Traitement automatique de la parole : contributions

Yannick Estève

24 août 2009

Table des matières

Table des figures

Liste des tableaux

Glossaire

BLEU Bilingual Evaluation Understudy

CER Confidence Error Rate

CMLLR Constrained Maximum Likelihood Linear Regression

CMU Carnegie Mellon University

CTrain ...

ESTER Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques

EPAC Projet ANR : Exploration de masse de documents audio pour l'extraction et le traitement de la parole conversationnelle

HMM Hidden Markov Model (modèle caché de Markov)

IRISA Institut de recherche en informatique et systèmes aléatoires.

IRIT Institut de Recherche en Informatique de Toulouse

LI Laboratoire d'informatique de l'Université de Tours

LIA Laboratoire d'informatique de l'Université d'Avignon et des Pays de Vaucluse

LIUM Laboratoire d'Informatique de l'Université du Maine

LMBB Language Model Back-off Behavior

MAP Maximum A Posteriori

MPE Minimum Phone Error

NCE Normalized Cross Entropy

PAP Probability *a posteriori*

SAT Speaker Adaptative Training

TAP Transcription automatique de la parole

WER Word Error Rate

Chapitre 1

Introduction

Ces travaux de recherche s'inscrivent dans le traitement automatique de la parole, en particulier au niveau des couches hautes de la transcription automatique. Dans cette partie introductive, il est certainement opportun pour situer mon travail de présenter brièvement le domaine du traitement automatique de la parole, son évolution, ses avancées, ses limites : ces points seront abordés dans la section qui suit.

Ensuite, il me semble intéressant de revenir sur la période qui précède mon recrutement comme Maître de conférences à l'Université du Maine. Il s'agira alors de résumer l'ensemble des travaux de recherche que j'ai effectués durant ma thèse et en période post-doctorale. Les travaux effectués depuis mon recrutement au Laboratoire d'Informatique de l'Université du Maine (LIUM) seront alors abordés : ils constitueront par la suite le sujet principal de ce document.

Les citations bibliographiques de cette partie introductive ne concernent que les publications auxquelles j'ai participé. La numérotation utilisée fait référence à la bibliographie présentée à la fin de ce chapitre. Pour le reste du document, les citations se réfèreront à la bibliographie générale présentée en fin de document.

1.1 Le traitement automatique de la parole

Le traitement automatique de la parole ne se limite certainement pas à la transcription automatique de la parole. Cette expression fait référence à l'ensemble des analyses, études, manipulations de la parole pouvant être effectuée par des processus automatiques. Se cache ainsi derrière cette expression, par exemple, l'ensemble des processus d'extraction automatique d'information provenant de la parole, à différents niveaux : détection de la parole dans un enregistrement audio, segmentation d'un enregistrement audio en zones de parole acoustiquement homogènes, identification du locuteur, détection du sexe du locuteur, détection de mots ou d'expressions (*word-spotting*), transcription automatique, etc.

Pour ma part, j'ai essentiellement travaillé dans le domaine de la reconnaissance automatique de la parole : d'abord dans un contexte applicatif de dialogue oral homme-machine, et ensuite en transcription automatique à très grand vocabulaire, en particulier pour le traitement d'émissions radiophoniques.

Lorsque j'ai commencé à travailler dans le domaine de la reconnaissance au-

tomatique de la parole, l'approche markovienne avait déjà supplanté les approches analytiques. Jusqu'à présent, rien n'a réussi à remettre en cause l'hégémonie de cette approche. Ce n'est donc pas la nature des systèmes de reconnaissance automatique de la parole qui a évolué ces onze dernières années, mais plutôt le contexte général de développement et d'exploitation de ces systèmes.

La performance et la robustesse des systèmes markoviens de reconnaissance de la parole, par essence probabilistes, sont très dépendants de la taille et de la diversité des données disponibles pour l'apprentissage des modèles acoustiques et linguistiques qui les composent. Or, c'est justement la quantité des données numériques disponibles qui a explosé ces dernières années. En particulier, le développement extrêmement rapide d'Internet et des capacités d'acquisition numérique, de stockage et de diffusion met à notre disposition un ensemble de données multimédia qui peuvent devenir soit sources de connaissance, soit objets d'études. Parallèlement à cela, la puissance des ordinateurs, en terme de vitesse de calcul, de capacité mémoire, de stockage ou de vitesse de communication n'a cessé de croître. Cet ensemble de facteurs a permis le développement et l'utilisation de nouveaux algorithmes (apprentissage discriminant par exemple) et l'exploitation de systèmes de reconnaissance de la parole plus performants et plus rapides.

Mais il reste encore de nombreux progrès à faire : en réalité, malgré la multitude de données audio disponibles, très peu sont exploitables pour l'apprentissage des modèles acoustiques puisqu'il est encore nécessaire, pour cela, de disposer d'une transcription correcte et donc généralement réalisée par un annotateur humain. Or ce processus d'annotation manuelle est tellement coûteux que le nombre de corpus audio transcrits manuellement est finalement très rare. Comme une des limites les plus contraignantes des systèmes markoviens est le peu de disposition à une trop grande variabilité entre les données d'apprentissage et les données d'exploitation, le manque de données audio transcrites manuellement constitue un frein au développement de systèmes de reconnaissance de la parole appliqués à de nouveaux types de documents audio ou d'environnements sonores.

Nous reviendrons, dans ce mémoire, sur ces différentes problématiques. Je présenterai également certains travaux qui dépassent le cadre de la reconnaissance de la parole mais qui restent du domaine du traitement automatique de la parole. Ces travaux exploitent les résultats fournis par des systèmes de reconnaissance automatique de la parole pour différent types d'analyse, comme la caractérisation et détection de la parole spontanée, ou encore l'identification nommée du locuteur.

1.2 1998-2003 : modélisation du langage, LIA, CNET

J'ai commencé mes travaux de recherche en 1998, au cours de mon stage de DEA au Laboratoire d'Informatique de l'Université d'Avignon (LIA), sous la direction de Thierry Spriet, Maître de conférences en informatique, et Marc El Bèze, Professeur. Le sujet de ce stage, la "désambiguïsation des homophones" est, onze ans plus tard, toujours d'actualité malgré les différents apports de la communauté scientifique du domaine.

Ces travaux, qui concernaient principalement le domaine de la modélisation du langage et en particulier la combinaison de modèles n-gram avec des grammaires

régulières, ont donné lieu à une publication dès 1999 (1) et ont été un des axes principaux de mes travaux de thèse (2). Cette thèse, financée par une bourse du Ministère de la Recherche, co-dirigée par le Professeur Renato De Mori (LIA) et Frédéric Béchet, alors Maître de conférences au LIA, a été menée dans le cadre d'une convention avec le CNET (qui deviendra durant cette thèse France Telecom R&D puis, plus tard, Orange Labs).

Jusqu'à mon recrutement comme Maître de conférences par l'Université du Maine en 2003, mes travaux de recherche ont été menés dans le cadre de cette convention entre le CNET et le LIA. Il s'agissait alors, dans un premier temps, de modélisation du langage pour la reconnaissance de la parole en contexte de dialogue. Ces travaux ont fait l'objet de plusieurs publications nationales (3; 4) et internationales (5; 6; 7; 8; 9; 10).

À la suite de ces travaux, et dans le cadre d'une nouvelle convention entre France Telecom R&D et le LIA, j'ai participé, comme post-doctorant, au développement d'une approche intégrée de décodage conceptuel des sorties d'un système de reconnaissance de la parole, toujours aux cotés de Frédéric Béchet et Renato De Mori. Il s'agissait de proposer une méthode d'interprétation sémantique, sous forme d'extraction de séquences de concepts propres à l'application de dialogue concernée, basée sur l'analyse d'un graphe de mots plutôt que sur l'analyse, forcément réductrice, de l'hypothèse de reconnaissance la plus probable (11). L'utilisation de mesures de confiance acoustiques, linguistiques, et conceptuelles fut aussi abordée pour l'interprétation sémantique (12). Ces travaux ont ensuite été poursuivis par le LIA et France Telecom R&D dans le cadre du projet européen LUNA qui s'est achevé en août 2009. Pour ma part, ayant été recruté à l'Université du Maine en 2003, j'ai arrêté de travailler sur ce sujet. Cet arrêt n'était que temporaire, puisque comme nous le verrons plus loin dans ce document, certains mes travaux de recherche m'amènent maintenant à me pencher de nouveau sur la problématique de l'interprétation sémantique des sorties d'un système de reconnaissance de la parole dans un contexte de dialogue.

1.3 2003-2009 : le traitement automatique de la parole au LIUM

Par la suite, ce mémoire traitera des travaux entrepris depuis mon arrivée au LIUM. Cette approche me semble la plus propice pour juger de mon potentiel pour diriger des recherches étant donné que c'est depuis mon recrutement en tant que Maître de conférences que je peux réellement participer aux choix, proposer et co-encadrer des actions de recherche.

Je suis arrivé au LIUM en septembre 2003. C'était une période propice pour développer une nouvelle thématique de recherche au sein du LIUM, puisque les travaux menés par le Professeur Paul Deléglise, sur la fusion d'informations audio et vidéo pour la reconnaissance de phonèmes par analyse des mouvements labiaux et du signal audio, venaient d'atteindre une évolution telle qu'il était alors nécessaire soit de collecter de nouvelles données de travail, soit de changer de thématique pour continuer à proposer des travaux de qualité qui puissent être reconnus par la communauté.

1.3.1 Reconnaissance de la parole

Après réflexion, nous avons décidé de nous investir dans le développement d'un système de reconnaissance automatique de la parole. En particulier, nous avons profité de la préparation de la campagne d'évaluation ESTER sur les systèmes de transcription automatique d'émissions radiophoniques en français pour nous y inscrire en tant que participants. En particulier, cela nous a permis de pouvoir récupérer l'ensemble des données audio et textuelles distribuées par les organisateurs et de pouvoir développer notre système.

Cette décision permettait de fédérer nos compétences, au niveau de la modélisation acoustique, de la modélisation du langage et des algorithmes de décodage, tout en nous permettant d'acquérir des données qui allaient devenir le matériau d'étude privilégié de nos travaux pendant quelques années. Ainsi, nous avions un objectif à court terme (la participation à la campagne ESTER), des données (ce qui dans le domaine du traitement de la parole est vital et très précieux) et ... nos compétences et notre motivation. L'objectif de la participation à la campagne ESTER était également de s'insérer plus fortement dans la communauté nationale du traitement automatique de la parole.

Notons également qu'à ce moment-là nous avons eu l'opportunité, Paul Deléglise et moi-même, de co-encadrer la thèse de Julie Mauclair. Ce n'était pas une situation idéale pour débuter une thèse, mais malgré ce contexte difficile où tout était à construire, elle a réussi à développer ses travaux sur les mesures de confiance, à publier dans des conférences internationales et à soutenir sa thèse en trois ans.

En 2004, un an après mon recrutement, Sylvain Meignier était également recruté comme Maître de conférence à l'Université du Maine. Ses compétences, en particulier en segmentation et regroupement en locuteur, étaient complémentaires de celles alors en place au LIUM, et furent d'un renfort particulièrement important pour mener à bien le projet de développement d'un système de transcription automatique de la parole.

Nous reviendrons plus longuement dans ce document sur les travaux sur ce système, présentés dans (13), (14) et (15), mais on peut dire d'ors et déjà que la stratégie mise en place en 2003 fut bénéfique. Le LIUM est actuellement un acteur visible et reconnu au niveau national dans la communauté du traitement automatique de la parole ; il commence également à se développer au niveau international, grâce à ses publications bien entendu, mais aussi à la diffusion de logiciels sous licence libre et à la mise en place de collaborations internationales.

Mes apports dans le domaine de la transcription automatique ne se limitent pas à la participation au développement de systèmes à l'état de l'art performants. J'ai également participé à différentes propositions pour en améliorer les performances. J'ai ainsi travaillé sur les mesures de confiance (16; 17) : ces résultats nous ont ensuite permis de travailler sur la combinaison de systèmes en collaboration avec le LIA (18), mais aussi avec l'IRISA (19) : ces travaux ont débouché sur le projet ANR ASH qui regroupe le LIA, l'IRISA et le LIUM et dont je suis le coordonateur. J'ai également participé aux travaux d'Antoine Laurent sur la phonétisation automatique (20; 21; 22), ainsi qu'aux travaux de Richard Dufour sur le post-traitement des hypothèses de reconnaissance pour corriger des erreurs spécifiques à l'aide d'approches spécifiques (23).

1.3.2 Identification nommée du locuteur

Ayant à disposition un système de segmentation et regroupement en locuteurs et un système de transcription automatique de la parole, Sylvain Meignier et moi avons naturellement fait converger nos travaux vers l'exploitation conjointe des sorties de ces systèmes.

Un système de segmentation et regroupement en locuteurs permet de découper automatiquement un enregistrement audio en portions acoustiquement homogènes. Ces portions, les segments, sont regroupés automatiquement en fonction des locuteurs : chaque locuteur potentiel est identifié par une étiquette anonyme, et est associé à un ensemble de segments.

L'idée d'aller chercher dans les transcriptions automatiques les noms des locuteurs, en particulier lorsque les fichiers audio traités sont des enregistrements d'émissions radiophoniques où naturellement les différents locuteurs sont présentés, ou se présentent, était dans l'air depuis quelques temps. Nous avons le mérite d'avoir été les premiers à mettre en œuvre et proposer une méthode automatique qui permet d'associer à des locuteurs, jusqu'alors anonymes et représentés par de simples étiquettes discriminantes, des prénoms et des noms de famille (24; 17). Cette approche est appelée 'identification nommée du locuteur'. Ce nom ne relève pas la particularité de la méthode qui n'utilise aucune connaissance acoustique ou linguistique a priori sur les locuteurs.

Nous travaillons encore sur cette approche qui est l'objet de la thèse de Vincent Jousse, co-encadrée par Sylvain Meignier, Christine Jacquin (MCF, LINA) et Béatrice Daille (PR, LINA). Ces travaux sont maintenant également soutenus par plusieurs chercheurs du LIUM. Notre approche a été améliorée (25; 26; 27) et s'avère très performante par rapport à d'autres approches qui ont depuis été proposées par d'autres laboratoires (28).

1.3.3 Parole conversationnelle

Dans le cadre du projet ANR EPAC dont j'assure la coordination, une partie des travaux du LIUM ces trois dernières années s'est orientée vers le traitement de la parole conversationnelle. Ces travaux ont consisté à construire un corpus de transcriptions manuelles de 100h de parole conversationnelle. Ce travail a été effectué par Thierry Bazillon qui, dans le même temps, prépare une thèse en linguistique sur le codage de la parole conversationnelle que je co-encadre avec Daniel Luzzati, Professeur en linguistique. Durant ce travail, nous avons pu mesurer l'apport de la transcription automatique pour assister la transcription manuelle en terme de gain de productivité (29; 30), établir un inventaire des outils informatiques pouvant aider à la transcription manuelle, mesurer leur interopératibilité (31) et proposer ainsi un état des lieux sur la problématique de la transcription manuelle de la parole spontanée (32).

Les 100h d'enregistrements audio à partir desquelles les transcriptions manuelles ont été produites proviennent d'un ensemble de plus de 1700h d'enregistrements d'émissions radiophoniques. Or, ces émissions ne sont pas constituées uniquement de parole conversationnelle. Sachant que nous étions particulièrement intéressés par de la parole non préparée, peu présente lors des émissions d'actualités par exemple,

nous avons développé une approche permettant de caractériser et d'extraire automatiquement les segments de parole spontanée (33; 34). Nous reviendrons dans ce mémoire sur le projet EPAC pour en décrire plus précisément le contenu.

1.4 Structuration du document

Ce mémoire est composé de deux parties principales, pour répondre à la double attente d'un mémoire d'habilitation de recherche. Dans un premier temps, je présente un bilan de mes travaux de recherche les plus récents, parmi ceux que je viens de résumer dans la section précédente. Dans la seconde partie, je montre comment j'ai participé à l'administration de la recherche à travers des responsabilités dans des projets de recherche labellisés et j'expose les encadrements de stagiaires et co-encadrements de doctorants que j'ai effectués.

Les travaux de recherche que j'ai choisi de mettre en valeur dans la première partie sont d'abord les travaux concernant le développement d'un système de transcription automatique, qui constitueront le chapitre 2 de ce document (le chapitre 1 étant cette partie introductive). Le chapitre 3 est consacré aux travaux sur les mesures de confiance. Les propositions sur l'utilisation de modèles spécifiques pour la correction d'erreurs spécifiques de transcription automatique sont exposées dans le chapitre 4. Les travaux sur l'identification nommée du locuteur sont présentés dans le chapitre 5. Le chapitre 6 est consacré au traitement de la parole conversationnelle. Dans le chapitre 7, j'expose mes premiers travaux dans le domaine de la traduction automatique (35; 36) avec le Professeur Holger Schwenk. Enfin, dans le chapitre 8 je reviens sur les participations du LIUM à différentes campagnes d'évaluation, en transcription automatique de la parole comme en traduction automatique.

Dans la seconde partie, je présenterai le projet Parole du LIUM et reviendrai sur mes responsabilités au sein de l'évolution scientifique du projet. Je présenterai également mon travail de coordination du projet ANR EPAC et ma responsabilité scientifique au sein du projet ANR PORT-MEDIA. Je présenterai également des projets de recherche qui débutent cette année. D'abord le projet ANR ASH, que je vais coordonner, puis les projets en traduction automatique auxquels je participe : le projet européen EuroMatrixPlus et le projet ANR COSMAT.

Enfin, j'exposerai une conclusion sur l'ensemble de mes travaux pour ensuite indiquer l'orientation probable de mes contributions futures.

Bibliographie personnelle

[1] Alexis Nasr, Yannick Estève, Frédéric Béchet, Thierry Spriet, and Renato De Mori, "A language model combining n-grams and stochastic finite state automata," in *Eurospeech*, Budapest, Hongrie, 1999, vol. 5, pp. 2175–2178.

[2] Yannick Estève, *Intégration de sources de connaissances pour la modélisation stochastique du langage appliquée à la parole continue dans un contexte de dialogue oral homme-machine*, Ph.D. thesis, Université d'Avignon et des Pays de Vaucluse, 2002.

[3] Yannick Estève, Frédéric Béchet, and Renato De Mori, "Sélection dynamique de modèles de langage dans une application de dialogue," in *JEP*, Aussois, France, 2000, pp. 185–188.

[4] Frédéric Béchet, Yannick Estève, and Renato De Mori, "Modèles de langage hiérarchiques pour les applications de dialogue en parole spontanée," in *TALN*, Tours, France, 2001, pp. 327–332.

[5] Yannick Estève, Frédéric Béchet, and Renato De Mori, "Dynamic selection of language models in a dialog system," in *ICSLP*, Pékin, Chine, 2000, pp. 214–217.

[6] Frédéric Béchet, Yannick Estève, and Renato De Mori, "Tree-based language model dedicated to natural spoken dialogs systems," in *ISCA TRW on Adaptation methods for speech recognition*, Sophia-Antipolis, France, 2001.

[7] Yannick Estève, Frédéric Béchet, Alexis Nasr, and Renato De Mori, "Stochastic finite state automata triggered by dialogue states," in *Eurospeech*, Aalborg, Danemark, 2001, pp. 725–728.

[8] Yannick Estève, Christian Raymond, and Renato De Mori, "On the use of structures in language models for dialogue, specific solutions for specific problems," in *ISCA TRW on Multi-modal dialogue in mobile environments*, Kloster Irsee, Allemagne, 2002.

[9] Renato De Mori, Yannick Estève, and Christian Raymond, "On the use of structures in language models for dialogue," in *ICSLP*, Denver, Colorado, USA, 2002, pp. 929–932.

[10] Yannick Estève, Christian Raymond, Frédéric Béchet, Renato De Mori, and David Janiszek, "On the use of linguistic consistency in automatic speech

recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 746–756, 2003.

[11] Yannick Estève, Christian Raymond, Frédéric Béchet, and Renato De Mori, "Conceptual decoding for spoken dialog systems," in *Eurospeech*, Genêve, Suisse, 2003, pp. 3033–3336.

[12] Christian Raymond, Frédéric Béchet, Renato De Mori, Géraldine Damnati, and Yannick Estève, "Automatic learning of interpretation strategies for spoken dialogue systems," in *ICASSP*, Montréal, Canada, 2004, pp. 929–932.

[13] Yannick Estève Paul Deléglise and Bruno Jacob, "Systèmes de transcription automatique de la parole et logiciels libres," *Traitement Automatique des Langues*, vol. 45, no. 2, 2004.

[14] Paul Deléglise, Yannick Estève, Sylvain Meignier, and Teva Merlin, "The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news," in *Interspeech*, Lisbonne, Portugal, 2005.

[15] Paul Deléglise, Yannick Estève, Sylvain Meignier, and Teva Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ?," in *Interspeech*, Brighton, Royaume-Uni, 2009.

[16] Julie Mauclair, Yannick Estève, and Paul Deléglise, "Automatic detection of well recognized words in automatic speech transcription," in *LREC*, Gênes, Italie, 2006.

[17] Julie Mauclair, Sylvain Meignier, and Yannick Estève, "Indexation en locuteur : utilisation d'informations lexicales," in *JEP*, Dinard, France, 2006.

[18] Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Julie Mauclair, "System combination by driven decoding," in *ICASSP*, Honolulu, Hawaii, USA, 2007.

[19] Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Guillaume Gravier, "Generalized driven decoding for speech recognition system combination," in *ICASSP*, Las Vegas, Nevada, USA, 2008.

[20] Antoine Laurent, Teva Merlin, Sylvain Meignier, Yannick Estève, and Paul Deléglise, "Combined systems for automatic phonetic transcription of proper nouns," in *LREC*, Marrakech, Maroc, 2008.

[21] Antoine Laurent, Sylvain Meignier, Yannick Estève, and Paul Deléglise, "Combinaison de systèmes pour la phonétisation automatique de noms propres," in *JEP*, Avignon, France, 2008.

[22] Antoine Laurent, Teva Merlin, Sylvain Meignier, Yannick Estève, and Paul Deléglise, "Iterative filtering of phonetic transcriptions of proper nouns," in *ICASSP*, Taïpei, Taiwan, 2009.

[23] Richard Dufour and Yannick Estève, "Correcting ASR outputs : specific solutions to specific errors in french," in *IEEE Workshop on Spoken Language Technology*, Goa, Inde, 2008.

[24] Julie Mauclair, Sylvain Meignier, and Yannick Estève, "Speaker diarization : about whom the speaker is talking ?," in *IEEE Odyssey*, San Juan, Porto Rico, USA, 2006.

[25] Vincent Jousse, Christine Jacquin, Sylvain Meignier, Yannick Estève, and Béatrice Daille, "Etude pour l'amélioration d'un système d'identification nommée du locuteur," in *JEP/TALN*, Avignon, France, 2008.

[26] Vincent Jousse, Simon Petitrenaud, Sylvain Meignier, Yannick Estève, and Christine Jacquin, "Automatic named identification of speakers using diarization and ASR systems," in *ICASSP*, Taïpei, Taiwan, 2009.

[27] Vincent Jousse, Sylvain Meignier, Christine Jacquin, Simon Petitrenaud, Yannick Estève, and Béatrice Daille, "Analyse conjointe du signal sonore et de sa transcription pour l'identification nommée de locuteur," *Traitement Automatique des Langues*, vol. 50, no. 1, 2009.

[28] Yannick Estève, Sylvain Meignier, Paul Deléglise, and Julie Mauclair, "Extracting true speaker identities from transcriptions," in *Interspeech*, Anvers, Belgique, 2007.

[29] Thierry Bazillon, Yannick Estève, and Daniel Luzzati, "Transcription manuelle vs assistée de la parole préparée et spontanée," in *JEP*, Avignon, France, 2008.

[30] Thierry Bazillon, Yannick Estève, and Daniel Luzzati, "Manual vs assisted transcription of prepared and spontaneous speech," in *LREC*, Marrakech, Maroc, 2008.

[31] Thierry Bazillon, Yannick Estève, and Daniel Luzzati, "Le codage des corpus oraux," in *Catcod*, Orléans, France, 2008.

[32] Thierry Bazillon, Yannick Estève, and Daniel Luzzati, "La parole spontanée : transcription et traitement," *Traitement Automatique des Langues*, vol. 49, no. 3, 2008.

[33] Vincent Jousse, Yannick Estève, Frédéric Béchet, Thierry Bazillon, and Georges Linarès, "Caractérisation et détection de parole spontanée dans de larges collections de documents audio," in *JEP*, Avignon, France, 2008.

[34] Richard Dufour, Vincent Jousse, Yannick Estève, Frédéric Béchet, and Georges Linarès, "Spontaneous speech characterization and detection in large audio database," in *13-th International Conference on Speech and Computer - SPECOM*, Saint-Pétersbourg, Russie, 2009.

[35] Holger Schwenk and Yannick Estève, "Data Selection and Smoothing in an Open-Source System for the 2008 NIST Machine Translation Evaluation," in *Interspeech*, Brisbane, Australie, 2008.

[36] Holger Schwenk, Yannick Estève, and Sadaf Abdul-Rauf, "The LIUM Arabic/English Statistical Machine Translation System for IWSLT 2008," in *International Workshop on Spoken Language Translation*, Hawaii, USA, 2008.

Première partie

Travaux de recherche

Chapitre 2

Le système de transcription automatique du LIUM

Sommaire

2.1 Introduction

Avoir à sa disposition un système de transcription automatique de la parole (TAP) complet permet à un laboratoire qui souhaite mener des travaux en traitement automatique de la parole de travailler sur un large spectre de problématiques. Ainsi, un système complet, qui inclut également la détection de la parole, la segmentation en zones acoustiquement homogènes, la classification en locuteurs, etc. offre un éventail important d'objets d'étude, allant de la paramétrisation du signal audio à la post-édition des sorties du système, en passant par les algorithmes de décodage, la modélisation acoustique, la modélisation du langage, la construction de dictionnaires phonétisés, etc.

Avoir à sa disposition des sorties d'un système de TAP est également utile ou nécessaire pour des chercheurs qui ne travaillent pas dans le domaine du traitement automatique de la parole. Par exemple, certaines tâches peuvent être facilitées par l'utilisation d'un système de TAP :

- l'annotation de corpus oraux : la TAP peut accélérer le développement de corpus oraux transcrits et annotés, comme nous avons pu le montrer avec Thierry Bazillon dans (1; 2), tout en en réduisant les coûts ; en utilisant des mesures de confiance qui permettent d'évaluer automatique la pertinence d'une hypothèse de reconnaissance, il est possible de séparer les hypothèses comportant peu d'erreurs des autres : nous avons illustré cela durant la thèse de Julie Mauclair (3) ;
- l'indexation de document oraux : la TAP s'avère nécessaire pour la manipulation de quantité très importantes de données, impossibles à gérer manuellement ; c'est le cas de la recherche d'information au sein d'archives radiophoniques (ou audiovisuelles) : comment retrouver rapidement un enregistrement précis traitant d'un sujet donné parmi des milliers d'heures de parole si les notes d'archivages concernant cet enregistrement sont perdues ou imprécises ? l'utilisation d'un système de TAP permet de générer des transcriptions qui, même imparfaites, peuvent constituer une base de départ suffisante pour des manipulation de ce type, comme dans le projet CallSurf impliquant, entre autres, EDF et le LIMSI dans le cadre du traitement d'enregistrements d'appels téléphoniques auprès du centre d'appels d'EDF (4) ;
- l'extraction de sous-corpus spécifiques : certains travaux de recherche peuvent porter sur l'étude d'événements oraux particulier (dialogues, débats, lecture, interactions interviewer/interviewé,etc. ; chacun de ces types de corpus oraux est rare en version annotée manuellement, alors que de grands corpus oraux hétérogènes sont susceptibles de contenir des parties intéressantes pour l'étude de ces événements. Durant le stage de Master Recherche Vincent Jousse, puis durant la thèse de Richard Dufour, nous avons par exemple travaillé à la caractérisation et la détection de la parole spontanée dans de grands corpus audio (5; 6)

D'autre tâches peuvent certainement nécessiter ou être facilitées par l'utilisation d'un système de TAP. En règle générale, un système de TAP peut aider à la manipulation de corpus oraux pour lesquels il n'existe pas de transcriptions, que ceux-ci sont nécessaires, et qu'une certaine marge d'erreurs peut être tolérée.

Il est donc important, pour un laboratoire qui ambitionne de réaliser des travaux scientifiques de qualité dans le domaine du traitement automatique de la parole, de disposer d'un système performant et complet de TAP. Or, le développement d'un tel outil est long, fastidieux et coûteux. Le simple fait de vouloir conserver un système de TAP à l'état de l'art demande des efforts très importants et gourmands en ressources humaines. Heureusement, il existe plusieurs systèmes de TAP, plus ou moins aboutis et performants, dans le monde du logiciel libre. Nous avons d'ailleurs présenté dans (7) un ensemble, non exhaustif, de logiciels libres ou disponibles pour développer un système de TAP. Le LIUM a choisi d'utiliser l'un des plus anciens et des plus performants systèmes probabilistes de TAP, toujours maintenu, et diffusé sous licence libre depuis 2000 : le décodeur CMU Sphinx.

2.2 Le projet CMU Sphinx

Le projet SPHINX a vu le jour en 1986 à Carnegie Mellon University (CMU) et une première description précise du système a été présentée dans (8). Il s'agissait à l'époque de développer un système de reconnaissance de la parole continue, à grand vocabulaire et indépendant du locuteur. Grand vocabulaire signifiait un vocabulaire contenant au moins 1000 mots. Ce projet était financé par la NSF (National Science Foundation) et la DARPA (Defence Advanced Project Agency).

Dépasser les contraintes de l'indépendance au locuteur, du traitement de la parole continue en opposition au traitement de mots isolés et l'augmentation de la taille vocabulaire étaient les principaux objectifs du projet. Il fut un des premiers systèmes de TAP à utiliser des phonèmes en contexte (*triphones*) comme unité sous-lexicale modélisée par des modèles de Markov (*HMM : Hidden Markov Model*). Ce système de TAP est l'ancêtre des décodeurs Sphinx actuels et a surtout permis de montrer la faisabilité de la tâche.

2.2.1 Les différents décodeurs Sphinx

Quatre familles de décodeurs Sphinx sont maintenant disponibles sous licence de type BSD. Ce type de licence est très permissif puisqu'il permet une utilisation commerciale des décodeurs et n'est pas contaminant car il n'impose pas une licence particulière aux logiciels dérivés. Ceci explique certainement le succès croissant du projet auprès du monde de la recherche et des entreprises commerciales.

