 1996]. Elle consiste à mettre à profit le maximum d'informations sur les données afin de réduire les faiblesses de certaines mesures de confiance à l'aide des autres. Ainsi, il faut que les techniques de fusion permettent de gérer tous les

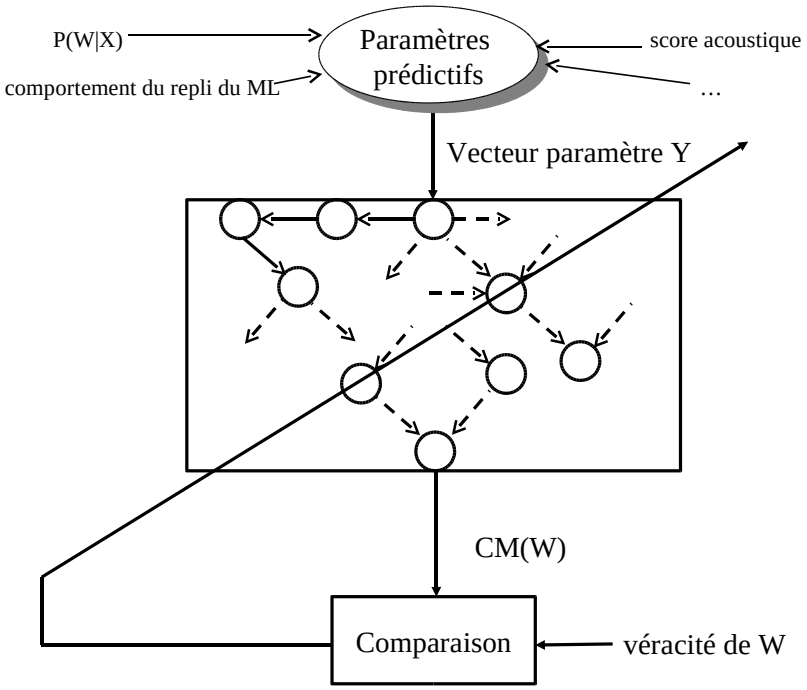


FIG. 3.5 – Exemple de réseau de neurones permettant la combinaison de paramètres en une mesure de confiance

cas possibles comme l'incertitude, la redondance, les conflits ou les incohérences entre les différentes mesures de confiance. Un des cadres de la fusion de données est la théorie des probabilités.

Les méthodes les plus utilisées pour la fusion de données ont tout d'abord été envisagées sous l'angle probabiliste. Les informations qui doivent être fusionnées sont délivrées par des experts et sont des vraisemblances issues de la modélisation des observations par des distributions de probabilités conditionnelles.

L'un des inconvénients majeurs de cette technique réside dans l'exigence de la connaissance parfaite des probabilités, et plus particulièrement de la probabilité a priori. Enfin, la notion d'ignorance sur un fait n'est pas prise en compte. L'ignorance se traduit par une égalité des probabilités a priori. Cela peut engendrer des incohérences selon la modélisation des hypothèses utilisées.

Indices de confiance

Pour pallier ce problème, on peut s'appuyer sur des indices de confiance [Leray 2000]. Pour fusionner des données, trois informations sont ici disponibles : l'expert, la classe et l'observation. L'expert nous apporte sa connaissance avec un taux de confiance α dans son propos. L'indice de confiance de classe β est en quelque sorte l'expérience que l'on a du modèle expert. Par exemple : sait-il bien évaluer la pertinence d'un mot, et avec quelle probabilité ? L'observation est ici la mesure ce confiance $CM(w_i)$ sur le mot w_i.

Estimation de l'indice d'expert α

Pour le calcul de l'indice d'expert α on peut appliquer l'une des techniques suivantes :

- La moyenne arithmétique des β.
- Le complément à 1 de la probabilité d'erreur : si un taux d'erreur pour l'expérience courante est disponible, la croyance de l'expert en est le complémentaire [Rahman 1999].
- En se basant sur le calcul de la variance : l'indice de confiance d'expert repose sur la variance normalisée de la classe la plus probable [Leray 2000].

Estimation de l'indice de classe β

β peut s'obtenir d'après la matrice de confusion *Correct|Incorrect*(voir tableau 3.1). En effet, β peut être vu comme le complément à 1 du coût fait en choisissant une classe erronée. Cela correspond aux valeurs normalisées des diagonales de la matrice. Par la suite, on notera C comme étant la classe correcte et NC la classe incorrecte.

	Classe attendue (%)	
Classe obtenue (%)	Correct	Incorrect
Correct	P(C|C)	P(NC|C)
Incorrect	P(C|NC)	P(NC|NC)

TAB. 3.1 – Estimation de β grâce à une matrice de confusion Classe attendue/ Classe obtenue

Ici, on a :

$$\beta_C = \frac{P(C|C)}{P(C|C) + P(NC|C)} \text{ et } \beta_{NC} = \frac{NC|NC}{P(C|NC) + P(NC|NC)} \tag{3.17}$$

Pour fusionner k mesures de confiance données par k experts, on définit alors un seuil au dessus duquel l'hypothèse w_i associée à la mesure de confiance $CM_j(w_i)$ donnée par l'expert j (j allant de 1 à k) sera considérée comme correcte et en-dessous duquel elle sera considérée comme incorrecte. On aura alors la formule pour chaque mot w :

$$\text{Coût}_j(w) = \min\{\{CM_j(w) * (1 - \beta_{j_{correct}})\}, \{(1 - CM_j(w)) * (1 - \beta_{j_{incorrect}})\}\} \tag{3.18}$$

Le score de confiance final sera attribué par l'expert qui maximise :

$$\alpha_j * (1 - \text{Coût}_j(w)) \tag{3.19}$$

D'autres théories telles que la théorie de l'évidence et la théorie des possibilités servent à fusionner des données. Ces théories ne sont pas utilisées dans ce manuscrit mais sont utilisées dans des travaux de fusion d'informations [Janez 1996].

3.7 Évaluation des mesures de confiance

Pour évaluer les mesures de confiance, il existe de nombreuses métriques [Siu 1999]. Nous allons détailler successivement les métriques basées sur la courbe DET, les métriques de précision et de rappel, le Confidence Error Rate et l'entropie croisée normalisée.

3.7.1 Detection Error Tradeoff (DET)

La courbe DET permet d'évaluer la capacité d'une mesure de confiance pour l'acceptation/rejet d'une hypothèse. Pour chaque seuil α on a :

$$\text{décision} = \begin{cases} \text{hypothèse acceptée} & \text{si } CM(\text{hypothèse}) \geq \alpha \\ \text{hypothèse rejetée} & \text{sinon} \end{cases} \tag{3.20}$$

Le test de l'équation 3.20 conduit à deux sortes d'erreurs :

- Les erreurs de rejet à tort, où l'hypothèse est considérée comme incorrecte alors qu'elle est correcte
- Les erreurs de fausse acceptation, où l'hypothèse est considérée comme correcte alors qu'elle est incorrecte

Sur l'ensemble d'un corpus d'évaluation, deux taux peuvent ainsi être calculés :

- Le taux de faux rejet :

$$FR = \frac{\text{Nombre d'hypothèses rejetées à tort}}{\text{Nombre d'hypothèses totales}}$$

- Le taux de fausse acceptations :

$$FA = \frac{\text{Nombre d' hypothèses acceptées à tort}}{\text{Nombre d'hypothèses totales}}$$

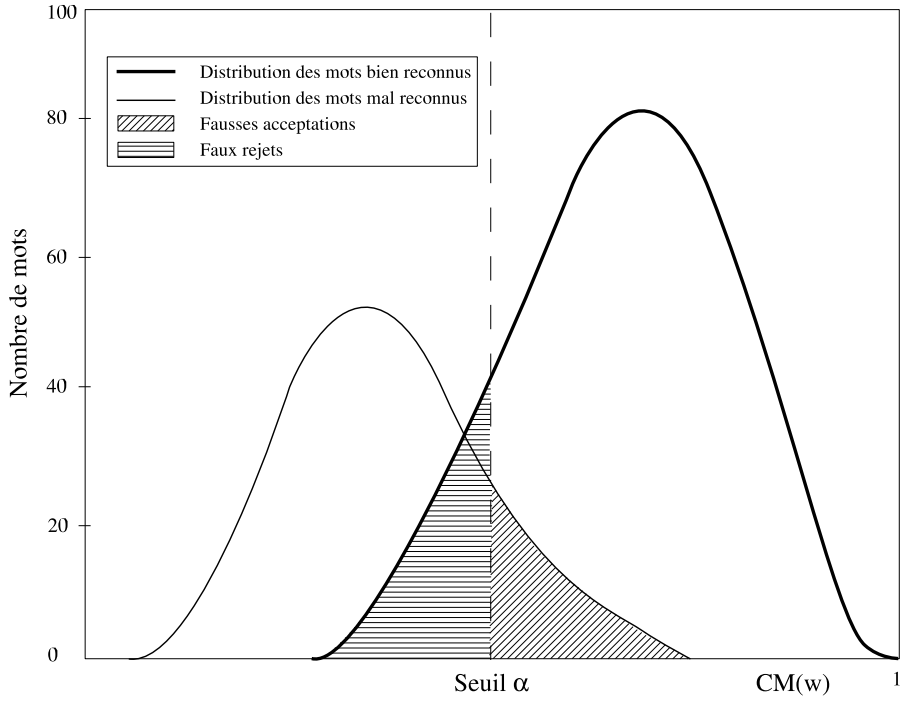


FIG. 3.6 – Distributions des mesures de confiances sur les hypothèses de reconnaissances

Les distributions de la figure 3.6 calculées sur un ensemble d'hypothèses correctes et un ensemble d'hypothèses incorrectes permettent de visualiser les valeurs des taux d'erreur de faux rejet (FR) et de fausse acceptation (FA), aires qui sont définies par le seuil α. Lorsque l'on déplace le seuil vers le haut, le taux de fausses acceptations diminue tout en augmentant le taux de faux rejets. Lorsque l'on abaisse le seuil, c'est le taux de faux rejets qui diminue, entraînant une augmentation du taux de fausses acceptations.

A chaque seuil α, on peut calculer un couple de valeurs (FR,FA) qui détermine un point de fonctionnement particulier du système.

Utilisée dans [Rose 2001, Benayed 2003, Duchateau 2002a] et [Henández-Ábrego 2000, Sun 2003], la courbe DET (voir figure 3.7) permet de visualiser ces différents points de fonctionnement. Cette courbe décrit le taux de faux rejet en fonction du taux de fausse acceptation en faisant varier le seuil d'acceptation/rejet. Elle peut également être appelée courbe ROC (Receiver Operating Characteristic) On peut ainsi voir les performances du système à différents points de fonctionnement et plus particulièrement au point où les deux taux d'erreurs sont égaux (EER : Equal Error Rate). Le taux d'égale erreur est souvent utilisé comme point de comparaison entre différents systèmes [Siu 1999].

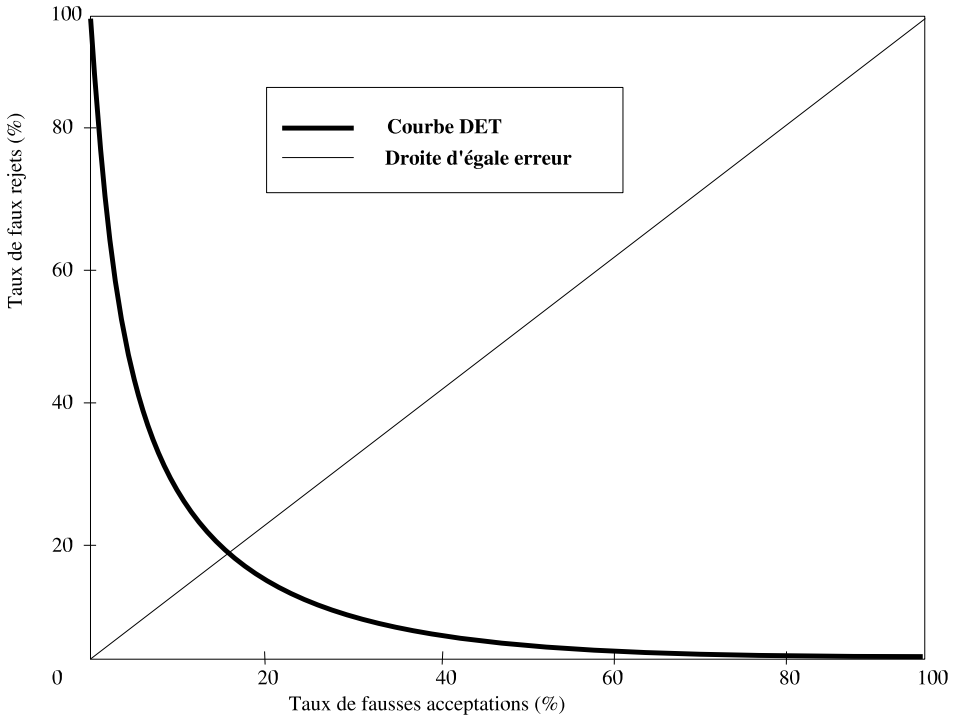


FIG. 3.7 – Exemple de courbe DET

Cette courbe apporte donc une mesure qualitative du pouvoir discriminant des mesures de confiance en mesurant l'écart entre les deux types d'erreurs de classification possible. Plus la courbe est près de l'origine, plus la mesure est discriminante.

3.7.2 Précision/Rappel

Les métriques de précision et de rappel sont empruntées à la RI (Recherche d'Informations) pour évaluer les performances d'un système [Cox 2002].

Le rappel mesure la capacité du système à sélectionner les hypothèses pertinentes.

$$\text{Rappel}(acceptation/rejet) = \frac{\text{Nombre de mots correctement acceptés/rejetés}}{\text{Nombre total de mots réellement corrects/incorrects}} \quad (3.21)$$

La précision mesure la capacité du système à rejeter les hypothèses non pertinentes.

$$\text{Précision}(acceptation/rejet) = \frac{\text{Nombre de mots correctement acceptés/rejetés}}{\text{Nombre total de mots acceptés/rejetés}} \quad (3.22)$$

La combinaison de ces métriques par moyenne harmonique s'appelle la f-mesure :

$$F = \frac{2 * \text{Précision} * \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{3.23}$$

Les métriques précision, rappel, FA et FR sont intéressantes car elles permettent en fonction de l'application visée de choisir un point de fonctionnement. Toutefois, il peut être utile de s'appuyer sur une mesure de confiance unique afin d'évaluer les performances d'une mesure de confiance et pouvoir en comparer plusieurs. La f-mesure et le taux EER tentent de répondre à ce problème. Cependant, la f-mesure (ainsi que les métriques de précision et de rappel) ne prend pas en compte la pertinence des mots et fonctionne sur un mode binaire : le mot est soit correct, soit incorrect. Le taux d'égale erreur dépend fortement de la probabilité *a priori* des classes Correct/Incorrect. Les métriques suivantes d'entropie croisée normalisée et de CER proposent une alternative à ces différents critères.

3.7.3 Confidence Accuracy(CA) et Confidence Error Rate(CER)

Dans [Kemp 1997, Schaaf 1997], les auteurs utilisent la métrique suivante :

$$CA = \frac{\text{Nombre d'étiquettes correctement assignées}}{\text{Nombre total d'étiquettes}} \tag{3.24}$$

La mesure complémentaire utilisée dans [Wessel 1998] et [Bouwman 2000]est :

$$CER = \frac{\text{Nombre d'étiquettes incorrectement assignées}}{\text{Nombre total d'étiquettes}} \tag{3.25}$$

La décision d'étiqueter un mot comme étant correct ou de l'étiqueter incorrect dépend d'un seuil qui est optimisé sur un corpus de développement. Il doit être minimal pour CER et maximal pour CA. Avec ce seuil, la mesure de confiance peut être évaluée sur un corpus de test.

Toutes les métriques proposées auparavant offrent uniquement un choix binaire pour classer les mots. Or les scores de confiance proposés ne sont pas soit 0 soit 1. Une métrique intéressante pour prendre en compte la continuité des scores de confiance est l'entropie croisée normalisée.

3.7.4 Entropie Croisée Normalisée

Pour comparer et déterminer l'efficacité de mesures de confiance, l'entropie croisée normalisée (NCE) [Siu 1999] est une des métriques les plus utilisées [Maison 2001, Duchateau 2002a, Evermann 2000]. Cette métrique est utilisée lors des évaluations NIST pour évaluer la qualité

des mesures de confiance. La NCE est en fait une estimation de l'information additionnelle que la mesure de confiance apporte à l'hypothèse de reconnaissance.

$$NCE = \frac{H_{max} + \sum_{Wcorrects} \log_2(m(W)) + \sum_{Wincorrects} \log_2(1 - m(W))}{H_{max}} \tag{3.26}$$

où $H_{max} = -n \log_2(p_C) - (N - n) \log_2(1 - p_C)$,

$n =$ le nombre de mots reconnus qui sont corrects,

$N =$ le nombre total de mots reconnus,

$p_C =$ la probabilité moyenne qu'on mot reconnu soit correct$(= n/N)$, et

$m(W) =$ la mesure de confiance associée au mot W

Plus la NCE se rapproche de 1 et plus la mesure de confiance peut prédire si le mot est correct ou non. L'entropie croisée normalisée est donc une métrique mathématiquement formulée et dédiée à la tâche d'évaluation des mesures de confiance.

3.8 Conclusion

Pour le choix d'une mesure de confiance efficace, nous avons vu qu'il s'agit de sélectionner attentivement les critères à utiliser. Pour les mesures dérivant des rapports de vraisemblances ou des probabilités *a posteriori*, la distance entre les scores de l'hypothèse choisie et les hypothèses compétitrices est associée à la fiabilité de ces mesures. Si l'hypothèse choisie est proche des autres, le score de confiance sera proche de 0 (hypothèse peu fiable) et si la différence est grande, le score sera proche de 1 (hypothèse fiable). La bonne modélisation des hypothèses compétitrice et la prise en compte des techniques d'élagage du graphe de mot sont donc indispensables. Afin de limiter les influences de la modélisation ou de l'élagage, la combinaison d'une mesure de confiance basée sur la probabilité *a posteriori* d'un mot avec une autre mesure ne dépendant pas de ces facteurs semble incontournable. La mesure résultante sera alors plus efficace pour détecter les zones erronées de l'hypothèse.

Dans la partie suivante, différentes applications utilisant ce système seront décrites. Ces applications sont un terrain d'expérimentation pour évaluer l'apport de mesures de confiance dans le processus de traitement de la parole. Un premier chapitre abordera les mesures de confiance proposées pour cette évaluation. Les deux chapitres suivants détailleront les expérimentations menées dans le cadre de la transcription automatique et de l'identification du locuteur.

Deuxième partie

Mesures de confiance et applications

Chapitre 5

Mesures de confiance proposées

Sommaire

Le système de RAP du LIUM décrit précédemment est la base de plusieurs applications en traitement de la parole comme celles décrites au chapitre 2. Pour que le système puisse évaluer la qualité des hypothèses de reconnaissance qu'il fournit dans ces applications, nous avons étudié, élaboré ou amélioré trois mesures de confiance. Les mesures de confiance présentées dans ce chapitre relèvent de différentes parties du système de reconnaissance. En effet, comme nous l'avons vu précédemment, le système lui-même détient diverses informations sur la pertinence d'un phonème, d'un mot ou plus généralement, d'une hypothèse. Ici, toutes les mesures de confiance sont calculées au niveau du mot. L'idée est de combiner des mesures de confiance provenant de deux parties différentes du SRAP, les modèles acoustiques et le modèle de langage. Ces deux mesures sont ensuite comparées à une mesure souvent utilisées dans la littérature, la probabilité *a posteriori* d'un mot.