Le projet CMU Sphinx a commencé à distribuer ses décodeurs sous licence libre en 2000. Les quatre branches actuelles des décodeurs Sphinx sont les suivantes :

- Sphinx 2 : cette version a été développée au début des années 90 pour proposer des solutions afin d'obtenir un système de TAP fonctionnant en temps réel sur des micro-ordinateurs de l'époque, en particulier pour des tâches de dialogue oral homme/machine. Ce décodeur a la particularité de fonctionner avec des modèles de Markov semi-continus pour la modélisation acoustique et permet de changer dynamiquement, en fonction de l'état du dialogue par exemple, les modèles de langage utilisés (9). Cette branche n'est plus développée, mais constitue la base du projet PocketSphinx ;
- Sphinx 3 : cette version a pour objectif d'obtenir la meilleure précision de reconnaissance possible. Il s'agit d'un décodeur qui a longtemps été le décodeur phare de la famille des décodeurs Sphinx (10). Il utilise des modèles de Markov continus. Deux sous-branches majeures du décodeur Sphinx 3 ont longtemps coexisté : un décodeur lent (*flat*) et un décodeur rapide (*lextree*). La différence majeure entre les deux provient de la gestion acoustique inter-mot de l'algorithme de recherche. Dans le décodeur lent, de vrais phonèmes en contexte (triphone + position du phonème dans le mot) sont utilisés en fin de mot, alors que dans la version rapide une approximation de la modélisation du phonème en fin de mot est effectuée qui accélère le traitement mais dégrade les performances. La version *flat* est dix fois plus lente que la version *lextree* pour une précision de reconnaissance sensiblement meilleure. Le décodeur Sphinx 3 est toujours en développement : les décodeurs *flat* et *lextree* ont par

exemple récemment été unifiés au sein d'un seul et même outil, pendant que de nouveaux ajouts permettent d'améliorer encore les performances et la vitesse d'exécution, comme présenté par exemple dans (11).
- Sphinx 4 : la version 4 de Sphinx est une réécriture complète d'un décodeur en Java décrit dans (12). Elle est le fruit d'une collaboration entre CMU, Sun Microsystems, Mitsubishi Electric Research Labs et Hewlett Packard. L'objectif était de développer un décodeur au moins aussi performant que le décodeur Sphinx 3. Mais Sphinx 4 n'est pas une copie de Sphinx 3 : d'un point de vue génie logiciel il a été conçu différemment et de façon beaucoup plus modulaire. Sphinx 3 et Sphinx 4 utilisent les mêmes modèles acoustiques et modèles de langage.
- PocketSphinx : le développement des applications embarquées et la montée en puissance des appareils concernés (téléphone portable, PDA, ...) ont entraîné la nécessité un système de TAP le moins gourmand possible en ressources afin de pouvoir être utilisé avec ce type de matériel. La version PocketSphinx est la réponse du projet Sphinx à ce besoin (13). Le projet Sphinx ne partait pas d'une feuille blanche puisque le décodeur Sphinx 2 était une solution déjà bien avancée d'un décodeur utilisé dans un contexte de ressources mémoire et CPU limitées. PocketSphinx est une réelle spécialisation de Sphinx 2 dans le domaine de l'informatique embarquée qui a permis de diviser pratiquement par 10 le temps d'exécution de PocketSphinx par rapport à Sphinx 2 pour une précision de reconnaissance comparable.

2.2.2 Les boîtes à outils du projet Sphinx

Le projet Sphinx propose également des outils pour estimer des modèles de langage n-gram et des modèles acoustiques. Ainsi, l'outil "CMU-Cambridge Statistical Language Modeling Toolkit" décrit dans (14) est disponible en version open source depuis 1994. Il comporte néanmoins quelques lacunes par rapport à d'autres boites à outils plus récents qui ne font pas partie du projet, comme SRILM (15). Les outils de CMU conserve toutefois quelques avantages comme un usage de mémoire beaucoup plus modéré.

La boîte à outils pour estimer les modèles acoustiques est, elle, actuellement irremplaçable. Ces outils, regroupés sous le nom de SphinxTrain permettent d'estimer des modèles acoustiques pour l'ensemble des décodeurs Sphinx. Les modèles acoustiques pour les décodeurs Sphinx 3 et 4 sont compatibles, alors que les modèles pour Sphinx 2 et PocketSphinx, semi-continus, sont compatibles entre eux mais incompatibles avec les modèles, continus, des autres décodeurs.

2.3 Architecture générale du système du LIUM

Le développement d'un système complet de TAP ne se réduit pas à simplement à un décodeur. Ce dernier s'intègre dans un chaîne de traitements et nécessite un ensemble de bases de connaissance : dictionnaire de mots avec prononciation, modèles de langage, modèles acoustiques. Le rôle du décodeur consiste à déterminer, parmi

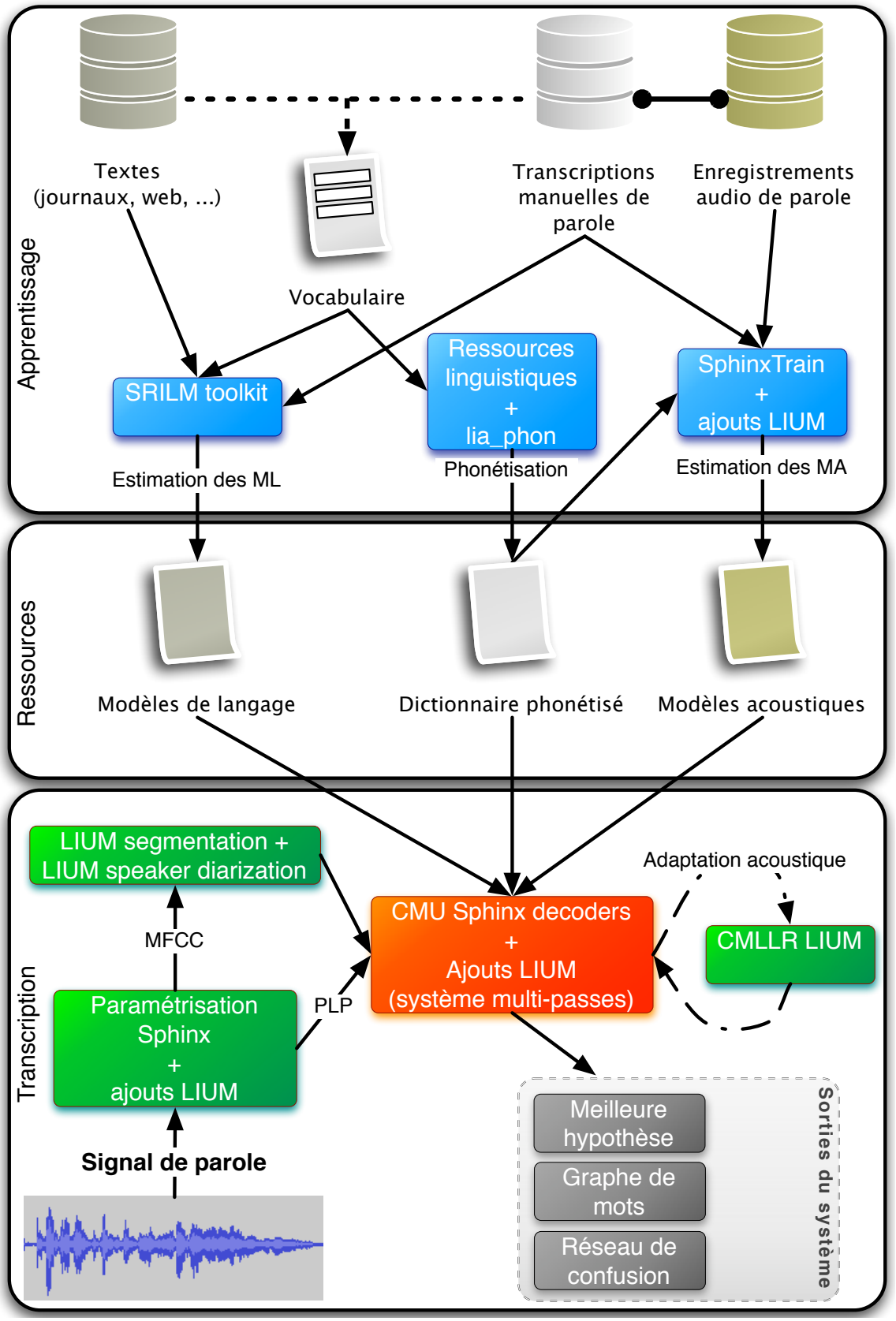

FIG. 2.1 – Architecture générale du système de transcription automatique du LIUM, de l'apprentissage à l'utilisation

toutes les séquences de mots W possibles, la séquence de mots $\hat{W}$ la plus probable en fonction de la séquence d'observations acoustiques X :

$$\hat{W} = \arg\max_{W} P(W|X) \tag{2.1}$$

Ce qui en appliquant la règle de Bayes devient :

$$\hat{W} = \arg\max_{W} \frac{P(X|W)P(W)}{P(X)} \tag{2.2}$$

Or, comme nous considérons que les observations acoustiques X, pour un intervalle temporel donné, ne changent pas, alors la valeur de $P(X)$ est la même quelque soit la séquence de mots W et n'influe donc pas dans la recherche de $\hat{W}$.

En pratique, les décodeurs actuels doivent ainsi déterminer la séquence de mots $\hat{W}$ en fonction de l'équation suivante :

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \tag{2.3}$$

Dans cette équation interviennent les connaissances acoustiques, qui permettent de délivrer la probabilité $P(X|W$ à travers les modèles acoustiques et le dictionnaire de prononciations, et les connaissances linguistiques qui délivrent la probabilité $P(W)$.

Le développement d'un système de TAP consiste donc d'une part à construire les bases de connaissances, et d'autre part à mettre en place la chaîne de traitements plus ou moins complexe en fonction du temps de calcul attendu, du niveau de qualité de transcription à atteindre et bien entendu de la puissance de calcul disponible.

La figure 2.1 résume l'architecture général du système de TAP du LIUM : les deux parties, la construction des ressources (l'apprentissage) et la transcription y sont représentées, pendant que les bases de connaissances créées pendant la phase d'apprentissage et utilisées lors de la transcription y sont mises en évidence.

Le système de TAP présenté ici est le système développé pour la participation du LIUM à la campagne d'évaluation ESTER 2 dont la période de test s'est déroulée en novembre 2008 (16). Nous reviendrons plus longuement sur les campagnes ESTER 1 et ESTER 2 dans la section 7.1. Ce qu'il est important de retenir ici c'est que système du LIUM présenté dans cette partie a été développé pour transcrire des émissions radiophoniques en français.

Ainsi, une grande partie des données d'apprentissage, en partie les données pour l'apprentissage des modèles acoustiques, sont très majoritairement des données fournies par les organisateurs de la campagne.

2.3.1 Apprentissage

Nous aborderons dans cette section la phase d'apprentissage des modèles acoustiques et linguistiques, ainsi que la constitution du dictionnaire et la phonétisation des mots du vocabulaire.

Données d'apprentissage

Les corpus d'apprentissage utilisés par le LIUM sont les corpus fournis par les organisateurs de la campagne ESTER 2, augmentés d'autres données. Les corpus ESTER 2 sont répartis comme suit :

- 200h d'enregistrements d'émissions radiophoniques transcrites manuellement (provenant principalement de radios françaises, mais contenant un peu d'enregistrement de radio africaines francophones) ;
- 40h d'enregistrements d'émissions radiophoniques de radio africaines francophones avec transcriptions manuelles rapides ;
- les articles du journal Le Monde de 1987 à 2006.

L'ensemble des participants ont également eu à disposition 40h d'enregistrements radiophoniques transcrits manuellement issus du projet EPAC et contenant plutôt de la parole conversationnelle. Nous reviendrons sur le projet EPAC, dans lequel le LIUM a eu un rôle important, dans le chapitre 5 et la section 8.2.

Nous avons également eu accès aux données French Giga Word Corpus, qui regroupe un très grand nombre des dépêches AFP (Agence France Presse) et AWP (informations financières) pour les années 1990 et 2000.

Enfin, le LIUM a enrichi ces corpus en récoltant des données sur le web. Il s'agit là de corpus textuels destinés à améliorer les modèles de langage :

- les archives du journal L'Humanité de 1990 à 2007 ;
- les articles du site Libération disponibles en 2007 ;
- les articles du site L'internaute disponibles en 2007 ;
- les articles du site Rue89 disponibles en 2007 ;
- les articles du site Afrik.com disponibles en 2007.

En résumé, nous disposions d'environ 240h d'enregistrements audio transcrits manuellement pour l'apprentissage des modèles acoustiques. Cela correspond à un corpus textuel proche de la tâche visée d'environ 3,3 millions de mots pour estimer les modèles de langage. L'ensemble des articles du journal Le Monde et du French Giga Word Corpus représentent environ 1 milliard de mots, alors que les données provenant du web représente 80 millions de mots.

Le tableau 2.1 présente la répartition des mots du corpus d'apprentissage en fonction de leur origine. On constate que la quantité de données spécifiques à la tâche visée et utilisées pour l'apprentissage des modèles de langage représente moins de 0,4% de l'ensemble des données.

Tab. 2.1 – Nombre de mots dans le corpus d'apprentissage en fonction de la source du sous-corpus

	Transcriptions manuelles d'émissions radiophoniques	Presse écrite et dépêches	Web
Nombre de mots	3,3M	1,0G	80M

Vocabulaire

La constitution du vocabulaire est une étape très importante et très délicate du développement d'un système de TAP. En effet, il sera impossible au système de pouvoir reconnaître un mot qui ne se trouve pas dans le corpus ; de plus, une erreur de transcription automatique due à un mot hors vocabulaire engendre plusieurs erreurs. Nous avons mesuré, dans nos expériences, que la présence d'un mot hors vocabulaire implique en moyenne la présence de 42% de mots erronés parmi les 3 mots précédents le mot hors vocabulaire, et 78% de mots erronés parmi les 3 mots suivants. Erreurs auxquelles s'ajoute l'erreur sur le mot hors vocabulaire lui-même. Ces résultats ont été obtenus avec un système ayant un taux d'erreurs global moyen de 23% sur le corpus de test de la campagne ESTER 1, dont le type de parole est proche de celui de la campagne ESTER 2.

Ainsi, le taux de mots hors vocabulaire a une influence importante sur les performances finales du système en terme de précision et doit être minimisé.

Ayant la responsabilité de construire le vocabulaire du système du LIUM pour sa participation à ESTER 2, j'ai choisi de suivre l'approche proposée dans (17). En résumé, cette approche consiste à :

1. estimer autant de modèles unigrams que nous avons de sources d'apprentissage (voir plus haut),
2. calculer les coefficients d'interpolation entre ces unigrams qui permettent de construire un modèle unigram de perplexité minimale sur le corpus de développement (ici le corpus de développement de la campagne ESTER) ; le calcul des coefficients d'interpolation utilise l'algorithme EM (18),
3. construire le modèle de langage unigram avec les coefficients d'interpolation calculés lors de l'étape précédente,
4. extraires les N mots les plus probables du modèle unigram, N étant la taille du vocabulaire visée.

Nous avons fixé arbitrairement la taille du vocabulaire à 120 000 mots. Nous pensions qu'il s'agissait d'une augmentation raisonnable de la taille du vocabulaire par rapport à celui de notre système précédent qui était limité à 65 500 mots.

Le tableau 2.2 indique les taux de mots hors vocabulaire calculés sur les corpus de développement et de test de la campagne ESTER 2. À titre indicatif, même si ces résultats ne peuvent pas être directement comparables, le taux de mots hors vocabulaire obtenu sur le corpus de test ESTER 1 lors de notre participation à cette campagne était de 1,18%, pour un vocabulaire de 65500 mots construits avec une méthode plus simple qui se justifiait car nous ne disposions alors que deux sources textuelles d'apprentissage : des transcriptions manuelles d'émissions radiophoniques et des articles du journal "Le Monde".

Au début de cette section consacrée à la construction du vocabulaire, j'ai affirmé que cette étape était très importante et très délicate. Nous avons vu que son importance était due à son impact sur le nombre d'erreurs de transcription. Le point délicat vient du fait que cette étape est la première d'une séquence de traitements (phonétisation, estimation des modèles de langage, etc.). Revenir sur cette étape implique de reprendre l'ensemble de ces traitements, ce qui s'avère très coûteux en temps de calcul, mais aussi en ressource humaine.

TAB. 2.2 – Taux de mots hors-vocabulaire du système du LIUM composé de 120 000 mots calculé sur les corpus de développement et de test de la campagne ESTER 2

Corpus ESTER	Taux de mots hors vocabulaire
Dev	0,66 %
Test	0,74 %

Phonétisation

Pour faire le lien entre le niveau lexical et le niveau acoustique, il est nécessaire d'associer à chaque mot du vocabulaire une ou plusieurs séquences d'unités acoustiques de base (19). Il s'agit généralement, et c'est le cas dans notre système, d'une séquence de phonèmes : nous utilisons un jeu de 35 phonèmes du français.

Pour obtenir ces séquences de phonèmes , nous disposons du lexique phonétisé BLDEX, (20) qui contient 450.000 formes fléchies de mots générées à partir de 50.000 formes canoniques. Malgré le nombre important de mots contenus dans BDLEX, ce lexique ne permet pas de couvrir l'ensemble de notre vocabulaire, d'autant plus que BDLEX ne contient aucun nom propre.

Aussi, nous utilisons un outil de phonétisation automatique, LIA_PHON (21), pour les mots de notre vocabulaire qui sont absents de BDLEX. À ce sujet, il faut noter que BDLEX et LIA_PHON n'utilisent pas le même jeu de phonèmes. Nous avons choisi de conserver les phonèmes utilisés par LIA_PHON.

Ces phonétisations ne sont pas exemptes d'erreurs : une vérification manuelle a été effectuée à diverses reprises depuis que nous avons développé notre premier système de TAP. En particulier, les mots les plus fréquents et les mots les plus mal reconnus ont été vérifiés. Ce travail ingrat permet de diminuer sensiblement le nombre d'erreurs. En conséquence, lorsque nous construisons un nouveau vocabulaire, nous procédons au traitement suivant :

1. si le mot existait déjà dans le vocabulaire du système de TAP précédent, nous conservons la ou les phonétisations déjà utilisées ;
2. sinon, si le mot existe dans BDLEX, nous utilisons la ou les phonétisation proposées par BDLEX ;
3. sinon, nous utilisons LIA_PHON.

Malheureusement, nous savons que cette approche n'est pas infaillible, et en particulier nous avons pu constater que la phonétisation des noms propres était un problème récurrent. J'ai participé aux travaux de thèse d'Antoine Laurent sur ce sujet : nous avons proposé une méthode (22; 23) qui utilise un algorithme d'alignement automatique de boucles de phonèmes sur des portions de signal audio correspondant à des noms propres afin d'extraire des phonétisations de ces noms propres. Ces phonétisations ne sont pas forcément les phonétisations exactes qu'un expert humain proposerait, mais elles prennent en compte les distorsions dues aux imperfections des modèles acoustiques. Cette approche, évaluée sur le corpus de

test de la campagne ESTER 1, permet de diminuer le taux d'erreurs sur les noms propres sans modifier de façon notable le taux d'erreurs global sur l'ensemble des mots. Mais les mots phonétisés par cette approche ne doivent pas être réutilisés avec la phonétisation proposée dans le cas où de nouveaux modèles acoustiques sont estimés : comme exprimé au-dessus, cette phonétisation est dépendante de la modélisation acoustique utilisée lors de l'alignement phonème/audio.

Le problème de la phonétisation des mots du vocabulaire est un problème entier sur lequel le LIUM travaille. À titre d'information puisque je n'ai pas participé à ces travaux, signalons que les travaux d'Antoine Laurent ont été étendus, en collaboration avec Paul Deléglise, en utilisant un système de traduction automatique pour la phonétisation des mots (24).

Modèles acoustiques

Je n'ai pas participé à la phase de construction des modèles acoustiques, mais il me semble indispensable dans la description du système de TAP du LIUM de les présenter en allant directement à l'essentiel.

Les modèles acoustiques utilisés par le système de TAP du LIUM, à base de modèles de Markov cachés, modélisent un jeu de 35 phonèmes du français, ainsi que 5 types de *filler*, c'est-à-dire d'éléments sonores qui ne sont pas des phonèmes constituant des mots : *silence, musique, bruit, inspiration, 'euh' prolongé*. Ces phonèmes sont modélisés en contexte : leur modélisation prend en compte leurs contextes phonémiques gauche et droit (*triphone*), ainsi que leur position dans le mot (*début, milieu, fin, isolé*).

Les paramètres acoustiques extraits du signal audio et traités au niveau de la modélisation acoustique sont au nombre de 39 : il s'agit de descripteurs issus d'une analyse du signal de type PLP (25) et d'un descripteur de l'énergie, ainsi que des dérivées et dérivées secondes de ces descripteurs.

Notre système de TAP dispose de différents ensembles de modèles acoustiques : chacun de ces modèles est spécialisé en fonction du genre du locuteur (*homme/femme*) et de la taille de la bande passante (*téléphone/studio*). Ces spécialisations ont été obtenues à partir d'une adaptation de type MAP (26) au niveau des moyennes, des co-variances et des poids.

En particulier, ce système étant un système multi-passes, nous pouvons distinguer deux familles de modèles acoustiques en fonction de la passe lors de laquelle ils sont utilisés :

1. en première passe, les modèles sont composés de 6500 états partagés, chaque état étant modélisé par une mixture de 22 gaussiennes ;
2. en seconde passe et par la suite, les modèles acoustiques sont composés de 7500 états, toujours modélisés par une mixture de 22 gaussiennes. Ces modèles ont été estimés à partir d'un apprentissage de type SAT (27) combiné à un apprentissage discriminant de type MPE (28). De plus, une matrice de transformation CMLLR (29) a été calculée pour chaque locuteur et appliqué sur les paramètres acoustiques de chacun des locuteurs respectifs.

Modèles de langage

Les modèles de langage guident le décodage acoustique afin de retenir comme hypothèses de reconnaissance les hypothèses acoustiques concurrentielles dont les probabilités linguistiques sont les plus élevées. Les modèles de langage du système de TAP du LIUM sont, comme pour la très grande majorité des systèmes de TAP, des modèles de langage n-grams. Nous utilisons des modèles 3-grams dans les premières passes de décodage, et des modèles 4-gram pour les dernières passes.

La responsabilité de la construction des modèles de langage du système de TAP m'incombe, ce qui va de pair avec la construction du vocabulaire. Après différentes expériences qui ont confirmé des résultats présents dans la littérature scientifique du domaine, j'ai choisi d'estimer des modèles de langage n-gram utilisant la technique de *discounting* dite de Kneser-Ney modifié (30; 31) avec interpolation des n-grams d'ordres inférieurs. Une autre des caractéristiques de nos modèles n-grams, c'est qu'aucun *cut-off* n'a été appliqué : tous les n-grams observés dans le corpus d'apprentissage, même une seule fois, sont pris en compte. Généralement, un *cut-off* est appliqué dans l'optique de réduire la taille du modèle de langage, mais également pour éliminer des coquilles ou des séquences de mots erronées (à cause d'une faute d'orthographe par exemple). Des expériences que j'ai menées pour mesurer l'influence du *cut-off* sur la valeur de la perplexité d'un modèle de langage ont montré que, effectivement, l'utilisation d'un *cut-off* permet d'améliorer la perplexité d'un modèle de langage. J'ai ensuite mené des expériences similaires pour mesurer l'influence du *cut-off* sur le taux d'erreurs sur les mots d'un système de TAP : les résultats montraient une dégradation d'un taux d'erreurs sur les mots lors de l'usage d'un *cut-off*, même le plus faible. Nos modèles de langage ont donc été estimés sans utilisation de *cut-off*.

Deux types de radios différents étaient visés lors de la campagne ESTER 2 : des radios françaises et deux radios africaines : TVME et Africa 1. La radio Africa 1 comporte un grand nombre de locuteurs avec de forts accents étrangers, ce qui n'est pas le cas de TVME où, hormis quelques prononciations de noms propres arabes, les locuteurs, peu nombreux, s'expriment dans un français relativement neutre. Par crainte de générer un trop grand nombre de prononciations inappropriées dans le dictionnaire de phonétisation pour le traitement de la radio Africa 1, nous avons décidé de diminuer le nombre de mots dans le dictionnaire pour cette radio. Nous avons donc développé deux types de modèles : des modèles pour la radio Africa 1, et des modèles pour les autres. Ce découpage concerne également les modèles acoustiques.

Tab. 2.3 – Répartition du nombre de mots et de segments dans les données de développement de la campagne ESTER 2 en fonction de deux types de radios

	Radios francaises + TVME	Africa 1	Total
Nombre de mots	41653	25887	67540
Nombre de segments	1801	1165	2966

TAB. 2.4 – Répartition du nombre de mots et de segments dans les données de test de la campagne ESTER 2 en fonction de deux types de radios

	Radios francaises + TVME	Africa 1	Total
Nombre de mots	62901	16155	79056
Nombre de segments	5232	1260	6492

Le tableau 2.4 présente la répartition sur le corpus de développement ESTER 2 du nombre de mots et du nombre de segments en fonction du type de radio. Le tableau 2.5 présente cette répartition sur le corpus de test ESTER 2. Nous pouvons nous rendre compte que la répartition est déséquilibrée entre le corpus de test et le corpus de développement : la radio Africa 1 a moins de poids dans le corpus de test, ce qui n'était pas attendu d'après les informations données par les organisateurs au début de la campagne. De plus, étonnamment, pour un nombre de mots relativement proche, le nombre de segments du corpus de test est deux fois plus important que sur le corpus de développement.

Les données d'apprentissage sont les données textuelles décrites en section 2.3.1 : ces données sont l'objet d'un travail ingrat mais fondamental de normalisation : cette tâche consiste à nettoyer les corpus, à faire en sorte qu'un mot ne s'écrive pas de différentes façons, à mettre sous forme de mots certains symboles, ..., avant de commencer à les manipuler. Chaque corpus d'apprentissage a été utilisé pour estimer un modèle n-gram. Ensuite, sur le corpus de développement approprié, les coefficients d'interpolation ont été optimisés avec l'algorithme E.M. afin de maximiser la mesure de perplexité du modèle interpolé sur ce corpus. Ces manipulations ont été réalisées à l'aide du *SRILM toolkit* (15).

Comme nous l'avons observé en section 2.3.1, la quantité de données spécifiques à la tâche visée et utilisées pour l'apprentissage des modèles de langage représente moins de 0,4% de l'ensemble des données. Or, quelque soit le modèle de langage interpolé, ces données pèseront pour environ 40% de la probabilité de ce modèle : le coefficient d'interpolation linéaire du modèle de langage estimé à partir des transcriptions manuelles d'émissions radiophoniques à pour chacun des modèles n-grams du système une valeur proche de 0,4. Ce constat ouvre quelques pistes de réflexion : faut-il tenter de collecter encore plus de données du domaine ? encore plus de données issues du web ou de la presse écrite ? faut-il travailler à l'adaptation des données au domaine visé ? La réponse englobe certainement ces différents aspects, mais les chiffres évoqués au-dessus peuvent aider à hiérarchiser les solutions en privilégiant, à mon avis, le travail de collecte et d'adaptation de données spécifiques au domaine.

Les tableaux 2.6 et 2.7 montrent les valeurs de perplexités des modèles 3-grams et 4-grams des configurations "radios françaises" et "Africa1". Pour chaque modèle, deux valeurs de perplexité, *ppl* et *ppl1* sont présentées : *ppl* correspond à la valeur perplexité obtenue en prenant en compte les mots et les étiquettes de début et de fin de phrase, alors que *ppl1* ne prend pas en compte ces étiquettes de début et fin de phrase. La valeur *ppl1* est la valeur la plus intéressante car elle ne dépend pas

de la segmentation du texte. Or, la segmentation qui est utilisée lors d'un décodage est une segmentation automatique différente d'une segmentation manuelle. De plus, la segmentation manuelle est subjective, ce qui semble corroborer par le nombre de segments manuels du corpus de test deux fois plus important que pour le corpus de développement.

Ainsi, on remarque que pour les radios françaises la perplexité *ppl1* augmente entre le corpus de développement et le corpus de test, alors que la perplexité *ppl* baisse. Cette baisse s'explique certainement par le nombre deux fois plus important, plus qu'il y a le double de segments, d'étiquettes de début et fin de phrase dans le corpus de test, ces étiquettes ayant des probabilités élevées et tendant dont à baisser la valeur de la perplexité.

Ce qui est plus étonnant, c'est le fait que la perplexité *ppl1* baisse entre le corpus de développement et le corpus de test pour la radio Africa 1 : cela signifie que le modèle de langage dédié au traitement de cette radio est plus pertinent sur le corpus de test que sur le corpus de développement. Ceci est plutôt original, puisque les coefficients d'interpolation ont été optimisés sur le corpus de développement, et sera confirmé par l'évaluation en taux d'erreurs sur les mots du système de transcription utilisant ce modèle : le taux d'erreurs sur les mots sera plus élevé sur le corpus de développement que sur le corpus de test pour la radio Africa 1. Sans surprise, les modèles 4-gram sont plus performants en valeur de perplexité que les modèles 3-gram correspondants.

Tab. 2.5 – Nombre de n-grams dans les modèles de langage

Configuration des modèles	1-grams	2-grams	3-grams	4-grams
Radios françaises + TVME	121K	29M	162M	376M
Africa 1	101K	27M	156M	370M

Tab. 2.6 – Perplexités des modèles de langage du système de TAP du LIUM sur les données de développement et de test de la campagne d'évaluation ESTER 2 sur les radios françaises et TVME

	Développement		Test	
LM	ppl	**ppl1**	ppl	**ppl1**
trigram	104	**127**	98	**143**
quadrigram	91	**110**	86	**125**

Le tableau 2.8 montre que la différence, en terme de perplexité, entre l'utilisation de modèles de langage spécifiques aux types de radios plutôt que de modèles généraux est minime. Nous verrons toutefois que cette stratégie de spécialisation et de réduction de risque en limitant de taille du vocabulaire pour Africa 1 s'est avérée utile au moment de l'évaluation du système complet de TAP.

Tab. 2.7 – Perplexités des modèles de langage du système de TAP du LIUM sur les données de développement et de test de la campagne d'évaluation ESTER 2 sur la radio Africa 1

	Développement		Test	
LM	ppl	**ppl1**	ppl	**ppl1**
trigram	103	**127**	82	**116**
quadrigram	88	**108**	70	**98**

Tab. 2.8 – Comparaison des perplexités des modèles de langage spécialisés par type de radios sur l'ensemble des données ESTER 2 par rapport à des modèles de langage généraux

	Développement	Test
LM	ppl1	ppl1
trigram général	129	138
trigrams spécialisés	127	137
quadrigram	111	120
quadrigrams spécialisés	109	119

2.3.2 Transcription

Nous venons de voir que le travail de constitution des bases de connaissances (modèles acoustiques, dictionnaire phonétisé, modèles de langage) constitue une tâche relativement complexe faisant appel à des algorithmes d'apprentissage automatique, à des choix parfois arbitraires, à des contraintes sur la disponibilité des données, mais également des outils. Ces bases de connaissances sont primordiales pour obtenir de bonnes performances de reconnaissance de la parole et leur construction est élaborée de façon minutieuse.

Dans cette section, nous allons décrire comment sont exploitées ces connaissances dans le système de TAP du LIUM.

Système de segmentation et de regroupement en locuteurs

Dans un premier temps, il est nécessaire de segmenter le signal audio en zones de parole et de non-parole. Par la suite, les outils de segmentation permettent de dégager des segments acoustiquement homogènes, jusqu'à segmenter le signal audio en fonction des interventions des locuteurs : le but recherché est d'obtenir un segment pour chaque intervention d'un locuteur, en fonction des conditions acoustiques, et d'être capable de pouvoir regrouper tous les segments d'un même locuteur. Ces regroupements permettent de mettre en ?uvre plus efficacement des adaptations acoustiques en fonction du locuteur.

L'outil utilisé par le LIUM est un outil développé en interne par Sylvain Meignier. Cet outil a terminé premier lors de la campagne ESTER 2 en obtenant la valeur de

diarization error rate la plus basse (10,6%).

Système de transcription multi-passes

Sans compter la phase de segmentation et de regroupement en locuteurs, le système de TAP du LIUM est un système qui comporte cinq passes. La notion de passe est assez subjective et consiste ici en l'utilisation d'un algorithme de recherche utilisant les données de la passe précédente et pouvant proposer une nouvelle hypothèse de reconnaissance.

La figure 2.2 décrit l'enchaînement des cinq passes du système de TAP :

1. La première passe consiste en un traitement utilisant la version 3.7 du décodeur rapide de Sphinx 3 appliquée sur des paramètres acoustiques PLP ; cette passe utilise un modèle de langage trigramme et des modèles acoustiques généralistes, seulement adaptés au sexe du locuteur (homme/femme) et aux conditions acoustiques (studio/téléphone).

2. La seconde passe utilise de nouveau la version 3.7 du décodeur rapide de Sphinx 3 appliqué sur les mêmes paramètres acoustiques PLP, mais une matrice de transformation CMLLR a été calculée de façon à adapter les paramètres acoustiques aux modèles acoustiques. Ces modèles ont été estimés en utilisant les méthodes SAT et MPE.