Les modèles acoustiques fournissent un score de vraisemblance pour un mot contenu dans le dictionnaire de phonétisation et retenu lors du processus de décodage. Ce score acoustique peut être assimilé à une mesure de confiance sur le mot donnant une indication sur sa pertinence. Cette mesure acoustique est détaillée dans le paragraphe 5.1.

Le modèle de langage donne une indication sur la pertinence de la séquence de mots que le système vient de décoder. Une mesure basée sur le comportement du modèle de langage sera expliquée dans le paragraphe 5.2.

Enfin, grâce au graphe de mots fourni par le système, une mesure de confiance peut être estimée avec le calcul de la probabilité *a posteriori* d'un mot.

5.1 Mesure acoustique (AC)

Cette mesure de confiance est calculée à partir des scores de vraisemblance du modèle acoustique contraint par l'arbre lexical et par le modèle de langage utilisés lors du processus de reconnaissance ainsi que des scores du modèle acoustique utilisé sans contraintes (boucle de phonèmes) [Young 1994, Raymond 2004] (voir section 3.4.1).

5.1.1 Normalisation

Le critère de décision acoustique $AC(w)$ donnée dans l'équation 3.7 du paragraphe traitant des critères acoustiques (section 3.4.1) ne respecte pas la propriété d'appartenance à l'intervalle $[0, 1]$ décrite en début de section 3.1 et propre aux mesures de confiance. Afin que ce critère respecte cette condition et soit considérée comme une mesure de confiance, nous proposons une normalisation qui permettra de rester dans l'intervalle $[0, 1]$ grâce à une transformation de type sigmoïdal [Cornuéjols 2002]. Cette transformation est présentée dans la formule suivante :

$$m_{ac}(w) = \frac{\exp\left(\frac{AC(w)-\mu}{\sigma}\right) + a}{\exp\left(\frac{AC(w)-\mu}{\sigma}\right) + 1} \tag{5.1}$$

où μ and σ sont respectivement la moyenne et l'écart-type des mesures acoustiques initiales sur les mots d'un corpus de développement. Pour s'approcher de la deuxième propriété de la section 3.1 aui indique que la moyenne d'une mesure de confiance idéale doit approcher le taux de mots émis bien reconnus $t_{correct}$, nous nous servons du même corpus de développement et obtenons pour a :

$$a = 2 * t_{correct} - 1$$

Si la distribution initiale des mesures est symétrique par rapport à μ, cette simple transformation suffit à obtenir la propriété de l'appartenance à l'intervalle $[0, 1]$.

5.2 Mesure de confiance basée sur le comportement du repli du Modèle de Langage (LMBB)

La mesure de confiance que nous proposons ici est basée sur le comportement du mécanisme de repli d'un modèle de langage n-gramme, qui a déjà été utilisé différemment dans [Uhrik 1997]. Nous appelons $LMBB$ (Language Model Back-off Behavior) cette méthode.
Pour un mot donné, il s'agit de prendre en compte l'ordre du n-gramme le plus élevé associé à ce mot. Par exemple, si la séquence de mots "... il est temps de ..." est reconnue en utilisant un modèle de langage quadrigramme et que le quadrigramme [il est temps de] a été conservé lors de l'apprentisage du modèle de langage, alors le mot 'de' sera associé à l'ordre 4. Par contre, si ce quadrigramme n'a pas été conservé, mais que le trigramme [est temps de] l'a été, alors le mot 'de' sera associé à l'ordre 3. De même, ce mot pourrait être associé à l'ordre 2 ou à l'ordre 1 le cas échéant, et même à l'ordre 0 dans le cas peu courant où les mots hors-vocabulaire peuvent être traités.
Un phénomène bien connu en reconnaissance de la parole est la propagation des erreurs : lorsqu'un mot est mal reconnu, les mots qui l'entourent sont souvent affectés par des erreurs. Dès lors, il semble très intéressant d'intégrer dans la mesure de confiance linguistique d'un mot des informations concernant son voisinage. En supposant que le comportement du mécanisme de repli d'un modèle de langage est un bon indicateur de la fiabilité de ce modèle, nous proposons de prendre également en compte l'ordre associé aux deux mots voisins du mot visé (le voisin de gauche et celui de droite). Chaque mot reconnu est alors associé à trois valeurs d'ordre de n-gramme.

Afin de ne pas distinguer un nombre de triplet trop important qui seraient difficiles à bien modéliser sans grande quantité de données d'apprentissage, nous ne prendrons pas les valeurs réelles des ordres associés aux mots voisins mais leur position relative par rapport à l'ordre associé au mot visé : plus grand (+), plus petit (-) ou égal (=). Ceci permet de réduire le nombre de classes possibles en passant d'un nombre de triplet égal à $(n+1)^3$ à un nombre de classes égal à $9(n+1)$. Les classes estimées sont des classes de comportement qui regroupe des mots ayant un niveau de connaissance gauche et droite similaire.

Pour illustrer ce propos, prenons la séquence de mots "... il est temps de lire ce livre..." et supposons :
- que le quadrigramme [il est temps de] et le trigramme [est temps de] n'existent pas dans le modèle de langage, alors que le bigramme [temps de] existe : le mot 'de' est associé à l'ordre 2 ;
- que le quadrigramme [est temps de lire] n'existe pas dans le modèle de langage alors que le trigramme [temps de lire] existe : le mot 'lire' est associé à l'ordre 3 ;
- que le quadrigramme [temps de lire ce] existe dans le modèle de langage : le mot 'ce' est associé à l'ordre 4.

La classe de comportement du mot 'lire' sera alors associée à l'étiquette (-,3,+), car le mot 'lire' est associé à l'ordre 3, son voisin de gauche est associé à un ordre inférieur (-) de valeur 2 et son voisin de droite est associé à un ordre supérieur (+) de valeur 4.

Pour chaque classe et sur un ensemble de transcription donné, on peut calculer le taux d'erreur de reconnaissance des mots composant chacune de ces classes. Ce taux d'erreur est le rapport entre le nombre de mots $n_{err}(cl)$ mal reconnus (substitutions ou insertions) contenus dans une classe cl sur le nombre de mots $n_{mots}(cl)$ qui composent cette classe. Ainsi, pour un mot w associé à la classe cl, la valeur $m_{lmbb}(w)$ donnée par la mesure de confiance LMBB se calcule à partir d'un corpus d'apprentissage avec la formule :

$$m_{lmbb}(w) = 1 - \frac{n_{err}(cl)}{n_{mots}(cl)} \tag{5.2}$$

Le nombre de mots de la classe $n_{mots}(cl)$ est choisi pour ne pas être égal à zéro.

La figure 5.1 montre l'existence d'une corrélation entre comportement du mécanisme de repli du modèle de langage et taux d'erreur. La classe 1 correspond à la fusion des classes (x,1,y) et (x,0,y) où x et y sont une des étiquettes '-', '+', '='. Ces résultats ont été calculés sur le corpus d'apprentissage décrit ultérieurement dans la section 5.4.1.

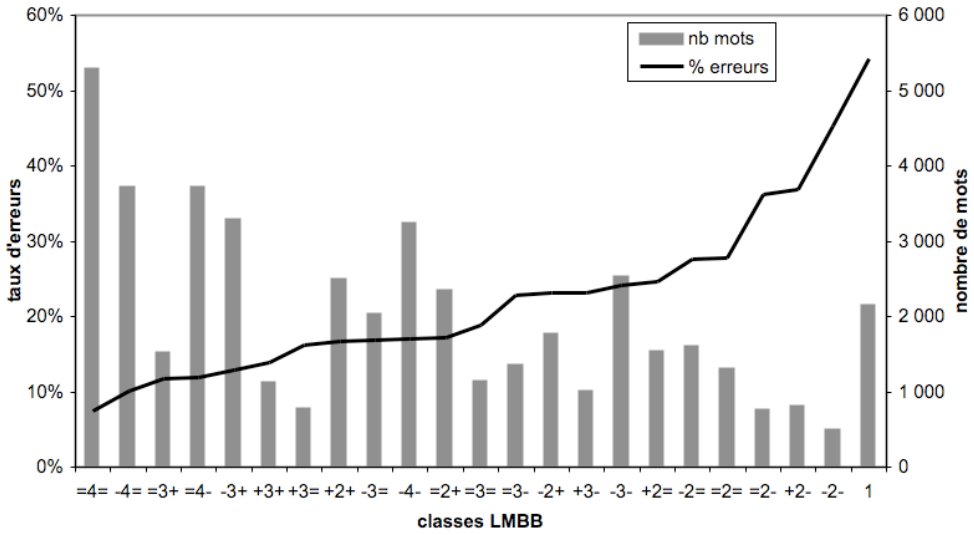


FIG. 5.1 – Taux d'erreur, répartition des mots transcrits et classes LMBB

Dans [Uhrik 1997], les auteurs utilisent également une mesure basée sur le comportement du repli du modèle de langage. Cette mesure est estimée pour un mot w par les formules :

$CM(w) = 1,0$ si w dérive d'un trigramme
$CM(w) = 0,8$ si w dérive de deux bigrammes
$CM(w) = 0,6$ si w dérive d'un seul bigramme
$CM(w) = 0,4$ si w dérive d'un unigramme précédé d'un bigramme
$CM(w) = 0,3$ si w dérive de deux unigrammes
$CM(w) = 0,2$ si w dérive d'un unigramme mais w_{i-1} n'existe pas dans le modèle de langage
$CM(w) = 0,1$ si w est inconnu

(5.3)

La mesure finale sur un mot prend en compte les mesures calculées auparavant pour les contexte gauche et droit du mot :

$$CM'(w_i) = \min\{\{CM(w_{i-2}) * CM(w_{i-1}) * CM(w_i)\}, \{CM(w_{i-1}) * CM(w_i) * CM(w_{i+1})CM(w_i) * CM(w_{i+1}) * CM(w_{i+2})\}\} \quad (5.4)$$

Cette expérience a été menée sur un modèle de langage trigramme. Pour pouvoir comparer les deux mesures de confiance, nous avons étiqueté les mots dérivant d'un quadrigramme comme dérivant d'un trigramme. Les résultats sont détaillés au paragraphe 5.4.3.

5.3 Probabilité *A Posteriori* (PAP)

Les probabilités *a posteriori* peuvent être calculées à partir des listes des N meilleures hypothèses comme dans l'article [Stolcke 1997], à partir des graphes de mots comme proposé dans les articles [Wessel 1998, Wessel 1999, Wessel 2000, Wessel 2001] ou encore à partir des réseaux de confusion [Mangu 2000, Evermann 2000]. En fait, la probabilité *a posteriori* d'un mot est le ratio entre la probabilité *a priori* d'un mot et la somme des probabilités *a priori* de toutes les autres hypothèses alternatives. Ces probabilités *a priori* sont une combinaison des scores fournis par les modèles acoustiques et linguistique.
Dans les listes des N meilleures hypothèses la probabilité *a posteriori* d'un mot est le rapport de la somme des probabilités *a priori* des occurrences de ce mot à une position donnée parmi les N hypothèses, sur la somme de toutes les probabilités *a priori* des mots situés à la même position, incluant celles des occurrences du mot courant.
Dans les approches basées sur les graphes de mots ou les réseaux de confusions, la probabilité *a posteriori* est la généralisation de l'approche précédente où la segmentation en mots et la profondeur de l'espace de recherche sont mieux pris en considération.

Ici, nous utilisons un réseau de confusion basé sur la technique utilisée dans [Mangu 2000] pour calculer les probabilités *a posteriori* des mots. Cette technique permet de regrouper dans le graphe de mots tous les arcs associés au même mot avec différentes prononciations et les mots qui sont en concurrence. L'estimation du réseau de confusion est détaillée au paragraphe 3.3.3 du chapitre 3 sur l'état de l'art des mesures de confiance. La mesure de confiance basée sur les probabilités *a posteriori* des mots sera notée PAP.

Les probabilités *a posteriori* des mots issues du réseau de confusion peuvent directement être utilisée en tant que mesure de confiance. Cependant, il apparaît comme montré sur la figure 5.2 que la probabilité *a posteriori* d'un mot tende à sous-estimer la pertinence réelle de ce mot. Certains mots ayant une probabilité *a posteriori* élevée (entre 0,97 et 1) sont eux surestimés alors que ces mots ne sont pas forcément corrects. Cela est dû au fait que les réseaux de confusion ne représentent pas toutes les hypothèses en compétition et qu'une partie de la "masse" totale des probabilités à distribuer parmi les mots est manquante. Les hypothèses en compétition ne sont pas toutes conservées à cause de la technique d'élagage nécessaire pour éviter de compromettre l'alignement final du réseau de confusion 3.3.3. Plus le graphe est grand,

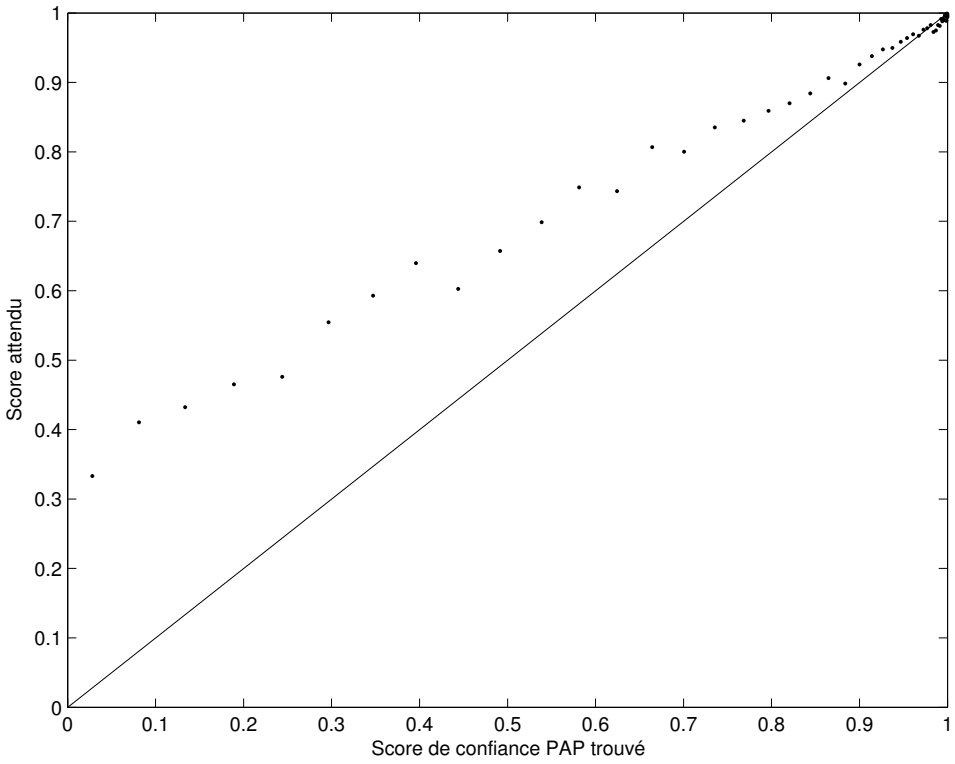


FIG. 5.2 – Répartition des moyennes de 50 échantillons de scores de confiance sur le corpus CTrain.

plus cette “masse” sera éparpillée. Pour lisser ces probabilités, une transformation linéaire par morceau est appliquée à la probabilité *a posteriori* d’un mot. Cette transformation est appelée mapping. Cette technique est par exemple utilisée dans [Evermann 2000] en utilisant un arbre de décision.

Pour calculer cette transformation, on dispose d’un corpus CTrain qui est le corpus d’apprentissage des mesures de confiance (voir paragraphe 5.4.1). Ce corpus est divisé en cent échantillons comprenant chacun 458 mots ayant une probabilité *a posteriori* calculée à partir d’un réseau de confusion. On effectue la moyenne des probabilités sur chacun des échantillons et on cherche là où les morceaux de droites reliant chacune de ces moyennes et minimisant les moindres carrés.

Le nombre de droites à utiliser pour séparer les échantillons est optimisé en termes de NCE sur le corpus CTrain. Sur le tableau 5.1 apparaissent les taux NCE calculés lorsque la transformation est constituée de 2, 4 ou 8 droites sur les corpora CTrain et Test. Les transformations constituées de 4 et 8 droites obtiennent des résultats similaires maximisant la NCE sur les corpora CTrain et Test. Cette métrique d’évaluation a été décrite dans le paragraphe 3.7.

	NCE	
Nombre de droites	**CTrain**	**Test**
2	0,299	0,290
4	**0,300**	**0,293**
8	0,300	0,293

TAB. 5.1 – Comparaison de différents mapping avec 2, 4 ou 8 droites en termes de NCE (entropie croisée normalisée) sur le corpus CTrain et le corpus de test

Pour la transformation à 8 droites, il existe un risque de sur-apprendre cette transformation. En conséquence, seules 4 droites sont utilisées avec pour équations :

$$\begin{aligned}
&\text{Pour } x \leq 0,2101, y = 0,8809x + 0,3181 \\
&\text{Pour } x \leq 0,4248, y = 1,0103x + 0,2396 \\
&\text{Pour } x \leq 0,6152, y = 1,0052x + 0,1602 \\
&\text{Pour } x > 0,6152, y = 0,6354x + 0,3566
\end{aligned} \tag{5.5}$$

5.4 Évaluation des mesures de confiance

5.4.1 Corpus d'apprentissage

Le corpus utilisé pour estimer les mesures de confiance utilisées dans ce manuscrit est composé de 4h de parole extraites de la campagne ESTER ainsi que de leur transcription manuelle. Ce corpus est noté CTrain. Les radios sont les mêmes que celles incluent dans le corpus d'apprentissage des modèles acoustiques et de langage mais les données sont différentes ; il n'y a pas d'intersection entre les deux corpora. Les transcriptions manuelle et automatique (fournie par notre système), permettent de calculer les différents paramètres des mesures de confiance utilisées. Pour la mesure acoustique, il s'agit de l'estimation des paramètres μ, σ and a de l'équation 5.1. Pour la mesure LMBB, on calcule le score de confiance d'un mot à partir du taux d'erreur moyen obtenu par les mots de sa classe de comportement sur le corpus CTrain (voir figure 5.1 et équation 5.2).

5.4.2 Corpus de test

Le corpus de test utilisé pour les mesures de confiance est le corpus de test officiel d'ESTER composé de 10h d'émissions de radio, comme le montre le tableau 1.1 du paragraphe 1.7 décrivant le système de RAP du LIUM.

5.4.3 Résultats en terme de NCE, de CER et de courbes DET

Pour les trois mesures AC, LMBB et PAP décrites précédemment, leur NCE et leur taux de CER ont été mesurés (ces métriques sont décrites au paragraphe 3.7). Pour calculer le CER de chaque mesure, on fait varier un seuil λ au dessus duquel les mots sont considérés comme étant corrects et en dessous duquel ils sont considérés comme étant incorrects. On évalue ensuite la classification réelle pour calculer le taux CER comme étant le nombre d'étiquettes incorrectement assignées sur le nombre total d'étiquettes. Pour le CER, le seuil pour lequel ce taux est le plus faible sur le corpus CTrain est conservé pour calculer le CER sur le corpus de test. Les tableaux 5.2 et 5.3 montrent ces résultats.