3. La troisième passe permet de pallier les approximations inter-mots faites par le décodeur s3.7 lors du calcul des scores acoustiques des phonèmes : afin d'accélérer fortement le traitement, les scores des phonèmes situés en fin de mots ne sont pas calculés en utilisant leur véritable contexte droit, mais à partir d'une approximation grossière. Ceci permet de diviser par dix le temps de calcul mais a une incidence négative sur le taux d'erreurs final. En utilisant le graphe de mots généré lors de la seconde passe comme espace de recherche figé, il est possible de corriger ces imprécisions inter-mots en utilisant, puisqu'il est connu *a priori*, le vrai contexte droit des phonèmes en fin de mot. Ce sont les mêmes modèles acoustiques et linguistiques que lors de la seconde passe qui sont utilisés, avec bien entendu l'application de la même matrice de transformation CMLLR sur les paramètres acoustiques.

4. La quatrième consiste à recalculer à l'aide d'un modèle 4-gram les scores linguistiques des mots du graphe de mots générés en passe 3.

5. Enfin, la passe 5 transforme le graphe de mots issu de la quatrième passe en un réseau de confusion. La méthode de consensus (32) est alors appliquée qui permet d'obtenir l'hypothèse de reconnaissance finale, avec pour chaque mot des probabilités *a posteriori* utilisables comme mesures de confiance.

Le tableau 2.9 montre l'évolution du taux d'erreurs sur les mots au fur et à mesure des passes de traitement sur l'ensemble du corpus de test ESTER 2 en utilisant les modèles acoustiques et linguistiques de la configuration "radios françaises". Entre la première passe et la dernière passe, le taux d'erreurs sur les mots chute, en relatif, de 29,15% en passant de 27,1% à 19,2%.

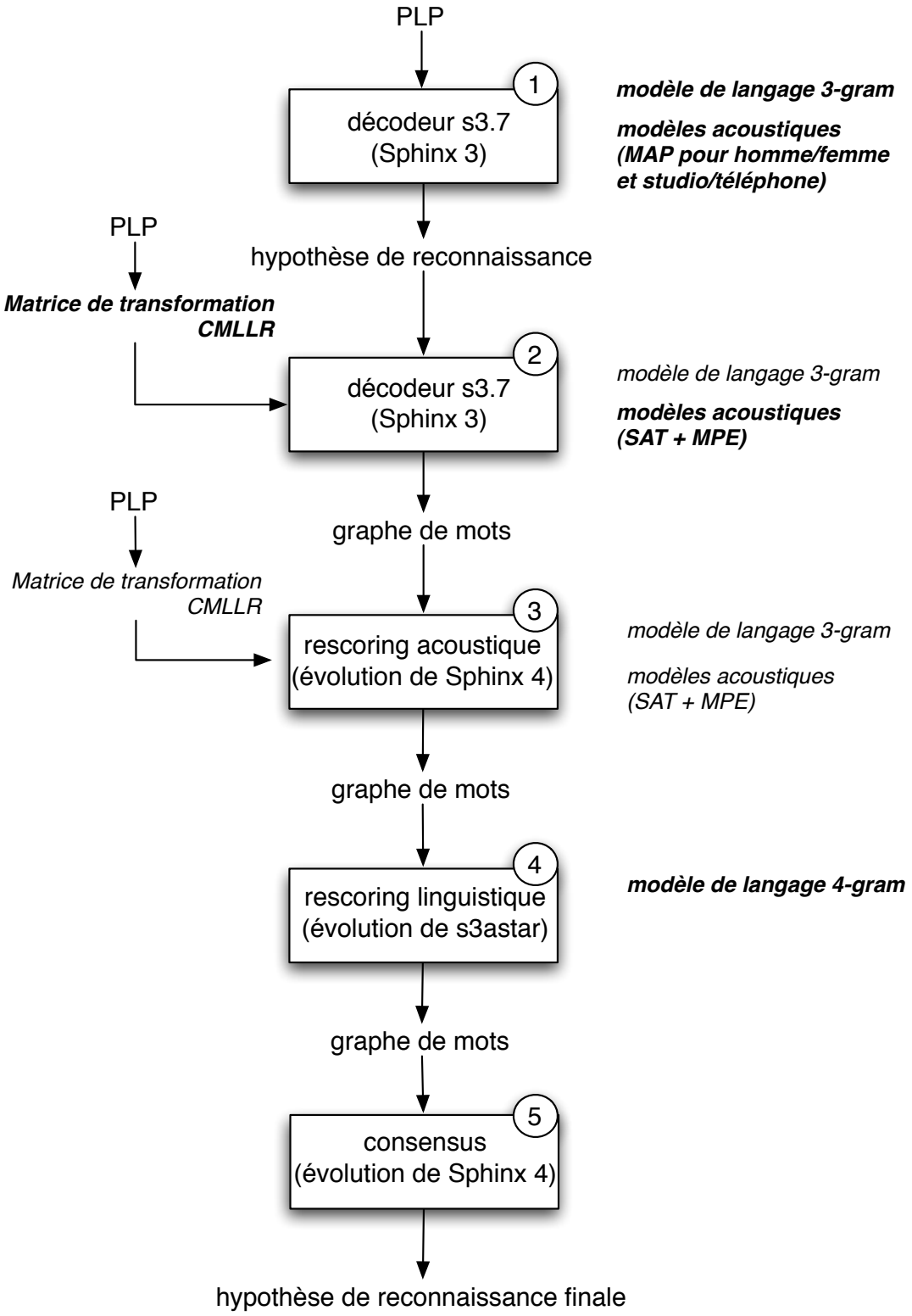

FIG. 2.2 – Description des cinq passes du système de transcription automatique du LIUM

TAB. 2.9 – Evolution du taux d'erreurs sur les mots en fonction de la passe de décodage du système de TAP sur l'ensemble du corpus de test ESTER 2 en utilisant la configuration "radios françaises"

Passe	Taux d'erreurs sur les mots
1. Modèles acoustiques généralistes, modèle trigram	27,1 %
2. Adaptation acoustique	22,5 %
3. Rescoring acoustique de graphe de mots	20,4 %
4. Rescoring de graphe de mots avec modèle quadrigram	19,4 %
5. Consensus	**19,2** %

2.4 Ajouts logiciels et améliorations et du LIUM

Le système de TAP du LIUM doit beaucoup au projet CMU Sphinx. Mais, comme en témoignent les résultats présentés dans le tableau 2.9, les améliorations et ajouts logiciels développés au LIUM permettent de réels gains en termes taux d'erreurs. En effet, seule la première passe a été obtenue à l'aide d'un décodeur Sphinx sans modification. Toutes les autres passes de décodage, mais aussi l'apprentissage, intègrent des modifications ou des ajouts lourds effectués au LIUM.

En voici une liste non exhaustive qui reprend les ajouts les plus importants :

- une modification de l'adaptation MAP des modèles acoustiques ;
- l'apprentissage des modèles acoustiques par les méthodes SAT et MPE ;
- la construction d'une matrice de transformation CMMLR et sa prise en compte dans le décodeur ;
- le décodeur acoustique de graphes de mots qui permet d'obtenir une aussi bonne précision pour le calcul des scores acoustiques au niveau inter-mot qu'au niveau intra-mot ;
- le décodeur linguistique de graphes de mots qui permet de recalculer les scores linguistiques avec des modèles d'ordre supérieur à 3 (nous n'utilisons actuellement que des 4-grams, mais l'outil est en théorie capable de manipuler des modèles de langage d'ordre plus élevé) ;
- un outil qui transforme un graphe de mots en réseau de confusion, ...

Dans (33), nous détaillons les gains, au niveau du taux d'erreurs sur les mots, apportés par les différents ajouts du LIUM.

La mise en place de la séquence de traitements que constituent les cinq passes de décodage n'a pas été réalisé sans tâtonnement ni de nombreuses expériences intermédiaires. Par exemple, une amélioration du taux d'erreurs de l'hypothèse de reconnaissance fournie par une passe intermédiaire n'implique pas une amélioration du taux d'erreurs de l'hypothèse de reconnaissance finale.

Le taux d'erreurs final de notre système sur le corpus de test ESTER 2 présenté dans le tableau 2.9, égal à 19,2%, est moins bon que le meilleur taux d'erreurs que notre système est capable d'obtenir. En utilisant les configurations "radios françaises" et "Africa 1", et en utilisant également des probabilités au niveau des variantes de prononciations, notre système obtient un taux d'erreurs de 18,1%. Mieux,

lors de la phase de test de la campagne ESTER 2, nous avons également expérimenté l'usage de modèles de langage continus, à base de réseaux de neurones, pour recalculer les scores linguistiques de listes de *1000-best* (les 1000 meilleures hypothèses de reconnaissance pour chaque segment audio traité). Avec ces modèles neuronaux de type 5-gram, notre système a obtenu le taux d'erreurs officiel de 17,8% lors de la campagne ESTER 2 : ces modèles seront bientôt intégrés dans le décodeur linguistique de graphes de mots.

2.5 Liens entre le LIUM et le projet *open source* CMU Sphinx

Puisque nous utilisons les outils du projet CMU Sphinx et que les évolutions que nous avons apportées à ce projet peuvent être intéressantes pour la communauté, nous avons contacté l'équipe de CMU chargés de maintenir et développer le projet, en particulier les personnes suivantes : Dr. Alex Rudnicky (CMU), Arthur Chan (ex-CMU, actuellement chez *BBN Technologies*), Dr. Evandro Gouvea (ex-CMU, actuellement chez MERL : *Mitsubishi Electric Research Laboratories*), David Huggins-Daines (CMU) et Bhisksha Raj (*Associate Professor* à CMU), ce dernier étant actuellement en charge du projet.

Le projet CMU Sphinx étant devenu un projet *open source* international, nous avons accepté en 2006 de faire partie de l'équipe de développement du projet. Pour le moment, seuls quelques ajouts que j'avais apportés dans l'outil *cmuslmtk* d'estimation des modèles de langage ont été intégrés dans la version distribuée sur le site web de CMU Sphinx. Ces ajouts permettent à l'outil d'utiliser la technique de *discounting* dite de Kneser-Ney modifié. De plus, les outils de segmentation et regroupement en locuteurs développés au LIUM par Sylvain Meignier ont été réécrits complètement en Java de façon à s'harmoniser avec les outils de Sphinx, et en particulier Sphinx 4.

Nous sommes actuellement en cours de portage de certains de nos outils (apprentissage acoustique SAT + MPE, adaptation CMLLR, décodeurs acoustique et linguistique de graphe de mots) dans les versions canoniques de Sphinx 3 et 4.

Enfin, en marge de la prochaine conférence ICASSP en 2010 à Dallas (USA), un *workshop* satellite sera organisé par le *Speech Group* de CMU : *CMU Sphinx Users and Developpers Workshop*. Je suis un membre du comité scientifique de ce *workshop*.

2.6 Conclusion

Nos travaux concernant le développement d'un système de transcription automatique de la parole compétitif et à l'état de l'art représentent une part importante du travail que nous avons effectué ces dernières années. On pourrait penser qu'il ne s'agit pas véritablement d'un travail de recherche. Pourtant, c'est un travail indispensable pour disposer d'un système de TAP permettant d'effectuer des travaux de recherche reconnus par la communauté scientifique nationale et internationale de la reconnaissance de la parole. De plus, ce travail sur notre système de TAP a

trouvé un écho dans la communauté puisqu'il a fait l'objet d'une publication dans une revue francophone importante (7) et de deux publications dans des conférences majeures du domaine (34; 33). Au sein de l'équipe Parole du LIUM, l'ensemble des travaux de recherche utilise une version du système de TAP du LIUM. Ainsi, toutes les publications scientifiques du LIUM ces dernières années font référence à cet outil et n'aurait pas pu avoir lieu sans lui.

La majeure partie du développement logiciel est à attribuer à Paul Deléglise qui s'est énormément investi dans ce projet. Pour ma part, j'ai essentiellement contribué au niveau des outils dits 'linguistiques', que ce soient lors de la phase d'apprentissage ou lors de la phase de décodage, ainsi qu'à de très nombreuses expérimentations du système.

Dans le même temps, je suis très actif sur la valorisation de nos travaux sur le système de TAP, par exemple à travers la création de corpus dans le cadre du projet ANR EPAC que nous évoquerons dans le chapitre 5, ou à travers la mise à disposition sous licence *open source* d'une grande partie de nos outils, en particulier dans le cadre du projet CMU Sphinx.

Les résultats obtenus par le LIUM lors de participations à des campagnes d'évaluation nationales ou internationales, que nous évoquerons dans le chapitre 7 avec plus de détails, assurent une visibilité grandissante du LIUM dans la communauté.

Cette valorisation et cette reconnaissance nous ont fait entrer dans un cercle vertueux qui nous permet d'obtenir des contrats auprès d'organisme de recherche comme l'ANR, mais aussi auprès d'entreprises privées. Ces contrats nous permettent d'augmenter notre puissance de travail et ainsi augmenter notre visibilité.

Dans les années à venir, le système de TAP sera certainement encore amélioré, mais le travail d'ingénierie à faire pour rester au niveau de l'état de l'art sera probablement moins important que celui qu'il a fallu effectuer pour atteindre notre niveau de compétitivité actuel.

Dans les prochains chapitres, nous traiterons de différents travaux de recherche auxquels j'ai participé : hormis le chapitre 6 qui traite de traduction automatique, tous ces travaux utilisent d'une manière ou d'une autre le système de TAP qui a été décrit ici.

Chapitre 3

Mesures de confiance et applications

Sommaire

3.1 Introduction

Les systèmes de transcription automatique commettent inévitablement des erreurs. Ils sont donc utilisés pour des applications qui tolèrent cette imprécision,

comme par exemple les applications de dialogue oral homme/machine lorsque le nombre et le type d'erreurs autorisent une interprétation sémantique satisfaisante de la hypothèse de reconnaissance, ou encore les applications d'indexation de document audio lorsque la qualité ou le type de ces documents sont suffisamment bien traités par les systèmes de TAP.

Pour ces applications qui doivent manipuler des données non fiables, il peut être très utile lorsque cela est possible d'avoir la possibilité d'exploiter des informations additionnelles concernant la fiabilité de ces données. Puisque cette fiabilité est liée à la distance qui existe entre les données à traiter et les connaissances acoustiques et linguistiques du système de TAP, mais aussi à l'algorithme de recherche utilisé par ce dernier, aux heuristiques d'élagage de l'espace de recherche choisis afin d'accélérer le traitement, *etc.* ... il revient généralement au système de TAP de proposer lui-même un auto-diagnostic concernant ses hypothèses de reconnaissance au travers des mesures de confiance. Cet auto-diagnostic est habituellement établi au niveau du mot.

Les mesures de confiance étaient l'objet principal de la thèse de Julie Mauclair, dirigée par Paul Deléglise et dont j'étais co-encadrant. Dans son mémoire de thèse (35), Julie Mauclair décrit un état de l'art très complet sur les mesures de confiance.

Dans ce chapitre, nous présenterons nos contributions. Dans un premier temps, après avoir brièvement décrit la mesure de confiance généralement utilisée dans les systèmes de TAP, à savoir la probabilité *a posteriori* d'un mot, nous présenterons une proposition de normalisation de mesure de confiance acoustique, ainsi qu'une nouvelle mesure de confiance basée sur le comportement, lors du traitement de reconnaissance, du repli (*back-off*) du modèle de langage. Nous montrerons ensuite comment il est possible de fusionner différentes mesures de confiance et présenterons quelques résultats expérimentaux. Enfin, dans les deux dernières sections seront abordées deux utilisations des mesures de confiances que nous avons expérimentées : une utilisation qui concerne la constitution automatique de corpus d'apprentissage de modèles acoustiques, et une autre qui concerne la combinaison de système de TAP.

3.2 Mesures de confiance

La mesure de confiance $CM(h)$ associée à une hypothèse de reconnaissance h appartient à l'intervalle $[0;1]$ et peut être interprétée comme étant la probabilité que l'hypothèse soit correcte.

Soit $\bar{W} = w_1 w_2 ... w_k$ la séquence de mots proposée par un système de TAP comme hypothèse de reconnaissance. Chaque mot w est associé à une mesure de confiance $CM(w)$. Une mesure de confiance idéale possède trois propriétés :

1. ses valeurs sont comprises entre 0 et 1,
2. elle doit être égale à 0 si w est incorrect et égale à 1 si w est correct,
3. en conséquence des deux premières propriétés, la moyenne des valeurs fournies par la mesure de confiance pour l'ensemble des mots w de l'hypothèse $\bar{W}$ est égale au taux d'erreurs des mots de l'hypothèse $\bar{W}$ en excluant les erreurs

de suppression puisqu'une mesure de confiance ne s'applique que sur les mots émis.

3.2.1 Probabilité *a posteriori*

Pour chaque mot w d'une hypothèse de reconnaissance, un système de TAP peut calculer sa probabilité *a posteriori* $P(w|X)$, où X correspond aux observations acoustiques traitées par le système de TAP pour obtenir w.

Comme nous l'avons vu dans le chapitre précédent dans l'équation (2.3), les systèmes de TAP omettent de normaliser le score d'une hypothèse de reconnaissance par la probabilité $P(X)$ qui est la probabilité d'apparition des observations acoustiques, considérée comme constante.

Ainsi, le score d'un mot délivré par un système de TAP n'est pas la probabilité $P(w|X)$, mais la valeur de $P(X|w)P(w)$: l'hypothèse la plus vraisemblable du point de vue des connaissances du système de TAP est connue, mais on ne sait pas à quel point elle est correcte. Le score délivré par le système de TAP est ainsi inadéquat comme mesure de confiance : pour obtenir une mesure de confiance, il faut déterminer $P(X)$. En théorie, on a :

$$P(X) = \sum_{hyp} P(hyp).P(X|hyp) \tag{3.1}$$

où les scores de toutes les hypothèses hyp possibles pour l'ensemble des observations acoustiques X sont sommés.

En pratique, l'énumération de toutes les hypothèses hyp possibles n'est pas réalisable, et l'on restreint ces hypothèses à celles qui ont été conservées lors de l'élaboration de l'espace de recherche par le décodeur : cette approximation peut-être basée sur une liste des N meilleures hypothèses (36; 37), sur un graphe de mots (38; 39; 37), ou plus généralement maintenant sur un réseau de confusion (32; 40; 41).

3.2.2 Mesure de confiance acoustique normalisée

Les modèles acoustiques d'un système de TAP fournissent un score de vraisemblance pour chaque mot du dictionnaire retenu lors du décodage. Il est possible, à partir de ce score acoustique, de créer une mesure de confiance acoustique en prenant en compte le biais acoustique introduit lors du choix d'un mot par les contraintes lexicales et phonétiques issues du modèle de langage et du dictionnaire de prononciation. Cette mesure acoustique, nommée $AC(w)$, est estimée en calculant la différence de la vraisemblance logarithmique entre le score acoustique d'un mot et le meilleur score acoustique qui pourrait être obtenu sans contrainte lexicales ou phonétiques, à l'aide d'un décodage acoustico-phonétique pur (42; 43) :

$$AC(w) = \frac{1}{N_f(w)}(logP(X|\lambda_C) - logP(X|\lambda_L)) \tag{3.2}$$

où :

– $logP(X|\lambda_C)$ est le score acoustique donné par le modèle acoustique contraint par le dictionnaire de phonétisation et le modèle de langage

– $logP(X|\lambda_L)$ est le score acoustique obtenu par un décodage acoustico-phonétique non contraint
– $N_f(w)$ est le nombre de trames qui composent le mot.

Or, $AC(w)$ est une mesure qui ne peut être utilisée en l'état comme mesure de confiance puisque qu'elle ne respecte pas la propriété d'appartenance à l'intervalle $[0;1]$. Afin que cette mesure respecte cette condition, nous avons proposé une normalisation qui permet de rester dans l'intervalle $[0;1]$ grâce à une transformation sigmoïdale (44). Cette transformation permet d'obtenir la mesure de confiance acoustique normalisée $m_{ac}(w)$ d'un mot w :

$$m_{ac}(w) = \frac{exp(\frac{AC(w)-\mu}{\sigma}) + a}{exp(\frac{AC(w)-\mu}{\sigma}) + 1} \tag{3.3}$$

où μ et σ sont respectivement la moyenne et l'écart-type des mesures acoustiques calculés sur un corpus de développement. Afin de se rapprocher d'une mesure de confiance idéale dont la moyenne des valeurs des mesures de confiance associées à une hypothèse de transcription soit égale aux taux d'erreurs (hors suppression) sur les mots de cette hypothèse, nous pouvons utiliser le corpus de développement pour calculer a :

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

où $t_{correct}$ est le taux de mots corrects sur le corpus de développement.

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

Nous avions fait cette proposition afin de pouvoir combiner cette mesure de confiance acoustique avec d'autre mesures de confiance, et en particulier la mesure présentée par la suite et basée sur des critères propres au comportement du modèle de langage utilisé durant le décodage.

3.2.3 Mesure de confiance *LMBB*

Les systèmes de TAP utilisent généralement des modèles de langage n-grams avec repli : lorsque le décodeur souhaite connaître la probabilité d'apparition d'un n-gram et que la séquence de mots correspondante n'a pas été observée dans le corpus d'apprentissage, le modèle utilisera le n-gram d'ordre inférieur (45), *i.e.* le (n-1)-gram pour calculer une probabilité. Cette réduction de la taille de l'historique est effectuée de manière récurrente jusqu'à obtention d'un n-gram observé : inévitablement, si tous les mots du vocabulaire ont été observés dans le corpus d'apprentissage, il existera au moins le 1-gram correspondant. Dans le cas contraire, soit le modèle de langage contient un item pour gérer les mots inconnus et le 1-gram de cet item sera utilisé, soit le modèle de langage utilise un vocabulaire fermé différent du vocabulaire du système de TAP (ce qu'il vaut mieux éviter), et il faut prévoir une valeur constante pour gérer ce cas extrême (zéroton).

En faisant l'hypothèse que la nécessité du recours au repli implique une dégradation de la précision du modèle de langage et donc une dégradation des performances du

décodage, nous avons proposé une mesure de confiance basée sur le comportement du repli lors du décodage d'un mot hypothèse. En faisant également l'hypothèse que les erreurs ont tendance à se propager lors du traitement, cette mesure prend également en compte le comportement du repli pour les mots hypothèses voisins du mot hypothèse ciblé.

Existant

Une proposition basée sur ce principe avait déjà été faite en 1997 dans (46), mais avec une approche différente. Dans (46), les auteurs proposent d'affecter une valeur arbitraire en fonction du comportement du repli (*back-off* en anglais). Ainsi, (46) propose, pour un modèle 3-gram, les valeurs de confiance $VC(w)$ suivantes pour le mot hypothèse w :

- $VC(w) = 1,0$ si w dérive d'un 3-gram
- $VC(w) = 0,8$ si w dérive de deux 2-grams
- $VC(w) = 0,6$ si w dérive d'un 2-gram
- $VC(w) = 0,4$ si w dérive d'un 1-gram et d'un 2-gram
- $VC(w) = 0,3$ si w dérive de deux 1-grams
- $VC(w) = 0,2$ si w dérive d'un 1-gram, mais le mot précédent n'existe pas dans le modèle de langage
- $VC(w) = 0,1$ si w est inconnu

La mesure de confiance $CM(w_i)$ du mot hypothèse w_i proposée par (46) se calcule alors en utilisant la formule suivante qui prend en compte les voisins gauche et droit du mot w_i :

$$\begin{aligned} CM(w_i) = \; & min\{VC(w_{i-2})VC(w_{i-1})VC(w_i), \\ & VC(w_{i-1})VC(w_i)VC(w_{i+1}), \\ & VC(w_i)VC(w_{i+1})VC(w_{i+2})\} \end{aligned}$$

Proposition

À partir des mêmes hypothèses que (46), nous proposons de créer des classes de comportements du repli du modèle de langage, que nous appelons classes LMBB (*Language Model Back-off Behavior*). Nous associerons chaque mot hypothèse proposé par le système à une classe LMBB afin de déterminer une valeur de confiance pour ce mot.

Les classes LMBB Une classe LMBB est construite à partir de trois informations :

1. l'ordre du n-gram le plus élevé qui a été utilisé par le modèle de langage pour attribuer une probabilité au mot hypothèse visé,
2. une comparaison entre l'ordre du n-gram le plus élevé qui a été utilisé pour son voisin de gauche et l'ordre du n-gram le plus élevé qui a été utilisé pour le mot hypothèse,
3. et une comparaison similaire entre le mot hypothèse et son voisin de droite.

Le résultat de la comparaison entre les ordres les plus élevés utilisés pour le décodage du mot hypothèse et chacun de ses voisins contigus est compris parmi les trois possibilités suivante :

1. égalité (=) : l'ordre du n-gram utilisé pour le mot voisin est le même que celui du mot hypothèse ciblé,
2. infériorité (−) : l'ordre du n-gram utilisé pour le mot voisin est plus petit que celui du mot hypothèse ciblé,
3. supériorité(+) : l'ordre du n-gram utilisé pour le mot voisin est plus grand que celui du mot hypothèse ciblé.

L'utilisation du résultat de cette comparaison est préférable à l'utilisation de l'ordre exact du n-gram utilisé pour le décodage d'un mot voisin car nous souhaitons limiter le nombre de classes LMBB. Ainsi, pour un modèle n-gram, nous créerons $9(n+1)$ classes LMBB au lieu de $(n+1)^3$ si nous n'avions pas effectué cette généralisation.

Les valeurs de confiance affectées aux classes LMBB Pour chacune des classes LMBB, nous proposons de calculer une valeur de confiance qui lui sera associée. Cette valeur de confiance sera estimée à partir de l'analyse des taux d'erreurs des mots d'un corpus d'apprentissage de ces valeurs de confiance en fonction de leur classe LMBB. Ainsi, à partir des sorties du système de TAP obtenues sur un corpus dédié à cet usage et dont on dispose de transcriptions manuelles, on calcule le taux d'erreurs de reconnaissance des mots qui composent chacune des classes LMBB. Ce taux d'erreurs 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.

Pour un mot w associé à la classe LMBB cl, la valeur m_{lmbb} fournie par la mesure de confiance LMBB se calcule à partir d'un corpus d'apprentissage avec la formule :

$$m_{lmbb}(w) = 1 - \frac{n_{err}(cl)}{n_{mots}(cl)} \tag{3.4}$$

Si $n_{mots}(cl) = 0$, alors $m_{lmbb}(w) = 0$

La figure 3.1 montre l'existence d'une corrélation entre le comportement du repli du modèle de langage et le taux d'erreurs de reconnaissance. Ces résultats confirment l'hypothèse de départ qui supposaient cette corrélation. Ils ont été calculés sur 4h d'enregistrement du corpus de développement de la campagne ESTER 1 : nous appellerons ces données $CTrain$. La classe 1 correspond à la fusion des classes (x,1,y) où x et y sont une des étiquettes −, +, ou =.

3.3 Evaluation des mesures de confiance

Plusieurs métriques existent pour évaluer les mesures de confiance (47). Nous présenterons ici quelques résultats utilisant les métriques CER *Confidence Error Rate* et NCE *Normalized Cross Entropy*.

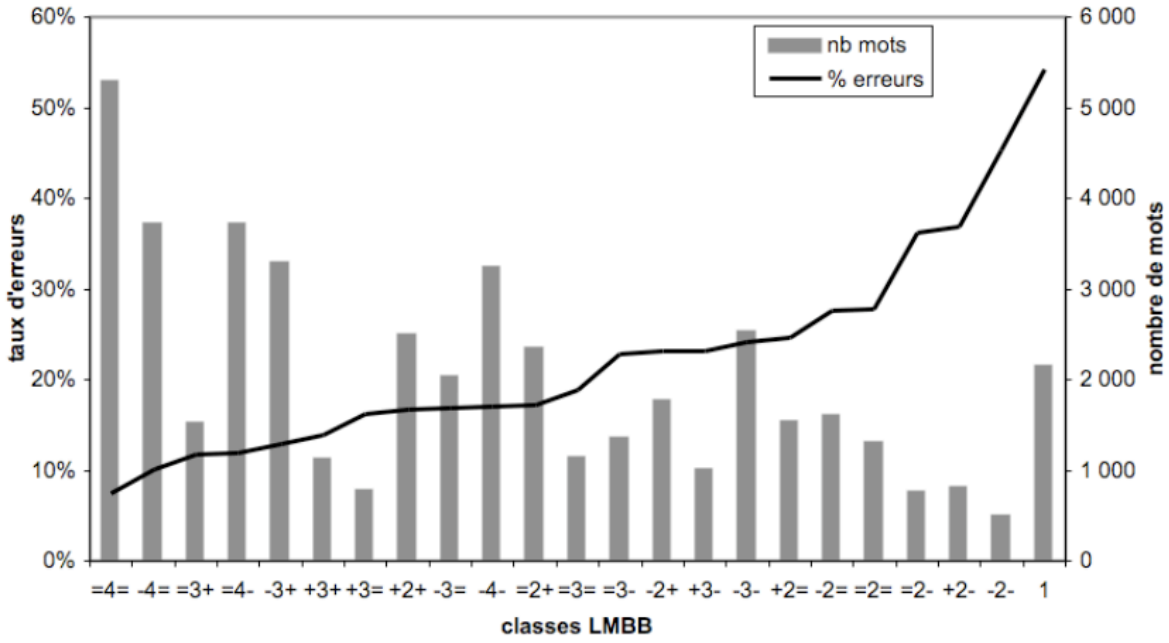

FIG. 3.1 – Taux d'erreurs et répartition des mots transcrits sur la moitié du corpus de développement de ESTER 1 en fonction de la classe LMBB

3.3.1 *Confidence Error Rate*

La mesure CER, utilisée par exemple dans (39), permet de mesurer la capacité d'une mesure de confiance pour l'acceptation/rejet d'une hypothèse. Pour un seuil α, on attribue une étiquette au mot w en fonction de la mesure de confiance $CM(w)$:

- *acceptation* : si $CM(w) > \alpha$,
- *rejet* sinon.

Elle se calcule à partir de la formule suivante :

$$CER = \frac{Nombre\ d'etiquettes\ incorrectement\ assignees}{Nombre\ total\ d'etiquettes} \tag{3.5}$$

Les mesures de confiance ne sont pas utilisées uniquement pour offrir un choix binaire : les scores de confiance étant compris entre 0 et 1, il peut être plus utile, en fonction de l'application visée, d'utiliser une autre mesure de confiance pour prendre en compte la continuité de ces scores. Un métrique plus adaptée est l'entropie normalisée croisée.

3.3.2 Entropie Normalisée Croisée

L'entropie normalisée croisée (NCE) est une des métriques les plus utilisées. Elle est en particulier la métrique utilisée lors des campagnes d'évaluation NIST pour évaluer la qualité des mesures de confiance. La NCE est une estimation de l'information additionnelle que la mesure de confiance apporté à l'hypothèse de reconnaissance. Cette métrique est calculée à partir de la formule suivante :

$$NCE = \frac{H_{max} + \sum_{w\ corrects} \log_2(m(w)) + \sum_{w\ incorrects} \log_2(1 - m(w))}{H_{max}} \quad (3.6)$$

où :

- $H_{max} = -n \log_2(p_c) - (N - n) \log_2(1 - p_c)$,
- n est le nombre de mots reconnus qui sont corrects,
- N est le nombre total de mots reconnus,
- p_c est la probabilité moyenne qu'un mot reconnu soit correct ($= n/N$),
- $m(w)$ est la mesure de confiance associé au mot w.

Plus la valeur de la NCE se rapproche de 1, plus les prédictions de la mesure de confiance sont fiables.

3.3.3 Résultats expérimentaux

Le tableau 3.1 compare les résultats, en termes de NCE, entre notre mesure LMBB et la mesure décrite dans (46) et présentée plus haut : ces deux mesures sont basées sur les mêmes hypothèses quant à l'impact de l'utilisation du repli par un modèle de langage lors d'un décodage et la prise en compte du voisinage d'un mot, mais les solutions proposées sont différentes. Comme on peut le constater, notre mesure est très significativement plus performante que la mesure présentée en 1997 dans (46), cette dernière, avec une NCE négative, n'apportant pratiquement aucune information additionnelle.

Cette expérience a été menée avec notre système de TAP en utilisant un modèle 4-gram : la mesure dans (46) étant limitée à des 3-grams, nous avons considéré les 4-grams comme des 3-grams. Le corpus $CTrain$ correspond à 4h du corpus de développement de la campagne ESTER 1, et le corpus $Test$ est le corpus de test officiel de cette campagne d'évaluation.

Tab. 3.1 – Entropies normalisées croisées des mesures de confiance étudiées obtenues sur le corpus ESTER 1

	NCE	
Mesure de confiance	CTrain	Test
Mesure présentée dans (46)	-1,702	-1,713
LMBB	0,080	0,063

Le tableau 3.2 présente les résultats en terme de NCE des différents mesures de confiance présentées plus haut. En particulier, ce sont les mesures de confiance que nous avons tenté de fusionner, comme nous le verrons dans la suite de ce document.

Nous pouvons constater que la probabilité *a posteriori* est la mesure qui obtient la valeur de NCE la plus élevée. En particulier, lorsque les valeurs données par cette mesure de confiance sont *mappées*, c'est-à-dire qu'elles subissent une transformation linéaire par segments afin d'optimiser les résultats en terme de NCE, comme cela est par exemple fait dans (40) : cette mesure est notée MAP(PAP).

Tab. 3.2 – Entropies normalisées croisées des mesures de confiance seules obtenues sur le corpus ESTER 1

	NCE	
Mesure de confiance	CTrain	Test
AC	0,019	0,023
LMBB	0,080	0,063
PAP	0,182	0,187
MAP(PAP)	0,300	0,293

Le tableau 3.3 compare ces mêmes mesures de confiance en terme de CER. Nous pouvons noter que la hiérarchie est conservée, et que, logiquement, le *mapping* de la probabilité *a posteriori* ne modifie pas son CER : la technique de *mapping* modifie les valeurs de la mesure de confiance mais n'en modifie par l'ordre. Dès lors, seule la valeur du seuil α qui permet d'accepter ou rejeter un mot sera modifiée, mais les résultats en terme d'acceptation/rejet seront identiques pour une mesure de confiance *mappée* et cette même mesure qui ne l'est pas.

Tab. 3.3 – *Confidence Error Rate* (CER) des mesures de confiance seules obtenues sur le corpus ESTER 1

	CER	
Mesure de confiance	CTrain (%)	Test (%)
référence	15,09	19,23
AC	15,08	22,31
LMBB	15,11	22,23
PAP	13,56	18,64
MAP(PAP)	13,56	18,64

Pour ces deux métriques d'évaluation, la probabilité *a posteriori*, confirme son efficacité et constituera par la suite la mesure de confiance de base à laquelle nous nous comparerons.

3.4 Fusion de mesures de confiance

Durant la thèse de Julie Mauclair (35), nous avons expérimenté plusieurs méthodes de fusion, comme la théorie des probabilités ou l'interpolation linéaire, pour combiner les mesures de confiance à notre disposition afin d'en améliorer les performances globales.

La méthode la plus efficace a été l'interpolation linéaire. De plus, il est à noter que la mesure de confiance acoustique n'apportait pratiquement aucune information permettant une amélioration de la mesure de confiance finale en terme de NCE ou

de CER : elle n'est donc pas prise en compte dans la combinaison qui se limite donc à enrichir la probabilité *a posteriori* (PAP) d'informations sur le comportement du modèle de langage.

Le tableau 3.4 présente les résultats en terme de NCE de la mesure de confiance résultant d'une interpolation linéaire entre la PAP et la mesure LMBB. Les différentes variantes proviennent de l'usage d'une fonction de *mapping* :

- aucun *mapping* n'est effectué : $PAP/LMBB$,
- le *mapping* est effectué après combinaison : $MAP(PAP/LMBB)$,
- le *mapping* est effectué avant combinaison sur la PAP et sur la mesure LMBB : $MAP(PAP)/MAP(LMBB)$
- le *mapping* est effectué avant combinaison uniquement sur la PAP : $MAP(PAP)/LMBB$,

Tab. 3.4 – Entropies normalisées croisées des fusions de mesures de confiance obtenues sur le corpus ESTER 1

	NCE	
Mesure de confiance	CTrain	Test
PAP / LMBB	0,277	0,284
MAP(PAP/LMBB)	0,296	**0,299**
MAP(PAP) / MAP(LMBB)	0,301	0,294
MAP(PAP) / LMBB	**0,304**	0,296
MAP(PAP)	0,300	0,293

Nous pouvons constater que la combinaison de la PAP avec la mesure LMBB, à condition d'utiliser une technique de *mapping*, permet d'obtenir de meilleurs résultats, en termes de NCE, que la mesure de base, à savoir la PAP.

Toutefois, les gains sont plus significatifs en termes de CER, comme le montre le tableau 3.5. La ligne de ce tableau nommée *référence* correspond aux résultats obtenus en considérant que tous les mots proposés par le système de TAP sont retenus. Cela correspond aux taux d'erreurs sur les mots émis (qui diffèrent du *WER*, taux d'erreurs sur les mots, par la non prise en compte des suppressions pour compter les erreurs). La baisse absolue du CER apportée par la PAP par rapport à la *référence*, vaut $0,59$,alors que la combinaison de la mesure LMBB avec la PAP permet de réduire d'encore de $0,34$, en absolu, la CER.

Ainsi, la mesure qui semble la plus intéressante est la mesure $MAP(PAP/LMBB)$ qui obtient les meilleurs résultats quelque soit la métrique utilisée.

3.5 Filtrage de données pour l'apprentissage de modèles acoustiques

La taille de leur corpus d'apprentissage est déterminante pour la robustesse des modèles acoustiques. Or, il s'agit de données coûteuses à collecter. Nous avons expérimenté (3; 48) l'usage des mesures de confiance afin de construire des corpus d'apprentissage de modèles acoustiques de façon totalement automatique (49).

TAB. 3.5 – *Confidence Error Rate* (CER) des fusions de mesures de confiance obtenues sur le corpus ESTER 1

	CER	
Mesure de confiance	CTrain (%)	Test (%)
référence	15,09	19,23
PAP / LMBB	13,17	**18,27**
MAP(PAP/LMBB)	13,17	**18,27**
MAP(PAP) / MAP(LMBB)	13,43	18,48
MAP(PAP) / LMBB	13,23	18,28
PAP	13,56	18,64

Nous avons principalement utilisé la mesure de confiance $PAP/LMBB$ pour filtrer les sorties automatiques obtenues avec notre système de TAP sur des enregistrements audio. Ces enregistrements font partie des 1500 heures d'enregistrements fournis par les organisateurs de la campagne d'évaluation ESTER1 sans transcription manuelle. Le sous-ensemble utilisé pour cette étude correspond à 558 heures d'enregistrements, composées de 58 heures de bande étroite (téléphone) et de 500 heures de bande large (studio). Nous appellerons *CorpusAuto* ce corpus de 558 heures.

3.5.1 Méthode de filtrage

Pour filtrer les hypothèses de reconnaissance, nous avons fixé une durée minimale de segment audio à accepter : seules les séquences de mots retenues dont le signal audio associé dépasse 4 secondes sont conservées. Ceci permet de garantir un nombre de phonèmes en contexte suffisamment important afin de ne pas sur-apprendre des phonèmes de début et fin de segment lors de l'estimation des modèles acoustiques.

Notre méthode de filtrage consiste donc à ne conserver que les séquences de mots dont le signal audio associé dépasse 4 secondes et dont la mesure de confiance de chacun des mots dépasse un certain seuil α défini empiriquement. Ce seuil est défini en fonction de la quantité de données que nous souhaitons conserver à partir des sorties automatiques de transcription.

La figure 3.2 indique le taux de mots incorrects, sur le corpus de test de ESTER 1, après filtrage en fonction du taux de mots rejetés, ce taux de rejet variant ici entre 75% et 95%. Le taux de mots incorrects est bas puisque qu'il ne dépasse pas 3,7%.

La méthode de filtrage *sans rattrapage de mots* correspond à la méthode présentée au-dessus. Pour obtenir un taux d'erreurs sur les mots émis, on ne conserve seulement qu'un peu plus de 10% du corpus initial. Nous avons alors proposé de relâcher la contrainte du seuil unique en raffinant le filtrage. Sur une fenêtre glissante de quatre mots, on choisit de conserver les mots sous deux conditions :

1. au moins un mot doit avoir un score de confiance supérieur à un seuil α_1,
2. les autres mots doivent tous avoir un score de confiance supérieur à un seuil α_2, avec $\alpha_2 \leq \alpha_1$

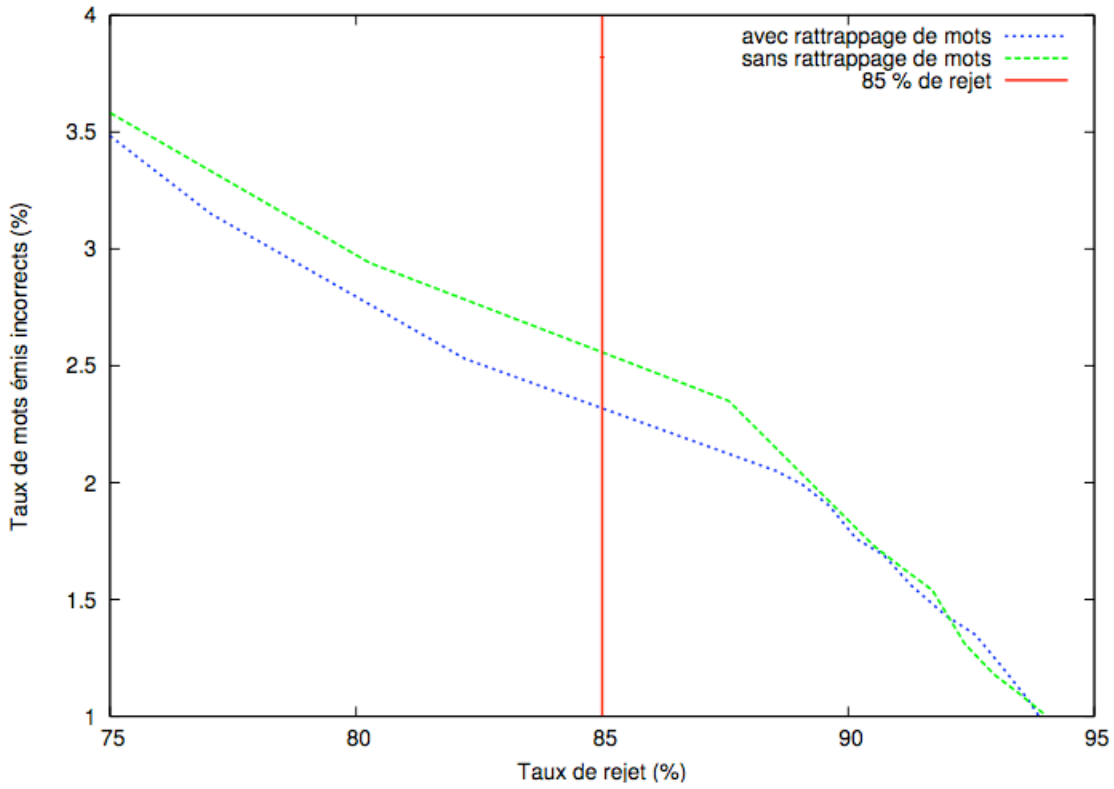

FIG. 3.2 – Taux d'erreurs et répartition des mots transcrits sur le corpus de test de ESTER 1 en fonction de la classe LMBB

La contrainte de 4 secondes est conservée pour l'acceptation d'une séquence de mots.

Cette méthode permet de conserver des mots ayant un score de confiance plus faible, ce qui signifie que si ces mots sont corrects, il peuvent fournir des informations supplémentaires aux modèles acoustiques. En effet, avec cette méthode de filtrage nous tentons d'éviter de réinjecter dans le corpus d'apprentissage des informations déjà connues des modèles acoustiques.

Dans la figure 3.2, les résultats de cette méthode de filtrage sont représentés par la courbe nommée *avec rattrapage de mots*. Pour cette courbe, la valeur de α_2, optimisée sur $CTrain$, est fixée à 0,5 pendant que α_1 varie entre 0,5 et 1.

TAB. 3.6 – Taux de mots retenus erronés en fonction de la méthode de filtrage employée pour des segments de parole de plus de 4 secondes sur le corpus de test de ESTER 1

Méthode de filtrage	Taux de mots retenus erronés
sans rattrappage de mots	2,63%
avec rattrappage de mots	2,31%

Le tableau 3.6 précise les taux de mots erronés obtenus sur le corpus de test de ESTER 1 lorsque α_2 vaut 0,5 et α_1 vaut 0,96. Le seuil α_1 a été choisi afin de garantir l'extraction d'environ 80 heures de données filtrées, ce qui correspond à la taille du corpus d'apprentissage initial que nous souhaitons doubler, provenant des 558 heures faisant partie du corpus additionnel *CorpusAuto* présenté plus haut. Comme en l'absence de transcriptions manuelles nous ne pouvons pas évaluer le taux de mots incorrects obtenu après filtrage sur *CorpusAuto*, nous utilisons le corpus de test ESTER 1 en espérant qu'il en donne une bonne approximation.

3.5.2 Résultats expérimentaux

Nous avons utilisé notre méthode de filtrage avec rattrapage de mots sur le corpus additionnel *CorpusAuto* en utilisant la mesure de confiance PAP/LMBB décrite dans la section précédente. Ceci dans le but d'augmenter la taille du corpus d'apprentissage des modèles acoustiques.

Le tableau 3.7 montre que les données filtrées retenues concernent aussi bien les données téléphoniques que les données de qualité studio. Bien entendu, cette classification bande étroite/bande large sur les données additionnelles a été effectuée de manière automatique. Les données téléphoniques sont proportionnellement moins importantes parmi les données ajoutées que dans le corpus d'apprentissage initial. Cela est dû au fait que ces données étant plus difficiles à transcrire, la mesure de confiance ne leur a pas attribué un score suffisamment élevé pour être retenues en grande quantité. Il faut également observer que les transcriptions automatiques de données téléphoniques ont généralement un taux d'erreurs sur les mots bien plus élevé que les transcriptions automatiques de données de qualité studio.

Tab. 3.7 – Répartition du corpus d'apprentissage en fonction de la bande passante

Corpus d'apprentissage	Bande étroite (téléphone)	Bande large (studio)
Apprentissage ESTER 1 : 80h (Ω)	8h	72h
(Ω) + 86h	11h	155h

Les résultats du système de TAP, en terme de taux d'erreurs sur les mots (WER) sur le corpus de test ESTER 1, en fonction de la quantité des données filtrées et ajoutées dans le corpus d'apprentissage des modèles acoustiques sont présentés dans le tableau 3.8. Le taux d'erreurs de référence, sans ajout de données dans le corpus d'apprentissage des modèles acoustiques, est de 23,7% pour un nombre de mots dans les transcriptions de référence égal à 114000. Si l'on considère qu'un résultat est statistiquement significatif lorsqu'il est en dehors de l'intervalle de confiance à 95%, comme proposé dans (50), alors un gain est significatif si le WER obtenu est en dessous de 23,45%.

Comme nous pouvons le constater, le WER obtenu par le simple ajout de 86 heures filtrées dans le corpus d'apprentissage ne permet pas ici d'atteindre un gain significatif. Nous pensons que cela est dû au nombre d'états partagés qu'utilisent les

HMM des modèle acoustiques. Leur nombre initial, à savoir 5500, a été optimisé en fonction de la taille du corpus d'apprentissage de départ, corpus que l'on nomme Ω dans ce tableau : pour une taille de corpus d'apprentissage plus élevée, il semble que les modèles acoustiques *saturent* et ne sont pas en mesure d'intégrer de nouvelles connaissances.

En augmentant le nombre d'états partagés, les modèles acoustiques deviennent plus précis lorsque la taille du corpus d'apprentissage augmente en conséquent : c'est ce qui se passe ici où avec 10000 états partagés au lieu de 5500 et en doublant la taille du corpus d'apprentissage, le système de TAP obtient un gain statistiquement significatif de 0,5% en absolu sur le WER.

Tab. 3.8 – Taux d'erreurs sur les mots en fonction de la taille du corpus d'apprentissage ajouté et du nombre d'états partagés utilisés dans les modèles acoustiques

Corpus d'apprentissage Corpus d'apprentissage	Nombre d'états partagés	WER
Apprentissage ESTER 1 : 80h (Ω)	5500	23,7%
(Ω) + 86h 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) + 11h filtrées (autres fichiers + filtrage AC)	5500	23,4%
(Ω) + 28h filtrées (autres fichiers + filtrage AC)	5500	23,3%

Enfin, il est à noter que d'autres expériences préliminaires, utilisant d'autres données audio provenant également des enregistrements d'audio brut de ESTER 1, nous avaient montré qu'il est possible de réduire significativement le WER en injectant seulement 28 heures de données filtrées, voire uniquement 11h. Il s'agit, dans le tableau 3.8 des deux dernières lignes avec la mention *autres fichiers + filtrage AC*.

Pour cela, nous avions utilisé la mesure de confiance acoustique normalisée décrite précédemment qui, par la suite de notre étude, avait été éliminée en raison de ses mauvaises performances en CER et NCE. Cela indique qu'il est possible que ces métriques d'évaluation ne soient pas les plus pertinentes quand il s'agit de filtrer des transcriptions automatiques pour collecter du corpus aligné audio/phonèmes dans le but d'estimer des modèles acoustiques. Mais peut-être que cela provient simplement des fichiers audio traités.

Il serait ainsi intéressant de reprendre ces expériences pour lever ces interrogations. Or, le temps de calcul nécessaire pour mener ces expériences est très important et le système de TAP du LIUM a largement évolué : il faudrait donc recommencer l'ensemble de ces expériences, ce qui nécessite plusieurs semaines de travail et de calculs, avec pour objectif de déterminer si la mesure de confiance acoustique normalisée est préférable à d'autres mesure *a priori* plus efficaces.

3.6 Combinaison de systèmes

Depuis quelques années, de nombreux travaux ont visé la combinaison de systèmes de TAP. Plusieurs approches ont été proposées qui diffèrent par la méthode de partage d'informations, les informations partagées et/ou par le niveau auquel la combinaison de systèmes se réalise dans de la chaîne de traitements. Certains chercheurs ont présenté des travaux pour lesquelles la combinaison se fait au niveau du traitement acoustique (51; 52), mais la plupart des combinaisons se réalisent généralement *a posteriori* à partir des hypothèses proposées par les différents systèmes de TAP à combiner, soit par un système de vote de type ROVER (53), soit en fusionnant des réseaux de confusion (40).

3.6.1 L'approche par DDA (*Driven Data Algorithm*) pour combiner deux systèmes

Le Laboratoire d'information d'Avignon (LIA) a développé un algorithme de recherche pour son système de TAP (54), nommé *Speeral*, afin de pouvoir utiliser des transcriptions imparfaites (initialement il s'agissait de sous-titres) associées à un enregistrement audio dans le but d'exploiter ces transcriptions incorrectes et souvent incomplètes pour améliorer la transcription automatique de cet enregistrement (55). Nous avons travaillé avec le LIA et l'IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires) sur l'application de cette technique à la combinaison de systèmes de TAP (56; 57).

Speeral *Speeral* a la spécificité d'être basé sur l'algorithme de recherche A*. Le décodage dépend donc d'une fonction d'estimation $F(h_n)$ qui estime la probabilité de l'hypothèse de reconnaissance h_n arrivant au nœud n de l'espace de recherche à l'aide de la formule suivante :

$$F(h_n) = g(h_n).p(h_n) \tag{3.7}$$

où $g(h_n)$ est la probabilité du chemin partiel qui compose l'hypothèse du début de l'espace de recherche jusqu'au nœud n et $p(h_n)$ est une fonction *sonde* qui estime la probabilité du chemin partiel le plus probable entre le nœud n et la fin de l'espace de recherche. Dans *Speeral*, cet espace de recherche correspond à un graphe de phonèmes.

DDA pour combiner deux systèmes de TAP Le décodeur *Speeral* génère des mots hypothèses au fur et à mesure de l'exploration du graphe de phonèmes, les meilleures hypothèses à l'instant t étant développées en fonction de $F(h_n)$. Le principe de l'algorithme DDA appliqué à la combinaison de systèmes consiste à intégrer dans le calcul de $F(h_n)$ les informations fournies par un autre système de TAP, le décodeur *Speeral* fonctionnant alors dans une relation maître/esclave avec l'autre système de TAP qui n'est alors considéré que comme un système de TAP auxiliaire. Plus précisément, c'est la partie linguistique de $g(h_n)$ qui prend en compte ces informations.

Le système auxiliaire fournisse des séquences de mots comme hypothèses de reconnaissance, alors que la génération d'hypothèses de reconnaissance par *Speeral* se produit en parcourant un graphe de phonèmes : il est nécessaire pour chaque mot décodé de trouver un point de synchronisation avec les transcriptions du système auxiliaire. Ces points de synchronisation sont trouvés en alignant l'hypothèse de *Speeral* avec la transcriptions fournie, cet alignement minimisant la distance d'édition entre elles.

Ce processus d'alignement permet d'identifier, dans la transcription auxiliaire H_{aux}, la meilleure séquence de mots qui correspond à l'hypothèse courante en cours de construction H_{cur}. Cette sous-séquence de mots qui apparaissent dans H_{aux} et H_{cur} est prise en compte dans le score linguistique utilisé dans *Speeral* à travers deux valeurs :

1. un score de correspondance $\theta(w_i)$ qui est simplement le nombre de mots identiques dans l'alignement entre H_{aux} et H_{cur},
2. les scores de confiance $\phi(w_i)$ des mots w_i, ces scores de confiance étant fournis par le système auxiliaire.

Avec l'algorithme DDA appliqué à la combinaison de deux systèmes, que l'on nommera *DDA-2* le score linguistique L utilisé pour calculer g de la formule (3.9) est calculé à partir de la formule suivante :

$$L(w_i|w_{i-2}w_{i-1}) = P(w_i|w_{i-2}w_{i-1})^{1-\beta}.\alpha(w_i)^{\beta} \tag{3.8}$$

où :

- $P(w_i|w_{i-2}w_{i-1})$ est la probabilité 3-gram initiale du modèle de langage de *Speeral*,
- β est un *fudge factor* empirique optimisé sur un corpus de développement,
- $\alpha(w_i)$ est le score de confiance du mot hypothèse w_i. Ce score est calculé à l'aide de la formule suivante :
  - si $\theta(w_i) > 0$ alors $\alpha(w_i) = \phi(w_i).\frac{\theta(w_i)}{\lambda}$ et $\beta = 0,6$
  - sinon $\beta = 0$ et $\alpha(w_i)^{\beta} = 1$

  où λ est la taille de la fenêtre d'analyse utilisée pour le calcul de l'alignement.

3.6.2 Résultats expérimentaux pour *DDA-2*

L'approche DDA-2 a été expérimentée en combinant les sorties du système de TAP de l'IRISA (58), *Irene*, avec *Speeral*, et en combinant également les sorties du système de TAP du LIUM avec *Speeral*. Les scores de confiance fournis par le système de TAP du LIUM sont des scores calculés avec la mesure PAP/LMBB.

Chacune des combinaisons a été effectuée soit avant l'adaptation acoustique des modèles acoustiques de *Speeral*, soit après. Ces expériences ont été menés sur 3 fichiers d'une heure chacun du corpus ESTER 1.

Le tableau 3.9 montre les taux d'erreurs sur les mots obtenus avec DDA-2, ainsi que les taux d'erreurs sur les mots des transcriptions proposées initialement par les différents systèmes de TAP seuls.

Comme nous pouvons le constater, quelque soit la combinaison effectuée, celle-ci offre de meilleurs résultats que les résultats du meilleur des systèmes qui compose

TAB. 3.9 – Taux d'erreurs sur les mots en fonction des systèmes combinés ou non par DDA-2 sur 3 heures du corpus de ESTER 1 avant (P1) ou après (P2) adaptation des modèles acoustiques de *Speeral*, le système du LIA

Système/Combinaison	France Inter	France Info	RFI
LIA	21,1	22,2	**24,6**
LIUM	**18,5**	**18,9**	25,6
IRISA	21,4	21,8	25,6
DDA-2 IRISA-P1	19,6	19,3	23,5
DDA-2 IRISA-P2	18,7	18,7	22,2
DDA-2 LIUM-P1	17,8	18,1	22,4
DDA-2 LIUM-P2	**17,2**	**17,8**	**21,5**

cette combinaison. Dans l'absolu, ce sont les combinaisons entre le système du LIUM et *Speeral*, après adaptation des modèles acoustiques du décodeur du LIA. Ceci semble logique puisque le système du LIUM est le système qui, globalement, obtient les meilleurs résultats seuls, sauf sur le fichier de la radio RFI. D'ailleurs, il est intéressant de noter qu'alors que le système du LIUM et celui de l'IRISA obtiennent le même taux d'erreurs sur ce dernier fichier, après combinaisons avec *Speeral* c'est celle qui utilise les sorties du système du LIUM qui obtient les meilleurs résultats. Ceci peut s'expliquer soit par le type d'erreurs engendrées par le systèmes (des erreurs de suppressions par exemple aideront moins au niveau de la combinaison), soit par la qualité des mesures de confiance. Nous n'avons pas encore mené d'étude pour expliquer précisément ce phénomène.

3.6.3 L'approche par DDA (*Driven Data Algorithm*) pour combiner plusieurs systèmes

L'approche par DDA a ensuite été généralisé de façon à combiner plus de deux systèmes. On notera *DDA-n* la combinaison de n systèmes : *n-1* systèmes auxiliaires et le système principal, ici *Speeral*.

Dans ce cas, toutes les transcriptions des systèmes auxiliaires sont d'abord traitées indépendamment pour mettre en œuvre la synchronisation entre chaque transcription et l'hypothèse de reconnaissance de *Speeral* et calculer le score de correspondance $\theta(w_i)$. Le score linguistique final est calculé à travers une combinaison des différents scores linguistiques à l'aide de la formule suivante :

$$L(w_i|w_{i-2}w_{i-1}) = P(w_i|w_{i-2}w_{i-1})^{1-\beta}.\frac{1}{N}\sum_{k=0}^{N}\alpha_k(w_i)^{\beta_k} \tag{3.9}$$

où :

- α_k est le score de confiance du mot hypothèse w_i fourni par le système k,
- β est la moyenne pondérée des valeurs β_k optimisées sur un corpus de développement,
- N est le nombre de systèmes auxiliaires.

3.6.4 Résultats expérimentaux pour *DDA-n*

Cette approche a été expérimentée pour combiner les trois systèmes évoqués précédemment (le système du LIA, celui de l'IRISA et celui du LIUM). L'approche ROVER étant très souvent utilisées pour combiner les sorties de systèmes de TAP, nous comparons l'approche DDA à cette approche. La méthode ROVER (53) consiste à aligner les différentes hypothèses et à procéder à un vote mot à mot pondéré par les scores de confiance donnés par les différents systèmes.

Le tableau 3.10 montre que la combinaison des trois systèmes par la méthode DDA (ligne *DDA-3*) donne significativement de meilleurs taux d'erreurs sur les mots que la méthode ROVER-3 (les transcriptions du système du LIA étant celles obtenus sans utiliser les transcriptions auxiliaires) sur chacun des fichiers audio utilisés pour l'expérience. Nous pouvons également noter que l'ajout d'un système auxiliaire permet d'améliorer les résultats : les résultats de DDA-3 sont meilleurs que ceux obtenus par la meilleure combinaison par DDA-2.

Enfin, si on ajoute les transcriptions obtenues par DDA-3 aux combinaisons des deux systèmes auxiliaires pour réaliser un vote par ROVER, les résultats sont globalement améliorés.

Tab. 3.10 – Taux d'erreurs sur les mots en fonction des systèmes combinés ou non par DDA-2 sur 3 heures du corpus de ESTER 1 avant (P1) ou après (P2) adaptation des modèles acoustiques de *Speeral*, le système du LIA

Système	France Inter	France Info	RFI
Meilleur système seul	18,5	18,9	24,6
ROVER-3	17,1	18,2	22,5
meilleur DDA-2	17,2	17,8	21,5
DDA-3	16,7	17,0	20,6
DDA-3+ROVER	16,0	16,4	20,7

Le tableau 3.11 montre également un phénomène intéressant : si nous étions capable de trouver l'hypothèse de reconnaissance qui minimise le taux d'erreurs sur les mots (taux *Oracle*) à partir des transcriptions données par les trois systèmes de TAP seuls, nous obtiendrons résultats inscrits sur la ligne *ORACLE-3*. Or, si nous remplaçons parmi ces transcriptions les sorties de *Speeral* utilisé seul par ses sorties obtenues par DDA-3, le taux d'erreurs *Oracle* obtenu est meilleur. Cela prouve qu'à la différence de la méthode ROVER dont la combinaison est réalisée *a posteriori*, il est possible, en utilisant DDA, de modifier l'espace de recherche du système de TAP auxiliaire et de récupérer certaines hypothèses parfois correctes qui auraient été éliminées en raison des heuristiques locales d'élagage effectuées pour limiter le temps de calcul.

TAB. 3.11 – Taux d'erreurs sur les mots en fonction des systèmes combinés ou non par DDA-2 sur 3 heures du corpus de ESTER 1 avant (P1) ou après (P2) adaptation des modèles acoustiques de *Speeral*, le système du LIA

Système	France Inter	France Info	RFI
ORACLE-3	10,3	10,5	14,5
ORACLE DDA-3+ROVER	9,8	10,0	13,6

3.7 Conclusion

Ce chapitre résume les travaux que nous avons d'abord réalisé durant la thèse de Julie Mauclair. Lors de ce travail, nous avons proposé une normalisation d'une mesure de confiance acoustique qui permet par exemple de pouvoir l'utiliser en combinaison avec d'autres mesures. Nous avons également proposé une mesure de confiance basée sur le comportement du repli du modèle de langage $n-gram$ lors du décodage : les principes sur lesquels repose notre mesure avaient déjà été exploités dans le passé (46) mais notre approche est très différente et améliore grandement ce qui avait été fait auparavant. Nous avons fusionné cette mesure, appelée mesure LMBB, avec la probabilité *a posteriori* qui est habituellement utilisée comme mesure de confiance. Le résultats, la mesure PAP/LMBB obtient des résultats légèrement meilleurs que la PAP seule après *mapping* en terme de NCE, et notre mesure est sensiblement meilleure en terme de CER sur ce que nous avons pu observer lors de nos expériences.

Nous avons utilisé la mesure de confiance PAP/LMBB pour filtrée des transcriptions automatiques afin d'augmenter le corpus d'apprentissage des modèles acoustiques de notre système de TAP : des gains statistiquement significatifs ont été relevés.

Enfin, nous avons continué d'exploiter les mesures de confiance lors du travail entrepris sur la combinaison de systèmes de TAP avec le LIA et l'IRISA. Il est tout de même important de rappeler que le LIA, et en particulier Georges Linarès dans le cadre de la thèse de Benjamin Lecouteux, a été le principal acteur de cette dernière partie du travail.

Ces travaux préliminaires sur la combinaison de systèmes nous ont amené à une réflexion plus générale sur cette problématique : avec le LIA et l'IRISA, nous avons proposé un projet blanc à l'ANR sur le domaine qui a été accepté en 2009. Ce projet s'appelle ASH (Attelage de systèmes hétérogènes) et j'en suis le coordonateur.

Chapitre 4

Identification nommée du locuteur

Sommaire

4.1 Introduction

Lorsque Sylvain Meignier a été recruté à l'Université du Maine comme maître de conférences, nous nous sommes concertés afin d'étudier quels travaux d'études pourraient nous permettre de mettre nos compétences en commun autour d'un projet fédérateur. Venant tous les deux du Laboratoire d'Informatique d'Avignon, nous avions déjà eu l'occasion de discuter auparavant de projets qui nous semblaient intéressants de développer lorsque l'occasion se présenterait.