	NCE	
Mesure	**CTrain**	**test**
Mesure de [Uhrik 1997]	-1,702	-1,713
AC	0,019	0,023
LMBB	0,080	0,063
PAP	0,182	0,187
MAP(PAP)	0,300	0,293

TAB. 5.2 – Comparaison des diverses mesures de confiance sur les données d'apprentissage et de test des mesures de confiance en termes d'entropie croisée normalisée (NCE)

En termes de NCE, la mesure sur le comportement du repli utilisée dans [Uhrik 1997] obtient un résultat négatif. Cette mesure n'indique pas si le mot est réellement correct ou non contrairement à notre mesure LMBB.

Au vu des résultats positifs en NCE, toutes les mesures que nous avons estimées apportent une information sur la pertinence du mot.

La mesure PAP atteint le meilleur score, notamment après application de la technique du mapping. Par la suite l'application de la technique du mapping à une mesure ψ sera indiquée par la notation MAP(ψ).

Pour le CER, le taux référence est calculé comme étant le taux d'erreur des mots émis par le système de reconnaissance, soit le nombre d'insertions et de substitutions divisé par le nombre total de mots reconnus. En effet, on peut considérer que par défaut, le système estime que toutes ses hypothèses sont correctes. Il est égal à 15,09% pour CTrain et 19,23% sur le corpus de test. Les mesures AC et LMBB ont un CER plus élevé que le taux référence sur les deux corpora. Les mesures PAP et MAP(PAP) obtiennent le meilleur score, améliorant le score de la référence de plus de 3,06% en relatif sur le test. Seuls 13,56% des mots sont incorrectement étiquetés si on place le seuil de référence à 0,250 pour la mesure PAP et à 0,511 pour la mesure MAP(PAP). La technique du mapping conserve le même CER car les scores des mots ont subit

	CER		
Mesure	**CTrain** (%)	**test** (%)	**seuil**
référence	**15,09**	**19,23**	-
AC	15,08	22,31	0,603
LMBB	15,11	22,23	0,552
PAP	13,56	18,64	0,250
MAP(PAP)	13,56	18,64	0,511

TAB. 5.3 – Comparaison des diverses mesures de confiance sur les données d'apprentissage et de test des mesures de confiance en termes de Confidence Error Rate (CER)

une transformation croissante linéaire par morceaux qui n'affecte pas la répartition par rapport au seuil optimal. Le seuil opimal sur CTrain subit lui-même la transformation linéaire pour devenir le seuil optimal sur le corpus de test.

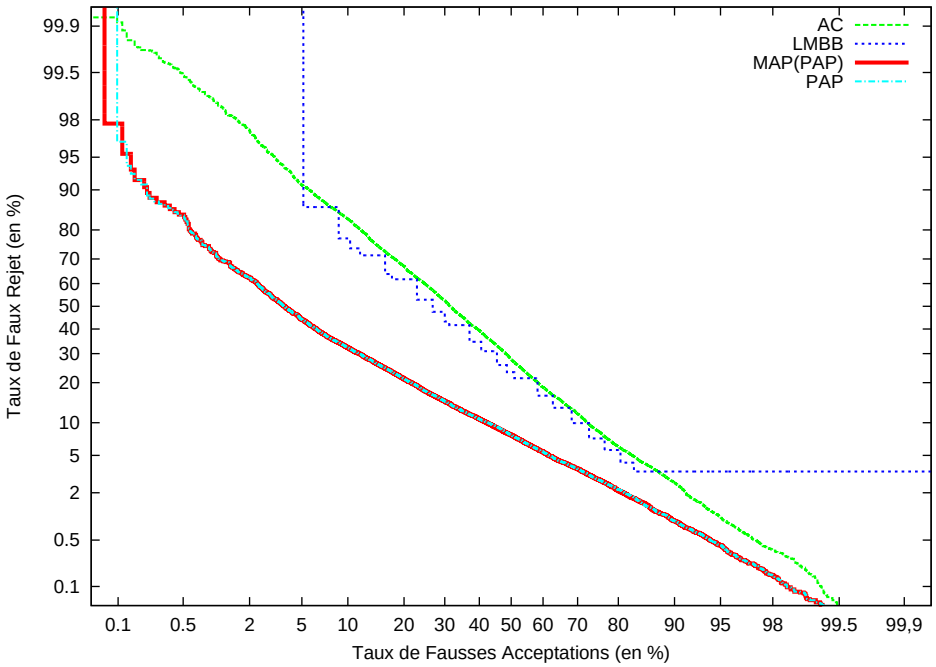


FIG. 5.3 – Courbe DET des mesures de confiance sur le corpus CTrain

Les figures 5.3 et 5.4 montrent les courbes DET des différentes mesures proposées sur les corpora CTrain et de test. La courbe LMBB est en escaliers car le score de confiance des mots est indiqué pour chaque classe de mots. Les courbes PAP et MAP(PAP) sont les plus performantes car elles allient les scores acoustiques et linguistiques et prennent en compte le niveau de confusion du système de transcription pour le choix de ce mot. Le score linguistique contenu dans la mesure LMBB diffère du score donné par le modèle de langage et prend en

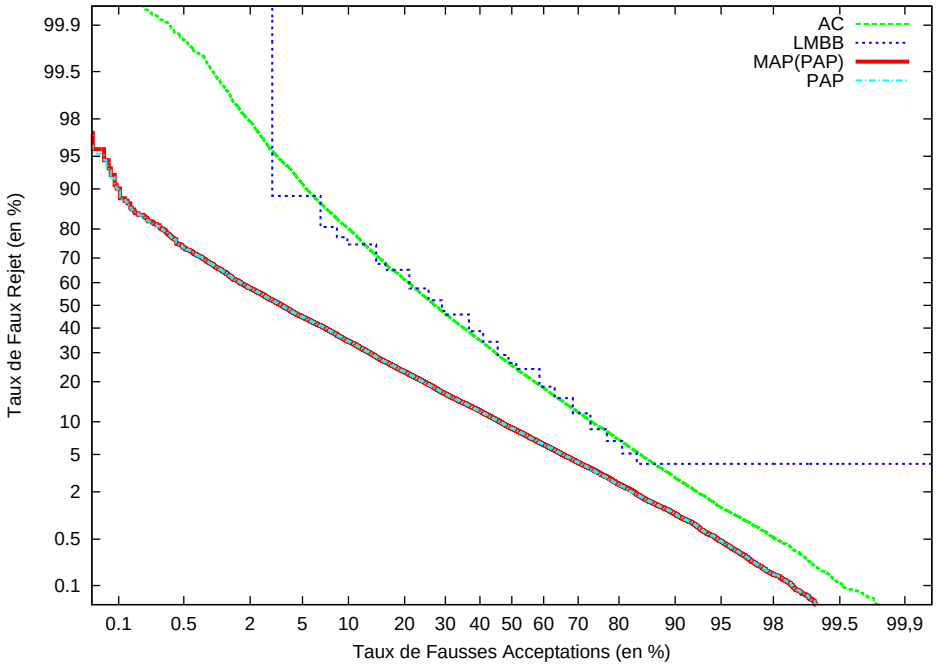


FIG. 5.4 – Courbe DET des mesures de confiance sur le corpus de test

compte le mot dans son contexte. Combiner la mesure PAP ou MAP(PAP) avec le score LMBB semble intéressant afin d'apporter une information nouvelle sur le niveau de connaissance du modèle de langage.

5.5 Fusion des mesures de confiance

Après avoir décrit et élaboré quelques mesures de confiance, différentes combinaisons ont été mises en œuvre et testées sur le corpus de test afin d'améliorer les performances pour déterminer la pertinence d'un mot.

5.5.1 Théorie des probabilités

Les différentes mesures de confiance (AC, LMBB, PAP et MAP) sont considérés comme des experts en classification *Correct/Incorrect*. Pour la classe obtenue, un seuil λ doit être fixé pour décider si le mot est considéré comme correct (le score de confiance du mot est au-dessus de λ) ou non (le score de confiance du mot est en-dessous de λ). λ est défini comme étant le seuil pour lequel les mesures de confiance ont le taux CER le plus faible (voir tableau 5.3). Pour chaque expert, une matrice de confusion *Correct/Incorrect* est calculée (voir le tableau 5.4). Les

indices β des classes sont calculés de la façon suivante :

$$\beta_C = \frac{P(C|C)}{P(C|C) + P(NC|C)} \text{ et } \beta_{NC} = \frac{NC|NC}{P(C|NC) + P(NC|NC)} \tag{5.6}$$

où C correspond à la classe *Correct* et NC à la classe *Incorrect*.

Expert AC	Classe attendue (%)	
Classe obtenue (%)	Correct	Incorrect
Correct	84,818	15,007
Incorrect	0,072	0,109
$\beta_{AC_{Correct}}$	84,966%	
$\beta_{AC_{Incorrect}}$	60,241%	

Expert LMBB	Classe attendue (%)	
Classe obtenue (%)	Correct	Incorrect
Correct	84,884	15,116
Incorrect	0	0
$\beta_{LMBB_{Correct}}$	84,388%	
$\beta_{LMBB_{Incorrect}}$	0%	

Expert PAP	Classe attendue (%)	
Classe obtenue (%)	Correct	Incorrect
Correct	81,032	9,704
Incorrect	3,852	5,412
$\beta_{PAP_{Correct}}$	89,305%	
$\beta_{PAP_{Incorrect}}$	58,418%	

Expert MAP	Classe attendue (%)	
Classe obtenue (%)	Correct	Incorrect
Correct	81,025	9,701
Incorrect	3,859	5,416
$\beta_{MAP_{Correct}}$	89,308%	
$\beta_{MAP_{Incorrect}}$	58,396%	

TAB. 5.4 – Estimation des indices de classes β pour chacun des experts AC, LMBB, PAP et MAP

Les indices de confiance d'expert α sont calculés en ajoutant les indices de confiance de classes β. Pour chaque mot w, la mesure de confiance finale est celle donnée par l'expert maximisant la formule :

$$\alpha_j * (1 - \text{Coût}_j(w)) \tag{5.7}$$

où la fonction de coût est donnée à l'équation 3.18 du chapitre 3. Des résultats comparant la combinaison issue de la théorie des probabilités avec des combinaisons linéaires sont présentés par la suite.

5.5.2 Combinaisons linéaires

Différentes combinaisons linéaires ont été testées entre les mesures AC, PAP et LMBB telles que la simple interpolation ou encore la technique du mapping. Les coefficients d'interpolation sont optimisés sur le corpus CTrain.

La première transformation testée est l'interpolation linéaire. La combinaison retenue maximisant la NCE s'écrit : 0,3LMBB+0,7PAP. Elle est notée PAP/LMBB. Elle maximise la NCE par rapport à d'autres interpolations linéaires avec des coefficients différents.

Ensuite, la technique du mapping est appliquée à cette fusion, elle est notée par la suite MAP(PAP/LMBB). Les équations des droites pour cette transformation sont :

$$\begin{aligned}
&\text{Pour } x \leq 0,1674 \; y = 0,7441x + 0,0414 \\
&\text{Pour } x \leq 0,1692 \; y = 1,4211x - 0,0529 \\
&\text{Pour } x \leq 0,1769 \; y = 0,4290x - 0,301 \\
&\text{Pour } x > 0,1769 \; y = 0,8383x + 0.1865
\end{aligned} \tag{5.8}$$

Un autre type de combinaison est testé entre MAP(LMBB) et MAP(PAP). La meilleure combinaison de ce type est l'interpolation : 0,1MAP(LMBB)+0,9MAP(PAP). La mesure notée MAP(LMBB) est obtenue avec 2 droites d'équation :

$$\begin{aligned}
&\text{Pour } x \leq 0,6 \; y = x + 0,645 \\
&\text{Pour } x > 0,6 \; y = 1,0422x - 0,0309
\end{aligned} \tag{5.9}$$

Seules deux droites sont utilisées car aucune amélioration en termes de NCE n'est notée en utilisant quatre droites. Les équations de droites utilisées pour la mesure MAP(PAP) sont décrites équation 5.5. Ceci est dû au fait que les classes de comportement LMBB sont déjà calculées à partir du taux d'erreur sur les mots qui les composent.

Enfin, la dernière combinaison est obtenue en calculant le mapping de la PAP et en cherchant la meilleure interpolation avec LMBB sur chacune des quatre droites utilisées dans l'estimation du mapping. Elle est notée : Interpolation MAP(PAP)&LMBB. Cela donne les équations :

$$\begin{aligned}
&\text{Pour } x \leq 0,5032 \;\; 0,89MAP(PAP) + 0,11LMBB \\
&\text{Pour } x \leq 0,6688 \;\; 0,8MAP(PAP) + 0,2LMBB \\
&\text{Pour } x \leq 0,7785 \;\; 0,5MAP(PAP) + 0,5LMBB \\
&\text{Pour } x > 0,7785 \;\; 0,95MAP(PAP) + 0,05LMBB
\end{aligned} \tag{5.10}$$

Aucune combinaison linéaire ne conserve la mesure AC dans son calcul. Cette mesure n'apporte aucune information supplémentaire à la combinaison des trois mesures car le score acoustique est déjà présent dans la mesure PAP.

5.5.3 Résultats en termes de NCE, CER et courbes DET

Le tableau 5.5 montre les taux NCE des meilleures combinaisons linéaires ainsi que la combinaison issue de la théorie des probabilités.

	NCE	
Mesure	**CTrain**	**test**
Théorie des probabilités Proba	0,195	0,231
PAP/LMBB	0,277	0,284
MAP(PAP/LMBB)	0,296	**0,299**
0,1MAP(LMBB)+ 0,9MAP(PAP)	0,301	0,294
Interpolation MAP(PAP) & LMBB	**0,304**	0,296

TAB. 5.5 – Comparaison de différentes combinaisons de mesures de confiance sur les données d'apprentissage et de test des mesures de confiance en termes d'entropie croisée normalisée (NCE)

Sans la technique du mapping, la meilleure fusion est la mesure PAP/LMBB. Après estimation du mapping sur cette mesure, les scores de confiance associés attestent un peu mieux du fait qu'un mot soit correct ou non. En estimant le mapping de chacune des mesures PAP et LMBB, la mesure 0,1MAP(LMBB)+ 0,9MAP(PAP) offre une légère amélioration en termes de NCE sur le corpus CTrain. La mesure offrant le meilleur taux de NCE sur CTrain est celle qui propose une interpolation différente pour LMBB et MAP(PAP) sur chaque morceau de droite utilisé pour le mapping. Sur le corpus de test, la meilleure mesure est MAP(PAP/LMBB).

	CER		
Mesure	**CTrain** (%)	**test** (%)	**seuil**
référence	**15,09**	**19,23**	-
PAP/LMBB	**13,17**	**18,27**	0,411
MAP(PAP/LMBB)	**13,17**	**18,27**	0,527
0,1MAP(LMBB)+ 0,9MAP(PAP)	13,43	18,48	0,548
Interpolation MAP(PAP) & LMBB	13,23	18,28	0,568
Théorie des Probabilités	13,62	18,58	0,528

TAB. 5.6 – Comparaison de différentes combinaisons de mesures de confiance sur les données d'apprentissage et de test des mesures de confiance en termes de Confidence Error Rate (CER)

En termes de CER, les meilleures mesures sont PAP/LMBB avant et après mapping sur les deux corpora. La mesure MAP(PAP/LMBB) est intéressante car elle offre de bonnes performances pour les deux mesures.

La figure 5.5 montre que les différentes combinaisons sont très proches en termes d'acceptation/rejet d'hypothèses. Le choix de la mesure de confiance à utiliser se fera en fonction de l'application visée.

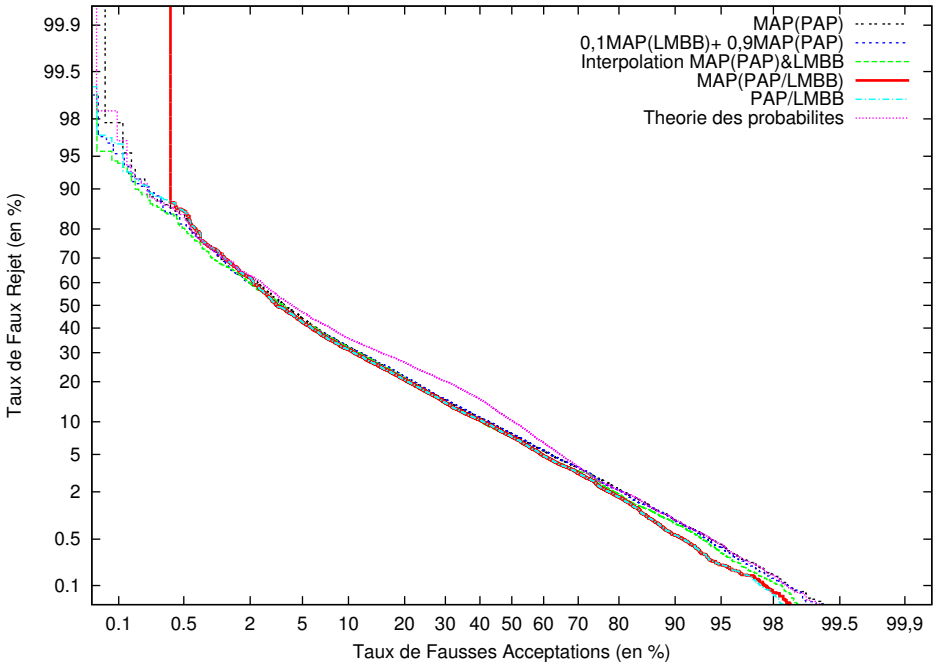


FIG. 5.5 – Courbe DET de différentes combinaison des mesures de confiance ainsi que de la meilleure mesure MAP(PAP) sur le corpus de test

Dans le prochain chapitre, nous utiliserons les mesures de confiance pour faire du filtrage de mots corrects. En effet, pour effectuer un apprentissage non-supervisé de modèles acoustiques, le but est de sélectionner sur un corpus additionnel des séquences de mots comportant le maximum de mots corrects en ayant peu de mots incorrects pouvant “bruiter” le corpus d’apprentissage. Ces séquences seront ajoutées au corpus d’apprentissage initial. D’après la figure 5.5, la meilleure mesure pour conserver le plus de mots corrects est la mesure PAP/LMBB. Elle a le même comportement que MAP(PAP/LMBB) avec une étape de calcul en moins. Le détail du choix des mots corrects à conserver sera expliqué dans le chapitre suivant.

5.6 Conclusion

Dans ce chapitre, trois mesures de confiances ont été proposées et élaborées. Chacune d’entre elles permet de repérer des zones erronées dans l’hypothèse fournie par un SRAP, qualité repérée grâce à leur taux positif en NCE. La mesure PAP allie les scores acoustiques et linguistiques du mot tandis que la mesure LMBB apporte des informations sur le mot dans son contexte. La meilleure combinaison de ces deux mesures avec la mesure AC ne fait plus intervenir cette mesure car elle est trop dépendante du score acoustique déjà contenu dans la

mesure PAP. La mesure AC n'apporte aucune information supplémentaire dans la combinaison. La mesure PAP/LMBB améliore le taux NCE par rapport aux mesures prises séparément et est la mesure la plus efficace en termes de Fausses Acceptations/ Faux Rejets. En fonction de l'application visée, le choix de la mesure de confiance à utiliser se fera sur les qualités recherchées. Certaines applications nécessitent que la mesure rejette plus de mots incorrects que de mots corrects ou vice et versa, d'autres encore nécessitent juste une mesure permettant de repérer les mots incorrects.

Dans le prochain chapitre, la meilleure mesure en termes de rejet de mots incorrects sera utilisée pour faire du filtrage de données. La mesure PAP/LMBB servira ainsi à repérer les séquence de mots corrects pour les injecter dans le corpus d'apprentissage des modèles acoustiques.

Ensuite, un chapitre sur l'identification du locuteur montrera comment utiliser d'autres types de scores de confiance pour associer un nom détecté dans une transcription au locuteur qui prononce un segment de parole.

Chapitre 6

Détection de séquences de mots corrects pour l'apprentissage non supervisé de modèles acoustiques

Sommaire

Ce chapitre présente l'utilisation de mesures de confiance pour la détection de mots corrects. Cette détection de zone de mots corrects va ici être appliquée à l'apprentissage non-supervisé de modèles acoustiques pour la reconnaissance automatique de la parole.

La transcription manuelle d'enregistrements audio est très coûteuse, ce qui limite la taille des corpus d'apprentissage des modèles où une transcription fidèle est nécessaire afin de permettre le bon alignement phonème/signal. Les performances d'un système de RAP découlent en grande partie de la qualité des modèles acoustiques. Ces modèles, d'essence statistique, sont d'autant plus robustes que leurs corpora d'apprentissage sont grands.

Une des méthodes utilisées pour augmenter à moindre coût la taille d'un corpus d'apprentissage consiste à ajouter des transcriptions automatiques [Lamel 2000, Deléglise 2005]. Le risque de cet ajout brut est d'introduire du "bruit" dans la modélisation dues aux erreurs de reconnaissance du système de RAP. Une des solutions est de filtrer ces données provenant des transcriptions automatiques afin de ne pas incorporer ces erreurs. Des mesures de confiance peuvent être utilisées afin de détecter les séquences de mots corrects et les incorporer dans le corpus d'apprentissage des modèles acoustiques.

Le premier paragraphe présentera les méthodes de filtrage de données utilisées dans la littérature. Ensuite, le corpus additionnel composé de transcriptions automatiques est décrit. La mesure de confiance permettant de détecter les mots corrects sera détaillée dans un troisième paragraphe. Pour apprendre les modèles acoustiques, il est nécessaire d'avoir des segments assez longs pour apprendre les phonèmes en contexte. Les séquences de mots corrects sont alors choisies grâce à une méthode détaillée dans le paragraphe 6.4. Enfin, les résultats de cette méthode permettant d'ajouter des séquences de mots corrects au corpus d'apprentissage des modèles acoustiques seront exposés.

6.1 Apprentissage non supervisé de modèles acoustiques

Transcrire manuellement un nouveau corpus requiert énormément de temps pour un annotateur humain (20 à 50 fois le temps réel selon le niveau de précision de l'annotation). Quand cela est possible, une solution consiste à utiliser les sous-titrages (*closed-captions*) [Lamel 2000, Lamel 2002, Chen 2004]. Plusieurs corpora comme les émissions télévisées sont sous-titrées : il suffit alors de transcrire automatiquement ces corpora à l'aide d'un SRAP. Seules les parties où l'hypothèse donnée par le système et les sous-titres sont en accord sont utilisées pour augmenter le corpus d'apprentissage. Les modèles acoustiques du SRAP sont alors réestimés avec ce nouveau corpus d'apprentissage. L'augmentation de volume améliore les performances en termes de taux d'erreur par rapport au premier système.

Les inconvénients de cette méthode sont que ces sous-titres ne contiennent pas d'informations sur le type de signal ni d'informations sur le locuteur, les tours de parole, ou sur l'environnement. Ce sont des annotations minimales sur la transcription orthographique et les hésitations, les répétitions n'y sont pas indiquées. Ces sous-titres peuvent contenir des insertions de mots, des suppressions ou des changement d'ordre de mots. Le corpus additionnel ne contiendra donc pas d'hésitations ou de bruit de bouche, fréquemment présents dans les données à transcrire, à moins d'incorporer un corpus spécifique annoté à la main.

Cette méthode permet tout de même de gagner 5% en relatif par rapport à l'ajout du corpus additionel brut (140h)[Lamel 2000]. Par rapport à un ajout d'un corpus transcrit manuellement, la méthode utilisant des sous-titres détériore les résultats de l'ordre de 10% en relatif.

Quand les sous-titres ne sont pas disponibles, une autre méthode proposée dans [Kemp 1999, Wessel 2001, Wessel 2005] permet d'augmenter les données du corpus d'apprentissage des modèles acoustiques. Ici, la probabilité *a posteriori* calculée à partir d'un graphe de mot est utilisée comme mesure de confiance afin d'estimer la pertinence de l'hypothèse fournie par le système de reconnaissance sur un corpus utilisé pour étoffer le corpus initial. Lorsque les hypothèses ont une mesure de confiance supérieure à un seuil d'acceptation, ces hypothèses viennent agrandir le corpus d'apprentissage. Les travaux de [Wessel 2001] s'appuient sur un corpus initial contenant peu d'heures d'apprentissage (entre 1,2h et 5,6h) et ajoutent des données filtrées avec des mesures de confiance en testant différents seuils d'acceptation. Le système de départ atteint un WER de 33,5%. Avec 5,6h dans le corpus initial, l'ajout de 32 heures filtrées apporte un gain absolu de 10,1%. Ajouter 72h brutes permet un gain de 8,7% en absolu.

Nos travaux consistent à employer une méthode similaire en utilisant des mesures de confiance pour incorporer de nouvelles données dans le corpus d'apprentissage des modèles acoustiques de notre SRAP.

6.2 Corpus additionnel

Le corpus additionnel dont on veut extraire des séquences de mots corrects est une partie des 1677 heures non transcrites fournies par la campagne d'évaluation ESTER (voir tableau 1.1). Il s'agit d'heures d'émissions radiophoniques provenant des mêmes radios dont sont issus les corpus d'apprentissage, de développement et de test utilisés par le SRAP du LIUM. 558 heures de ces 1677 heures ont été transcrites automatiquement grâce à ce SRAP décrit dans le chapitre 1.7. Elles comprennent 58h de bande étroite (qualité de prise de son via le téléphone) et 500h de bande large (qualité de prise de son en studio). Le but de l'expérience est de de vérifier

la pertinence d'un processus de filtrage de données grâce à une mesure de confiance. On se propose ici de doubler la taille du corpus initial des modèles acoustiques qui contient 80 heures ce qui correspond à un rejet de 85% des 558 heures.

6.3 Mesure de confiance utilisée

La mesure de confiance utilisée est la mesure PAP/LMBB décrite dans le chapitre II. Il s'agit d'une combinaison d'une mesure de confiance calculée à partir de la probabilité *a posteriori* d'un mot et d'une mesure de confiance estimée en fonction du repli du modèle de langage. Cette dernière est celle qui offre les meilleures performances en termes de Faux Rejets/Fausses Acceptations pour conserver le maximum de mots corrects (voir figure 5.5 du chapitre précédent).

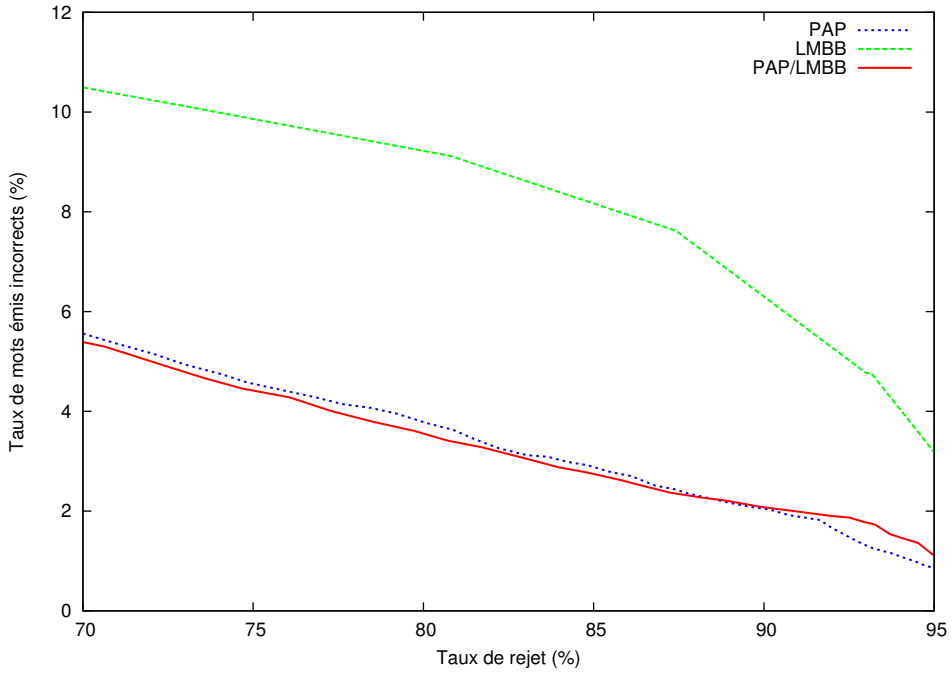


FIG. 6.1 – Taux de mots émis incorrects en fonction du taux de rejet pour trois mesures de confiance sur le corpus de test : PAP, LMBB et PAP/LMBB

La figure 6.1 montre la mesure de confiance PAP/LMBB ainsi que les mesures qui la compose pour les comparer en termes de taux de mots émis incorrects en fonction du taux de rejet. Ces courbes sont estimées sur le corpus de test.

La mesure PAP/LMBB permet d'obtenir un faible taux de mots émis incorrects en rejetant 85% du corpus, taux de rejet que l'on veut atteindre sur le corpus additionnel. Ce taux de rejet correspond à un seuil λ. Par la suite, ce seuil sera optimisé pour répondre à différentes

contraintes expliquées dans le paragraphe suivant. Au dessus de ce seuil, les mots seront conservés pour augmenter le corpus d'apprentissage des modèles acoustiques.

6.4 Choix des séquences de mots

Pour estimer les modèles acoustiques, des segment de parole de plus de 4 secondes sont conservés. En effet, il est nécessaire d'avoir des segments assez longs pour estimer les modèles des phonèmes en contexte : ces segments correspondent à des séquences de mots contigus.

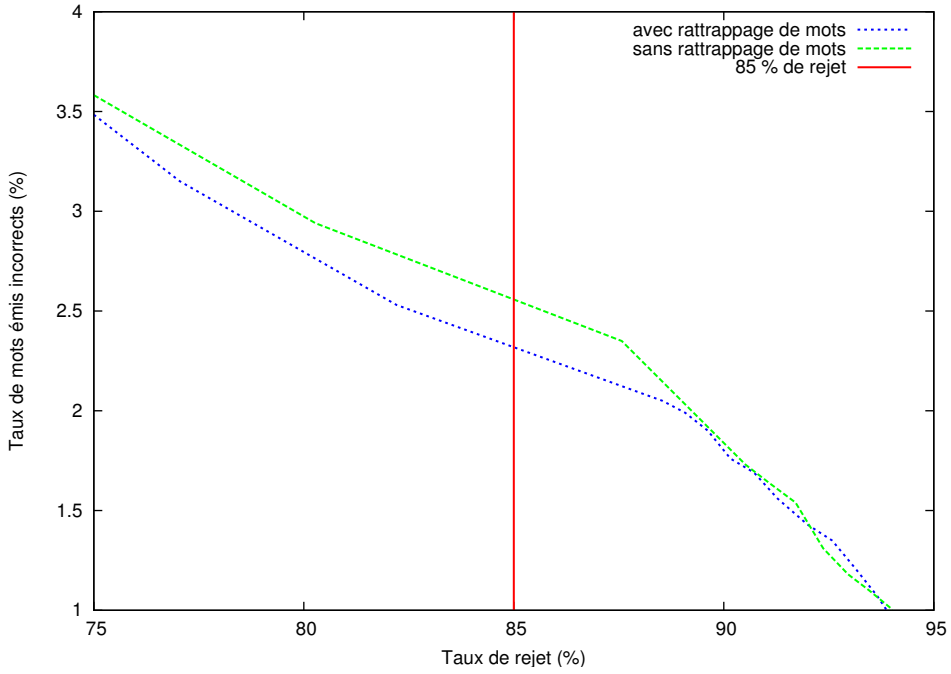


FIG. 6.2 – Taux d'erreur et de rejet sur le corpus de test de différents filtrages à l'aide d'un seuil sur le score de confiance des mots et pour une durée de séquence de mots supérieure à 4 secondes

La figure 6.2 détaille les différents taux d'erreur sur les mots émis par le SRAP et leurs taux de rejet en fonction d'un seuil sur le score de confiance des mots et pour une durée de séquence de mots conservés supérieure à 4 secondes. La courbe nommée "Sans rattrappage de mots" représente pour différents seuils λ_1 le comportement du filtrage quand seuls les mots au-dessus de ce seuil sont conservés.

Les scores de confiance élevés qui offrent des taux d'erreur sur les mots émis faibles ne permettent pas de conserver suffisament de mots dans le corpus. L'utilisation d'un unique seuil pour sélectionner les mots corrects amène à la sélection de mots isolés peu utiles pour

l'apprentissage des modèles acoustiques. Pour un taux d'erreur sur les mots émis de 2,5%, on ne conserve seulement qu'un peu plus de 10% du corpus.

Pour s'affranchir de ce problème, la contrainte du seuil est relâchée. Sur une fenêtre de quatre mots, on choisit de garder les mots sous deux conditions, au moins un mot doit avoir un score de confiance supérieur à λ_1 et les autres mots doivent avoir un score de confiance supérieur à un seuil λ_2. Le seuil λ_2 est optimisé sur le corpus CTrain. On conserve toujours les séquence de mots de plus de 4 secondes.

La courbe nommée " Avec rattrappage de mots" montre le comportement d'une méthode de filtrage où λ_2 est fixé à 0,5. On choisira le seuil λ_1 pour conserver 80h du corpus additionnel.

Pour un seuil λ_1 fixé, la première méthode "Sans rattrappage de mots" va rejeter beaucoup plus de mots que la méthode "Avec rattrappage de mots" pour un taux d'erreur similaire. Pour conserver environ 80 heures du corpus additionnel, le seuil principal est fixé à 0,96. Ceci correspond à un taux d'erreur sur les mots émis de 2,31% sur le corpus de test comme le montre la tableau 6.1. Avec la méthode "Sans rattrappage de mots", le taux d'erreur sur les mots émis est plus élevé et atteint 2,63%.

Méthode de filtrage	**Taux d'erreur sur les mots émis sur le corpus de Test**
Sans rattrappage de mots	2,63%
Avec rattrappage de mots	2,31%

TAB. 6.1 – Taux d'erreur sur les mots émis en fonction de la méthode de filtrage employée pour des séquences de mots de plus de 4 secondes

Cette méthode de rattrapage permet d'obtenir 86 heures supplémentaires à réinjecter dans le corpus d'apprentissage des modèles acoustiques avec peu de mots incorrects.

Le tableau 6.2 montre que la sélection des mots permet d'augmenter le corpus en bande large et en bande étroite.

Corpus d'apprentissage	**Bande étroite (téléphone)**	**Bande large (studio)**
Initial 80h (Ω)	8h	72h
Ω **+ 86h filtrées**	11h	155h

TAB. 6.2 – Répartition du corpus d'apprentissage en fonction de la bande passante

L'ajout de données permet d'augmenter le corpus initial de 3h de données en bande étroite et de 83h de données en bande large.

6.5 Résultats

Différentes expériences ont été réalisées pour le filtrage de données, notamment en changeant le nombre d'états partagés pour modéliser les phonèmes en contexte des HMMs des modèles acoustiques.

En effet, dans une architecture telle que celle employée dans le système du LIUM, les modèles partagent une palette de gaussiennes. Ce partage de paramètres limite les problèmes d'estimation liés à de petites tailles de corpus d'apprentissage et permet de réduire l'espace mémoire occupé par le stockage des modèles acoustiques.

Le système de reconnaissance initial atteint un taux d'erreur WER de 23,7%. Les modèles acoustiques du système initial sont estimés avec 5500 états partagés, nombre qui a été optimisé auparavant. Les résultats du tableau 6.3 montrent qu'en ajoutant 86h filtrées avec 6000 états partagés, les gains ne sont pas significatifs par rapport aux 23,5% obtenus en rajoutant 80h non filtrées. L'idée est donc de tenter d'accroître ce nombre d'états partagés pour essayer de modéliser avec plus de précision certaines séquences de phonèmes qui apparaissent dans le nouveau corpus.

Corpus d'apprentissage	**Nombre d'états partagés**	**Taux d'erreur (%)**
Initial 80h (Ω)	5500	23,7%
Ω + 80h non filtrées	5500	23,5%
Ω + 86h filtrées	6000	23,4%
Ω + 86h filtrées	7500	23,2%
Ω + 86h filtrées	8500	23,3%
Ω + 86h filtrées	10000	23,2%
Ω + 28h non-filtrées (2e filtrage)	5500	23,7%
Ω + 11h filtrées (2e filtrage)	5500	23,4%
Ω + 28h filtrées (2e filtrage)	5500	23,3%

TAB. 6.3 – Taux d'erreur (WER) pour plusieurs tailles de corpus d'apprentissage des modèles acoustiques et en modifiant le nombre d'états partagés

En ajoutant 2000 états partagés, une diminution significative du taux d'erreur est observée. En effet, pour un score de départ de 23,7% de WER et 114000 mots dans le corpus de test, un résultat est statistiquement significatif s'il est en dehors de l'intervalle de confiance à 95% [23,453-23,947] [Simonin 1998]. Ce taux d'erreur pour 7500 états partagés de 23,2% semble être un optimum car augmenter encore le nombre d'états partagés ne fait pas décroître ce taux. Il semble alors que le filtrage de données conserve quelques séquences de phonèmes qui permettent aux modèles acoustiques d'améliorer leur estimation. Certains contextes sont mieux modélisés et permettent au SRAP d'obtenir de meilleures performances en WER.

La deuxième partie du tableau montre un filtrage ajoutant beaucoup moins de données mais filtrées avec une autre mesure de confiance associant uniquement la mesure acoustique (AC) et la mesure LMBB [Mauclair 2006]. Il s'agit d'une étude préliminaire avant la mise en place du calcul de la probabilité *a posteriori* grâce aux réseaux de confusion. Cette étude montre tout de même qu'un ajout de seulement 28h de données apporte un gain relatif important de 1,69%. Il apparaît que l'ajout de données filtrées atteint un plafond et que tester avec encore plus de données (le corpus non transcrit d'ESTER comporte plus de 1500 h) risque de ne pas apporter de gain significatif car l'ajout concerne des séquences de mots avec un score de confiance élevé qui ont donc été bien modélisées.

Pour l'estimation de modèles acoustiques avec 166h heures dans le corpus d'apprentissage, les calculs ont nécessité 6 ordinateurs cadencés à 3 GHz pendant 3 jours.

6.6 Conclusion

Dans ce chapitre, les mesures de confiance ont été utilisées pour injecter des séquences de mots corrects dans le corpus d'apprentissage des modèles acoustiques. Les performances du SRAP atteignent un taux d'erreur mot de 23,2% soit 2,11% d'amélioration en relatif par rapport au système initial.

La méthode de sélection des séquence des mots corrects utilisée dans [Wessel 2005] et qui proposait un unique seuil d'acceptation est ici modifiée pour permettre un choix de séquence de mots plus intéressant.

L'ajoût de données filtrées atteint un plafond au niveau des performances autour de 35% de données ajoutées et la modification du nombre d'états partagés n'engendre pas une amélioration significative (le WER passe pour 28h filtrées de 23.4% avec 5500 états partagés à 23,2% pour 86h filtrées et 7500 états partagés). Il semble que les séquences de mots ajoutées au corpus initial sont des séquences de mots qui ont déjà un score de confiance élevé et qui sont donc déjà bien modélisées.