Nous avons décidé de travailler à l'identification des locuteurs d'un enregistrement audio, sans connaissance acoustique *a priori* sur ces locuteurs, en exploitation

les sorties d'un système de TAP et de système de segmentation et regroupement en locuteur.

Ce chapitre apporte une synthèse des travaux effectués au LIUM depuis 2005 sur ce sujet, et s'inspire en partie de l'article que nous avons publié dans TAL (27).

4.1.1 Problématique

Pour faciliter la recherche et l'accès à l'information, les grandes collections de données audio ont besoin d'être indexées. Les annotations manuelles sont coûteuses, particulièrement pour la transcription des paroles prononcées, les thèmes des documents ou les noms des locuteurs. Il paraît donc nécessaire de s'intéresser à la réalisation automatique de ces annotations.

Le système présenté dans ces travaux s'intéresse au cas de l'annotation des documents avec l'identité des locuteurs. Cette identité est composée du prénom et du nom du locuteur, ce couple étant appelé par la suite du document "nom complet".

La première étape consiste, à partir du signal acoustique, à segmenter le signal sonore en tours de parole. Ces derniers débutent lorsqu'un locuteur commence à parler et finissent lorsqu'un autre locuteur prend la parole ou qu'un intermède débute (jingle, chanson, publicité...).

Les différents tours de parole sont ensuite regroupés en classes contenant les segments produits par un même locuteur. Elles sont identifiées par des labels anonymes (locuteur 1, locuteur 2...). À ce stade, aucune connaissance *a priori* sur les locuteurs n'est utilisée. Toutefois, une détection du genre (homme ou femme) de chaque classe est réalisée. L'étape suivante consiste à transcrire automatiquement les tours de parole en mots. La transcription du document peut être, en plus, complétée par une annotation en entités nommées. Les entités nommées de type "personne" constituent une source d'information sur les locuteurs du document, mais elles ne permettent pas d'identifier directement qui parle et quand.

4.1.2 Solutions

Pour résoudre ce problème, il existe deux approches principales qui permettent d'attribuer un nom complet à un locuteur.

La première se fonde sur l'analyse exclusive de l'acoustique à partir de méthodes issues de la reconnaissance du locuteur. Par exemple, les systèmes proposés pour la tâche de suivi du locuteur de la campagne d'évaluation ESTER 1 (59) permettent, avec quelques modifications mineures du système de décision, d'identifier les locuteurs d'un document. Ces méthodes reposent sur des modèles de locuteurs appris à partir d'échantillon d'enregistrement de chaque locuteur cible à identifier. Une des difficultés est de collecter ces échantillons de voix pour que les systèmes de reconnaissance du locuteur aient de bonnes performances. Ils doivent être représentatifs des différentes conditions acoustiques rencontrées dans la collection de documents à annoter. Ils doivent idéalement avoir été enregistrés à la même période que les documents de la collection et en quantité suffisante (plusieurs minutes). De plus, ces systèmes sont amenés à traiter des collections susceptibles d'évoluer quotidiennement ; se pose alors le problème de l'ajout des nouveaux locuteurs.

La seconde approche (60; 61; 24; 63) propose d'extraire les noms complets des locuteurs présents dans la transcription enrichie. Le principe général consiste à déterminer si un nom détecté dans les transcriptions se rapporte à un locuteur du document, plus exactement à une classe contenant les segments d'un locuteur du document, ou bien à une personne qui ne parle pas dans le document. L'approche se base sur un système en deux étapes. Une première étape affecte les noms complets aux tours de parole proches. Puis dans une seconde étape, ces informations sont propagées au niveau des classes. Les documents traités dans ce type d'approche sont des enregistrements de journaux radiophoniques car, généralement, les locuteurs s'annoncent ou sont annoncés tout au long de l'enregistrement.

C'est cette dernière approche que nous avons explorée. Notre système utilise un arbre de classification sémantique pour déterminer à quel tour de parole se rapporte un nom complet : il s'agit de décisions locales prises au niveau d'un nom complet détecté dans un segment. Il s'agit ensuite d'exploiter ces décisions locales pour affecter ces noms complets aux labels anonymes associés aux classes de segments censées regrouper chacune les segments de parole d'un même locuteur.

4.2 Hypothèses de travail

Les méthodes proposées dans (60; 61; 62; 63) s'appuient toutes sur une transcription enrichie. Comme le décrit la figure 4.1, il est supposé que :

- le document est découpé en tours de parole,
- les tours de parole d'un même locuteur sont regroupés en classes identifiées par des labels anonymes (par exemple *locuteur1*, *locuteur2*, etc.),
- le genre (homme ou femme) peut être renseigné pour chaque classe : comme nous le verrons plus loin, cette information peut aider à l'affectation d'un nom complet, en particulier en fonction du genre du prénom lorsque celui peut être déterminé,
- la transcription en mots est disponible pour chaque tour de parole,
- les noms complets (prénom + nom) sont détectés dans la transcription.

L'hypothèse de travail majeure proposée initialement dans (60) suppose qu'un nom complet détecté dans un tour de parole permet d'identifier le tour courant ou un des tours de parole directement contigus (tour de parole suivant ou précédent). Les personnes n'intervenant pas dans le document ne présentent pas d'intérêt dans le cadre de l'identification nommée car seuls les locuteurs du document sont recherchés. Ces personnes seront donc considérées comme "autre", de même pour les locuteurs intervenant dans le document mais dans des tours de paroles non contigus.

La figure 4.2 illustre les quatre types d'affectations possibles pour un nom complet détecté dans un tour de parole.

Les méthodes d'identification nommée à partir d'une transcription enrichie ne sont pertinentes que si les locuteurs s'annoncent ou sont présentés. Elles sont bien adaptées aux enregistrements radiophoniques (journaux d'information par exemple) où le passage de parole est généralement indiqué en nommant le locuteur, mais le sont moins pour les enregistrements de réunion, par exemple. En effet, dans les journaux d'information les locuteurs se présentent ou annoncent le locuteur suivant, ils félicitent le précédent ou le suivant, concluent le reportage par leur nom, ... Il

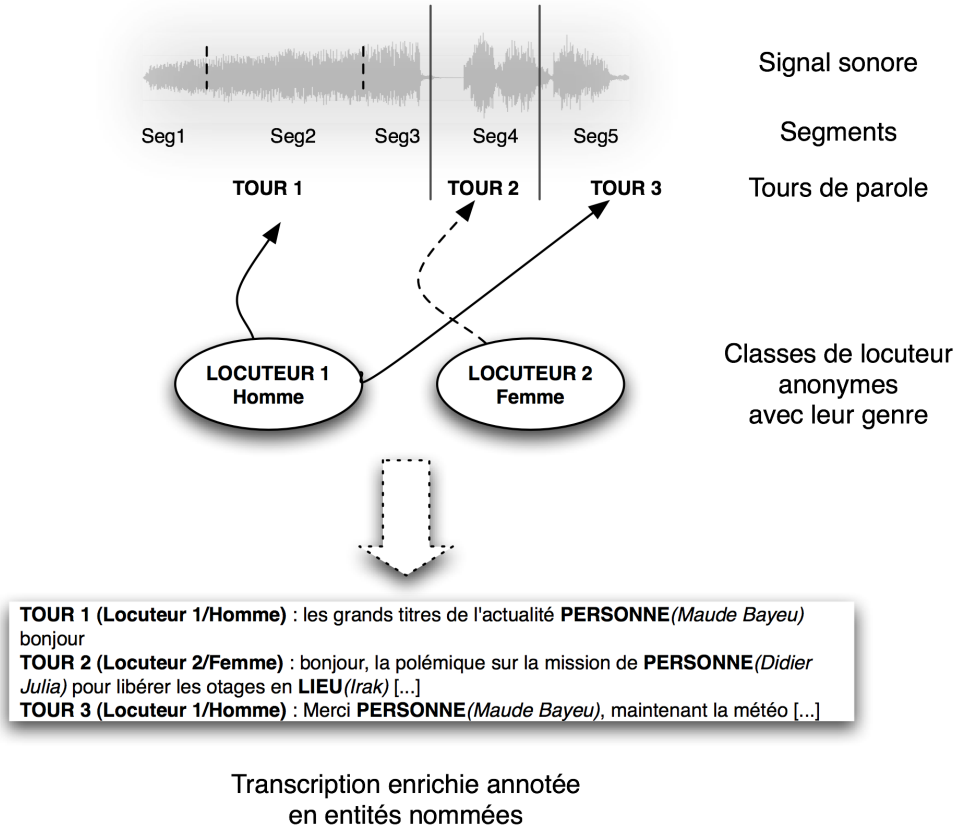

FIG. 4.1 – Informations disponibles dans une transcription enrichie

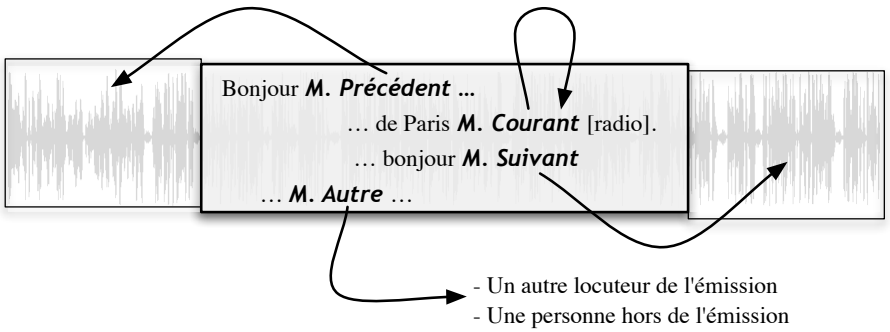

FIG. 4.2 – Principe de base des systèmes d'identification nommée basés sur une analyse conjointe

est donc possible, en prenant en compte des marqueurs spécifiques, de déterminer à quel tour de parole se rapporte un nom complet détecté. Une étude sur le corpus a montré que les noms complets des locuteurs d'émissions radiophoniques étaient, à de rares exceptions près, systématiquement présents dans la transcription (25).

Ces systèmes peuvent être utilisés dans le domaine de la recherche d'information multimédia pour rechercher, par exemple, des interventions de personnes précises dans des collections de documents. Deux cas d'utilisation peuvent être distingués : soit le système connaît les identités des locuteurs cibles et utilise au mieux cette information, soit il n'a aucune connaissance *a priori* sur ces identités.

4.3 Méthodes d'identification nommée à partir de transcription enrichie

Les premiers travaux ont été conduits historiquement dans (60) sur des journaux d'information en langue anglaise. Les auteurs ont été les premiers à montrer que le prénom et le nom d'un locuteur apparaissant dans un contexte lexical donné permettaient d'identifier de manière précise l'identité des locuteurs s'exprimant dans les tours de parole proches. Leur méthode repose sur l'utilisation de règles affectant les étiquettes *"tour courant"*, *"tour précédent"*, *"tour suivant"* aux noms complets détectés. Ces étiquettes ont été reprises dans l'ensemble des systèmes décrits ci-dessous. Les règles utilisées ont été définies manuellement après analyse d'un corpus. 12 règles sont utilisées pour désigner le locuteur courant, 34 pour le suivant et 6 pour le précédent. Cette méthode nécessite un traitement manuel du corpus : les règles sont décrites par un humain. Le temps de mise en place de telles règles peut être très long suivant la quantité de corpus à analyser. Dans les travaux décrits dans (60; 66), 150 heures ont été étudiées. Ces règles sont peu transposables d'un corpus à l'autre : il faut réécrire le jeu de règles pour l'utiliser sur des documents d'une autre langue par exemple. On notera aussi que les entités nommées sont étiquetées manuellement et que le système proposé n'est pas complet. Ces études se sont axées sur la première phase, où les noms complets sont attribués aux tours de parole voisins. La seconde phase, où les locuteurs de l'enregistrement sont nommés grâce à la propagation des décisions globales, n'est pas définie.

La méthode décrite dans (61) propose un système d'apprentissage automatique à base de n-grams pour attribuer une étiquette à un nom complet. Un modèle 3-grams est appris sur une fenêtre glissante de 5 mots autour du nom complet. Ce modèle permet ensuite d'attribuer les étiquettes *"tour courant"*, *"tour précédent"*, *"tour suivant"* ou *"autre"* aux noms complets détectés dans le document traité. À la différence de celle de (60), cette méthode a l'avantage d'utiliser un système d'apprentissage automatique pour attribuer les étiquettes relatives aux tours de parole.

Nous avions publié trois mois plutôt notre méthode s'appuyant sur un système d'apprentissage automatique (62). Elle repose sur l'utilisation d'un arbre de classification sémantique (SCT : Semantic Classification Tree (67)) pour attribuer les étiquettes *"tour courant"*, *"tour précédent"*, *"tour suivant"* ou *"autre"* aux noms complets détectés dans la transcription. D'une part, l'arbre de classification sémantique

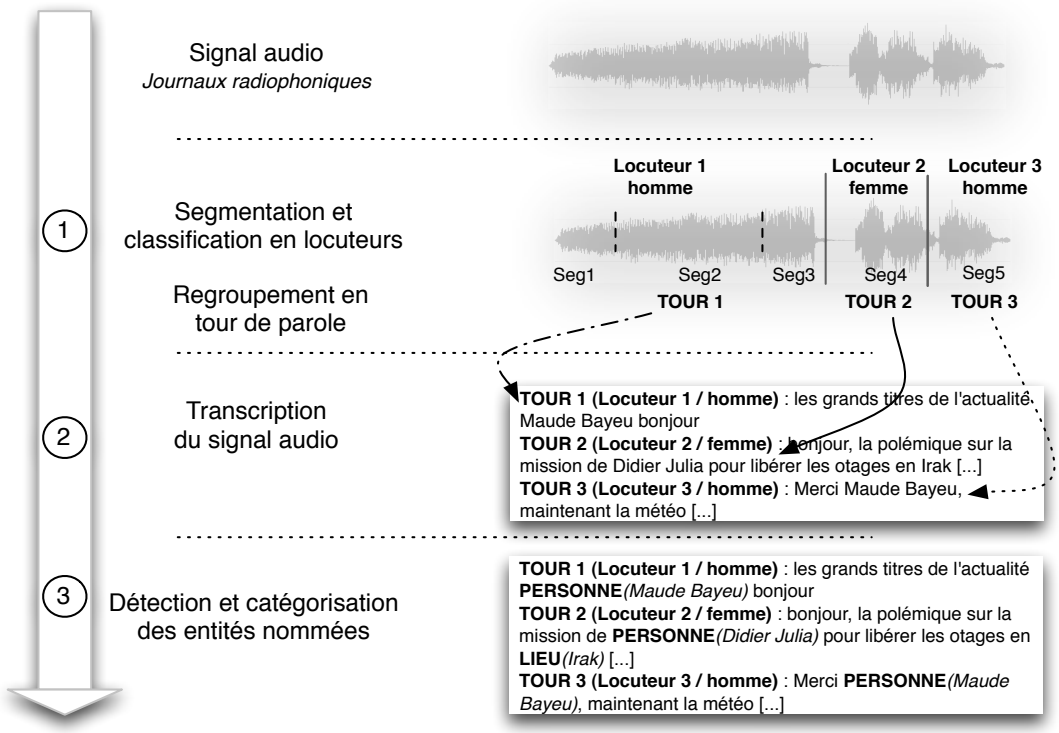

Fig. 4.3 – Description du système de transcription enrichie

utilise des expressions régulières pour l'attribution des étiquettes. Ces expressions modélisent le contexte lexical des noms complets détectés. D'autre part, l'arbre de classification sémantique utilise des questions globales conjointement aux expressions régulières portant sur la phrase. Notamment, la place du nom complet dans le tour de parole (début, fin, milieu) est un critère améliorant la qualité de la décision.

J'ai mené avec Paul Deléglise et Sylvain Meignier une étude comparative entre notre méthode basée sur les SCT et la méthode s'appuyant sur des n-grams qui a été publiée dans (28). Les résultats sont similaires pour les deux méthodes sur des transcriptions manuelles, en revanche le système basé sur les SCT est largement plus robuste aux erreurs de transcriptions et obtient de meilleurs résultats sur des transcriptions automatiques.

4.4 Système de transcription enrichie

La méthode d'identification nommée proposée s'appuie sur des documents préalablement transcrits et enrichis. Cette transcription nécessite de découper le document en segments qui seront ensuite classifiés en locuteurs. Ces segments, groupés en tours de parole, sont transcrits et les entités nommées sont annotées. La figure 4.3 illustre ces trois étapes.

Le système de TAP et le système de segmentation et regroupement en locuteurs

que nous utilisons dans nos expériences sont les systèmes du LIUM, que nous avons présentés dans le chapitre 2.

Dans l'optique d'un travail sur des transcriptions automatiques, nous avons choisi d'utiliser un étiquetage automatique en entités nommées au lieu de l'étiquetage manuel. L'outil utilisé est Nemesis (72), développé par l'équipe TALN (Traitement Automatique des Langues Naturelles) du LINA (Laboratoire d'Informatique de Nantes Atlantique).

Nemesis (72) est un système d'identification et de catégorisation d'entités nommées pour le français. Ses spécifications ont été élaborées à la suite d'une étude en corpus et s'appuient sur des critères graphiques et référentiels. Ces derniers ont permis de construire une typologie des entités la plus fine et la plus exhaustive possible, fondée sur celle de Grass (73). L'architecture logicielle de Nemesis se compose principalement de 4 modules (prétraitement lexical, première reconnaissance, apprentissage, seconde reconnaissance) qui effectuent un traitement immédiat des données à partir de textes bruts. L'identification des entités nommées est réalisée en analysant leur structure interne et leurs contextes gauche et droit immédiats à l'aide de lexiques de mots déclencheurs, ainsi que de règles de réécriture. Leur catégorisation s'appuie quant à elle sur la typologie construite précédemment. L'outil atteint environ 90% de précision et 80% de rappel sur des textes écrits en langage naturel.

Il est évident que les entités nommées de type "personne", qui permettent la détection des noms complets, sont primordiales pour le fonctionnement du système d'identification nommée proposé. Toutefois, d'autres entités nommées sont conservées, à savoir les lieux, les radios, et les organisations. Ces dernières permettent en effet de généraliser les informations utilisées par l'arbre de classification en remplaçant des informations spécifiques (les mots) par leurs catégories plus génériques. À noter que la liste des entités nommées retenues que nous venons d'énumérer est très proche de celle proposée dans le système de (61).

4.5 Architecture de notre approche

Notre méthode d'identification associe les étiquettes (*"tour courant"*, *"tour précédent"*, *"tour suivant"*) aux noms complets détectés dans la transcription *via* un arbre de classification sémantique. Les noms complets sont ainsi associés aux tours de parole correspondants avec les scores fournis par l'arbre. Ces informations sont agrégées dans les classes fournies par l'étape de segmentation et de classification en locuteurs. Enfin, pour chaque classe, un et un seul nom complet est sélectionné à partir des scores.

4.5.1 Décisions locales *via SCT*

Notre méthode utilise un arbre de décision binaire reposant sur le principe des arbres de classification sémantique (SCT — (67)), qui apprend automatiquement des règles lexicales à partir des noms complets détectés dans le corpus d'apprentissage. L'arbre permet d'attribuer l'étiquette *"tour courant"*, *"tour précédent"*, *"tour suivant"* ou *"autre"* à chaque nom complet détecté.

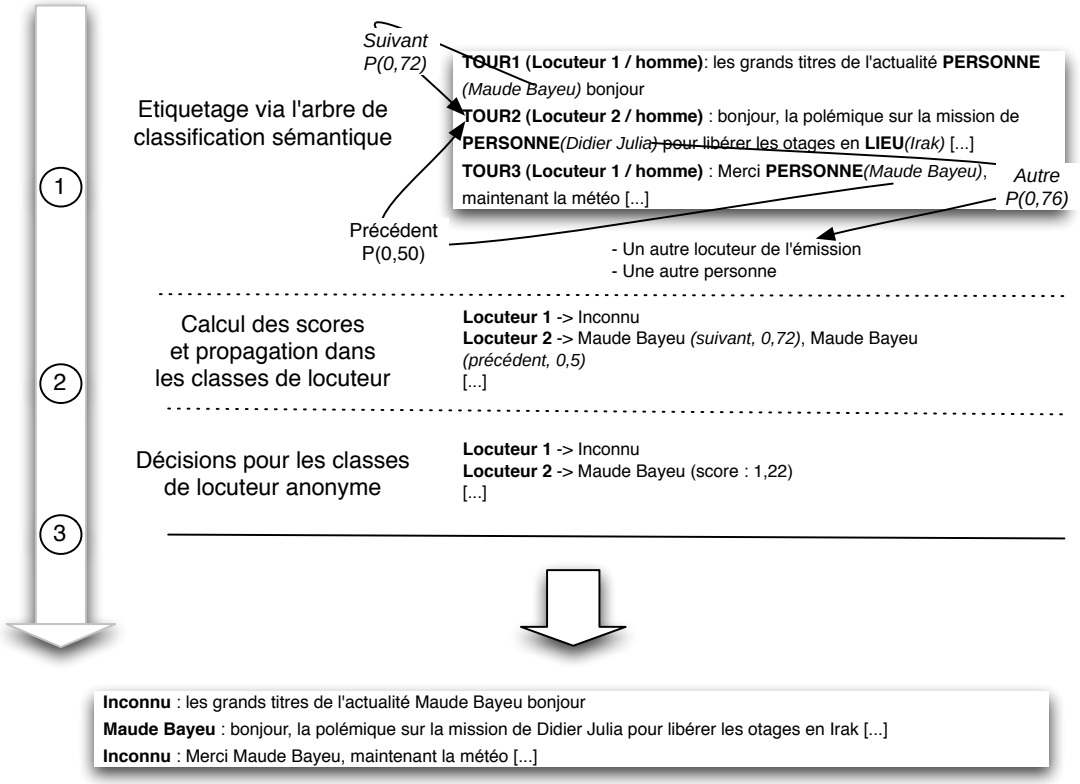

FIG. 4.4 – Description du système d'identification nommée

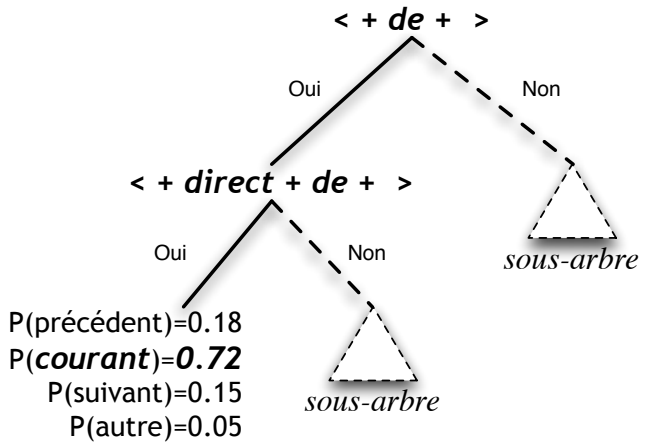

FIG. 4.5 – Arbre de classification sémantique

Les cinq mots à gauche et les cinq mots à droite d'un nom complet constituent le contexte lexical analysé par l'arbre de classification sémantique. Cette taille de contexte lexical a été optimisée expérimentalement à partir d'un corpus de développement. Les articles considérés comme non porteurs de sens pour la tâche d'identification nommée (*l', le, la, les, un, une, uns*) ont été éliminés du contexte analysé. En revanche, les prépositions comme *de, dans, à* ont été gardés puisqu'ils peuvent potentiellement concerner une information importante comme un lieu, une radio ou une émission. Les entités nommées de type "personne", "lieu", "radio" et "organisation" ont été remplacées par les catégories correspondantes. De plus, lors de la phase d'apprentissage, un poids beaucoup plus faible est attribué aux échantillons correspondant à l'étiquette *"autre"* ; en effet ce cas ne nous intéresse pas directement et sert uniquement à rejeter des noms complets. Il a été fixé expérimentalement à partir du corpus de développement.

Les arbres de classification sémantique intègrent pour chaque nœud une expression régulière. La suite d'expressions régulières activées depuis la racine jusqu'à une feuille de l'arbre permet de classer les contextes lexicaux en fonction des 4 étiquettes décrites précédemment. En complément des expressions régulières, l'arbre peut intégrer des questions globales comme la position du nom complet dans le tour de parole : la position correspond aux situations où le nom complet apparaît au début (dix premiers mots, soit la taille du contexte lexical), au milieu ou à la fin (dix derniers mots) d'un tour de parole. Une autre question globale permet de savoir si le tour de parole est considéré comme très court, à savoir plus petit que le contexte lexical utilisé par l'arbre (cinq mots à gauche du nom complet, cinq mots à droite).

La figure 4.5 représente un arbre de classification avec deux exemples d'expressions régulières ainsi que les scores associés aux étiquettes dans une des feuilles. Lorsqu'un exemple atteint une feuille de l'arbre, les scores correspondent aux probabilités de chacune de ces étiquettes. Ces scores sont établis lors de l'apprentissage de l'arbre et reflètent les distributions des étiquettes *"tour courant"*, *"tour précédent"*, *"tour suivant"* et *"autre"* pour les échantillons qui ont atteint cette feuille dans le corpus d'apprentissage.

Nous utilisons l'outil LIA_SCT développé par Frédéric Béchet au Laboratoire d'Informatique d'Avignon (74). Pour estimer l'arbre de classification, le corpus d'apprentissage est étiqueté automatiquement en entités nommées. L'étiquetage manuel en entités nommées disponible avec les transcriptions de référence n'est pas utilisé (nos expériences ont montré que nous obtenions de meilleures performances). De plus, les types d'entités nommées disponibles dans le corpus (étiqueté manuellement) et ceux utilisés par notre système sont difficilement compatibles. Pour cette phase d'apprentissage, la segmentation et la classification utilisées sont toujours celles fournies avec le corpus : elles ont été réalisées manuellement. Cela permet de s'affranchir des problèmes induits par une segmentation automatique qui, en cas d'erreurs, perturbe les tours de parole : ils n'appartiennent alors plus à un et un seul locuteur. Seul l'étiquetage en entités nommées est donc réalisé automatiquement pour l'apprentissage de l'arbre, le reste de la transcription est réalisé manuellement, elle ne contient donc pas d'erreurs.

Pour éviter les problèmes de surapprentissage, une expression régulière est retenue par l'arbre de classification sémantique lorsqu'elle a été rencontrée au moins 50 fois dans le corpus d'apprentissage. Le critère d'arrêt optimisé lors de l'apprentissage est le critère de Gini (67).

4.5.2 Système de décision globale

Dans l'étape précédente, à chaque nom complet détecté dans la transcription est associée une liste d'étiquettes (tour courant, ...) valuée par un score. Or, l'objectif du système est d'affecter à chaque classe de segments associés à un même locuteur un nom complet détecté dans la transcription. Ainsi, un système de décision globale détermine le nom complet attribué à chaque classe en exploitant les résultats des décisions locales, tout en prenant en compte des contraintes liées au regroupement des tours de parole en classes et au genre des locuteurs.

Nous prenons comme hypothèse que chaque classe fournie par le système de segmentation et de classification en locuteur contient des segments mono-locuteurs générés par un même locuteur. On suppose que les segments et les classes sont sans erreur. Cette hypothèse sera discutée dans la section 4.6.

Notations

$\mathcal{E} = \{e_1, \ldots, e_I\}$ correspond à l'ensemble des noms complets candidats pour nommer une classe. Ces candidats sont issus d'une liste des locuteurs possibles connue du système.

Soit $\mathcal{O} = \{o_1, \ldots, o_J\}$ les occurrences des noms complets détectés par Nemesis dans les transcriptions. $\mathcal{T} = \{t_1, \ldots, t_K\}$ désigne l'ensemble des tours de parole, et $\mathcal{C} = \{c_1, \ldots, c_L\}$ l'ensemble des classes à nommer. Les traitements vont permettre d'attribuer un nom complet issu de $\mathcal{E}$ aux classes de $\mathcal{C}$. Une classe est définie par $c_l = \{t_k \in T / \; c_l \textit{ est le locuteur de } t_k\}$, chaque classe c_l regroupant un ou plusieurs tours de parole t_k. Chaque tour de parole appartient à une et une seule classe.

Pour chaque occurrence d'un nom complet o_j (pour $j = 1, \ldots, J$) détecté dans un tour de parole t_k, on notera t_{k-1} (respectivement t_{k+1}) le tour de parole précédant (respectivement suivant) celui où il a été détecté. De la même manière, le score

$P(o_j, t_r)$ désigne la probabilité que o_j soit le locuteur du tour de parole t_r avec $r \in \{k-1, k, k+1\}$. Ces scores sont fournis par l'arbre de classification, ils correspondent respectivement à l'étiquette *"tour précédent"*, *"tour courant"*, *"tour suivant"* et *"autre"*.

Calcul des scores

Afin d'éviter l'utilisation d'informations inutiles ou bruitées, les scores $P(o_j, t_r)$ pour $r = k-1, k, k+1$ sont filtrés. Les trois seuils α_r , à partir desquels les scores $P(., t_r)$ sont pris en compte, sont fixés à partir d'un corpus de développement. Leurs valeurs sont fixées à $0,09$ pour $r = k-1$, $0,2$ pour $r = k$ et $0,2$ pour $r = k+1$. Si $P(o_j, t_r)$ est en dessous du seuil α_r, alors $P(o_j, t_r)$ est mis à 0. Les expériences sur le corpus de développement ont montré que cette précaution évitait l'accumulation de petites erreurs de l'arbre de classification (scores très faibles). De plus, les scores de l'étiquette *"autre"* ne permettant pas d'attribuer un nom complet à une classe, ils sont aussi mis à 0.

Pour l'assignation d'un nom complet e_i à une classe c_l donnée, nous calculons un score pour chaque nom complet e_i, dénoté $s_l(e_i)$, qui n'est autre que la somme des scores concernant les tours de parole de la classe c_l :

$$s_l(e_i) = \sum_{\{(o_j,t_r)|o_j=e_i,\, t_r \in c_l\}} P(o_j, t_r) \tag{4.1}$$

Processus de décision

Comme il a déjà été dit, nous faisons l'hypothèse que la segmentation et le regroupement en classes sont corrects. Le but est maintenant d'attribuer à chaque classe c_l un nom complet e_i.

Notre solution propose de réorganiser le partage des noms complets entre les classes anonymes. Elle s'effectue en deux étapes qui sont répétées jusqu'à ce que toutes les classes aient été nommées, ou jusqu'à ce qu'il n'y ait plus de candidat :

- tri des classes candidates pour un nom complet en fonction des scores,
- assignation du nom complet à la classe la plus probable.

Plusieurs stratégies peuvent être utilisées pour trier les classes c_l en concurrence pour un nom complet e_i donné. Prendre le score maximum $s_l(e_i)$ semble être la solution la plus naturelle. Mais si l'on veut pouvoir comparer ces scores, il faut les normaliser au préalable. Cela peut poser des problèmes, notamment dans le cas fréquent où un mauvais nom complet avec un score faible mais sans concurrent pourrait être affecté à une classe.

Afin d'éviter ce type d'erreurs, nous proposons d'utiliser un compromis, à savoir le produit des scores normalisés et non normalisés. Soit $\mathcal{D} = \{c_l \in \mathcal{C} |\ \forall e_i \in \mathcal{E}, s_l(e_i) = 0\}$ l'ensemble des classes non nommées, alors :

$$\mathrm{SC}_l(e_i) = \frac{s_l^2(e_i)}{\sum_{q=1}^{I} s_l(e_q)} \quad \text{si } c_l \notin \mathcal{D} \tag{4.2}$$

et

$$\mathrm{SC}_l(e_i) = 0 \ si \ c_l \in \mathcal{D}. \tag{4.3}$$

Lorsqu'un nom complet est le seul candidat pour une classe, son score reste inchangé. Lorsqu'il y a plusieurs candidats pour une classe, la normalisation permet de prendre en compte la contribution du score par rapport à l'ensemble des scores de la classe. Les scores en forte concurrence sont alors pénalisés. Cette normalisation est particulièrement efficace lorsque plusieurs noms complets potentiels ont des scores élevés.