6.7 Perspectives

Un autre type d'application des mesures de confiance pour la transcription automatique est l'aide à l'annotation manuelle. Pour l'aide à la transcription manuelle, il existe l'outil Transcriber développé par la DGA[Barras 2001]. Cette application permet de créer des tours de parole des différents locuteurs d'un document sonore, de les transcrire, d'enrichir la saisie de données... Elle permet également de corriger la meilleure hypothèse fournie par un SRAP. Mais

cette correction n'influe pas sur le comportement du SRAP et n'intègre pas d'informations spécifiques générées par le SRAP.

Pour améliorer les fonctionnalités d'un logiciel tel que Transcriber, les mesures de confiance ou encore les sorties du systèmes de RAP peuvent être utilisées. Chacun des mots de la transcription finale fournie par le SRAP du LIUM peut être associé à un score de confiance. Les mots en dessous d'un certain seuil peuvent être mis en exergue pour permettre à un annotateur humain de détecter les zones incorrectes afin de les corriger. Pour cette correction, les autres sorties du SRAP peuvent être utilisées. Un SRAP peut fournir un graphe de mot et un réseau de confusion. Ces sorties peuvent permettre à l'annotateur de connaître les mots en concurrence avec celui pour lequel une erreur a été détectée. L'annotateur pourra alors sélectionner un autre mot susceptible de correspondre à la transcription réelle.

Un exemple de détection de mots incorrects grâce aux mesures de confiance proposées dans ce manuscrit a été mis en place dans [Laurent 2006]. Dans une plate-forme de transcription de réunions, les mots qui ont un score de confiance inférieurs à un seuil sont mis en surbrillance afin d'être facilement repérables. Un annotateur humain peut alors corriger ces zones erronées de la transcription fournie par un SRAP. Aucune expérience quantitative ni qualitative n'a encore permis de mesurer l'impact de ces informations en ce qui concerne la correction des transcriptions automatiques.

Chapitre 7

Identification automatique des segments par nom de locuteurs

Sommaire

De grands ensembles de données audio sont actuellement disponibles mais la plupart d'entre elles n'ont pas de transcription enrichie. Obtenir de façon manuelle une transcription enrichie est très coûteux, particulièrement lorsqu'on cherche à indexer des informations spécifiques telles que le thème principal, les mots-clés, le nom du locuteur, la langue de l'intervenant... Seules des méthodes automatiques génèrent des transcriptions enrichies à moindre coût, mais leur taux d'erreur doit être suffisamment faible pour pouvoir les exploiter.

Dans ce chapitre, nous nous intéresserons uniquement au problème de l'indexation du locuteur. Comme nous l'avons vu dans le paragraphe 2.7.1 du chapitre sur les applications dans le domaine du traitement de la parole, le processus d'indexation automatique (ou *diarization*) résulte uniquement en segments étiquetés anonymement alors que le véritable nom du locuteur est une information importante pour l'enrichissement des transcriptions.

Pour associer le nom complet (prénom, nom) d'un locuteur aux segments issus de la *diarization*, il existe plusieurs méthodes :

- Les méthodes basées sur des informations purement acoustiques : elle reposent généralement sur une reconnaissance automatique du locuteur requérant des échantillons de leur voix pour l'apprentissage des modèles de locuteur [Bimbot 2004].
- Les méthodes basées sur des informations lexicales : elles extraient l'identité du locuteur directement depuis la transcription de l'émission. En effet, dans les émissions radiophoniques, les intervenants se présentent ou présentent le suivant, ils congratulent le précédant ou le suivant, concluent le reportage par leur nom... Le nom du locuteur et une indication sur le moment de son intervention dans l'émission sont souvent présents dans les mots prononcés durant une émission de radio et ces informations peuvent être utilisées pour identifier le locuteur avec son véritable nom. Aucun échantillon de voix n'est ici nécessaire, il suffit de disposer de la transcription. Dans ce manuscrit, nous nous intéresserons plus particulièrement à ces méthodes détaillées par la suite.

Dans de récents travaux [Canseco-Rodriguez 2005], menés sur des émissions radiophoniques en anglais, le LIMSI propose d'utiliser des patrons linguistiques extraits manuellement pour identifier le locuteur d'un segment avec son véritable nom. Durant une émission de radio, le locuteur annonce le nom de la personne qui est en train d'intervenir, qui va intervenir ou qui est déjà intervenu. Le nom prononcé permet d'étiqueter avec la véritable identité du locuteur, le segment courant, le suivant ou le précédent. Une erreur est comptabilisée lorsqu'une identité d'un locuteur a été incorrectement attribuée à un segment. Le taux d'erreur de ce processus basé sur des règles manuelles est d'environ 13% pour les transcriptions manuelles (18% pour les transcriptions automatiques).

Dans [Charad 2005b, Charad 2005a], des patrons linguistiques similaires sont utilisés pour l'indexation en locuteurs de journaux télévisés. Les données sont issues de TREC 2003 (quatre journaux télévisés). Les prédictions fournies par ces différentes règles sont ensuite propagées à tout le document par similarité acoustique. Une identité du locuteur a pu être attribuée à 53% (en termes de durée) du corpus annoté avec 82% d'attribution correcte.

[Tranter 2006] utilise des n-grammes comme patrons linguistiques. Ils sont appris sur un corpus d'apprentissage (données de Hub-4 1996/7 sur des émissions radiophoniques nord-américaines) étiqueté manuellement afin d'indexer des documents radiophoniques. Quand un nom est détecté, son contexte n-gramme est conservé (ici, n vaut 5). Chaque n-gramme qui apparaît plus d'un certain de fois est alors considéré comme une règle linguistique pouvant prédire la localisation du locuteur. Des expériences sont menées sur un corpus transcrit et segmenté de manière manuelle puis automatique (les données proviennent de la campagne NIST RT-04[14]). Les résultats sont calculés en termes rappel à 95% de précision (métriques décrites au chapitre 3 sur les mesures de confiances. Ici, le rappel correspond aux noms correctement étiquetés sur la somme des noms correctement étiquetés, des suppressions et des substitutions. La précision est le ratio entre les noms correctement étiquetés et la somme des noms correctement étiquetés, des insertions et des substitutions. Le rappel sur un corpus de test assez éloigné du corpus d'apprentissage est de seulement 30% pour des transcriptions manuelles et 20% sur des transcriptions automatiques.

Les règles utilisées par [Canseco-Rodriguez 2005] sont manuelles et le contexte utilisé dans [Tranter 2006] se limite à 5 mots. Notre étude propose alors d'apprendre automatiquement les régles d'attribution d'un nom de locuteur à un segment grâce à l'utilisation d'un arbre de classification sémantique en tenant compte d'un contexte plus grand que les 5 mots proposés précédemment.

Le processus utilisé ici est entièrement automatique et répond à deux questions :

1. quand un nom est détecté dans la transcription, à quel segment se réfère-t'il ?
2. à la fin du processus de segmentation, les segments sont étiquetés en fonction de leurs caractéristiques acoustiques avec une étiquette anonyme de locuteur. Étant donné un ensemble de segments ayant les mêmes caractéristiques acoustiques, quel est l'identité de leur locuteur ?

Pour répondre à la première question, une décision locale est proposée grâce à l'utilisation d'un arbre de classification sémantique (SCT) lors de la détection d'un nom dans la transcription afin d'attribuer ce nom à un segment contigu. Le SCT estime des probabilités d'association nom/segment et ces probabilités sont utilisées ici comme score de confiance. Cette attribution

[14] `http://www.nist.gov/speech/test/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf`

est ensuite propagée sur la durée totale de l'émission de radio. Les conflits sont résolus grâce aux scores de confiance provenant des SCT.

Notre étude permet d'évaluer la pertinence de la méthode proposée. En conséquence, les critères de confiance utilisés sont simples et seule une segmentation ainsi qu'une transcription manuelle des références sont utilisées par notre approche. Les erreurs de segmentation et de transcription automatiques réduisent les performances d'un système d'identification du locuteur basé sur le contexte lexical (voir résultats de [Canseco-Rodriguez 2005]). Ces erreurs ne sont pas prises en compte dans notre étude afin de dissocier leur impact de celui de la méthode elle-même. Les corpora utilisés pour l'apprentissage, le développement et l'évaluation proviennent de la campagne d'évaluation française ESTER [Galliano 2005] décrite dans l'introduction de ce manuscrit. Toutes les expressions régulières utilisées sont ici en français mais le processus étant automatique, la méthode est facilement adaptable à une autre langue.

Dans la suite de ce chapitre, les informations concernant le locuteur client de l'application seront présentées. La méthode employée utilisant un arbre de classification sémantique sera décrite en section 7.2. La première étape de la méthode propose d'associer un nom détecté à un segment contigu étiqueté anonymement puis la seconde étape propage le nom associé à un segment à tous les segments portant la même étiquette anonyme. Ces étapes seront notées respectivement décision locale et décision globale. Les premières expériences menées sur des corpora transcrits manuellement seront détaillées section 7.3.

7.1 Informations sur le locuteur

7.1.1 Identité cliente

Les locuteurs présents dans des émissions radiophoniques sont principalement des personnes publiques comme des journalistes, des politiciens, des artistes ou des sportifs. Cette population est facilement identifiable : leurs noms et prénoms sont bien connus, ils sont présents dans plusieurs émissions, et ils correspondent aux locuteurs principaux en termes de temps de parole. Ces locuteurs sont identifiés par leur nom complet dans les conventions de transcriptions d'ESTER et de LDC (Linguistic Data Consortium) pour les évaluations NIST : ce sont les locuteurs à identifier dans la tâche proposée (appelés locuteurs clients).
Une liste d'identités de locuteur est extraite à partir des transcriptions de référence. Plusieurs noms sont ainsi extraits à partir des corpora utilisés dans nos expériences en ne conservant que les noms des personnes publiques. Le procédé de détection du nom du locuteur est basé sur cette liste fermée. Nous avons choisi d'employer le nom complet pour éviter les fausses détections introduites par la méthode de détection : l'ambiguïté induite par l'utilisation de noms partiels

(seulement le prénom ou le nom) pose des problèmes que nous ne résoudrons pas dans notre étude.

7.1.2 Étiquetage des occurrences de noms

Un nom complet détecté dans un segment peut être associé à une des 4 étiquettes suivantes : *current*, *next*, *previous* et *other*. Elles sont données respectivement si le nom détecté se rapporte au locuteur du segment de parole courant, du segment suivant ou du segment précédent . Si ce n'est pas le cas, l'étiquette *other* est attribuée au nom détecté.

7.2 Méthode employée

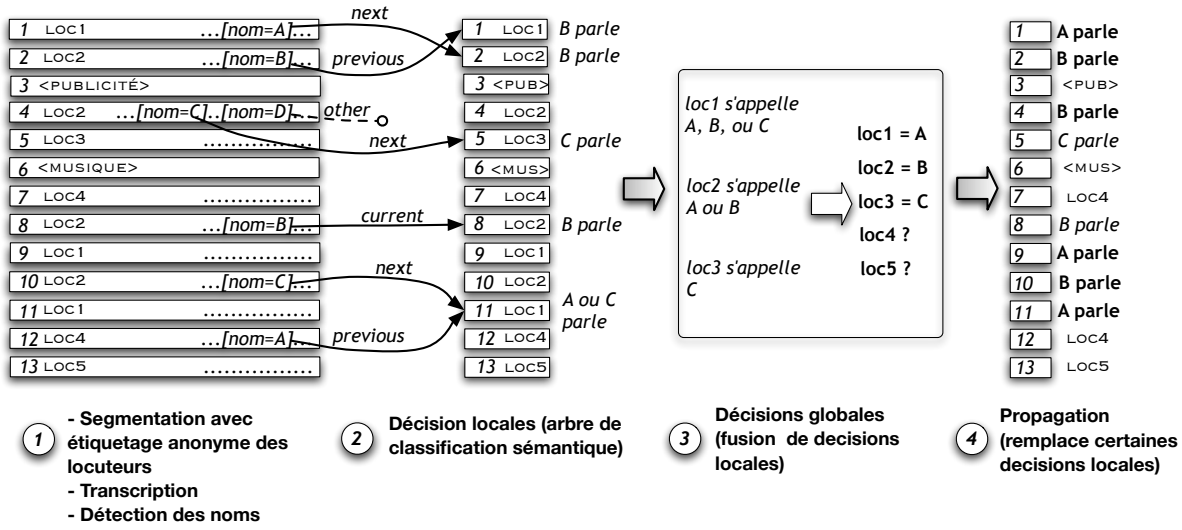


FIG. 7.1 – Identification du locuteur

Après un processus de segmentation, chaque locuteur est normalement associé à une étiquette anonyme. À partir d'un ensemble de segments et de leurs transcriptions, nous proposons d'associer un nom complet à cette étiquette anonyme en deux étapes (figure 7.1, partie ①) :

1. **Analyse du contexte lexical dans chaque segment de parole contenant un nom complet** (figure 7.1, partie ②) : cette étape traite chaque nom complet détecté dans la transcription d'un segment de parole. Elle détermine grâce à un classifieur, si ce nom se rapporte au locuteur précédent, courant, suivant ou à un autre locuteur : seuls les segments contigus du segment où un nom a été détecté dans la transcription peuvent être associés à ce nom. En outre, il est possible que des segments soient associés à des noms différents :

les processus d'association sont effectués sans coopération et peuvent fournir des résultats antagonistes sur un même segment. Plusieurs noms (ou un nom) ainsi que le score de confiance que le classifieur leur a associé sont conservés pour un même segment.

2. **Dénomination du locuteur**(figure 7.1, partie ③et partie ④) : la deuxième étape consiste à fusionner les hypothèses précédentes et choisir le nom qui a le score de confiance provenant du SCT le plus élevé pour assigner un nom complet à l'étiquette anonyme d'un locuteur (figure 7.1, partie ③) Ensuite cette assignation est propagée à tous les segments étiquetés avec cette même étiquette anonyme (figure 7.1, partie ④).

Ces deux étapes sont détaillées dans ce qui suit.

7.2.1 Analyse du contexte lexical

Quand un nom est détecté, le contexte lexical de la transcription est analysé par un arbre de classification sémantique (SCT) [Kuhn 1995]. La classification fournie associe l'étiquette la plus pertinente au nom détecté. Cette étiquette apporte des informations sur le locuteur des segments contigus.

Arbre de classification sémantique

Les SCTs sont souvent utilisés dans le traitement du langage naturel. Par exemple, ils sont employés pour des systèmes de dialogue [Kuhn 1995], pour l'estimation de modèles de langage hiérarchiques *n-grammes* [Estève 2001], ou pour la détection de noms propres inconnus [Béchet 2000]. Les SCTs comme les arbres de décisions décrits au chapitre 3 s'appuient sur les patrons linguistiques qui sont ses paramètres de décision permettant de classer le nom détecté parmi les étiquettes *previous*, *current*, *next* et *other*. Les SCTs reposent sur l'utilisation d'expressions régulières. Afin d'être analysés par le SCT, des couples sont créés comprenant une occurrence de nom complet ainsi que son contexte lexical. Le but est de classer ces couples avec les 4 étiquettes *previous*, *current*, *next* et *other* (voir les feuilles de la figure 7.2) en utilisant les expressions régulières.

Apprentissage de l'arbre

Pendant le processus de construction du SCT, chaque noeud est associé à une expression régulière contenant des mots et des caractères spéciaux (<, > et +). le < (resp. >) se rapporte au commencement (resp. la fin) d'une intervention d'un locuteur tandis que + se rapporte à n'importe quelle suite de mots. Par exemple, l'expression régulière < + *de* + > correspond à chaque intervention contenant le mot *de*, alors que < + *en direct* + *de* + > correspond à chaque

intervention contenant les mots *en direct* et *de* apparaissant dans cet ordre. La figure 7.2 montre une petite partie d'un tel arbre de classification.

Le processus de construction du SCT doit choisir pour chaque noeud l'expression régulière qui réduit au minimum un critère d'impureté. Ici le critère utilisé est le critère de Gini (voir l'article de [Breiman 1984] pour de plus amples détails). Un autre critère communément utilisé est l'entropie de Shannon.

Pour chaque niveau dans l'arbre, ce processus de construction ajoute un mot à l'expression régulière courante. Le critère d'impureté permet d'évaluer le degré de déterminisme associé à un noeud : plus le critère d'impureté est bas, plus la classification est déterministe.

Mesures de confiance utilisées

Chaque feuille de l'arbre donne une probabilité à un nom complet et son contexte lexical pour chaque étiquette possible (ici : *previous*, *current*, *next* et *other*). Ces probabilités sont apprises sur le corpus d'apprentissage utilisé pour la construction de l'arbre. Il s'agit de la probabilité pour le nom et son contexte de désigner un des segments contigus. Ainsi, cette probabilité définit à quel point le nom et son contexte doit être associé au segment courant, précédent, suivant ou à un autre segment. Cette probabilité est assimilée à un score de confiance sur l'appartenance du nom à une de ces quatre étiquettes.

Décisions locales

Pour une occurrence o de nom complet détectée dans un contexte lexical $W_s(o)$ associé à un segment de parole s, le SCT fournit la probabilité $P(t|W_s(o))$ pour chaque étiquette possible t de l'ensemble des étiquettes $T = \{previous, current, next, other\}$. L'étiquette $\delta(o) \in T$ est associée à l'occurrence du nom complet du segment de parole s qui maximise $P(t|W_s(o))$:

$$\delta(o) = \underset{t}{\text{argmax}}\, P(t|W_s(o)) \tag{7.1}$$

Dans notre approche, parmi les 4 étiquettes possibles pour $W_s(o)$, seule l'étiquette $\delta(o)$ est prise en considération pour la suite du processus. Conserver toutes les étiquettes dans la suite du processus n'améliore pas les performances. En outre, si plus d'une étiquette obtient une probabilité égale à $\max_t P(t|W_s(o))$, aucune décision locale n'est prise car aucune étiquette ne semble plus pertinente qu'une autre. Définissons la valeur $\Gamma(o)$ représentant la probabilité finale attribuée à l'occurrence o pour l'étiquette $\delta(o)$.

$$\Gamma(o) = P(\delta(o)|W_s(o)) \tag{7.2}$$

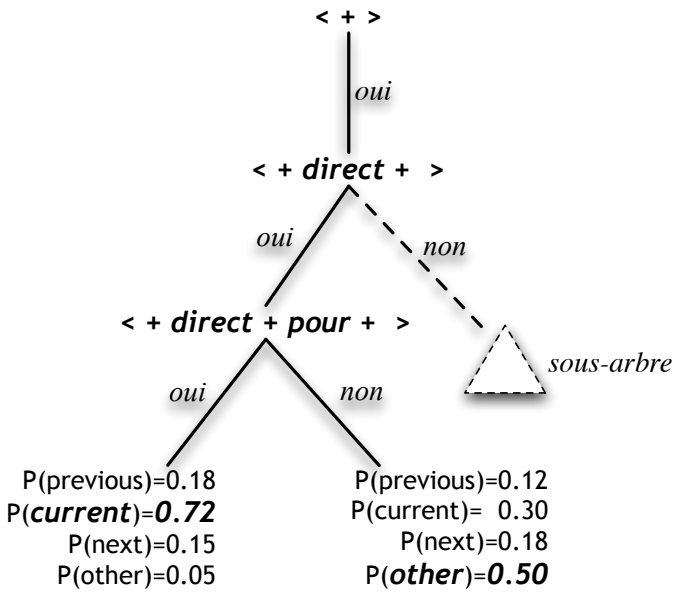


FIG. 7.2 – Exemple d'une partie d'un arbre de classification sémantique : à chaque feuille, une probabilité est associée à chaque étiquette.