Tous les noms complets possibles sont pris en compte *a priori* et triés en fonction de leur score $\mathrm{SC}_l(e_i)$. Premièrement, le nom complet avec le score maximum (noté e_i^*) est choisi, et si plusieurs classes sont associées au même e_i^*, alors ce nom complet sera assigné à la classe dont le score $SC_l(e_i^*)$ est maximum. Ensuite, tous les noms complets choisis sont supprimés des classes qui n'ont pas encore été nommées.

Un exemple concret est donné dans le tableau 4.1. Le nom complet "Jacques Derrida" a été assigné à trois classes différentes. Dans cet exemple, c_{13} a le meilleur score et "Jacques Derrida" devrait donc être affecté à c_{13} ; mais le score ne représente que 39% des scores totaux parmi tous les candidats possibles pour c_{13}, alors que le score pour c_{15} représente 79%. Finalement "Jacques Derrida" est assigné à c_{15} et d'autres noms complets seront attribués aux classes c_{13} et c_{14}.

Classe	nom complet e_i^*	$s_l(e_i^*)$	$\mathrm{SC}_l(e_i^*)$
c_{13}	Jacques Derrida	**8.58**	3.36
c_{14}	Jacques Derrida	1.67	1.09
c_{15}	Jacques Derrida	4.94	**3.88**

TAB. 4.1 – Exemple d'une assignation initiale multiple

Lors de l'itération suivante, les noms complets restants sont examinés de la même manière pour les classes restantes et ainsi de suite, jusqu'à ce que toutes les classes soient nommées ou que la liste des noms complets à attribuer soit vide. Le tableau 4.2 montre les résultats obtenus pour l'exemple précédent.

Classe	nom complet e_i^* (1ère itération)	2ème itération
c_{13}	Jacques Derrida (3.36)	**Nicolas Demorand** (0.99)
c_{14}	Jacques Derrida (1.09)	**Alexandre Adler** (0.30)
c_{15}	**Jacques Derrida** (3.88)	-
c_{16}	**Olivier Duhamel** (0.93)	-

TAB. 4.2 – Exemple du processus de décision avec deux itérations (décision en gras, scores entre parenthèses).

Prise en compte du genre

Le processus de décision précédent ne prend pas en compte une information disponible et qui peut être déterminante pour la validation du locuteur associé à une classe : le genre des locuteurs. En effet, pour chaque classe, cette caractéristique est disponible car d'une part, elle est déterminée de manière automatique lors des phases de segmentation et de classification (avec un taux d'erreurs inférieur à 5% sur

des données ESTER 1 phase II). D'autre part, le genre des noms complets extraits de la transcription peuvent être déterminés à travers celui de leur prénom associé. La comparaison de ces deux informations, obtenues de deux manières différentes, nous permet d'affiner le processus de décision. En cas d'incohérence, le couple prénom et patronyme n'est pas retenu et il est supprimé de la liste des candidats potentiels. Cependant, pour le cas des prénoms ambigus comme Dominique, l'entité nommée sera conservée.

Pour avoir la connaissance relative aux genres des prénoms, nous utilisons une base de données extraite du Web composée d'environ 20 000 prénoms. À chaque prénom est associé le nombre de fois où il a été attribué au genre féminin et au genre masculin depuis 1900 en France. Cette base de données ne semble pas exempte d'erreurs : par exemple le prénom Vincent apparaît 227180 fois comme prénom masculin et 373 fois comme prénom féminin. Le genre retenu correspondra au genre majoritaire. Si ce dernier a une fréquence inférieure à 75%, alors le genre est considéré comme indéterminé.

Au niveau du processus de décision, la comparaison des genres est directement incluse dans le calcul des scores (4.5.2, formule (4.1)) se traduisant par le filtre suivant : si le genre de l'occurrence o_j (et donc du nom complet e_i correspondant) et celui du tour de parole (et donc de la classe c_l à laquelle il appartient) sont différents, les scores $P(o_j, t_r)$ de la formule (4.1) ne sont pas pris en compte lors du calcul. Soit $g(e_i)$ et $g(t_r)$ les genres (féminin, masculin ou indéterminé) d'un nom complet e_i (ou d'un tour de parole t_r), alors : $(o_j = e_i,\ t_r(o_j) \in c_l$ et $g(e_i) \neq g(c_l)) \Rightarrow P(o_j, t_r) = 0$.

4.5.3 Evolution du système de décision globale

Initialement, j'avais proposé un système plus simple de décision globale qui consistait à, localement, ne prendre en compte que l'étiquette fournie par l'arbre de classification sémantique dont le score était le plus élevé. Globalement, pour chaque nom complet candidat pour une classe de segments, le score global était la somme des scores locaux qui associait ce nom à un tour de parole de la classe de segments : le nom retenu était celui dont la somme des scores était la plus élevée.

Cette méthode, très simple, donnait de relativement bons résultats et permettait déjà à notre approche d'être plus performante que l'approche à base de n-gram, tel que nous l'avons publié dans notre étude comparative dans (28).

Cependant, ce système de décision globale était perfectible et la nouvelle approche, présentée ci-dessus, permet d'obtenir de meilleurs résultats que notre approche précédente. Le mérite en revient à Simon Petit-Renaud, maître de conférences au LIUM.

Les derniers résultats expérimentaux obtenus, que nous allons présenter dans la section suivante, ainsi que quelques améliorations subtiles, sont le fruit du travail de Vincent Jousse, doctorant au LIUM, sous la direction de

`NON !! BEATRICE ET CHRISTINE JACQUIN`

Paul Deléglise et Sylvain Meignier. En particulier, ils montrent l'apport de leurs contributions à partir du système initial que Sylvain Meignier et moi-même avions conçu.

4.6 Évaluation du système proposé

4.6.1 Description des corpus

L'évaluation du système proposé est réalisée à partir d'émissions radiophoniques en français de la campagne ESTER 1 phase II (64; 75). La majorité de ces émissions contient essentiellement de la parole lue ou préparée, et peu de parole spontanée : 15% du corpus correspond à des interventions de personnes parlant au téléphone.

Les émissions proviennent de 5 radios françaises et de Radio Télévision Marocaine et durent de 10 à 60 min. Elles sont réparties en 3 corpus utilisés pour l'apprentissage de l'arbre de classification, le développement et l'évaluation du système. Le corpus de développement a été utilisé pour fixer les différents paramètres du système comme la taille du contexte lexical de l'arbre ou le poids donné aux échantillons lors de l'apprentissage (cf. 4.5.1).

Le corpus d'apprentissage contient 76h de données (75095 segments et 7416 tours de parole) dans lesquels 11292 noms complets sont détectés. 755 locuteurs différents interviennent dans ce corpus dont 40 qui n'ont pu être nommés. Le corpus de développement contient 30h (27149 segments et 2931 tours de parole) dans lesquels 4533 noms complets ont été détectés. 359 locuteurs différents interviennent dans ce corpus dont 38 qui n'ont pu être nommés. Le corpus d'évaluation contient 10h (10335 segments et 1082 tours de parole) dans lesquels 1541 noms complets ont été détectés. 213 locuteurs différents interviennent dans ce corpus dont 24 qui n'ont pu être nommés. 26,5% de ces locuteurs sont communs au corpus d'apprentissage seul et 28,4% sont communs aux corpus d'apprentissage et de développement. Ce découpage correspond au découpage de la campagne d'évaluation officielle ESTER 1 PHASE II 2005.

Les transcriptions fournies avec les corpus ont été créées pour l'évaluation des tâches de segmentation et de classification en locuteur, ainsi que pour la tâche de transcription. Les références proposées sont d'une grande qualité, les annotateurs ont essayé de nommer le maximum de locuteurs par des identifiants permettant d'en déduire leurs noms complets. Ces noms complets ont été extraits automatiquement et n'ont pas fait l'objet de validation manuelle approfondie.

Le tableau 4.3 montre la répartition *a priori* des 4 étiquettes calculée pour le corpus de test. L'étiquette *"autre"* est la plus fréquente et représente 79,5% des cas ; vient ensuite l'étiquette *"tour suivant"* avec 16,5%, tandis que les deux dernières étiquettes *"tour précedent"* et *"tour courant"* sont les moins fréquentes : environ 2% chacune.

4.6.2 Métriques utilisées

Le système d'identification nommée est évalué en comparant l'hypothèse générée par celui-ci à la référence distribuée avec le corpus. Cette comparaison met en évidence 5 cas (d'erreurs ou de succès) possibles relatifs aux situations suivantes :

– l'identité proposée est correcte (C_1) : le système propose une identité correspondant à celle indiquée dans la référence ;

	Évaluation
Tour précédent	2,0% (31)
Tour courant	2,0% (30)
Tour suivant	16,5% (255)
Autre	79,5% (1225)
Total	100% (1541)

TAB. 4.3 – Répartition des étiquettes sur le corpus d'évaluation, statistiques sur les noms complets (fréquence et effectif).

- erreur de substitution (S) : le système propose une identité différente de l'identité présente dans la référence ;
- erreur de suppression (D) : le système ne propose pas d'identité alors que le locuteur est identifié dans la référence ;
- erreur d'insertion (I) : le système propose une identité alors que le locuteur n'est pas identifié dans la référence ;
- il n'y a pas d'identité (C_2) : le système ne propose pas d'identité et la référence ne contient pas d'identité.

Une mesure de Précision et de Rappel peut être définie à partir des 5 cas d'erreurs :

$$P = \frac{C_1}{C_1 + S + I} \quad ; \quad R = \frac{C_1}{C_1 + S + D} \tag{4.4}$$

La précision et le rappel peuvent être synthétisés en calculant la F-mesure : $F = (2 \times P \times R)/(P + R)$.

Comme il a été proposé dans (61), nous complétons ces valeurs par un taux d'erreurs Err global également calculé à partir de ces 5 erreurs. Ce taux s'inspire du calcul du WER utilisé pour l'évaluation de la transcription. Il a l'avantage de mesurer la qualité des résultats du système d'identification nommée en une seule valeur, facilitant les comparaisons entre les systèmes par rapport aux mesures de précision et de rappel.

$$Err = \frac{S + I + D}{S + I + D + C_2 + C_1} \quad ; \tag{4.5}$$

Les erreurs peuvent être calculées en terme de durée ou en terme de nombre de locuteurs (classes correctement nommées). Pour une évaluation en durée, dans le cas où un locuteur parlant 90% du temps est correctement nommé et que les six autres locuteurs parlant seulement 10% du temps ne le sont pas, le système présentera un taux d'erreurs de 10%.

Pour une évaluation en terme de nombre de locuteurs, dans le même cas de figure, le système aura un taux d'erreurs de 87,5%. À noter que ce taux d'erreurs ne peut être calculé qu'avec une segmentation et une classification manuelles (donc identique à la référence).

D'un point de vue applicatif, la métrique exprimée en durée est préférable si les locuteurs considérés comme importants correspondent aux locuteurs s'exprimant

beaucoup. En revanche, si l'application cherche à nommer le plus possible de locuteurs, il est plus intéressant d'évaluer les performances en terme de nombre de locuteurs.

4.6.3 Protocole d'évaluation

Le système initial du LIUM décrit dans (68) est utilisé comme système de référence afin de mesurer les derniers apports évoqués plus haut. Ce système ne bénéficie pas du processus de décision décrit précédemment et de la prise en compte des genres. Il permet ainsi d'évaluer l'apport de ces deux modifications. Dans le système de référence, seule l'étiquette ayant la probabilité maximale est prise en compte pour chaque nom complet. Chaque nom complet détecté n'est donc propagé qu'à un et un seul tour de parole, avant d'être ensuite propagé au sein de la classe. Si plusieurs probabilités pour un même nom complet sont présentes au sein de la classe, elles sont additionnées pour donner le score du nom complet au sein de cette classe. La décision globale consiste ensuite à attribuer, pour une classe, le nom complet dont le score est maximal, sans prise en compte des informations des autres classes.

4.6.4 Évaluation du système avec transcriptions manuelles

Évaluation du système

Dans les expériences, il est supposé que le système connaît tous les noms complets susceptibles d'être des locuteurs. Le système de décision utilise cette connaissance pour rejeter les noms complets ne correspondant pas à des locuteurs recherchés. Dans la section 4.6.4, nous présenterons des résultats avec et sans cette connaissance *a priori*. Cette liste est constituée de 1008 noms complets de locuteur apparaissant dans les corpus d'apprentissage, de développement et d'évaluation. Cette connaissance est uniquement introduite dans le système de décision, bien que cela soit envisageable de l'introduire aussi au niveau du système automatique de transcription enrichie.

La comparaison entre le système de référence et le système présenté ici est effectuée sur des transcriptions et segmentations manuelles. C'est à dire qu'il n'y a pas d'erreurs de segmentation et de classification en locuteurs : toutes les frontières sont justes et tous les tours de parole appartiennent à la bonne classe. De même, la transcription en mots est sans erreur et tous les noms complets de locuteurs sont correctement transcrits. En revanche, la détection des entités nommées est faite en utilisant Nemesis ; elle comporte donc des erreurs.

Comme le montre le tableau 4.4, le système actuel a une précision plus faible d'environ 3 points en absolu, mais un rappel meilleur de plus de 12 points. Il est difficile de dire quel est le meilleur système à partir de ces deux valeurs. Le calcul d'un taux d'erreurs en durée ($ErrDur$) permet de clarifier la situation : le système proposé obtient 10 points d'erreurs de moins en absolu que le système de référence. En ce qui concerne le nombre de locuteurs identifiés, le nouveau système étiquette correctement deux fois plus de locuteurs que le système de référence. En effet, si l'on

	En durée				En nb de Locuteur
Système	**Rappel**	**Précision**	**F-mesure**	**ErrDur**	**ErrLoc**
Référence	70,7%	92,6%	0,80	26,6%	37,4%
Proposé	83,2%	89,7%	0,86	16,6%	19,5%

TAB. 4.4 – Comparaison système proposé et système de référence sur le corpus d'évaluation ESTER 1 phase II

Les résultats sont donnés en utilisant la transcription enrichie de référence.
***Rappel**, **Précision** et **F-mesure** calculés en en durée.*
***ErrDur** : Taux d'erreurs en durée.*
***ErrLoc** : Taux d'erreurs en nombre de locuteurs.*

prend en compte le nombre de locuteurs, le taux d'erreurs (*ErrLoc*) est d'environ 20% pour le nouveau système contre 40% pour celui de référence.

En conclusion le nouveau système obtient de meilleurs taux d'erreurs qu'ils soient mesurés en durée ou en nombre de locuteurs que le système initial.

Influence de la connaissance *a priori* des noms de locuteurs

Jusqu'ici, les résultats donnés utilisent une liste de locuteurs potentiels lors du processus de décision et d'évaluation. Le tableau 4.5 présente les résultats avec et sans connaissance *a priori* sur les noms complets cibles de l'application d'identification nommée. À noter que les systèmes proposés dans (61; 63) utilisent aussi cette connaissance. La liste contient l'ensemble des participants aux émissions des corpus d'ESTER. Ces personnes sont principalement des personnes publiques comme des journalistes, des politiciens, des artistes ou des sportifs. Cette population est 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.

Dans 4.6.1, nous avons pu noter que seulement 26,5% des locuteurs du corpus d'évaluation sont communs aux locuteurs du corpus d'apprentissage. Ceci s'explique par un éloignement temporel important des enregistrements. Les données d'apprentissage les plus récentes ont été enregistrées en juillet 2003 pour RTM, les autres radios, représentant la majorité des données, ont été enregistrées entre 1998 et 2000. Les données d'évaluation ont été enregistrées entre octobre et décembre 2004, à plus d'un an des données RTM et à plus de 4 ans des autres radios. En utilisant uniquement les noms complets des locuteurs présents dans le corpus d'apprentissage, l'évaluation porterait sur un nombre très faible de candidats (56 locuteurs sur les 213 présents dans le corpus d'évaluation). Nous avons choisi de nous placer dans le cadre où tous les locuteurs cibles sont connus.

En termes de durée, le taux d'erreurs *ErrDur* augmente d'environ 12 points si le système n'utilise pas la liste de noms complets cibles pour filtrer les décisions (cf. tableau 4.5). Les taux d'd'erreurs sont respectivement de 16,7% et 28,9% pour le système utilisant la liste de locuteurs et pour le système ne l'utilisant pas. En

	En durée				En nb de Locuteur
Noms complets	**Rappel**	**Précision**	**F-mesure**	**ErrDur**	**ErrLoc**
connus	83,2%	89,7%	0,86	16,7%	19,5%
inconnus	73,61%	74,27%	0,74	28,9%	27,4%

TAB. 4.5 – Résultats avec et sans connaissance *a priori* sur les noms complets, évaluation faite sur le corpus d'évaluation ESTER 1 phase II

Les résultats sont donnés en utilisant la transcription enrichie de référence.
***Noms complets connus** : le système de décision connaît les noms complets des locuteurs potentiels.*
***Noms complets inconnus** : le système de décision ne connaît pas les noms complets des locuteurs potentiels.*
***Rappel**, **Précision** et **F-mesure** calculés en durée.*
***ErrDur** : Taux d'erreurs en durée.*
***ErrLoc** : Taux d'erreurs en nombre de locuteurs.*

termes de nombre de locuteurs, la différence est d'environ 8 points en absolu en défaveur du système n'utilisant pas la liste. La liste contient 1008 noms complets, dont 213 présents dans le corpus d'évaluation. Sur les 1514 occurrences de noms complets détectés, 655 noms complets sont éliminés car ils ne font pas partie de la liste. Lorsque l'on n'utilise pas la liste des locuteurs, des noms de locuteurs qui auraient du être considérés comme "autre" se retrouvent propagés au sein des classes, augmentant ainsi les sources d'erreurs. Ces noms éliminés sont majoritairement des personnes citées dans le discours mais ne parlant pas dans l'enregistrement. Quelques erreurs dues à un mauvais étiquetage de Nemesis sont aussi évitées.

4.6.5 Vers un système entièrement automatique

Évaluation du système avec transcriptions automatiques

Les résultats présentés dans le tableau 4.6 correspondent aux expériences utilisant des segmentations et classifications en locuteurs automatiques ou manuelles ainsi que des transcriptions automatiques ou manuelles. Le système de référence et le système proposé ont dans tous les cas été évalués avec une détection des entités nommées automatique.

On constate que plus le système tend vers un système entièrement automatique, plus les performances se dégradent. Pour le système proposé le taux d'erreurs en durée *ErrDur* passe de 16,7% à 75,2%, étant ainsi multiplié par plus de 4,5.

Le système proposé est tributaire des erreurs de la transcription enrichie. Cette dégradation des performances provient autant des erreurs de segmentation et classification en locuteur que des erreurs de transcription. Concernant la classification et la segmentation automatique en locuteur, nous avons constaté que la segmentation automatique engendrait plus d'd'erreurs que la classification automatique. Actuellement, ce module a été développé pour minimiser le taux de DER qui est de 11,5%.

		En durée				En nb de Locuteur
Trans.	**Seg/Class.**	**R**	**P**	**F**	**ErrDur**	**ErrLoc**
Système présenté						
M	**M**	83,2%	89,7%	0,86	16,7%	19,5%
M	**A**	38,0%	58,2%	0,46	58,3%	-
A	**M**	31,0%	58,3%	0,40	62.8%	70,0%
A	**A**	18,4%	42,1%	0,26	75,2%	-
Système de référence						
M	**M**	75,7%	95,3%	0,84	22,5%	33,1%
M	**A**	27,8%	71,6%	0,40	66,0%	-
A	**M**	28,1%	76,9%	0,41	63,9%	74,0%
A	**A**	15,0%	75,7%	0,25	77,34%	-

TAB. 4.6 – Système proposé avec une transcription enrichie manuelle ou automatique sur le corpus d'évaluation ESTER 1 phase II

Trans. *: Transcription* ***M****anuelle ou* ***A****utomatique.*
Seg/Class. *: segmentation/classification manuelles ou automatiques.*
R, P, F *: rappel, précision et F-mesure calculés en durée.*
ErrDur *: Taux d'erreurs en durée.*
ErrLoc *: Taux d'erreurs en nombre de locuteurs.*

L'impact des erreurs de transcription est étudié dans la section suivante 4.6.5.

Influence de la qualité de la transcription

Les résultats précédents montrent une dégradation des performances lorsque la transcription est automatique. Cette dernière obtient un taux d'erreurs sur les mots de 20,5%. Le tableau 4.7 compare les résultats d'identification nommée obtenus en utilisant les transcriptions réalisées par le système du LIUM et celles réalisées par le système du LIMSI (76). La transcription du LIMSI correspond à la transcription générée par leur système durant la campagne d'évaluation ESTER 1 phase II en 2005 ; où ce système a obtenu les meilleurs résultats avec un taux d'erreurs sur les mots de 11,9%. Les transcriptions ont été générées à partir de segmentations et de classifications automatiques. Pour supprimer les erreurs de segmentation et de classification, les mots ont été replacés dans les segments de référence avant d'appliquer le système de détection d'entités nommées.

La transcription du LIMSI permet de réduire les taux d'erreurs $ErrDur$ et $ErrLoc$ d'environ 10 points. Elle contient notamment plus de noms complets correctement transcrits que celle du LIUM, ce qui a un impact non négligeable sur les résultats du processus d'identification nommée. En effet, alors que 1541 noms complets sont détectés dans les transcriptions de référence, seulement 970 sont détectés dans les transcriptions du LIUM contre 1192 dans les transcriptions du LIMSI.

	En durée				En nb de Locuteur
Transcription	**Rappel**	**Précision**	**F-mesure**	**ErrDur**	**ErrLoc**
LIUM	31,0%	58,3%	0.40	62,8%	70,0%
LIMSI	41,0%	65,1%	0.50	53,8%	59,5%

TAB. 4.7 – Comparaison des résultats avec deux systèmes de transcription différents sur le corpus d'évaluation ESTER 1 phase II

Rappel, ***Précision*** *et* ***F-mesure*** *calculés en en durée.*
ErrDur *: Taux d'erreurs en durée.*
ErrLoc *: Taux d'erreurs en nombre de locuteurs.*

4.7 Conclusion

La méthode d'identification des locuteurs que nous avons proposé consiste à extraire de transcriptions, manuelles ou automatiques, les identités des locuteurs. L'identification est réalisée à l'aide d'un arbre de classification sémantique qui attribue les prénoms et noms détectés dans la transcription aux locuteurs s'exprimant dans l'enregistrement. Nous avons montré que l'utilisation d'un arbre de classification était une solution robuste, aussi bien pour une utilisation sur des transcriptions manuelles que sur des transcriptions automatiques. Le choix des identités des locuteurs est reporté en fin de processus où tous les noms complets candidats sont mis en concurrence. Le problème de cette prise de décision globale a également été l'objet de travaux de la part du LIUM qui ont permis de faire évoluer notre approche.

Les expériences ont été réalisées sur des émissions radiophoniques en français issues de la campagne d'évaluation ESTER 1 phase II. Le système obtient de très bonnes performances pour le traitement de transcriptions manuelles, en revanche les performances se dégradent fortement lorsque l'identification est réalisée à partir de transcriptions obtenues de manière automatique. Nos travaux actuels se focalisent sur le traitement des transcriptions automatiques. Par la suite, nous aimerions travailler sur des collections de documents plutôt que de travailler sur les fichiers indépendamment les uns des autres. Enfin, nous pensons qu'il est possible d'exploiter les résultats de notre système d'identification nommée, malgré ses performances faibles, en mode automatique pour des applications particulières. Par exemple, pour l'extraction automatique, à partir de masses de données audio, d'enregistrements d'un locuteur particulier.

Sylvain Meignier et moi-même avons été les initiateurs de ce travail : nous avons proposé une méthode complète d'identification nommée basée sur des décisions locales et sur une prise de décision globale exploitant les résultats locaux. Nous avons proposé l'utilisation des arbres de classification sémantique pour résoudre le problème des décisions locales. Par la suite, nous avons réussi à fédérer autour de ce travail un certain nombre de chercheurs du LIUM et une thèse, celle de Vincent Jousse, lui est actuellement totalement dédiée. Ceci a permis de faire évoluer notre approche, de la comparer à d'autres, et d'intégrer ce travail dans des projets avec

d'autres partenaires, comme le projet régional MILES et le projet ANR EPAC.

Chapitre 5

Traitement de la parole conversationnelle

Sommaire

5.1 Transcription manuelle de la parole conversationnelle

5.1.1 Introduction

D'un point de vue énonciatif, la parole spontanée peut se définir comme un « énoncé conçu et perçu dans le fil de son énonciation » (Luzzati, 2004), c'est-à- dire un énoncé produit pour un interlocuteur réel par un énonciateur qui improvise ; cela implique que les corrections ne peuvent se traduire que par un prolongement

du message. La parole préparée (celle qu'emploient les journalistes présentant les informations radiophoniques ou télévisées) est une parole produite pour un interlocuteur plus ou moins fictif, par un énonciateur qui en possède la maîtrise, qui est capable de produire des énoncés qui n'ont plus à être repris ou corrigés, ou qui est capable de le masquer. De ce point de vue, on comprend qu'on puisse parler également de parole conversationnelle, non préméditée ou co- construite. D'un point de vue morphosyntaxique, la parole spontanée se caractérise par deux phénomènes saillants : on y trouve un grand nombre de disfluences (Adda-Decker et al., 2004) et le fenêtrage syntaxique y est particulier. Les fenêtres de cohérence syntaxique (Luzzati, 2004) y sont courtes (empan moyen inférieur à huit « mots »), elles ne sont pas nécessairement conjointes, et elles sont superposables. À l'inverse, la parole préparée tend vers l'écrit, avec des fenêtres de cohérence syntaxique parfois très longues (l'hypotaxe y est importante), conjointes et sans interjection. D'un point de vue phonologique, la parole spontanée se caractérise par deux phénomènes importants : la disparition des schwas (ou e muet, caduc, central...) et les phénomènes d'assimilation qui en découlent. À titre d'exemple, un mot comme « cheval », suite à la disparition du schwa et à une assimilation, se prononce désormais [Sfal]1, de façon généralement inconsciente pour les locuteurs, mais patente pour un système de reconnaissance automatique de la parole. C ?est pourquoi nous nous proposons dans un premier temps de confronter linguistiquement la parole spontanée à la parole préparée, afin d ?en faire ressortir les principales spécificités. Puis, après avoir proposé un état des lieux des corpus disponibles, nous nous intéresserons au traitement de la parole spontanée par le biais de diverses expériences, qui ont pour but d ?en optimiser la détection et la transcription à l ?aide d ?un système de reconnaissance automatique de la parole (RAP). Les résultats ainsi obtenus nous permettront notamment d ?envisager une typologie des erreurs commises par les systèmes de RAP, et de proposer quepistes en vue d ?améliorer leurs performances.

5.1.2 Parole spontanée vs. parole préparée

Huit critères (notamment morphosyntaxiques) permettent de caractériser la parole dite « spontanée », c'est-à-dire une parole altérée, variable en débit et en fluidité.

Elisions du schwa et assimilations

Tout d'abord l'élision du schwa, et les assimilations qui souvent en résultent. La réalisation (ou non) du schwa est en elle-même un problème complexe, sur lequel nombre de linguistes se sont penchés, qui revêt une importance particulière pour la parole spontanée car elle induit souvent des assimilations portant sur des

morphèmes ou structures parmi les plus fréquentes (pronom + verbe, de + nom notamment). En premier lieu, « je » + consonne sourde, qui devient [S] : des formes telles que « j'pense » ou « j'crois » deviennent respectivement « ch'pense » et « ch'crois ». Les mêmes arguments s'appliquent également à « de », et dans des proportions presque identiques : des données extraites du corpus ESTER nous ont permis de constater que des expressions comme « pas d'problème » ou « pas d'chance » reviennent à de nombreuses reprises dans la langue parlée et, toujours par effet d'assimilation, sont prononcées « pas t'problème » et « pas t'chance ». À un degré

moindre, on retrouve également l'élision du schwa avec « te », « se » ou « que » + consonne sonore (« on s'donne », « tu t'demandes » ou « qu'vous », prononcés « on z'donne », « tu d'demandes » ou « g'vous »). Enfin, le cas des formes « le », « me » et « ne » est un peu particulier : les nasales « m » et « n » ne varient pas au contact d'une consonne sourde ou sonore lorsque le « e » est élidé (« je m'fâche », « je n'crois pas »). Quant au « l », le fait que cette lettre soit une consonne liquide fait que par nature, elle se combine facilement avec d'autres consonnes ; ainsi, que ce soit au contact d'une sourde (« l'problème ») ou d'une sonore (« je l'vois bien »), sa prononciation n'est pas modifiée.

Autres élisions

Autres monosyllabes, « tu », « il(s) », « elle » et « vous » sont eux aussi souvent élidés : « t'as », « i' vient », « i' savaient pas », « e' va », voire « 'pouvez pas vous traîner » ou « zêt sûr », avec aspiration de « vous » par le verbe. Tout fonctionne en somme comme si aux personnes 1, 2, 4 et 5, le sujet était marqué par une enclise consonantique droite, et comme si aux personnes 3 et 6 demeurait surtout une opposition masculin / féminin (i / è). Outre ces cas spécifiques, la principale élision rencontrée concerne la vibrante « r » dans les mots à finale en « -bre », « -cre », « -dre », « -tre » ou « vre ». Cela est particulièrement flagrant lorsque le mot suivant commence par une consonne : en effet, l'immense majorité des locuteurs, dans un contexte spontané, ne dira jamais « à quatre pattes » mais plutôt « à quat'e pattes », séquence beaucoup plus simple à articuler dans un discours à débit relativement rapide. Dans le corpus ESTER (Galliano, 2005), on trouve ainsi : « novemb'e », « convainc'e », « descend'e », « peut-êt'e », « surviv'e »... « é » et « è » disparaissent parfois dans « c'était », « c'est-à-dire », « déjà » ou « écoutez », qui deviennent « s'tait », « c't-à-dire », « d'jà » ou « 'coutez ». Parfois, cela fait apparaître un schwa qui, accentué, passe du [@] au [9] : il en va ainsi du démonstratif « cette », parfois prononcé « c'te » « l », dans deux cas précis, peut également être élidée : « plus » ou « je lui » sont parfois réduits à « p'us' » ou « j'ui' » (notons que dans ce cas, le schwa de « je » est lui aussi élidé ». Pour terminer, nous mentionnerons quelques cas d'élisions isolés : « puis », « parce que » et « enfin » deviennent très souvent « p'is », « pac'e que » et « 'fin », avec un sens sans doute différent d'un emploi sans élision. L'expression « tout à l'heure » se transforme quelquefois en « t't à l'heure ». Enfin, certains mots commençant par « at- » ont tendance à voir cette séquence initiale disparaître : « attention » deviendra « 'tention » et « attendez », « 'tendez ».

Troncations

La troncation2 est un autre phénomène spécifique de la parole spontanée : c'est un mot que le locuteur commence à prononcer puis, pour diverses raisons (principalement le bégaiement ou l'hésitation), ne finit pas. Dans certains cas, le mot tronqué est ensuite complètement prononcé. Cela donne des séquences telles que celles-ci (la troncation est symbolisée ci-dessous par l'emploi de parenthèses) : « des idées ré() révolutionnaires » « il était aussi passionné d'avia() d'aviation » « et la première ém() émission » « elle ne sera pas premier secrétai() euh secrétaire » « et ça rebaisse

régulière() régulièrement » Cependant, il arrive que le mot tronqué ne soit pas repris ensuite : soit le locuteur poursuit alors son énoncé comme s'il n'y avait pas eu troncation (1), soit il le reprend partiellement (2 et 3) ou en totalité, créant ainsi une anacoluthe (4 et 5). (1) « alors auj() le starsystem s'est emparé de la télé » (2) « c'est t() vraiment une lettre très émouvante » (3) « c'est-à-dire que pou() sur un repas que vous vendez sept à huit euros » (4) « et il y a un truc s() il y a quelque chose de suspect » (5) « oui et c'est un k() ah oui oui et c'est un canadien »

Faux départs

L'anacoluthe nous amène à parler du faux départ, phénomène assez proche des deux derniers exemples que nous venons de mentionner, mais qui s'en distingue en désignant une interruption à l'intérieur d'un énoncé, et non à l'intérieur d'un mot. La conséquence est toutefois la même : l'apparition d'une rupture de syntaxe puisque le locuteur commence un énoncé qu'il ne finit pas pour y adjoindre un second : « ça a été lu et c'est () on a la photo » « il y a dix mille () mais c'était mal » « j'ai () on a essayé de récupérer tous les éléments » Il arrive en outre que l'on rencontre des « semi faux-départs », où le second énoncé est en fait le complément d'une partie du premier. Le locuteur corrige son propos initial, mais sans produire une phrase complète, ayant toujours à l'esprit le premier fragment prononcé : « je voulais vous dire aussi () passer un gros coup de gueule »

Répétitions

La répétition d'un même mot ou d'une même séquence de mots est aussi un signe patent d'un discours spontané. À nouveau, bégaiement et hésitations en sont les deux principaux moteurs. La répétition est parfois étroitement liée à la troncation (5) ; de même, elle peut parfois jouer sur deux mots très proches (6) : (5) « la l() la lettre de Guy Môquet » (6) « là c'était le le la le la l'accusation la plus grave »

Fenêtrage syntaxique

Toujours sur le plan morphosyntaxique, la parole spontanée se caractérise également par un phénomène remarquable : les fenêtres de cohérence syntaxique y sont courtes, pas nécessairement conjointes, et superposables

Morphèmes spécifiques

D ?un point de vue lexical, la parole spontanée se caractérise par l'emploi de morphèmes typiquement oraux tels que « euh » ou « ben » (et ses dérivés) (Luzzati, 1982). Extrêmement nombreux dans les corpus que nous avons constitués, leur rôle est pourtant parfois opaque. Si « euh » indique majoritairement l'hésitation (et est à ce titre souvent employé de façon répétée), les emplois de « ben » sont quant à eux beaucoup plus difficiles à cerner : forme oralisée de « bien », conjonction de coordination, adverbe... « banalisons le pain comme n'importe quel euh objet » « et évidemment l'état euh euh à gauche ou à droite » « ben c'est gentil mais » « eh ben ton installation est est impeccable » « qui euh ben qui va s'avérer être un un apprenti euh formidable » Notons que ces morphèmes et les analyses s ?y rattachant

ne sont pas spécifiques à la langue française : l ?anglais, par exemple, possède avec la forme well une expression sémantiquement proche de nos euh et ben français dans certaines de ses acceptions. Deborah Schiffrin s ?y est intéressée dans une étude sur les « discourse markers » (Schiffrin, 2001), terminologie plus globale que celle que nous employons ici et qui, outre well, recouvre également des formes comme and ou y ?know. De même, on pourra lire avec intérêt les travaux de Gisèle Chevalier (Chevalier, 2000), qui propose une étude des emplois de well en Acadien du sud-est du Nouveau-Brunswick, une variante du Français.

Phénomènes prosodiques

Enfin, nous terminerons cette étude des spécificités de la parole spontanée en évoquant quelques objets ayant trait à la prosodie : tout d ?abord, le mélisme (Caelen-Haumont, 2002a), désignant dans notre champ d ?études un allongement syllabique en fin de mot. Très caractéristique de l ?oral, où il se veut bien souvent être la marque d ?une hésitation, ce phénomène s ?est révélé particulièrement efficace lors d ?une expérience interne pour détecter des zones de parole spontanée dans de gros corpus audio (Jousse, 2008). Ensuite, les pauses, et plus précisément leur durée et leur fréquence, sont un autre aspect remarquable de la parole spontanée. En effet, si l ?on observe un corpus de parole préparée, et notamment journalistique, on s ?aperçoit que les pauses dans le flux de parole y sont généralement peu nombreuses, relativement brèves, et bien souvent liées à la respiration et/ou à la déglutition, plus qu'à un phénomène d'hésitation par exemple. À l ?inverse, les pauses dans un cadre spontané (interviews par exemple) sont en général beaucoup plus longues et nombreuses : d ?une part parce que les locuteurs ne bénéficient pas d ?un canevas (prompteur, notes) pour tisser leurs propos, et qu ?ils les conçoivent donc au fur et à mesure, ce qui demande des périodes de réflexions ; d ?autre part parce qu ?à l ?inverse d ?un journaliste, dont le métier sous-entend une réelle aisance pour s ?exprimer, les intervenants lors d ?interviews ou de témoignages ne sont pas toujours familiarisés avec ces exercices ; il en résulte souvent de longs « blancs », témoins de leurs hésitations. Le débit phonémique obéit également à cette dualité : dans le cas d ?un journal d ?informations, il varie généralement peu. Lors d ?un entretien, et pour les raisons que nous venons d ?évoquer, il arrive que le locuteur peine à enchaîner ses propos, s ?attarde, puis soudain accélère son flux de parole, au gré de ses idées ou de son état émotionnel. Enfin, pour clore cette analyse prosodique, nous aborderons l ?intonation. Le projet EPAC, qui est présenté en 4.1., nous a permis d ?envisager ce phénomène de manière concrète : en effet, l ?un des objectifs d ?EPAC est de fournir la transcription annotée d ?environ 100 heures de parole, majoritairement spontanée. Or, toutes les transcriptions que nous réalisons dans cette optique sont ponctuées, et naturellement cette ponctuation se base entre autres sur l ?intonation. Il est ainsi apparu clairement que l ?annotateur éprouvait beaucoup plus de difficultés à ponctuer des propos spontanés que des propos préparés. Certes, la rigueur syntactico-sémantique des propos journalistiques y est pour beaucoup, mais les intonations marquées apportent également au transcripteur des indications non négligeables. Or, celles-ci sont beaucoup moins transparentes lorsqu ?il s ?agit de parole spontanée, tout simplement parce que comme nous le disions plus haut, la parole est alors élaborée au fur et à mesure qu ?elle est énoncée. Ainsi, le locuteur

ne sait parfois pas où le segment de parole qu ?il a commencé se terminera, et ne peut donc y adjoindre une quelconque marque intonative. Ou encore, il arrive qu ?il le fasse, par exemple en adoptant une intonation descendante pour indiquer la fin d ?un énoncé, puis se ravise ensuite et complète celui-ci. Il est alors bien délicat de déterminer ce qui doit régir la décision de l ?annotateur : l ?intonation, ou la structure phrastique ?

5.1.3 Les corpus de parole conversationnelle

Historique

Se lancer dans l ?établissement d ?un corpus de parole spontanée sous-entend des possibilités d ?enregistrement et de traitement post-enregistrement importantes. Comment en effet envisager d ?analyser un objet dont on ne saurait avoir une quelconque trace ? A cet égard, l ?historique des corpus qui nous intéressent est étroitement lié aux avancées technologiques du vingtième siècle : Queneau considérait dans Bâtons, chiffres et lettres que « l ?usage du magnétophone a provoqué en linguistique une révolution assez comparable à celle du microscope avec Swammerdam », et bien que quelques travaux précurseurs sur le sujet n ?aient pu bénéficier d ?un tel support, il est incontestable que le fait de pouvoir « capturer » l ?oral en a radicalement modifié la perception. Et pourtant, avant même que ne naisse cette invention, Damourette et Pichon (Damourette et Pichon, 1911-1927), en s ?appuyant sur des conversations recueillies auprès d ?un médecin, d ?une institutrice ? avaient dessiné les premiers contours morpho-syntaxiques d ?une « langue orale » dont la communauté linguistique avait encore pourtant du mal à admettre l ?existence. Puis il y a eu Bally (Bally, 1929) et surtout Frei (Frei, 1929) qui s'est attaché à analyser les lettres non parvenues aux soldats de la grande guerre. Ces lettres étaient rédigées par des familles souvent peu familières de l'écriture, et le style employé était en conséquence très oralisé. Mais ce sont véritablement Gougenheim et ses collaborateurs (Gougenheim et al., 1964) qui, s ?appuyant sur 275 enregistrements sonores, ont révélé par effet de bord la véritable teneur de l ?oral spontané : souhaitant avant tout proposer un équivalent du « basic English » (Ogden, 1932), et ainsi favoriser l'apprentissage du français, ils ont effectué une étude quantitative du nombre d ?occurrences des formes. Il apparut alors que des mots comme « on », « hein » ou « ben » étaient parmi les plus utilisés de la langue française, ce qu'aucune grammaire de l'époque n'envisageait. Aujourd ?hui, certaines d'entre elles se refusent encore à considérer seulement leur existence. À l ?époque où Damourette et Pichon entreprirent leurs recherches, les micro- ordinateurs n ?existaient évidemment pas, et les machines à écrire en étaient à leurs balbutiements. Les transcriptions étaient donc réalisées « à la volée », ce qui, on l ?imagine aujourd ?hui, devait s ?avérer fort peu confortable. Les ordinateurs ont certes changé la donne, offrant la possibilité d ?avoir recours au traitement de texte, et ainsi de corriger, modifier et surtout sauvegarder sans peine ses travaux. Un grand pas a ensuite été franchi lorsqu ?il est devenu possible, d ?une part de transférer les données sonores vers un ordinateur (et d ?assurer ainsi leur pérennité), et d ?autre part d ?aligner le signal audio avec le texte de la transcription : il était alors possible, en quelques secondes, d ?écouter n ?importe quelle partie de l ?enregistrement, et de voir apparaître à l ?écran la transcription qui en

avait été faite. Cette synchronisation offre entre autres la possibilité de réécouter très facilement un extrait pour voir si les propos transcrits y correspondent, et ainsi de corriger rapidement une erreur ou une interprétation. Les logiciels d ?aide à la transcription qui proposent cette fonctionnalité sont aujourd ?hui très répandus, et nous allons nous arrêter sur quelques-uns des plus utilisés à l ?heure actuelle.

Les logiciels d'aide à la transcription

Il existe principalement trois logiciels utilisés pour la transcription orthographique d'un fichier son : TRANSCRIBER (Barras et al., 1998), PRAAT4 et WINPITCHPRO (Martin, 2003). Moins répandus pour des raisons diverses (outils payants, ergonomie discutable...), CLAN, EXMARALDA ou encore TRANSANA n'en méritent pas moins d'être cités ici, chacun offrant des possibilités intéressantes. Sur le fond, bien qu'aucun ne soit réellement optimisé pour transcrire de la parole spontanée à grande échelle, leur interface globale offre cependant la possibilité d'en gérer quelques aspects. TRANSCRIBER, logiciel avec une interface et des fonctionnalités simplifiées, est optimisé pour la transcription et l'annotation de gros corpus, mais ne propose que quatre niveaux d'annotation (texte, locuteurs, thème, bruits de fond éventuels) et aucune possibilité analytique. Malgré cela, la gestion des locuteurs est très satisfaisante, puisque l'on peut indiquer pour chacun d'entre eux des informations telles que leur sexe, le degré de spontanéité, le canal d'expression... Par ailleurs, un nombre important de balises est intégré pour représenter les événements sonores (bruit, respiration, toux, reniflement...), les prononciations particulières ou encore des particularités lexicales. Le gros inconvénient de TRANSCRIBER concerne la parole superposée qui, nous le verrons, est traitée de façon trop simplifiée pour pouvoir rendre compte de ce phénomène majeur dans la langue parlée. Pour ce genre de données, PRAAT s'avère nettement plus efficace, puisqu'il offre un grand nombre de tires indépendantes les unes des autres. Il est possible d'en assigner une à chaque locuteur, et ainsi de transcrire indépendamment leurs propos, tout en les alignant avec le signal. Le fichier de sortie correspondant (« textgrid ») offre la possibilité d'organiser la transcription suivant l'échelle temporelle ou bien par locuteur, ce qui se révèle fort pratique, pour des recherches lexicales par exemple. On peut toutefois regretter qu'il ne soit pas au format XML, standard aujourd'hui incontestable pour assurer l'échange et la compatibilité des données. Autres aspects dommageables, ce logiciel présente une interface assez austère, et n'offre qu'une gestion minimale des locuteurs : hormis leur nom, rien ne peut être indiqué dans l ?espace qui leur est attribué. À vrai dire PRAAT, bien moins efficace que TRANSCRIBER pour le traitement de gros corpus, est généralement privilégié pour des tâches spécifiques, notamment l'analyse de la prosodie, domaine dans lequel il se révèle très complet grâce à la possibilité d'intégrer à ce logiciel de nombreux modules complémentaires. WINPITCHPRO est pour sa part plus difficile d'accès, moins par son interface (plutôt intuitive) que par la richesse de ses fonctionnalités. Certes moins habile que TRANSCRIBER pour gérer les fichiers audio de grande taille, il permet des analyses très fines (96 niveaux d'annotation sont disponibles, soit autant de possibilités de codage) et à plusieurs niveaux : prosodie, phonologie... Par ailleurs, il traite les fichiers audio et vidéo, ce qui le distingue des deux outils précités et permet une synchronisation entre l'image, le signal et la transcription. Il est malgré tout regrettable

que cet outil ne fonctionne que sous Windows, et qu'il ne soit pas open-source.

Conventions d'annotation

L'un des problèmes qui se posent lorsque l'on entreprend d'effectuer une transcription est celui des conventions d'annotation à adopter : outre le texte lui- même, que veut-on représenter à l'écran, et surtout comment souhaite-t-on le faire ? Transcriber, par exemple, a son propre « manuel du transcripteur », indépendant du manuel d'utilisation, et qui passe en revue de nombreux aspects de la langue orale, en en proposant à chaque fois un codage14. Heureuse initiative qui permet aux utilisateurs de réaliser rapidement des transcriptions complètes et unifiées, d'autant que ni PRAAT ni WINPITCHPRO ne proposent ce genre de documentation. Dans la pratique, des conventions diverses sont nées au fil des projets ou des groupes de recherche qui ont vu le jour. Des initiatives telles que le Linguistic Data Consortium15 ou la TEI16 proposent également des conventions pour la transcription de la parole. Ainsi, si parfois ces codages se recoupent, il possibilités de représentation soient très nombreuses.

Comme on le voit ci-dessus, pour un seul phénomène (parmi bien d'autres), il existe au moins cinq codages différents. Il serait pourtant indispensable de s'orienter vers un codage unifié, ne serait-ce que pour permettre un échange, une compatibilité et une lecture des données plus simples. La TEI (Text Encoding Initiative), ensemble de recommandations pour coder des informations avec une nomenclature prédéfinie afin de pouvoir les échanger facilement ensuite, est une base difficilement contestable. Un chapitre y est consacré à la transcription de la parole. Très complet, il passe en revue tous les principaux phénomènes conversationnels et propose des solutions très intéressantes, pour la parole interactive et superposée notamment. La prosodie y est également considérée sous de nombreux aspects (vitesse d'élocution, volume sonore, intonation, rythme, qualité de la voix...). Cependant, ce format reste assez difficile à appréhender car, s'il permet de coder de très nombreux paramètres, il n'existe malheureusement pas d'interface qui les représente de façon intuitive à l'écran.

5.1.4 Transcription manuelle vs. transcription assistée : quel(s) gain(s) ?

Dans le but de quantifier le gain de temps qui pouvait être obtenu grâce à l'utilisation d'un système de reconnaissance automatique par rapport à une transcription réalisée entièrement à la main, nous avons mené l'expérience suivante : 24 segments d'environ 10 minutes ont été sélectionnés parmi les données non transcrites du corpus ESTER : 12 considérés comme étant de la parole spontanée (débats ou interviews), et 12 comme de la parole préparée (informations). Sur chacun de ces fichiers, une transcription manuelle et une transcription assistée ont été effectuées par le même transcripteur (suffisamment longtemps après pour que la seconde transcription ne soit plus influencée par la mémoire de la première). Cette transcription comportait trois niveaux : la segmentation en tours de parole et la transcription ; l'assignation des locuteurs ; la vérification orthographique. Pour chacune de ces étapes, un chronométrage à la minute a été effectué. Voici les principaux résultats que nous avons

obtenus :

Tab. 5.1 – Durée totale de la transcription (durées respectives des corpus : 2h08 et 2h10)

	Parole préparée	Parole spontanée
Transcription manuelle	17h36	19h33
Transcription assistée	8h31	15h44

Le tableau 6 montre que la transcription assistée induit un important gain de temps, surtout pour la parole préparée. Pour ce type de données, le temps nécessaire à la transcription est approximativement deux fois moins important lorsque le transcripteur est assisté. Lorsqu'il s'agit de parole spontanée, ce bénéfice est bien moindre.

Tab. 5.2 – Rapport entre la durée totale de la transcription et la durée totale des fichiers

	Parole préparée	Parole spontanée
Transcription manuelle	8,26	9,05
Transcription assistée	4,00	7,29

Étant donné que les fichiers spontanés et préparés représentent peu ou prou la même durée, le rapport entre celle-ci et le temps total nécessaire à la transcription (Tableau 7) est un élément qu'il est pertinent de prendre en compte : si l'on considère un segment de parole préparée de dix minutes, le transcripteur aura besoin d'environ quarante minutes pour transcrire le texte, assigner les locuteurs et vérifier l'orthographe, s'il s'appuie sur un fichier de transcription généré automatiquement. Si l'on réalise les mêmes tâches sur le même fichier, mais cette fois de façon entièrement manuelle, environ 83 minutes seront nécessaires, soit un temps de travail plus que doublé. La même expérience, mais cette fois avec un fichier de parole spontanée, montre qu'une transcription assistée demande 73 minutes de travail, chiffre qui est presque le double de celui obtenu dans les mêmes conditions avec la parole préparée. À l'inverse, la transcription manuelle (90 minutes) n'est cette fois pas beaucoup plus coûteuse en temps que la transcription assistée. Ainsi, s'il est indéniable qu'une transcription assistée est synonyme de gain de temps, ce dernier est beaucoup plus important lorsqu'il s'agit de parole préparée.

Tab. 5.3 – Transcription du texte et segmentation

	Parole préparée	Parole spontanée
Transcription manuelle	13h36	16h15
Transcription assistée	5h06	12h41

C'est lors de la tâche de transcription du texte (Tableau 8) que le gain le plus intéressant a été obtenu : sur de la parole préparée, une transcription manuelle nécessite environ 2,67 fois plus de temps qu'une transcription assistée (5h06 vs 13h36). Ce chiffre est très significatif, notamment s'il est comparé à celui obtenu avec la parole spontanée : pour une durée sensiblement équivalente, il chute à 1,28. Cet écart met en exergue le fait que les systèmes de reconnaissance automatique de la parole éprouvent des difficultés à traiter la parole spontanée, obligeant le transcripteur à effectuer par la suite beaucoup de corrections manuelles.

Tab. 5.4 – Assignation des locuteurs

	Parole préparée	Parole spontanée
Transcription manuelle	1h17	2h13
Transcription assistée	1h17	2h13

En ce qui concerne l'assignation des locuteurs (Tableau 9), il est surtout important de retenir qu'elle demande presque deux fois plus de temps quand la parole est spontanée. Cela peut s'expliquer relativement facilement : la parole spontanée, avec ses nombreux tours de parole, contraint le transcripteur à devoir souvent leur assigner un locuteur, quand bien même il peut n'y en avoir que deux différents dans un fichier. À l'inverse un segment de parole préparée contient souvent de nombreux locuteurs (journalistes, reporters, interviewés, speakers...), mais beaucoup moins de tours de parole dans la mesure où ceux-ci sont beaucoup plus longs. De plus, dans un segment spontané se trouve parfois de la parole superposée, et lorsque trois locuteurs ou plus sont susceptibles de prendre la parole, cela peut être long et difficile de déterminer qui parle réellement.

Tab. 5.5 – Correction orthographique

	Parole préparée	Parole spontanée
Transcription manuelle	2h43	1h05
Transcription assistée	2h08	0h51

Le minutage de la correction orthographique (Tableau 10) a permis d'observer un phénomène remarquable : si la différence spécifique entre transcription manuelle et assistée n'est certes pas très significative, celle entre parole préparée et spontanée l'est beaucoup plus. La raison en est fort simple : les segments de parole préparée contiennent essentiellement de l'information radiophonique ; or ce genre de données s'avère très riche en noms propres (reporters, interviewés, personnalités, villes...), dont les orthographes exactes ne peuvent être systématiquement connues de l'annotateur. Les rechercher peut donc être une tâche assez longue, notamment dans le cas de noms étrangers. Inversement, les fichiers de parole spontanée étant des interviews ou des débats, on y trouve très peu de noms propres car les thèmes abordés ne nécessitent en général qu'un faible emploi de ces entités nommées.

Tab. 5.6 – Correction orthographique

	Parole préparée	Parole spontanée
Transcription manuelle	16,95%	35,21%
Transcription assistée	15,83%	34,33%

Les dernières observations effectuées concernent le taux d'erreur mot (Tableau 11). Celui-ci a été mesuré à partir des sorties automatiques générées par le système LIUM RT, que nous avons comparées aux transcriptions manuelles puis assistées réalisées par l'annotateur. Les moyennes indiquées ci-dessus confirment ce que nous disions précédemment : le système de reconnaissance automatique du LIUM n'est pas aussi performant sur la parole spontanée que sur la parole préparée. À titre d'exemple, le taux d'erreur mot le plus élevé que nous ayons obtenu sur de la parole préparée était de 21,8%, alors qu'il s'est élevé à 53,4% avec la parole spontanée. Les différences observées entre les tâches manuelles et assistées peuvent être expliquées par le fait que le transcripteur n'a pas forcément transcrit le même texte à chaque fois : il est parfois difficile de percevoir clairement des phénomènes tels que les répétitions, les faux départs ou encore la parole superposée, et en conséquence leur transcription ne sera pas toujours identique, même lorsqu'elle est réalisée deux fois par la même personne.

5.1.5 Le corpus EPAC

5.2 Caractérisation et Détection de la parole spontanée

5.2.1 Introduction

Information Extraction (IE) from large audio databases requires to extract the structure of audio documents as well as their linguistic content. One part of this structuration process is for example to add punctuations and sentence boundaries to the automatic transcriptions of the speech segments detected : this segmentation process is very important for many tasks like speech summarization, speech-to-speech translation or the *distillation* task as defined in the GALE program (77). Adding this structure to the automatic transcripts is a very challenging task when processing spontaneous speech as this kind of speech is characterized by ungrammaticality and disfluencies. Moreover, in order to cluster some documents according to their contents or structuration, the presence of spontaneous speech segments should be an interesting descriptor. It is therefore useful to detect spontaneous speech segments at an early stage in order to adapt the ASR, as presented in (6) and structuration processes to this particular kind of speech. This is the goal of this study.

Spontaneous speech occurs in Broadcast News (BN) data under several forms : interviews, debates, dialogues, etc. The main evidences characterizing spontaneous speech are disfluencies (filled pause, repetition, repair and false start) and many

studies have focused on the detection and the correction of these disfluencies (78; 79) as pointed out by the recent NIST Rich Transcription Fall 2004 blind evaluation. All these studies show an important drop in performance between the results obtained on reference transcritptions and those obtained on automatic transcripts. This can be explained by the noise generated by ASR systems on spontaneous speech segments with higher Word Error Rate (WER) values than on *prepared* speech. Indeed high WER values are obtained by state-of-the-art ASR systems when transcribing data is likely to contain a lot of spontaneous speech like conversational speech or meeting recordings. In this study, we illustrate this link between WER and spontaneous speech.

In addition to disfluencies, spontaneous speech is also characterized by ungrammaticality and a language register different from the one that can be found in written texts (80). Depending on the speaker, the emotional state and the context, the language used can be very different. In this study we define *spontaneous speech* as *unprepared speech*, in opposition to *prepared speech* where utterances contain well-formed sentences close to those that can be found in written documents. We propose a set of acoustic and linguistic features for characterizing *unprepared speech*. The relevance of these features is estimated on an 11 hour corpus (French Broadcast News) manually labelled according to a level of spontaneity in a scale from 1 (clean, prepared speech) to 10 (highly disfluent speech, almost not understandable). We present an evaluation of our features on this corpus, describe the correlation between the Word-Error-Rate obtained by a state-of-the-art ASR decoder on this BN corpus and the level of spontaneity and finally propose a strategy that takes advantage of a global decision process to assign a level of spontaneity to a speech segment.

5.2.2 Spontaneous speech characterization

Levels of spontaneity

By defining spontaneous speech as *unprepared speech*, we follow a definition proposed by (81) that defined a spontaneous utterance as : "a statement conceived and perceived during its utterance". This definition illustrates the subjectivity of the classification prepared/spontaneous speech. Ideally, to annotate a speech corpus with labels representing the spontaneity of each speech segment, we would have to ask each speaker to annotate his own utterances. This is of course not feasible, however we followed this definition by defining an annotation protocol. This protocol is based on the perception given by a human judge thanks to a *level of spontaneity* for a given speech segment. Our approach was to manually tag a corpus of speech segments with a set of eight labels, each one corresponding to a spontaneity level : grade 1 stands for prepared speech, almost similar to read speech, and grade 8 stands for very disfluent speech, almost not understandable. This approach allows us to subjectively choose where the limit between spontaneous and prepared speech is placed. In the experiment we considered 3 classes : *prepared speech* corresponding to grade 1 ; *low spontaneity* corresponding to the grades 2 to 4 ; and *high spontaneity* corresponding to the grade 5 and over.

In fact, in this article, we focus particularly on the detection of the *high spontaneity* class of speech.

Two human judges have annotated a speech corpus by listening to the audio recordings. The corpus was cut into segments thanks to a state-of-art automatic segmentation and diarization process (71). No transcriptions were provided to the annotators. In order to evaluate inter-annotator agreement for this specific tagging task on the 3 classes presented above, we computed the Kappa coefficient of agreement (82) on one hour of Broadcast News. The coefficient obtained was very high : 0.852 — a value greater than 0.8 is usually considered as excellent (83).

Then, they have annotated the remaining corpus separately. One of the problems encountered was that spontaneous speech segments can occur everywhere, not only in conversational speech, in the middle of very *clean* utterances. Similarly even conversational speech can contain segments that can be considered as prepared speech. To take this into account, we decided to evaluate each segment independently : a spontaneous segment can be surrounded by many prepared ones.

The corpus obtained after this labelling process is made of 11 files containing French Broadcast News data from 5 different media (France Culture, France Inter, France Info, Radio Classique, RFI). The files were chosen for being likely to contain spontaneous speech according to the kind of radio show broadcast. The total duration is 11h37 for a total of 11821 segments (after removal of the non speech segments : music, jingles, ...). Among these segments, 3670 were annotated with the *prepared speech* label, 4107 with the *low spontaneity* label and 4044 with the *high spontaneity* label.

Acoustic and linguistic features

In parallel to the subjective annotation of the corpus presented in the previous section, we now introduce the features used to describe speech segments. We chose speech segments that are relevant to characterize the spontaneity of those, and on which an automatic classification process can be trained on our annotated corpus. This problem has been studied recently as a specific task from the Rich Transcription Fall 2004 blind evaluation which was focused on the detection of speech disfluencies. Some approaches use only linguistic features (79), both linguistic and prosodic features (84), or linguistic and more general acoustic features (85). These features will be associated with confidence measures given by ASR.

In this paper we use three sets of features : acoustic features related to prosody, linguistic features related to the lexical and syntactic content of the segments, and confidence measures. We combine them in order to characterize the spontaneity class of a speech segment : this task is different from the speech disfluency detection task as spontaneous speech segments do not necessarly contain disfluencies. For example, they can also be characterized by a high variation in the speech rate. The features used in this study are briefly presented in the next section.

Prosodic features The prosodic features used are related to vowel duration and phonetic rate, as presented below.

Duration : following previous work describing the link between prosody and spontaneous speech (86), we use two features : vowel duration and the lengthening of a syllable at the end of a word. The latter has been proposed in (87) and is

associated to the concept of *melism*. In addition to the average durations, their variance and standard deviation are also added as features in order to measure the dispersion of the durations around the average.

Phonetic rate : previous studies (87) have shown the correlation between the variations of speech rate and the emotional state of a speaker. Following this idea we use as feature an estimate of the speech rate by speech segment, in order to observe its impact on the spontaneity of the speech. We estimate the phonetic rate in two ways : the average and the variance of the phonetic rate on the whole segment, firstly including pauses and fillers, and secondly removing them.

Tables 5.7 and 5.8 respectively present the average on duration and the average on variance of vowels, phonetic rate and melisms between the three levels of spontaneity. These values were computed on the experimental data described in section 5.2.2.

	prepared	low spontaneous	high spontaneous
vowels	0.075	0.081	0.091
phonetic rate	0.078	0.081	0.087
melisms	0.082	0.094	0.11

TAB. 5.7 – Comparison of average on duration of vowels, phonetic rate and melisms according to the three classes of spontaneity.

	prepared	low spontaneous	high spontaneous
vowels	0.0018	0.0033	0.0071
phonetic rate	0.0017	0.0026	0.0046
melisms	0.0025	0.0051	0.0113

TAB. 5.8 – Comparison of average on variance of vowels, phonetic rate and melisms between the the three classes of spontaneity.

Results show that there is a correlation between these features and the level of spontaneity, whether at duration level or at variance level, which is higher on *high spontaneous* segments.

Linguistic features The main characteristic of spontaneous speech is the concept of *speech disfluencies*. They can be categorized as filled pause, repetition, repair and false start. A lot of studies have been focused on their description at the acoustic (86) or lexical level (88). We use two features representing them in the description of the speech segments :

- filled pause : the ASR lexicon contains several symbols, filler words, for representing filled pause in French, like *euh*, *ben* or *hum*. The number of occurences of all of them in a segment is the first feature.
- repetition and false start : we use here a very simple feature counting the number of 1-gram and 2-gram repetitions in a segment.

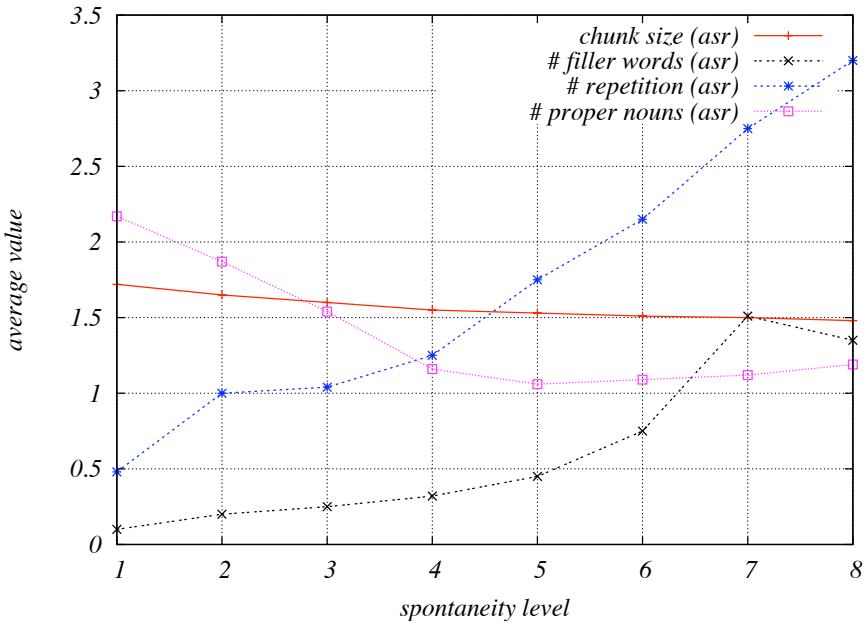

FIG. 5.1 – Linguistic feature average values according to the degree of spontaneity on the manually labeled corpus

As shown by (80) on BN data, spontaneous speech is also characterized at the linguistic level by other phenomenon than filled pause or repetition. Agrammaticality and language register are also very characteristic of unprepared speech. In order to capture this link between spontaneity on one side and lexicon and syntax on the other side, we apply to the transcriptions of audio segments a shallow parsing process including a POS tagging and a syntactic chunking process and use the following features to describe them :

- bags of n-grams (from 1 to 3-grams) on words, POS tags and syntactic chunk categories (noun phrase, prepositional group) ;
- average length of syntactic chunks on the segment, words and POS tags count.

Moreover, as presented in (2), a high number of occurrences of proper nouns in a speech segment can be informative to characterize prepared speech : this information is used in this work.