7.2.2 Dénomination du locuteur

Soit ψ un locuteur anonyme d'un segment de parole : il s'agit de trouver le vrai nom $N(\psi)$ de ce locuteur.
Chaque segment de parole est associé à un locuteur anonyme (par exemple dans la figure 7.1, le segment 1 est associé à SPK1, ainsi que les segments 9 et 11). En utilisant un arbre de classification sémantique sur les noms complets détectés dans la transcription des segments comme présenté dans la figure 7.2, on obtient, pour chacun des segments, une liste de noms correspondant à l'ensemble des locuteurs possibles (figure 7.1, partie ②).

Fusion des décisions prises par le SCT

Soit K, l'ensemble de tous les noms des locuteurs clients. Soit ν_ψ, l'ensemble des différents noms complets associés à au moins un segment prononcé par ψ grâce à une décision locale du SCT : ν_ψ est alors la liste des noms complets qui sont candidats à ψ et on a $\nu_\psi \subset K$. Soit la fonction $\nu(o)$ qui associe une occurrence o d'un nom complet n à ce nom complet. Nous avons : $\nu(o) = n$. Enfin, soit l'ensemble Ω_ψ des occurrences o qui réfèrent aux segments prononcés par ψ grâce aux décisions locales prises par le SCT.

Pour déterminer le nom $N(\psi)$ d'un locuteur ψ, nous proposons la formule suivante :

$$N(\psi) = \underset{n \in K}{\text{argmax}} \frac{\sum_{(\nu(o)=n) \wedge (o \in \Omega_\psi)} \Gamma(o)}{\sum_{o \in \Omega_\psi} \Gamma(o)} \quad (7.3)$$

$$= \underset{n \in K}{\text{argmax}} \sum_{(\nu(o)=n) \wedge (o \in \Omega_\psi)} \Gamma(o) \quad (7.4)$$

Ainsi, le nom complet qui sera associé à une étiquette anonyme de locuteur est le nom dont les occurrences maximisent la somme des scores donnés par le SCT quand ces occurrences se rapportent à des segments associés à cette même étiquette. Notons que, comme expliqué dans le paragraphe 7.2.1, seules les valeurs associées aux décisions locales valides sont conservées.

7.3 Expériences et résultats

7.3.1 Données

Corpora

Nous avons utilisé les données de la campagne ESTER 2005 pour expérimenter notre approche. Les données (voir tableau 7.1) comportent six radios différentes dont les émissions durent entre 10 et 60 minutes et sont décomposées en 3 corpora.

Le corpus d'apprentissage (*Train*) contient 73h de données (8547 segments) dans lesquels 3297 noms sont détectés. Le corpus de développement (*Dev*) contient 17h (2294 segments) et 920 noms. Le corpus de test (*Test*) contient 10h (1417 segments) et 507 noms, il correspond au corpus d'évaluation officiel d'ESTER. Il contient 2 radios non-présentes dans le corpus d'apprentissage et a été enregistré 15 mois après les autres corpora. En tout, 1007 noms différents ont été détectés dans les différents corpora.

	Train	***Dev***	***Test***
Durée totale (h)	73	17	10
Nombre d'émissions	150	26	18
Nombre de radios	5	5	6
Nombre de segments	8547	2294	1417
Nombre de noms complets détectés	3297	920	507

TAB. 7.1 – Détails sur les corpora : Apprentissage, Développemment & Test provenant de la campagne d'évaluation ESTER.

Le tableau 7.2 montre les probabilités *a priori* des 4 étiquettes sur ces corpora. L'étiquette la plus fréquente est l'étiquette *next* (entre 45% et 49% des corpora). En effet, le plus souvent, les différents locuteurs des émissions annoncent le locuteur qui va parler ensuite.

Étiquettes	***Train* (%)**	***Dev* (%)**	***Test* (%)**
Previous	14,3	12,6	18,6
Current	7,2	7,1	5,3
Next	46,0	45,3	49,3
Other	32,5	35,0	26,8

TAB. 7.2 – Détails sur les corpora : statistiques sur les différentes étiquettes applicables aux noms.

Pré-traitement des différents corpora

Les références (transcriptions enrichies) doivent être transformées et adaptées pour être utilisées avec un arbre de classification sémantique ainsi que pour évaluer les résultats expérimentaux. Les adaptations effectuées sont :

- La définition des 4 étiquettes de noms complets suppose que les locuteurs voisins sont différents du locuteur courant. Les segments contigus contenant le même locuteur sont donc fusionnés pour obtenir une segmentation basée sur les changements de locuteur et non pas sur les phrases (principalement séparées par du silence et des respirations) comme c'est le cas dans la transcription manuelle.

- Les informations concernant les 4 étiquettes doivent être accessibles durant les phases d'apprentissage et d'évaluation. Nous avons donc étiqueté automatiquement la référence en extrayant du flux audio les noms et prénoms des locuteurs. Chaque nom complet extrait est comparé au nom de locuteur associé au segment ainsi qu'aux noms de locuteurs associés aux segments voisins. Cette tâche étant automatisée, nous supposons qu'il n'y a pas d'erreur d'identification du locuteur.

- Dans la transcription de référence, les phrases contiennent plus d'informations que les transcriptions fournies par les systèmes automatiques. Les transcriptions sont ensuite normalisées pour être aussi proches que possible de celles délivrées par les systèmes de transcriptions automatiques.
  Par exemple, toutes les ponctuations sont enlevées ainsi que les majuscules.

- Les articles définis (*le, la, les*) et les articles indéfinis (*un, une, des*) sont enlevés des phrases. En effet, ceux-ci ne semblent pas porteurs d'information.

- Pour généraliser les exemples d'apprentissage pour la construction de l'arbre, chaque nom de locuteur est remplacé par une étiquette générique [FULLNAME].

- Le SCT construit les expressions régulières en tenant compte des mots appartenant aux contextes gauche et droit de l'occurence du nom détecté. Au plus 20 mots pour le contexte gauche et 20 pour le droit sont conservés, ce nombre ayant été fixé empiriquement sur le corpus de développement *Dev* pour maximiser le nombre de bonnes détections locales sur les 4 étiquettes.

Paramètres d'apprentissage du SCT

Les paramètres de l'arbre de classification sémantique sont améliorés sur le corpus de développement. Les principaux paramètres pour l'apprentissage sont le critère de Gini comme critère d'impureté et la taille des feuilles de l'arbre. L'expansion des branches est arrêté lorsque le critère de Gini ne diminue plus ou quand le noeud courant n'est associé qu'à moins de cinq séquences de mots. Le nombre de séquences de mots a été fixé empiriquement sur le corpus de développement.

7.3.2 Étiquetage des segments

	Train (%)	***Dev (%)***	***Test (%)***
Nombre de noms complets détectés	3297	920	507
Étiquetés	94,51	94,78	97,23
Correctement étiquetés sur l'ensemble des noms complets	88.25	76.49	68.76

TAB. 7.3 – Résultats des décisions locales obtenus grâce au SCT sur les différents corpora.
- Étiquetés : % de noms complets détectés pour lesquels une règle de décision locale propose un étiquetage other, current, previous ou next.
- Correctement étiquetés : % de noms complets correctement étiquetés

L'arbre de classification sémantique qui permet d'obtenir les résultats du tableau 7.3 a été construit à partir du corpus d'apprentissage. Le tableau montre les résultats des décisions locales prises sur chaque segment contenant un nom complet sur les corpora *Train*, *Dev* et *Test*.

La première colonne montre les résultats sur les données d'apprentissage utilisées comme des données de test. Les deuxième et troisième colonnes montrent les résultats sur *Dev* et *Test*.

94% des noms complets détectés sur *Dev* et 97% sur *Test* sont associés à une de nos 4 étiquettes. Les noms ne sont pas étiquetés quand l'arbre de décision ne permet pas une classification unique avec une étiquette plus probable que les autres. On conserve uniquement l'étiquette la plus probable car des expériences ont été menées en conservant toutes les étiquettes jusqu'au bout du processus mais celles-ci ont une performance moins importante.

Le taux d'étiquetage correct est d'environ 76,4% sur *Dev* et de seulement 68,7% sur *Test* : ces valeurs peuvent être considérées comme étant la précision de la décision locale sur chaque corpus.
Le taux moins élevé d'étiquetage correct sur le corpus de test (environ 8% de moins que sur *Dev*) peut être expliqué par la présence de deux nouvelles radios. Ces radios (RTM et une radio "surprise") sont absentes du corpus *Dev*. Les données du corpus de test ont en plus été enregistrées 15 mois après les autres. Comme la plupart des émissions enregistrées sont des émissions d'actualité, les intervenants autres que les journalistes de la radio ont changé en fonction des événements en cours. Les faits divers, les actualités des artistes diffèrent généralement d'un mois à l'autre. Les données des corpora de développement et d'apprentissage et plus particulièrement les noms de personnes publiques employés dnas ces corpora sont donc plus proches entre eux qu'avec ceux contenus dans les données de test.

7.3.3 Dénomination du locuteur

Après le processus de segmentation, les segments sont étiquetés avec une étiquette anonyme pour réprésenter le locuteur. Les décisions locales sur les segments sont alors fusionnées pour associer un véritable nom à tous les segments de parole prononcés par le même locuteur (voir paragraphe 7.2.2).

Méthode d'évaluation

L'entrée du système repose actuellement sur les transcriptions manuelles de référence : il n'y a pas d'erreurs sur la transcription, sur l'indexation en locuteur anonyme ou de segmentation parole/non parole. Les frontières des segments de référence et d'hypothèse sont les mêmes, seuls les noms des locuteurs diffèrent car la référence contient les véritables noms des locuteurs tandis que l'hypothèse contient des étiquettes anonymes.
Ici, seuls les locuteurs qui sont des personnes publiques (avec un nom dans la référence) sont les locuteurs clients. L'identité des autres locuteurs ne peut pas être trouvée. Il y a donc erreur

Locuteur	Dénomination	***Train*(%)**	***Dev*(%)**	***Test*(%)**
Client	Correcte	63,68	64,82	66,35
Client	Incorrecte	3,19	5,48	14,36
Client	Incorrecte (non nommé)	15,68	18,19	11,91
Non Client	Correcte (non nommé)	15,50	7,54	3,59
Non Client	Incorrecte	1,95	3,98	3,79
	Total	100	100	100

TAB. 7.4 – Dénomination du locuteur : résultats détaillés pour les différents corpora (les taux sont calculés en termes de durée).
- Locuteur : correspond aux 2 catégories de locuteurs de la référence, ceux qui sont les locuteurs clients de l'application (locuteurs publiques avec un nom complet) et les autres, non clients.
- Dénomination : correspond aux dénominations correctes et incorrectes. "Non nommé" correspond au cas où le processus ne propose pas de nom.

quand le processus donne à un locuteur non-client un nom complet et quand il ne donne pas de nom à un locuteur client (Table 7.4 lignes 2 & 5).

Le processus ne propose pas de nom à un locuteur client dans les cas où :

- aucune décision locale ne concerne un segment de ce locuteur client car aucune décision n'est prise pour toutes les occurrences détectées de ce locuteur ;

- ce nom n'est pas détecté dans la transcription.

Résultats

Le tableau 7.4 montrent les résultats en termes de durée pour la dénomination du locuteur comme c'est le cas lors des évaluations NIST [NIST 2004] pour la tâche d'indexation en locuteurs.

Pour les locuteurs clients, quand les noms hypothèses et ceux de la référence sont les mêmes, la dénomination est correcte (Table 7.4 ligne 1). Pour les locuteurs non-clients, quand le processus ne propose pas de noms pour eux, la dénomination peut être également considérée comme correcte (Table 7.4 ligne 4).

Quand le processus attribue un nom à un segment, il est correct dans 66% des cas sur le corpus de test. Il semble tout de même raisonnable de considérer comme étant correct de ne pas associer un nom complet (nom de locuteur client) à des segments de parole prononcée par des personnes qui ne sont pas publiques. Le processus de dénomination de locuteur atteint 72% de décisions correctes en terme de durée ($64.82\% + 7.54\%$) sur le corpus *Dev* et environ 70% ($66.35\% + 3.59\%$) sur *Test* (voir tableau 7.4).

La différence sur le taux de dénomination correcte entre les corpora *Dev* et *Test* est d'environ 2%, soit moins que les 8% observés pour les décisions locales correctes entre ces deux même corpus (voir tableau 7.3). Cela signifie que même s'il y a moins de décisions locales dans le corpus *Test*, ces décisions sont pertinentes pour trouver le nom complet d'un locuteur client à associer au segment.

7.4 Conclusion

Dans le contexte de la transcription enrichie, nous proposons une méthode totalement automatisée qui permet d'identifier les locuteurs par leur véritable identité. Cette identité est extraite directement depuis la transcription. Pour se focaliser sur la validité de la méthode, les transcriptions et la segmentation sont celles de référence.
Le procédé proposé est basé sur l'utilisation d'un arbre de classification sémantique qui permet d'étiqueter les occurences de noms détectées : cette première étape consiste à prendre des décisions locales qui associent une telle occurrence à un segment de parole. La segmentation automatique associe les segments à une étiquette anonyme pour spécifier le locuteur. Les résultats obtenus sont alors fusionnés pour associer un nom complet à tous les segments d'un même locuteur qui étaient anonymement annotés.
Les expériences sont menées sur des émissions radiophoniques françaises, fournies par la campagne d'évaluation ESTER. Environ 70% de la durée totale des émissions est correctement indexée en locuteur pour chacun des corpora de développement et d'évaluation. Sur le corpus d'évaluation, 18,15% de la durée totale des émissions est incorrectement indexée et aucune décision n'est prise pour 11,91%.
Le but principal est atteint : les résultats valident la méthode proposée de dénomination du locuteur à partir d'une segmentation et d'une transcription manuelle.

7.5 Perspectives

Les perspectives de travail s'orientent vers l'utilisation d'une segmentation et d'une transcription automatiques dans lesquelles des erreurs interviendront. Les mesures de confiance proposées au chapitre II s'insèreront dans le processus d'identification du locuteur. En effet, ces mesures permettent de déterminer les zones erronées dans la transcription founie, qualité intéressante à utiliser lorsque la transcription et la segmentation deviendront automatiques. Si les erreurs de transcription portent sur le nom complet détecté, les mesures de confiance permettront d'apporter une information sur la décision à prendre lors de l'étiquetage et la propagation de ce nom.

Une fois les erreurs de transcription et de segmentation détectées, les mesures de confiance peuvent également intervenir dans les décisions locales (attribution d'une étiquette à un segment contigu) et globales (propagation des décisions locales sur tout le document) concernant l'identification du locuteur. Actuellement, les scores provenant de l'arbre de décision sont utilisés comme mesures de confiance pour définir la meilleure étiquette possible pour une occurrence de nom complet détecté dans une transcription. Une des pistes est d'étudier un processus d'intégration des mesures de confiance proposées au chapitre II pour les incorporer dans les calculs permettant des prises de décisions locales et / ou globales.

Pour ce système d'identification du locuteur, un système de détection d'entités nommées sera également nécessaire afin d'extraire les noms de locuteur directement depuis la transcription.

Conclusion et perspectives

Le travail de thèse présenté dans ce manuscrit s'inscrit dans le cadre de la transcription enrichie de documents sonores.

Les expériences ont été menées sur les corpora fournis lors de la campagne d'évaluation ESTER consacrée au traitement d'émissions radiophoniques d'actualités francophones.

Le système de reconnaissance automatique de la parole du LIUM qui a atteint la seconde position de la campagne ESTER avec un score de 23,7% en termes de taux erreur mots est le système sur lequel se sont appuient les expérimentations.

Dans ce contexte, nous avons étudié et élaboré différentes mesures de confiance ainsi que quelques possibilités d'applications de ces mesures pour le traitement automatique de la parole.

Une mesure de confiance associée à une hypothèse de reconnaissance est une estimation de la fiabilité de cette hypothèse. Une bonne mesure de confiance peut ainsi permettre de repérer les zones de mots corrects/incorrects dans la transcription fournie par un SRAP.

Cette capacité de détection a été appliquée pour le filtrage de données pour l'apprentissage non supervisé de modèles acoustiques. Des scores de confiance ont également été utilisés pour l'identification du locuteur pour associer un nom complet détecté dans la transcription à un segment de parole.

1 Mesures de confiance utilisées

Les mesures de confiance proposées et élaborées sont issues du SRAP du LIUM. Elles sont de trois types :

- La première est basée sur une mesure acoustique déjà connue qui exploite la distance introduite par l'utilisation d'un modèle de langage et d'un dictionnaire de prononciations entre les scores acoustiques données par un modèle acoustique contraint et ce même modèle utilisé sans contrainte. Nous avons proposé une normalisation de cette mesure qui nous a permis de la combiner avec d'autres.
- La deuxième mesure (appelée LMBB) provient du modèle de langage et prend en compte son comportement de repli (*backoff*) lors du décodage par le SRAP qui a fourni l'hypothèse évaluée.
- La troisième s'appuie sur le calcul de la probabilité *a posteriori* d'un mot issue d'un réseau de confusion. Celui-ci représente l'espace de recherche du SRAP où les mots en compétition au même instant sont alignés. La probabilité *a posteriori* d'un mot regroupe intrinséquement les scores acoustiques et linguistiques du SRAP.

Ces trois mesures prises séparément apportent une information sur la pertinence d'un mot et permettent par exemple de repérer des zones erronées ou correctes dans l'hypothèse fournie par un SRAP.

Cependant, en les combinant, il est possible de proposer une nouvelle mesure de confiance plus performante. La meilleure combinaison de ces mesures, que nous appelons PAP/LMBB, ne fait pas intervenir la mesure basée sur les scores acoustiques d'un SRAP : cette dernière mesure semble contenir des informations redondantes avec celles contenues dans la probabilité *a posteriori* d'un mot. La mesure LMBB, elle, ne fait pas intervenir le score linguistique tel que celui contenu dans la probabilité *a posteriori* d'un mot mais apporte une information supplémentaire.

Nos expériences montrent que la mesure PAP/LMBB améliore les capacités des mesures prises séparément, en particulier pour détecter les zones de mots erronés et les zones de mots corrects.Nous avons utilisé les métriques classiques d'évaluation des mesures de confiance : la NCE (entropie croisée normalisée) qui mesure l'information additionnelle portée par la mesure, et le CER (*confidence error rate* qui évalue la capacité de la mesure à distinguer les mots incorrects des mots corrects à partir d'un seuil optimisé. La probabilité *a posteriori* (PAP) est la mesure de confiance la plus couramment utilisée. En termes de CER, la PAP obtient un taux de 18,64% sur le corpus de test officiel de la campagne ESTER, alors que la mesure PAP/LMBB que nous proposons permet de réduire ce taux à 18,27%. En termes de NCE, une procédure de mapping est généralement nécessaire pour optimiser la PAP : sur ce même corpus de test, sans mapping la PAP obtient une valeur de NCE de 0,187 tandis qu'avec mapping, elle obtient une valeur de NCE de 0,293. La mesure PAP/LMBB obtient sans mapping une valeur de 0,284 qui atteint 0,299 avec mapping. Ainsi, la mesure PAP/LMBB obtient de meilleurs résultats que la mesure de confiance de référence dans la littérature.

2 Applications

Cette faculté de détection des zones de mots erronés/corrects est utilisée dans une première application où des séquences de mots corrects sont recherchées afin d'être ajoutées au corpus d'apprentissage des modèles acoustiques d'un SRAP.

Collecte de corpus automatique Nous avons ajouté 86h de données audio transcrites automatiquement à un corpus initial de 80h d'enregistrements audio transcrits manuellement. Ces 86h sont le fruit d'un filtrage de 558h. Nous avons proposé une méthode de filtrage qui utilise les mesures de confiance et qui intègre des contraintes de temps et de double seuillage. En ajoutant ces 86h de données filtrées aux 80h du corpus d'apprentissage initial des modèles acoustiques, le taux d'erreur sur les mots est réduit de 0,5% en absolu (soit 2,11% en relatif) sur le corpus de test.

Identification nommée du locuteur Les scores de confiance provenant d'un arbre de classification sémantique au sujet de décisions locales ont également été utilisés dans un processus d'identification automatique du locuteur. Cette méthode s'appuie sur les transcription et les résultats d'une segmentation automatique du signal pour lesquels les segments sont étiquetés de manière anonyme pour spécifier leur locuteur.

Nos expériences, qui étaient destinées à étudier la faisabilité de la tâche, utilisent des transcriptions manuelles pour évaluer la méthode.

À partir d'une liste fermée de locuteurs "cibles" (principalement des hommes ou femmes publiques : journalistes, politiciens, sportifs, artistes, ...), des noms de personnes sont détectés dans les transcription du flux audio. À partir du contexte lexical, un arbre de classification permet d'associer localement chaque nom au segment précédent, suivant, courant ou à aucun segment. Cette association est guidée par une probabilité définissant la pertinence de classification à tel ou tel segment. Cette probabilité est utilisée ici comme score de confiance. Les noms associés aux segments d'un même locuteur jusque-là anonyme sont ensuite confrontés à l'aide des scores de confiances auxquels ces noms sont associés. Enfin, lorsque cela est possible, les étiquettes anonymes caractérisant les locuteurs et fournies par le système de segmentation sont associés au nom véritable de la personne qui parle.

Nos expériences montrent la faisabilité de notre approche, automatique, qui permet d'identifier correctement les locuteurs pour 70% du corpus de test en termes de durée.

3 Perspectives

Dans un premier temps, nos expériences sur la collecte automatique de corpus ont montré que les mesures de confiance peuvent aider à déterminer des zones fiables parmi les hypothèse de reconnaissance d'un SRAP. Ceci peut être très utile pour des applications ayant besoin d'utiliser les sorties d'un SRAP. Par exemple, les transcriptions automatiques comportent des erreurs : les mesures de confiance peuvent permettre d'aider un annotateur humain à repérer les zones où une transcription automatique comporte des erreurs. Cet annotateur humain pourrait corriger plus aisément ces erreurs si en plus de l'hypothèse la plus probable, le SRAP lui fournissait les différents mots en compétition avec les mots erronés. Ce type d'application a fait l'objet d'un stage de Master en 2006 utilisant le SRAP du LIUM et les mesures de confiance développées dans ce mémoire de thèse : ce stage devrait déboucher sur de nouveaux travaux de recherche dans le cadre d'une thèse qui étudiera, outre les aspects d'aide à la correction de transcription automatique, les possibilités d'auto-adaptation d'un SRAP en fonction des erreurs corrigées par un annotateur humain.

Nos travaux sur l'identification nommée du locuteur ont été effectués sur des transcriptions manuelles. Or, il est évident que les erreurs de transcription rendront plus difficiles la tâche d'identification nommée. Il faudra donc certainement adapter la méthode proposée dans ce mémoire en prenant en compte, par exemple, les mesures de confiance fournies par le SRAP pour chacun des mots hypothèses.

Les mesures de confiance sont intéressantes à utiliser tout au long du processus de transcription enrichie. Elles peuvent certes permettre l'amélioration de la transcription, mais elles apportent surtout une valeur ajoutée évidente pour l'exploitation des sorties d'un SRAP. Certaines composantes de la tâche de transcription enrichie, comme la détection de mots-clés, d'entités nommées ou la détection de thèmes sont donc à adapter pour prendre en compte les mesures de confiances du SRAP utilisé. Dans d'autres domaines, les mesures de confiance associées à des hypothèses de transcriptions peuvent aussi être très utiles, comme par exemple les systèmes de dialogue oral homme/machine pour lesquels les mesures de confiance peuvent aider le gestionnaire de dialogue dans sa gestion de la conduite de l'interaction avec l'usager.

Dans le cadre du projet régional MILES, le LIUM, en collaboration avec le LINA (Laboratoire d'Informatique de Nantes-Atlantique), va travailler à la mise en place et à l'expérimentation d'un moteur de questions/réponses sur des données audio. Ces travaux ne pourraient pas être menés efficacement sans l'utilisation de mesures de confiance : les apports de cette thèse seront donc manifestes dans les prochains travaux du LIUM en ce qui concerne le traitement automatique de la parole.

Bibliographie personnelle

Conférences d'audience internationale avec comité de sélection

– J. Mauclair, S. Meignier, Y. Estève (2006), *Speaker diarization : about whom the speaker is talking ?*, Dans : IEEE Odyssey 2006, 28-30 Juin 2006, San Juan (Porto Rico, USA).

– J. Mauclair, Y. Estève, S. Petit-Renaud, P. Deléglise (2006), *Automatic detection of well recognized words in automatic speech transcription*, Dans : LREC 2006, 24-26 Mai 2006, Gênes (Italie).

– J. Mauclair, J. Pinquier (2004), *Fusion of descriptors for speech / music classification*, Dans : EUSIPCO'2004, 06-10 Septembre 2004, Vienne (Autriche).

Conférences d'audience nationale avec comité de sélection

– J. Mauclair, S. Meignier, Y. Estève (2006), *Indexation en locuteur : utilisation d'informations lexicales*, Dans : JEP'06, 12-16 juin 2006, Dinard (France).

– J. Mauclair, Y. Estève, P. Deléglise(2006), *Probabilité a posteriori : amélioration d'une mesure de confiance en reconnaissance de la parole*, Dans : JEP'06, 12-16 juin 2006, Dinard (France).

– J. Mauclair, J. Pinquier (2004), *Fusion de paramètres en classification Parole/Musique*, Dans : JEP'04, Avril 2004, Fès (Maroc).

– J. Mauclair, J. Pinquier, R. André-Obrecht(2003), *Fusion de paramètres en classification Parole/Musique/Bruit*, Dans : RJC'03, Septembre 2003, Grenoble (France).

– J. Pinquier, J. Mauclair, J.L. Rouas, R. André-Obrecht (2003), *Détection de la parole et de la musique : fusion de deux approches*, Dans : GRETSI'03, 8-11 Septembre 2003, Paris (France).

Bibliographie

[Abberley 1999] Abberley D., Renals S., Robinson T. et Ellis D., The thisl sdr system at trec, dans *Text Retrieval Conference (TREC-8)*, 1999.

[Ajmera 2003] Ajmera J. et Wooters C., A robust speaker clustering algorithm, dans *Proc. of ASRU, Automatic Speech Recognition and Understanding*, pages 411–416, St. Thomas, U.S. Virgin Islands, Novembre 2003.

[Allauzen 2003] Allauzen A., *Modélisation liguistique pour l'indexation automatique de documents audiovisuels*, Thèse de doctorat, LIMSI, 2003.

[Allauzen 2004] Allauzen A. et Gauvain J.-L., Construction automatique du vocabulaire d'un système de transcription, dans *Proc. of JEP, Journées d'Etudes sur la Parole*, Fès, Maroc, Mai 2004.

[Anastasakos 1996] Anastasakos T., McDonough J., Schwartz R. et Makhoul J., A compact model for speaker adaptation training, dans *Proc. of ICSLP, International Conference on Spoken Language Processing*, pages 1137–1140, Philadelphie, USA, 1996.

[Barras 2001] Barras C., Geoffrois E., Wu Z. et Liberman M., Transcriber : development and use of a tool for assisting speech corpora production, *Speech Communication*, 33(1-2) :5–22, Janvier 2001.

[Barras 2006] Barras C., Zhu X., Meignier S. et Gauvain J., Multi-stage speaker diarization of broadcast news, *IEEE Transactions on Audio, Speech and Language Processing*, 14(5) :1505–1512, Septembre 2006.

[Baum 1972] Baum L., An inequality and associated maximization technique in statistical estimation of probabilistic functions of markov processes, *Inequalities*, 3 :1–8, 1972.

[Béchet 2001] Béchet F., LIA_PHON un système complet de phonétisation de texte, dans *Traitement Automatique Des Langues*, volume 42, pages 47–68, Hermès, 2001.

[Béchet 2000] Béchet F., Nasr A. et Genet F., Tagging unknown proper names using decision trees, dans *38th Annual Meeting of the Association for Computational Linguistics*, pages 77–84, Hong Kong, Chine, Octobre 2000.

[Ben 2004] Ben M., Betser M., Bimbot F. et Gravier G., Speaker diarization using bottom-up clustering based on a parameter-derived distance between GMMs, dans *Proc. of ICSLP, International Conference on Spoken Language Processing*, Jeju, Korea, Octobre 2004.

[Benayed 2003] Benayed Y., Fohr D., Haton J. et Chollet G., Confidence measures for keyword spotting using suport vector machines, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 588–1591, Hong Kong, Chine, Mai 2003.

[Bengio 2004] Bengio S. et Mariethoz J., The expected performance curve : a new assesment measure for person authentification, dans *Odyssey, The Speaker Language Recognition Workshop*, Tolède, Espagne, Mai 2004.

[Billi 1997] Billi R. et Lamel L., RailTel : Railway telephone services, *Speech Communication*, 23(1-2) :63–65, 1997.

[Bimbot 2002] Bimbot F., *Reconnaissance de la parole*, chapitre Reconnaissance automatique du locuteur, pages 85–120, Traité IC2, Hermès, 2002.

[Bimbot 2004] Bimbot F., Bonastre J.-F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska D. et Reynolds D. A., A tutorial on text-independent speaker verification, *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, 4 :430–451, 2004.

[Boite 2000] Boite R., Bourlard H., Dutoit T., Hancq J. et Leich H., *Traitement de la parole*, Collection électricité, Presses Polytechnique et Universitaires Romandes, 2000.

[Bonastre 2003] Bonastre J., Morin P. et Junqua J., Gaussian Dynamic Warping (gdw) method applied to text-dependant speaker detection and verification, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 2013–2016, Genêve, Suisse, Septembre 2003.

[Bourlard 1994] Bourlard H. et Morgan N., *Connectionist Speech Recognition : a hybrid approach*, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, USA, 1994.

[Bouwman 2000] Bouwman G., Boves L. et Koolwaaij J., Weighting phone confidence measures for automatic speech recognition, dans *COST249 Worshop on Voice Operated Telecom Services*, pages 59–62, Ghent, Belgique, 2000.

[Breiman 1984] Breiman L., Friedman R., Olshen R. et Stone C., *Classification and Regression Trees*, Wadsworth, 1984.

[Burges 1998] Burges C., *A tutorial on support vector machines for pattern recognition*, volume 2, Data Mining and Knowledge Discovery, 1998.

[Caelen 1981] Caelen J. et Caelen G., Indices et propriétés dasn le projet ARIAL II, dans *Séminaire GALF*, 1981.

[Calliope 1989] Calliope, *La Parole et Son Traitement Automatique*, Collection technique et scientifique des télécommunications, Masson, 1989.

[Canseco-Rodriguez 2005] Canseco-Rodriguez L., Lamel L. et Gauvain J.-L., A comparative study using manual and automatic transcriptions for diarization, dans *Proc. of ASRU, Automatic Speech Recognition and Understanding*, San Juan, Porto Rico, USA, Novembre 2005.

[Cardeilhac 1995] Cardeilhac F. et Palisson F., Système de navigation carin, dans *SIA/FIEV/EQUIP'AUTO*, Paris, France, Octobre 1995.

[Chan 2004] Chan A., Sherwani J., Ravishankar M. et Rudnicky A., Four-level categorization scheme of fast GMM computation techniques in large vocabulary continous spech recognition systems, dans *Proc. of ICSLP, International Conference on Spoken Language Processing*, pages 689–692, Jeju, Korea, Octobre 2004.

[Charad 2005a] Charad M., Moraru D., Ayache S. et Quénot G., Speaker identity indexing in audio-visual documents, dans *CBMI, Content-Based Multimedia Indexing*, Riga, Latvia, Juin 2005a.

[Charad 2005b] Charad M. et Quénot G., Approche par patrons linguistiques pour la détection automatique de l'identité du locuteur : Applications à l'indexation par le contenu des journaux télévisés, dans *CORESA, Compression et Représentation des Signaux Audiovisuels*, Rennes, France, Novembre 2005b.

[Charlet 2001] Charlet D., Mercier G. et Jouvet D., On combining confidence measures for improved rejection of incorrect data, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Aalborg, Danemark, Septembre 2001.

[Chen 2004] Chen L., Lamel L. et Gauvain J.-L., Lightly supervised acoustic model training using consensus network, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Montréal, Canada, Mai 2004.

[Chen 1996] Chen S. et Goodman J., *An Empirical Study of Smoothing Techniques for Language Modelling*, Morgan Kaufmann Publishers, San Fransisco, USA, 1996.

[Chen 1998] Chen S. et Gopalakrishnan P., Speaker, environement and channel change detection and clustering via the bayesian information criterion, dans *DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, Landsdowne, USA, Février 1998.

[Cohen 1981] Cohen J., Segmenting speech using dynamic programming, *J. Acoust. Soc. Am.*, 69(5) :1430–1438, Mai 1981.

[Cornuéjols 2002] Cornuéjols A. et Miclet L., *Apprentissage artificiel : concepts et algorithmes*, Eyrolles, 2002.

[Cox 2002] Cox S. et Dasmahapatra S., High-level approaches to confidence estimation in speech recognition, *IEEE Transactions on Speech and Audio Processing*, 10(7), 2002.

[Crépy 1997] Crépy H., Marcadet J. et Waast C., Dictée à grand vocabulaire en français : IBM VoiceType 3.0, un produit de la recherche, dans *1ères JST FRANCIL*, pages 19–23, 1997.

[Davis 1952] Davis K., Biddulph R. et Balashek S., Automatic recognition of spoken digits, *J. Acoust. Soc. Amer.*, 24 :637–642, 1952.

[Delacourt 2000a] Delacourt P., *La segmentation et le regroupement par locuteurs pour l'indexation de document audio*, Thèse de doctorat, ENST-Eurocom, Juillet 2000a.

[Delacourt 2000b] Delacourt P. et Wellekens C., Distbic : a speaker-based segmentation for audio data indexing, *Special Issue of Speech Communication on Accessing Information in Spoken Audio*, 32(1-2) :111–126, Septembre 2000b.

[Deléglise 2005] Deléglise P., Estève Y., Meignier S. et Merlin T., The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbonne, Portugal, Septembre 2005.

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

[Derouault 1993] Derouault A. M., Keppel E., Fusi S., Marcadet J. C. et Janke E., The IBM Speech Server Series and ots application in Europe, dans *Joint ESCA-NATO/RSG.10 tutorial and research workshop on applications of speech technology*, Lautrach, Allemagne, Septembre 1993.

[Digalakis 1995] Digalakis V., Rtischev D. et Neumeyer L., Speaker adaptation unsing constrained estimation of gaussian mixture, *IEEE, Trans. Speech and Audio Processing*, 3 :357–366, 1995.

[Dolmazon 1997] Dolmazon J., Bimbot F., Adda G., El Beze M., Caerou J., Zeiliger J. et Adda-Decker M., Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale, dans *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 13–18, Avignon, Frannce, Avril 1997.

[Doshita 1965] Doshita S., *Study on the analysis and recognition of japanese speech sounds*, Thèse de doctorat, Université de Kyoto, Japon, 1965.

[Dubois 1992] Dubois D. et Prade H., On the relevance of non-standard theories of uncertainty in modeling and pooling expert opinions, *Reliability Engineering and System Safety*, 36 :95–107, 1992.

[Dubois 1994] Dubois D. et Prade H., La fusion d'informations imprécises, *Traitement du Signal*, 11(6) :95–107, 1994.

[Duchateau 2002a] Duchateau J., Demuynck K. et Wambacq P., Confidence scoring based on backward language models, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 221–224, Orlando, USA, Mai 2002a.

[Duchateau 2002b] Duchateau J. et Wambacq P., Unconstrained versus constrained acoustic normalisation in confidence scoring, dans *Proc. of ICSLP, International Conference on Spoken Language Processing*, volume 3, pages 1617–1620, Denver, USA, Septembre 2002b.

[Estève 2003] Estève Y., Raymond C., De Mori R. et Janiszek D., On the use of linguistic consistency in systems for human-computer dialogues, *IEEE Transactions on Speech and Audio Processing*, 11(6) :746–756, 2003.

[Estève 2001] Estève Y., Béchet F., Nasr A. et De Mori R., Stochastic finite state automata language model triggered by dialogue states, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 1, pages 725–728, Aalborg, Denmark, Septembre 2001.

[Evermann 2000] Evermann G. et Woodland P., Posterior probability decoding, confidence estimation and system combination, dans *Speech Transcription Workshop*, 2000.

[Falavigna 2002] Falavigna D., Gretter R. et Riccardi G., Acoustic and word lattice based algorithms for confidence scores, dans *Proc. of ICSLP, International Conference on Spoken Language Processing*, pages 1621–1624, Denver, USA, Septembre 2002.

[Federico 1998] Federico M. et De Mori R., *Language modelling*, Academic Press, 1998.

[Franco 1992] Franco H., Cohen M., Morgan N., Rumelhart D. et Abrash V., Hybrid neural network/hidden markov model continuous-speech recognition, dans *Proc. of ICSLP, International Conference on Spoken Language Processing*, pages 915–918, 1992.

[Freund 1995] Freund Y. et Schapire R., A decision-theoric generalization of on-line learning and an application to boosting, dans *Computational Learning Theory : Eurocolt'95*, pages 23–37, New-York, USA, 1995.

[Fu 2005] Fu Y. et Du L., Combination of multiple predictors to improve confidence measure based on local posterior probabilities, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 93–96, Philadelphie, USA, Mars 2005.

[Fujisaki 1987] Fujisaki H., Overview of the japanese natinal project on advanced man-machine interface through spoken language, dans *European Conference on Speech Technology*, Edimbourg, Écosse, Septembre 1987.

[Fürgen 2006] Fürgen C., Kolss M., Paulik M., Stüker S., Schultz T. et Waibel A., Open domain speech translation : from seminar and speeches to lectures, dans *Proc. of JEP, Journées d'Etudes sur la Parole*, pages 281–288, Dinard, France, 2006.

[Gales 1997] Gales M., Maximum likelihood linear transformation for HMM-based speech recognition, Rapport technique, Cambridge University Engineering Department, Mai 1997.

[Gales 1998] Gales M., Maximum likelihood linear transformations for hmm-based speech recognition, *Computer Speech and Language*, 12 :75–98, 1998.

[Gales 1992] Gales M. et Young S., An improved approach to hidden markov model decomposition of speech and noise, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 233–296, San Fransisco, USA, Mars 1992.

[Galliano 2005] Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J. et Gravier G., The ESTER phase II evaluation campaign for the rich transcription of french broadcast news, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbonne, Portugal, Septembre 2005.

[Garofolo 2004] Garofolo J., Laprun C. et Fiscus J., The rich transcription 2004 spring meeting recognition evaluation, dans *RT 2004 spring meeting recognition workshop*, Montréal, Canada, Mai 2004.

[Gauvain 1994] Gauvain J.-L. et Lee C., Maximum a posteriori estimation for multivariate gaussain mixture observations of markov chains, *IEEE, Trans. Speech and Audio Processing*, 2(2) :291–298, Avril 1994.

[Gerlach 1993] Gerlach M. et Onken R., Speech input/output as interface devices for communication between aircraft pilots and the pilot assistant system "cassy", dans *Joint ESCA-NATO/RSG.10 tutorial and research workshop on applications of speech technology*, Lautrach, Allemagne, Septembre 1993.

[Gillick 1997] Gillick L., Ito Y. et Young J., A probabilistic approach to confidence estimation and evaluation, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 879–883, Munich, Allemagne, Avril 1997.

[Gish 1991] Gish H., Siu M.-H. et Rohlicek R., Segregation of speakers for speech recognition and speaker identification, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 873–877, Toronto, Canada, Mai 1991.

[Golub 1989] Golub G. et Van Loan C., *Matrix Computations*, Johns Hopkins University Press, 1989.

[Gravier 2004] Gravier G., Bonastre J.-F., Galliano S. et Geoffrois E., The ESTER evaluation campaign of rich transcription of french broadcast news, dans *LREC, Language Evaluation and Resources Conference*, Lisbonne, Portugal, Mai 2004.

[Guo 2004] Guo G., Huang C., Jiang H. et Wang R.-H., A comparative study on various confidence measures in large vocabulary speech recognition, dans *Proc. of the fourth International Symposium on Chinese Language Processing*, Hong Kong, Chine, 2004.

[Hazen 2002] Hazen T., Seneff S. et Polifroni J., Recognition confidence scoring and its use in speech understanding system, *Computer speech and language*, 16(1) :49–67, 2002.

[Henández-Ábrego 2000] Henández-Ábrego G. et Mariño J., Contextual confidence measures for continuous speech recognition, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1803–1806, Istanbul, Turquie, Juin 2000.

[Hersher 1972] Hersher M.-B. et Cox R.-B., An adaptative isolated word speech recognition system, *Speech Commnucation and Processing*, 1972.

[Higgins 1991] Higgins A., Bahler L. et Porter J., Speaker verifiaction using randomized phrase prompting, *Digital Signal Processing*, 1 :89–106, 1991.

[Janez 1996] Janez F., *Fusion de sources d'informations définies sur des référentiels non exhaustifs différents*, Thèse de doctorat, Université d' Angers, Novembre 1996.

[Jelinek 1976] Jelinek F., Continuous speech recognition by statisical methods, *IEEE Proceedings*, 64(4) :532–556, 1976.

[Jelinek 1997] Jelinek F., *Statistical methods for speech recognition*, MIT Press, Cambridge, USA, 1997.

[Jiang 2005] Jiang H., Confidence measures for speech recognition : a survey, *Speech Communication Journal*, 45 :455–470, 2005.

[Kamppari 2000] Kamppari S. et Hazen T., Word and phone level acoustic confidence scoring, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turquie, Juin 2000.

[Kemp 1997] Kemp T. et Schaaf T., Estimating confidence using word lattices, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Rhodes, Grèce, 1997.

[Kemp 1999] Kemp T. et Waibel A., Unsupervised training of a speech recognizer : Recent experiments, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 2725–2728, Budapest, Hongrie, Septembre 1999.

[Kuhn 1995] Kuhn R. et De Mori R., The application of semantic classifciation trees to natural language understanding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5) :449–460, 1995.

[Lamel 2000] Lamel L., Gauvain J.-L. et Adda G., Lightly supervised acoustic model training, pages 150–154, 2000.

[Lamel 2002] Lamel L., Gauvain J.-L. et Adda G., Lightly supervised and unsupervised acoustic model training, *Computer Speech and Language*, 16 :115–129, 2002.

[Lamel 1999] Lamel L., Rosset S., Gauvain J. et Bennacef S., The LIMSI ARISE system for train travel information, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 501–504, Phoenix, USA, 1999.

[Laurent 2006] Laurent A., *Évaluation en reconnaissance de la parole*, Rapport de DESS, Université du Maine, Juin 2006.

[Lee 2001] Lee C.-H., Statistical confidence measures and their applications, dans *Proc. of ICSP*, pages 1021–1028, Daejon, Corée du Sud, Août 2001.

[Lee 1989] Lee K.-F., Hon H.-W., Hwang M.-Y., Mahajan S. et Reddy R., The sphinx speech recognition system, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 445 – 448, Mai 1989.

[Leggetter 1995] Leggetter C. et Woodland P., Maximum likelihood linear regression for speaker adaptation of continuous density hidden markiv models, *Computer Speech and Language*, 9 :171–185, 1995.

[Leray 2000] Leray P., Zaragoza H. et d'Alché Buc F., Pertinence des mesures de confiance en classification, dans *RFIA*, Paris, France, Février 2000.

[Lesser 1975] Lesser W.-R., Organization of the hearsay II speech understanding system, *IEEE Transactions ASSP*, 23 :11–23, 1975.

[Lowerre 1976] Lowerre B.-T., The harpy speech recognition system, Rapport technique, Carnegie Mellon University, 1976.

[Maison 2001] Maison B. et Gopinath R., Robust confidence annotation and rejection for continuous speech recognition, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, Mai 2001.

[Mangu 2000] Mangu H., Brill E. et A. S., Finding consensus in speech recognition : Word error minimization and other applications of confusion networks, *Computer Speech and Language*, 14(4) :373–400, 2000.

[Mariani 2002a] Mariani J., *Analyse, Synthèse et Codage de la Parole*, Traité IC2, Hermès, 2002a.

[Mariani 2002b] Mariani J., *Reconnaissance de la Parole*, Traité IC2, Hermès, 2002b.

[Mauclair 2006] Mauclair J., Estève Y., Petit-Renaud S. et Deléglise P., Automatic detection of well recognized words in automatic speech transcriptions, 2006.

[Meignier 2002] Meignier S., *Indexation en locuteurs de documents sonores : segmentation d'un document et appariement d'une collection*, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, France, Novembre 2002.

[Meignier 2006] Meignier S., Moraru D., Fredouille C., Bonastre J.-F. et Besacier L., Step-by-step and integrated approaches in broadcast news speaker diarization, *Computer Speech and Language*, 20(2-3) :303–330, 2006.

[Merlin 2004] Merlin T., *AMIRAL, une plateforme générique pour la reconnaissance automatique du locuteur-de l'authentification à l'indexation*, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, France, Novembre 2004.

[Metze 2000] Metze F., Kemp T., Schaaf T., Schultz T. et Soltau H., Confidence measure based language identification, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turquie, Juin 2000.

[Moreau 2001] Moreau N., *Détermination d'un degré de confiance en reconnaisssance automatique de la parole par estimation de rapports de vraisemblance au niveau des trames acoustiques*, Thèse de doctorat, France Télécom R&D - DIH/IPS, Lannion, Juin 2001.

[Moreau 2000] Moreau N., Charlet D. et Jouvet D., Confidence measure and incremental adaptation for the rejection of incorrect data, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turquie, Juin 2000.

[Moreno 2001] Moreno P., Logan B. et B. R., A boosting approach for confidence scoring, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 2109–2112, Aalborg, Danemark, Septembre 2001.

[Mostefa 2006] Mostefa D., Hammon O. et Choukri K., Evaluation of automatic speech recognition and speech language translation within TC-STAR : Results from the first evaluation campaign, dans *LREC, Language Resources and Evaluation Conference*, Mai 2006.

[NIST 2004] NIST, Fall 2004 rich transcription (RT-04F) evaluation plan, August 2004, `http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf`.

[NIST04] NIST04, NIST speaker recognition 2004 evaluation plan, 2004, http ://www.nist.gov/speech/tests/spk/2004.

[Néel 2002] Néel F. et Minker W., *Reconnaissance de la parole*, chapitre Reconnaissance automatique du locuteur, pages 18–216, Traité IC2, Hermès, 2002.

[Pallett 1996] Pallett D., Fiscus J., Fisher W. et Garofolo J., 1995 HUB-3 NIST multiple microphone corpus benchmark tests, dans *Proc. of ARPA Speech Recognition Workshop*, New York, USA, Février 1996.

[Panaget 1998] Panaget F., Le générateur de langue naturelle de l'agent dialoguant ARTIMIS, *Traitement automatique des langues*, 39(2) :107–126, Décembre 1998.

[Pastor 1993] Pastor D. et Gulli C., D.I.V.A.(DIalogue Vocal pour Aeronef) performances in simulated aircraft cockpit, dans *Joint ESCA-NATO/RSG.10 tutorial and research workshop on applications of speech technology*, Lautrach, Allemagne, Septembre 1993.

[Rabiner 1989] Rabiner L., Tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77(2) :257–285, 1989.

[Rabiner 1993] Rabiner L. et Juang B.-H., *Fundamentals of speech recognition*, Prentice Hall, 1993.

[Rahim 1997] Rahim M. G. et Lee C.-H., Discriminative utterance verification for connected digit recongnition, *IEEE Trans. on Speech and Audio Processing*, 5(3) :266–277, 1997.

[Rahman 1999] Rahman A. et Fairhust M., A novel confidence-based framework for multiple expert decision fusion, dans *BMVC*, volume 2, pages 205–213, 1999.

[Ravishankar 1997] Ravishankar M., Bisiani R. et Thayer E., Sub-vector clustering to improve memory and speed performance of acoustic likelihood computation, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 1, pages 151–154, Rhodes, Grèce, Septembre 1997.

[Raymond 2005] Raymond C., *Décodage conceptuel : co-articulation des processus de transcription et compréhension dans les systèmes de dialogue*, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, France, Décembre 2005.

[Raymond 2004] Raymond C., Béchet F., De Mori R., Damnati G. et Estève Y., Automatic learning of interpretation strategies for spoken dialogue systems, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 425–428, Montréal, Canada, 2004.

[Reynolds 2003] Reynolds D., Andrews W., Campbell J., Navratil J., Peskin B., Adami A., Jin Q., Klisacek D., Abramson J., Mihaescu R., Godfrey J., Jones D. et Xiang B., The SuperSID project : exploiting high-level information for high-accuracy speaker recognition, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 784–787, Hong Kong, Chine, Mai 2003.

[Rose 2001] Rose R. C., Yao H., Riccardi G. et Wright J., Integration of utterance verification with statistical language modeling and spoken language understanding, *Speech Communication*, 34 :321–331, 2001.

[Rudnicky 1999] Rudnicky A., Thayer E., Constantinides P., Tchou C., Shern R., Lenzo K., Xu W. et Oh A., Creating natural dialogs in the Carnegie Mellon Communicator System, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 4, pages 1531–1534, Budapest, Hongrie, Septembre 1999.

[Sadek 1997] Sadek M., Bretier B. et Panaget F., ARTIMIS : Natural dialogue meets rational agency, dans *Proc. of IJCAI*, pages 1030–1035, Nagoya, Japon, 1997.

[San-Segundo 2001] San-Segundo R., Pellom B., Hacioglu K., Ward W. et Pardo J., Confidence measures for spoken dialogue systems, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, Mai 2001.

[Schaaf 1997] Schaaf T. et Kemp T., Confidence measures for spontaneous speech recognition, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 875–878, Munich, Allemagne, Avril 1997.

[Simon 1983] Simon J., *La reconnaissance des formes par algorithmes*, Dunod, 1983.

[Simonin 1998] Simonin J., Delphin-Poulat L. et Damnati G., Gaussian Density Tree Structure in a Multi-Gaussian HMM-Based Speech Recognition System, dans *Proc. of ICSLP, International Conference on Spoken Language Processing*, 1998.

[Siroux 1995] Siroux J., Guyomard M., Jolly Y., Multon F. et Rémondeau C., Speech and tactile-based GEORAL system, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 18–21, Madrid, Espagne, 1995.

[Siu 1999] Siu M. et Gish H., Evaluation of word confidence for speech recognition systems, *Computer Speech and Language*, 13(4) :299–319, Octobre 1999.

[Siu 1997] Siu M., Gish H. et Richardson F., Improved estimation, evaluation and applications of confidence measures for speech recognition, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 2, pages 831–834, Rhodes, Grèce, Septembre 1997.

[Steinbiss 1993] Steinbiss V., Ney H., Haed-Umbach R., Iran B.-H., Essen U., Kneser R., Oerder M., Meier H.-G., Aubert X., Dugast C., Geller D., Hollerbauer W. et Bartosik H., The philips research system for large-vocabulary continuous-speech recognition, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 2125–2128, Berlin, Allemagne, Septembre 1993.

[Stemmer 2002] Stemmer G., Steidl S., Nöth E., Niemann H. et Batliner A., Comparison and combination of confidence measures, dans Sojka P., Kopecek I. et Pala K., rédacteurs, *Text, Speech and Dialogue, Proceedings of the Fifth International Conference on Text, Speech, Dialogue - TSD 2002* , volume 2448 de *Lecture Notes in Artificial Intelligence* , pages 181–188, Springer-Verlag, 2002.

[Stern 1986] Stern P., Eskenasi M. et Memmi D., An expert system for speech spectrogram reading, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, 1986.

[Stolcke 1997] Stolcke A., Konig Y. et Weintraub M., Explicit word error minimization in N-best list rescoring, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 1, pages 163–166, Rhodes, Grèce, 1997.

[Sukkar 1996] Sukkar R., Setlur A., Rahim M. et Lee C.-H., Utterrance verification of keywords strings using word-based minimum verification error(wb-mve) training, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, USA, Mai 1996.

[Sun 2003] Sun H., Zhang G., Zheng F. et Xu M., Using word confidence measure for oov words detection in a spontaneous spoken dialog system, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 2, pages 2713–2716, Genêve, Suisse, Septembre 2003.

[Tranter 2006] Tranter S. E., Who really spoke when ? finding speaker turns and identities in broadcast news audio, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 1013–1016, Toulouse, France, Mai 2006.

[Tsiporkova 2000] Tsiporkova E., Vanpoucke F. et Van Hamme H., Evaluation of various confidence-based rejection strategies for isolated word rejection, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turquie, Juin 2000.

[Tubach 1970] Tubach J.-P., *Reconnaissance automatique de la parole : Étude et réalisation fondées sur les niveaux acoustique, morphologique, syntaxique*, Thèse d'état, Université de Grenoble, France, 1970.

[Uhrik 1997] Uhrik C. et Ward W., Confidence metrics based on n-gram language model backoff behaviors, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Rhodes, Grèce, Septembre 1997.

[Vergyri 2000] Vergyri D., Use of word level side information to improve speech recognition, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1823–1826, Istanbul, Turquie, Juin 2000.

[Vicens 1969] Vicens P., *Aspects of speech recognition by computer*, Thèse de doctorat, Université de Stanford, USA, 1969.

[Viterbi 1967] Viterbi A., Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*, 13(3) :260–269, Avril 1967.

[Waibel 2004] Waibel A., Steusloff H. et Stiefelhagen R., CHIL computers in the human interaction loop, dans *International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbonne, Portugal, 2004.

[Weintraub 1997] Weintraub M., Beaufays F., Rivlin Z., Konig Y. et Stolcke A., Neural-network based measures of confidence for word recognition, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 887–890, Munich, Allemagne, Avril 1997.

[Wessel 1999] Wessel F., Macherey K. et Ney H., A comparison of word graph and n-best list based confidence measures, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 315–318, Budapest, Hongrie, Septembre 1999.

[Wessel 1998] Wessel F., Macherey K. et Schlüter R., Using word probabilities as confidence measures, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 225–228, Seattle, USA, Mai 1998.

[Wessel 2005] Wessel F. et Ney H., Unsupervised training of acoustic models for large vocabulary continous speech recognition, *IEEE Transactions on Speech and Audio Processing*, 13 :23–31, 2005.

[Wessel 2001] Wessel F., Schlüter R., Macherey K. et Ney H., Confidence measures for large vocabulary continuous speech recognition, *IEEE Transactions on Speech and Audio Processing*, 9(3) :288–298, Mars 2001.

[Wessel 2000] Wessel F., Schlüter R. et Ney H., Using posterior probabilities for improved speech recognition, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 1587–1590, Istanbul, Turquie, Juin 2000.

[Willsky 1976] Willsky A. et Jones H., A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems, *IEEE Transcations on Automatic Control*, 21(1) :108–112, Février 1976.

[Woods 1976] Woods W., Bates M., Brown G., Bruce B., Cook C., Klovstad J., Mak-houl J., Nash-Webber B., Schwartz R., Wolf J. et Zue V., Speech understanding systems, Rapport technique, B.B.N. Cambridge, USA, 1976.

[Young 1994] Young S., Detecting misrecognitions and out-of-vocabulary words, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 21–24, Adélaide, Australie, Avril 1994.

[Young 1997] Young S., Adda-Decker M., Aubert X., Dugast C., Gauvain J., Kershaw D., Lamel L., Leeuwen D., Pye D., Robinson A., Steeneken H. et Woodland P., Multilingual

large vocabulary speech recognition : the european SQALE project, *Computer Speech and Language*, 11(1) :73–89, Janvier 1997.

[Zhang 2001] Zhang R. et Rudnicky A., Word level confidence annotation using combinations of features, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, pages 2105–2108, Aalborg, Danemark, Septembre 2001.

[Zue 1989] Zue V., Glass J. et Seneff S., Acoustic segmentation and phonetic classification in the Summit system, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, 1989.

Résumé

Ce travail de thèse se place dans le contexte de la campagne d'évaluation ESTER (Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophonique).

L'objectif de ce travail de thèse consiste dans un premier temps à proposer des indicateurs ou mesures de confiance permettant de distinguer les zones correctes ou erronées au sein des hypothèses de reconnaissance fournies par un système de transcription automatique de la parole (STAP).

Dans un second temps, deux types d'applications utilisant des mesures de confiance sont appréhendées :

- la collecte de corpora de transcriptions automatiques fiables alignés sur de la parole enregistrée, par exemple pour augmenter la taille des corpora d'apprentissage disponibles pour l'estimation de modèles acoustiques ;
- l'identification nommée automatique du locuteur, qui consiste à extraire des transcriptions automatiques des noms de locuteurs et à les associer aux étiquettes anonymes utilisées par un système d'indexation.

Trois mesures de confiance seront proposées, une mesure de confiance dérivée des sorties acoustiques du STAP, une mesure de confiance basée sur le repli du modèle de langage et une mesure de confiance provenant de la probabilité a posteriori d'un mot. A l'aide de métriques d'évaluation de mesures de confiance, ces trois mesures sont comparées et la meilleure combinaison des trois est calculée. Cette combinaison permet d'accroître les performances de chacun des trois mesures dans la détection des zones erronées ou correctes.

La première application vise à accroître les performances du STAP utilisé par le LIUM en augmentant de manière non supervisée les données d'apprentissage des modèles acoustiques. Ainsi, les zones de parole ayant un degré de confiance élevé dans un deuxième corpus transcrit automatiquement sont prélevées et ajoutées au corpus d'apprentissage initial transcrit manuellement. Grâce à cet ajout, les performances du SRAP sont significativement améliorées en termes de taux d'erreur sur les mots.

Enfin, dans le cadre de l'identification du locuteur d'un document sonore, des scores de confiance sont utilisés pour déterminer le nom du locuteur directement à partir de la transcription. Environ 70% de la durée totale des émissions est correctement indexée en locuteur sur un corpus de test.

Mots-clés: Mesures de confiance, Fusion, Reconnaissance automatique de la parole, Traitement automatique de la parole, Apprentissage non-supervisé, Identification du locuteur