Figure 5.1 shows the correlation between the level of spontaneity assigned to the speech segment of our corpus and the linguistic features presented. Although these numbers are obtained on automatic transcripts with a high WER on the most spontaneous segments, there is a clear increase for both disfluency features between clean and spontaneous speech. Although the variation of the average chunk size is limited, figure 5.1 shows a reduction of this size, from 1.7 down to 1.4 words on average per chunk between clean speech and very spontaneous one.

Confidence measures Confidence measures are computed scores that expressed reliability of recognition decisions made by ASR system. These scores could be used to characterize spontaneity of speech segments. Indeed, as we have seen, automa-

tic speech recognition systems have more difficulties to well recognize spontaneous speech segments than prepared speech segments. Table 5.9 presents the comparison of the confidence measures average and variance between the three classes of spontaneity obtained on the labeled corpus. These confidence measure were provide by our ASR system described below.

	prepared	low spontaneous	high spontaneous
average	0.91	0.88	0.82
variance	0.021	0.026	0.036

TAB. 5.9 – Comparison of the confidence measures average and variance according to the speech category.

It appears that there is a gradation of results, depending of the level of spontaneity : average on values decreases when speech becomes more spontaneous, while variance value increases. Confidence measures seems to be a good indicator of class of spontaneity. Thus, confidence measures will be associated with acoustic and linguistic features to improve detection and classification of speech segments.

5.2.3 Automatic detection of spontaneous speech segments

To automatically extract the acoustic and linguistic descriptors to categorize speech segments according to class of spontaneity, we used the LIUM ASR system. This section firstly briefly presents it.

Then we present the classifier used to combine these descriptors to categorize speech segments.

Entire automatic speech recognition system

The LIUM ASR system was developed to participate to the French ESTER 2 evaluation campaign on Broadcast News automatic transcription systems. This ASR system includes a speaker diarization system.

Diarization The speaker diarization system is an internal tool developed at LIUM. It is composed of an acoustic BIC-based segmentation followed by a BIC-based hierarchical clustering. Each cluster represents a speaker and is modeled with a full covariance Gaussian. Viterbi decoding is used to adjust the segment boundaries using GMMs for each cluster. Music and jingle regions are removed using Viterbi decoding with 8 GMMs, for music, jingle, silence, and speech (with wide/narrow band variants for the latter two, and clean/noised/musical background variants for wide-band speech). Gender and bandwidth are then detected using 4 gender- and bandwidth-dependent GMMs. Speech segments are then limited to 20 s by splitting overlong segments using a GMM-based silence detector. This system, completed by a CLR-based clustering phase, obtained the best diarization error rate during the ESTER 2 evaluation campaign on december 2008. This evaluation campaign compared ASR systems on French Broadcast News recordings.

Speech recognition system The LIUM automatic speech recognition system is based on the CMU Sphinx system. The tools distributed in the CMU Sphinx open-source package, although already reaching a high level of quality, can be supplemented or improved to integrate some state-of-art technologies. It is the solution LIUM has adopted to develop its own ASR system, by building on this base and gradually extending it to bring it to new performance levels.

Features The transcription decoding process is based on multi-pass decoding using 39 dimensional features (PLP with energy, delta, and double-delta). Two sets of features are computed for each show, corresponding to broadband (130 Hz - 6800 Hz) and narrowband (440 Hz - 3500 Hz) analysis.

Decoding The decoding strategy involves 5 passes :

1. After speaker diarization, the first pass uses a trigram language model and generic acoustic models (one for each of the four gender/band conditions — female/male + studio/telephone).
2. The best hypotheses generated by pass (1) permit to compute a CMLLR transformation for each speaker. Decoding (2), using SAT and Minimum Phone Error (MPE) acoustic models and CMLLR transformations, generates word-graphs.
3. In the third pass, the word-graphs are used to drive a graph-decoding with full 3-phone context with a better acoustic precision, particularly in inter-word areas. This pass generates new word-graphs.
4. The fourth pass consists in recomputing with a quadrigram language model the linguistic scores of the updated word-graphs of the third pass.
5. The last pass generates a confusion network from the word-graphs and applies the consensus method to extract the final one-best hypothesis (32).

Acoustic models Acoustic models for 35 phonemes and 5 kinds of fillers are trained using a set of 240 hours from the ESTER 1 & 2 training corpus, plus 40 hours of transcribed French broadcast news provided by the EPAC project. Models for pass (1) are now composed of 6500 tied states. Models for passes (2) to (5) are composed of 7500 and are trained in a MPE (28; 89) framework applied over the SAT-CMLLR models. Each state is modeled by a mixture of 22 diagonal Gaussians. Both decoding passes employ tied-state word-position 3-phone acoustic models which are made gender- and bandwidth-dependent through MAP adaptation of means, covariances and weights. (28; 89) framework applied over the SAT-CMLLR models.

Vocabulary and Language models Data used to build the linguistic models are of three kinds :

1. Manual transcriptions of broadcast news. They correspond to the transcription of the data used to train the acoustic models. We have also used manual transcriptions of conversations from the PFC corpus (90) ;

2. Newspaper articles : in addition to 19 years of "Le Monde" newspaper corpus, we also use articles from another French newspaper, "L'Humanité", from 1990 to 2007, and the French Giga Word Corpus ;
3. Web resources drawn from "L'Internaute", "Libération", "Rue89", and "Afrik.com".

To build the vocabulary, we generate a unigram model as a linear interpolation of unigram models trained on the various training data sources listed above. The linear interpolation was optimized on the ESTER 2 development corpus in order to minimize the perplexity of the interpolated unigram model. Then, we extract the 122k most probable words from this language model.

The trigram and quadrigram models are trained with the SRILM toolkit using the modified Kneser-Ney discounting method no cut-off were applied. The models are composed of 121k unigrams, 29M bigrams, 162M trigrams, and 376M quadrigrams.

Phonetic transcriptions for the vocabulary are taken from the BDLEX database, or generated by the rule-based, grapheme-to-phoneme tool LIA_PHON (21) for words not in the database.

Classification

The features presented in the previous section are evaluated on our labeled corpus with a classification task : labeling speech segments according to the three classes of spontaneity : *prepared speech*, *low spontaneity* or *high spontaneity* label. The classification tool used is *icsiboost*[1], an open source tool based on the AdaBoost algorithm like the *Boostexter* software (91). This is a large-margin classifier based on a boosting method of *weak* classifiers. The weak classifiers are given as input. They can be the occurrence or the absence of a specific word or n-gram (for the linguistic features) or a numerical value (discrete or continuous : for the acoustic features or the confidence measures). At the end of the training process, the list of the selected classifiers is obtained as well as the weight of each of them in the calculation of the classification score for each speech segment to process.

This classification process, taking into consideration the acoustic and linguistic descriptors presented in section 5.2.2, plus some other descriptors as the duration of speech segments and the number of recognized words propose a categorization of the speech segments according the three class of spontaneity. Each segment is processed individually.

5.2.4 Probabilistic contextual model for global decision

Intuitively, we can feel that it should be rare to observe a *high spontaneous* speech segment surrounded by two prepared speech segments. Our previous approach, presented in (6), takes only into consideration the descriptors which are extracted from within the targeted segment, without taking into consideration information about surrounding segments. In order to improve our approach, we propose to take into account the nature of the contiguous neighboring speech segments. It implies that the categorization of each speech segment from an audio file has an impact on the

[1] http ://code.google.com/p/icsiboost

categorization of the other segments : the decision process becomes a global process. We have chosen to use a statistical classical approach by using a maximum likelihood method.

General approach

Let be s_i a tag of the segment i, with $s_i \in \{$"*high spontaneity*", "*low spontaneity*","*prepared*"$\}$. We define $P(s_i|s_{i-1}, s_{i+1})$ as the probability of observing a segment i associated to the tag s_i when the previous segment is associated to the tag s_{i-1} and the next segment is associated to the tag s_{i+1}. Let be $c(s_i)$ the confidence measure given by the AdaBoost classifier on choosing the tag s_i for the speech segment i according to the values of the descriptors extracted from this segment. S is a sequence of tags s_i associated to the sequence of all the speech segments i (only one tag by segment). The global decision process consists in choosing the tag-sequence hypothesis $\bar{S}$ which maximizes the global score obtained by combining $c(s_i)$ and $P(s_i|s_{i-1}, s_{i+1})$ for each speech segment i detected on the audio file. The sequence $\bar{S}$ is computed by using the following formula :

$$\bar{S} = \arg\max_{S} \quad c(s_0) \times c(s_n) \times \prod_{i=2}^{n-1} \times c(s_i) \times P(s_i|s_{i-1}, s_{i+1}) \tag{5.1}$$

where n it the number of speech segments automatically detected in the recording file.

Finite-State Machine paradigm

In practice, to resolve the equation (5.1) we have projected the problem into the Finite-State Machine (FSM) paradigm by using weighted finite-state transducer, as presented in (92). For our experiments, we used the AT&T FSM toolkit[2].

To do that, we have to represent the model containing the probabilities $P(s_i|s_{i-1}, s_{i+1})$ for all the 3-tuples (s_{i-1}, s_i, s_{i+1}) in a transducer representation. This FSM will be called M. Figure 5.2 shows the topology used to represent all these probabilities using the FSM formalism. As we can see in this figure, input values of the transducer are used to represent 3-tuples (s_{i-1}, s_i, s_{i+1}), while output values correspond to the effective tag-sequence : the global cost of a hypothesis tag-sequence, according to the contextual tag model, is the cost of the path of which the output values corresponds to this tag-sequence.

We have chosen to handle the costs by using the tropical semi-ring : the cost value of a 3-tuple (s_{i-1}, s_i, s_{i+1}) corresponds to the values of $-logP(s_i|s_{i-1}, s_{i+1})$. Notice that in this paper, *null* input or output values are represented by the symbol '#' (sharp).

To apply Mod to the search space representing all the hypotheses, we had to represent these hypotheses by using a FSM formalism compatible with Mod. Figure 5.3 shows by an example that the FSM representing all the possible tag-sequence hypotheses in an audio file is built in order to make compositional its output values and its topology with the input values and the topology of Mod. The FSM representing

[2]http ://www.research.att.com/ fsmtools/fsm/

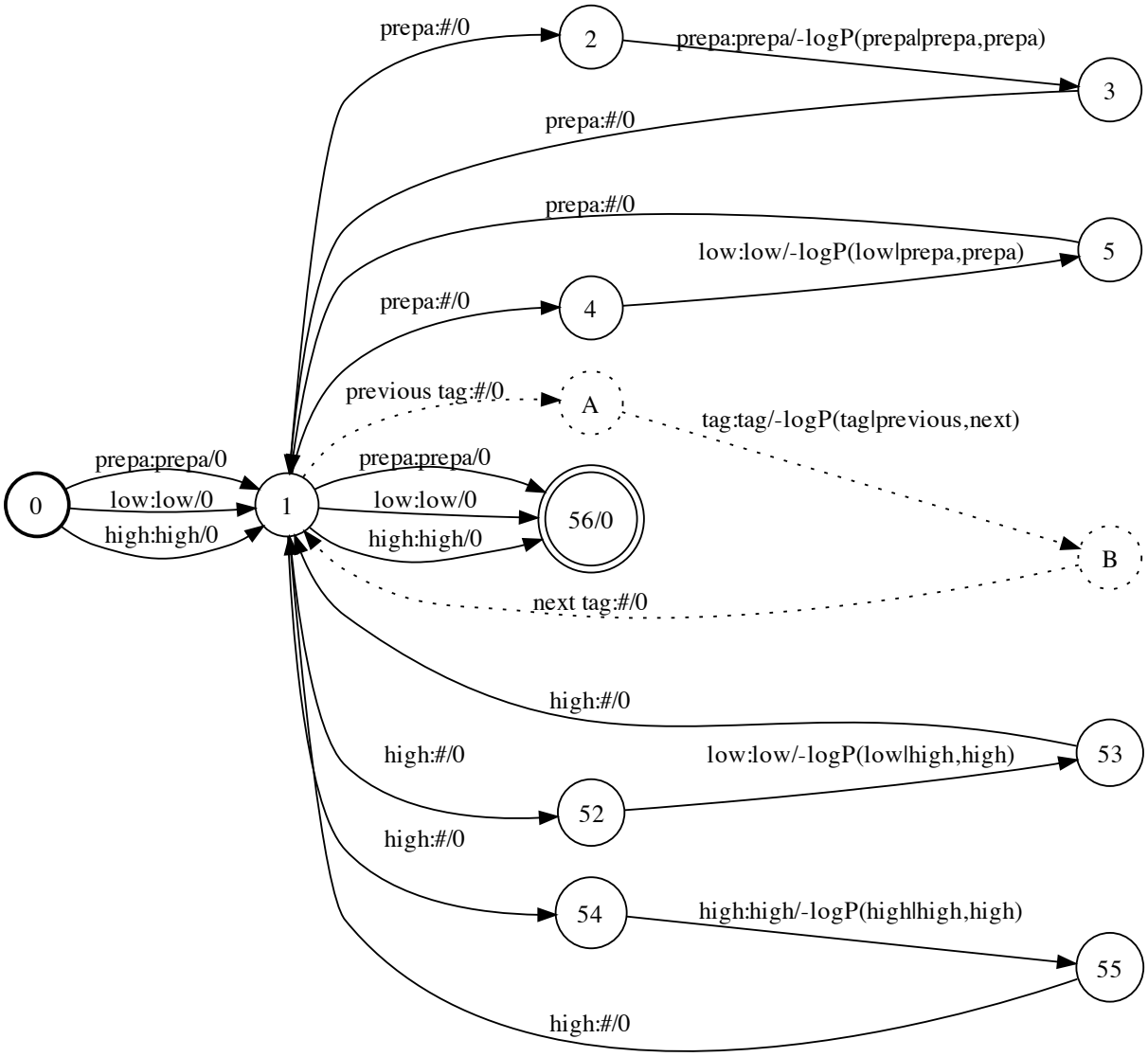

FIG. 5.2 – Transducer *Mod* modeling all the contextual probabilities $P(s_i|s_{i-1}, s_{i+1})$

FIG. 5.3 – Topology of the transducer *Hyp* representing all the hypotheses

the hypotheses will be called Hyp. The costs in Hyp are the confidence measure values $c(s_i)$ given by the AdaBoost classifier for each of the three possible tags (*prepa*, *low*, and *high* labels for respectively *prepared speech*, *low spontaneous* and *high spontaneous* classes of spontaneity).

In order to reduce the number of paths in Hyp, we have taken into consideration the fact that some parts of paths can be factorized : as all the 3-tuples $(*, s_i, s_{i+1})$ will be followed by 3-tuples $(s_i, s_{i+1}, *)$, we have used this property during the generation of the topology of Hyp. This can be observed in the figure 5.3 just before the dotted boxes which means that the FSM has to be continued.

Then, to apply the contextual tag-model represented by Mod to the hypotheses represented by Hyp, we make the following composition of transducers : $Hyp \circ Mod$.

Last, as we use the tropical semi-ring to handle FSM costs to find the minimum cost path in $Hyp \circ Mod$, this means that an approximated Viterbi decoding is used : when multiple paths are identically labelled, the tropical semi-ring selects only the minimum cost path. The output values of the minimum cost path corresponds to the final tag-sequence hypothesis.

Due to the very small number of tags, and due to the average number of speech segments detected in an audio file, this computation (FSM composition + Viterbi decoding) is very fast (less than 6 seconds of computation time in a 2007 Apple MacBook Pro laptop to process the 11 files corresponding to 11h37 of speech and 11821 segments).

Of course, an optimized tool, in term of processing time, to compute the best tag-sequence hypothesis with respect to the formula 5.1 could be implemented without using the FSM formalism. But this latter allowed us to make our experiments without having to develop a specific tool for a very sufficient computation time.

5.2.5 Experiment

The experimental corpus (as described in 5.2.2) is made of 11 audio files from radiophonic recording. For the experiments, we used the *Leave One Out* method : 10 files used for training, 1 for the evaluation and this process is repeated until all files have been evaluated.

ASR performances

The acoustic and linguistic features used as descriptors to characterize the spontaneous speech are issued from the LIUM ASR system described in section 5.2.3. Table 5.10 presents the results in terms of word error rate (WER) and normalized cross entropy (NCE) of this ASR system on the experimental data. These data were not included in the training or development corpus of the models used in the ASR system. The WER is the classical metric to evaluate ASR systems, while the NCE is usually used to evaluate the confidence measures provided by an ASR system.

Table 5.10 shows that the global performances of the ASR, with a WER of 15% and a NCE of 0.331 are very good for French Broadcast News processing. As it was expected, more the speech is fluent, more the WER is low : from 10.1% for speech segments manually annotated as "prepared" until 28.5% for "high spontaneous"

speech category	# segments	WER	NCE
prepared	3670	10.1%	0.358
low spontaneous	4107	18.4%	0.315
high spontaneous	4044	28.5%	0.237
all	11821	15.0%	0.331

TAB. 5.10 – Performances of the ASR system according to speech category in terms of WER and NCE. The number of segments according to speech category is also included

speech segments. It is interesting to notice that the correlation between subjective annotation on spontaneous level and the WER obtained by an ASR system.

These results about the performances of the ASR system help to provide a context for the results presented below about spontaneous speech detection.

Last, it is interesting to notice that the computation time of the LIUM ASR system for this task is about 10x real time, including phone alignments and confidence measure computation. Notice too that such computation is very easily distributable, for example by using one CPU in a cluster by file to process.

Automatic categorization and detection of spontaneous speech

In order to measure the information provided by the different kinds of descriptors and the gain provided by the use of a probabilistic contextual model for global decision, five conditions were evaluated :

- Linguistic features only on reference transcription *ling(ref)*
- Linguistic features only on automatic transcription *ling(asr)*
- Acoustic features only on automatic transcription *acou(asr)*
- All features on automatic transcription *all(asr)*
- Use of a probabilistic contextual model *all(asr)* results : *all + global(asr)*

Table 5.11 presents the detection results (in terms of precision and recall) for each spontaneity class. As we can see the detection performance on the *low spontaneity* segments is low, this is not suprising as these segments can be easily misclassified as *prepared speech* one one side or *high spontaneity* on the other side.

As we can see the drop between the performance achieved on the reference transcriptions using linguistic features and the automatic transcriptions, due to ASR errors, is compensated by the acoustic features that are more robust to ASR errors : the use of a classifier based on all the acoustic and linguistic features extracted automatically (*all(asr)*) improves performances. In comparison to the use of linguistic features coming from manual transcriptions, by merging acoustic and linguistic features extracted from the ASR outputs, we obtain better results whatever the class of spontaneity or the metric used, except in terms of recall for prepared speech.

By examining the results of the *global+all(asr)* condition, we observe that the probabilistic contextual tag model applied on the *all(asr)* condition allows to significantly improve the performance of the classification whatever the class of spontaneity or the metric used.

prepared speech					
Feat.	ling(ref)	ling(asr)	acou(asr)	all(asr)	all+global(asr)
Prec.	56	53.0	56.3	57.8	62.1
Recall	64.1	61.8	58.3	61.7	64.2
low spontaneous					
Feat.	ling(ref)	ling(asr)	acou(asr)	all(asr)	all+global(asr)
Prec.	43.8	40.7	44.0	45.5	49.2
Recall	37.7	31.7	41.3	40.5	44.2
high spontaneous					
Feat.	ling(ref)	ling(asr)	acou(asr)	all(asr)	all+global(asr)
Prec.	65.2	58.0	59.7	65.5	**69.3**
Recall	65.9	62.4	61.6	68.8	**74.6**

TAB. 5.11 – Precision and recall in the classification of the speech segments according to 3 categories : *prepared speech*, *low spontaneity* and *high spontaneity*

In fact, in this article we are particularly interested on the detection of *high spontaneous* speech segments. By accepting all the propositions of the classification, our method allows to achieve a 69.3% precision for high spontaneous speech detection with a 74.6% recall measure, as presented in table 5.11. More precisely, 83.5% of high spontaneous detection errors are due to confusion between low and high spontaneous speech.

But, according to the application targeted by using this high spontaneous speech segment detection, it can be necessary to get a better precision. Using the scores $c(s_i)$ given by the classifier combined with the probabilities $P(s_i|s_{i-1}, s_{i+1})$ provided by the contextual tag model, it is possible to filter the proposition by applying a threshold to the value of $c(s_i) \times P(s_i|s_{i-1}, s_{i+1})$.

Figure 5.4 presents the detection performance obtained by changing the threshold on classification score for *high spontaneous* segments : we can see that our system could be more accurate (precision increase) when we take less decisions (recall decrease). This possibility of thresholding can adapt the use of the classification method by finding the best compromise between recall and precision for the targeted application.

5.2.6 Conclusion

We propose a set of acoustic and linguistic features that can be used for characterizing and detecting spontaneous speech segments from large audio databases. To better define this notion of unprepared speech, a set of speech segments representing an 11 hour corpus (French Broadcast News) has been manually labelled according to a level of spontaneity : the correlation between the Word-Error-Rate and the level of spontaneity obtained by LIUM state-of-the-art ASR decoder on this BN corpus is presented.

The acoustic and linguistic features are evaluated in order to characterize and detect spontaneous speech segments : the combination of acoustic and linguistic fea-

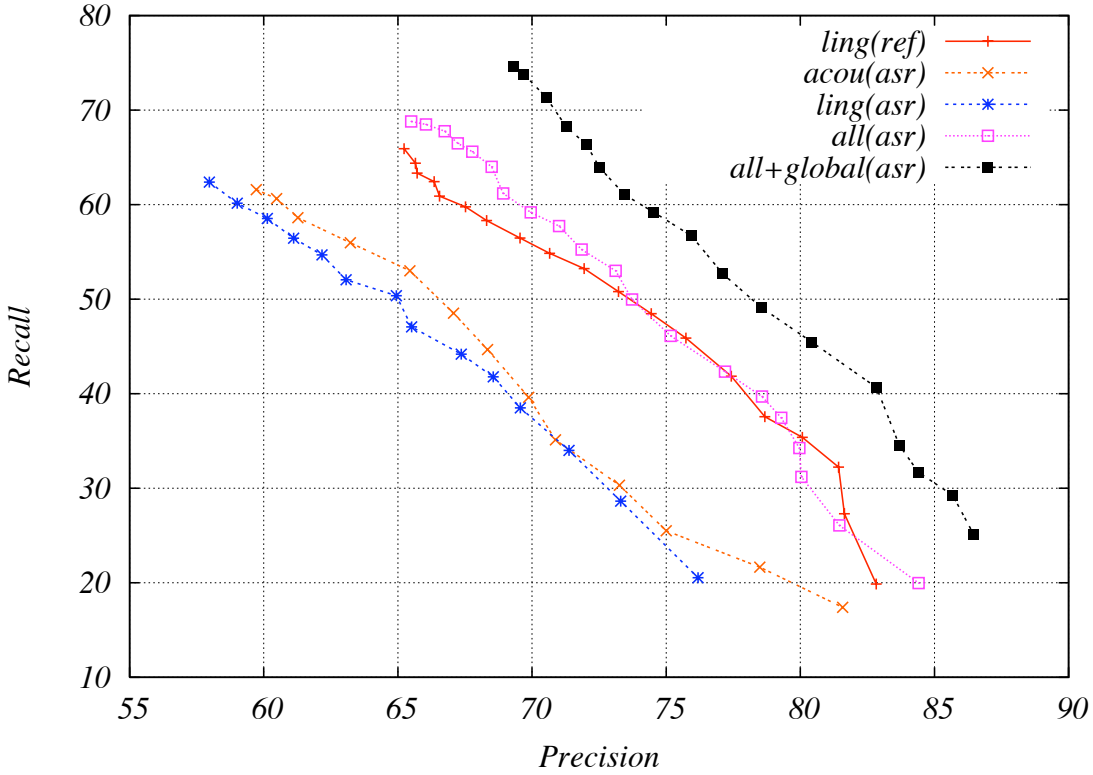

FIG. 5.4 – Detection performance of high spontaneous segments according to a varying threshold on the classification score

tures extracted from ASR outputs obtains a better precision and a better recall than the linguistic features extracted from the reference transcriptions alone. Moreover, using a probabilistic contextual tag-sequence model to globalize the classification process allows a better 74.6% precision in the detection of high spontaneous speech with a 69.3% recall measure, and 83.5% of high spontaneous speech detection errors are due to confusion between low and high spontaneous speech.

By applying a threshold on the scores obtained during the classification process, the high spontaneous speech detection precision can reach 85%, but with a recall equals to 25%. Although the classification task of labeling speech segments according to the spontaneity level is hard — even for human annotators — much progress has been made in the automatic detection since our previous work (6), mainly due to an improvement of the ASR system, but especially due to the addition of the probabilistic contextual tag model allowing a more global classification process.

This spontaneous speech detection provides very useful piece of information which can be used by various applications : speech recognition, for example, by developing specific methods for minimizing the word error rate on this kind of speech ; but also, for example, by providing additional information to automatic structuration or classification of collections of audio documents in large audio database.

5.3 Transcription automatique de la parole conversationnelle

5.3.1 Relevé, classement et analyse des principales erreurs des systèmes de reconnaissance automatique

Si l'on sait aujourd'hui que les systèmes de reconnaissance automatique sont moins performants sur la parole que l'on appelle spontanée, il n'en reste pas moins que la parole « préparée » est elle aussi source d'erreurs, bien que celles-ci soient en nombre nettement inférieur, et surtout appartiennent à des catégories bien précises. Nous allons donc tenter de proposer, exemples à l'appui, un classement et une analyse des erreurs commises par le système de reconnaissance LIUM RT.

Homonymes / paronymes

La principale difficulté éprouvée par le système de reconnaissance automatique est de traiter les phénomènes d ?homonymie et de paronymie. Ceux-ci sont particulièrement importants en français, où les monosyllabes homophones sont beaucoup plus nombreux que dans d'autres langues, et où la combinaison de syncopes et d'assimilations produit une morphologie liée particulièrement ambiguë. Concernant les homonymes, nombreux sont en effet les cas où la suite de phonèmes perçue par le système est la bonne, mais sans la transcription orthographique idoine. En voici quelques exemples : « là je viens d'ouvrir » : l'âge vient d'ouvrir « affirment elles avoir interpellé » : affirmaient l'avoir interpellé « proches hein » : prochain « chevauchement de compétence » : chevauchent Mende compétences « sont là statiques i(l)s bougent pas » : sont lasse Tati qui bougent pas Comme en atteste ce relevé, la distinction parole préparée / parole spontanée n'est pas forcément la cause des erreurs du

système : dans les exemples 2 et 9, l'élision du « l » appartient certes au domaine du spontané, et il est à peu près certain que la prononciation du phonème correspondant aurait évité les confusions qui résultent de son élision. Néanmoins, si l'on considère l'exemple 3, qui est issu d'un flash d'informations, aucune altération phonologique n'apparaît, ce qui n'empêche pas le système de proposer une séquence, acoustiquement exacte, mais sémantiquement erronée. Ce type d'erreurs, difficilement évitable à l'heure actuelle, est donc susceptible d'apparaître quel que soit le contexte langagier dans lequel on se trouve. Il est cependant indéniable que la langue française elle-même joue un rôle important dans l'apparition de ces confusions : contrairement à ses homologues anglaise, allemande ou espagnole, elle est d'une très grande richesse homonymique, allant des monosyllabes (foi/fois/foie/Foix ; lait/les/lais/laie...) aux vers holorimes (Gal, amant de la reine, alla tour magnanime / galamment de l'arène à la tour Magne à Nîmes). Cette singularité, qui passerait volontiers pour un charmant idiotisme, devient dans le domaine de la reconnaissance automatique de la parole un insoluble casse-tête... Par rapport aux autres langues latines notamment, elle repose sur le fait que le français a opéré au cours de son histoire une réduction syllabique massive, qui aboutit à un nombre d'autant plus considérable de monosyllabes homophones qu'on inclut les formes fléchies et la morphologie liée. Le tableau 12 met par exemple en regard les différentes graphies de la séquence [tã] et leurs traductions en italien, en l'occurrence toutes fondées sur les mêmes étymons latins. A cela s'ajoute, notamment pour les verbes, que le français marque la personne non plus à droite du verbe par une désinence (comme le latin ou l'italien) mais à gauche du verbe, par un « pronom » susceptible d'être modifié et disjoint (eux, qui parlent, sont...), représenté (moi qui ai/a), ou même réduit (je suis, tu es, vous êtes deviennent [Sshi ou Shi, te ou tE, zEt])

« e » ouvert / « e » fermé

Ensuite, et c ?est là sans doute le n ?ud du problème, il existe en français une confusion parfois totale entre le « e » ouvert et le « e » fermé. Théoriquement, la phonétique voudrait par exemple que la forme verbale « j'ai » se prononçât [ZE], puisque composée de la séquence « ai ». Toutefois, nombreux sont les cas dans lesquels le son produit est un « e » fermé, ce qui donne la séquence [Ze] (j'ai mis / gémir). Cette ambivalence est notamment très délicate à gérer pour les formes de l'imparfait, parfois presque impossibles à distinguer de celles du passé composé ou de l'infinitif (l'enfant aimait sauter dans l'eau / l'enfant aimé sautait dans l'eau). De même, entre autres mots outils monosyllabiques, déterminants et pronoms sont systématiquement sources de confusions (je l'ai / geler, les faits / l'effet, des faits / défaire...). Enfin, cette ambivalence induit des erreurs de structure. L'ambiguïté phonologique transforme la morphologie et fait dérailler la syntaxe : « j ?ai été » : j ?étais « le papa c'est une » : le pas passé une « c ?est cool » : s ?écoule « traîner » : Trénet « vous demandez » : vous demandait De même, il arrive parfois que LIUM RT assimile certaines séquences sonores à des suites de lettres, toujours phonétiquement identiques ou très proches : « et ça » : SA « et euh » : et E « j ?ai j ?ai » : g g « c'était » : CT Inversement, il se peut que le système ne reconnaisse pas un sigle et le transcrive sous forme de mots : « MSA » : mais ça Notre corpus le montre : ces erreurs ne sont pas toutes dues à l'emploi de la parole spontanée, et

bon nombre d'entre elles proviennent d'extraits contenant de la parole préparée, ou s'y appliqueraient volontiers.

Assimilations

Cela dit, il est effectivement des spécificités de la parole spontanée qui sont source d'erreur d'interprétation du logiciel de reconnaissance automatique, et notamment l'assimilation. Cette variation phonétique, entraînant la modification de la prononciation d'une consonne sourde au contact d'une consonne voisine sonore (ou l'inverse), est d'autant plus fréquente dans la parole spontanée qu'elle est très souvent provoquée par la disparition d'un schwa, caractéristique récurrente de ce type de discours. Et le système, souvent peu entraîné à ce genre de phénomène, ne sait pas toujours déduire le mot ou la séquences de mots exacts à partir de sa prononciation « assimilée », d'où un nombre important d'erreurs potentielles, tant les possibilités d'interférence entre consonnes sont nombreuses. La plus fréquente est certainement celle confrontant le « d » (qui est une consonne sonore) à une consonne sourde, dans la séquence « de + nom ou verbe », où le « e » est élidé, contraignant ainsi le son « d » à devenir « t ». Nous avons eu l'occasion d'en relever plusieurs occurrences lors de nos expériences : « envie d(e) passer » : vite passé « pas d(e) sanitaires » : patte sanitaire « coup d(e) fil » : coûte fils Dans chacun de ces trois exemples, le système LIUM RT, ne percevant ni la consonne « d » (puisqu'elle est prononcée « t ») ni la voyelle « e » (puisqu'elle est élidée), est incapable de générer la structure prépositionnelle introduite par « de ». En lieu et place de celle-ci, il propose donc une suite de mots rigoureusement exacte phonétiquement, mais incohérente contextuellement, comme il le faisait pour les autres cas d'homonymies que nous avons vus précédemment.

Répétitions, faux départs, troncations

Par ailleurs, outre l'assimilation, d'autres spécificités de la parole spontanée posent régulièrement problème aux systèmes de reconnaissance automatique : les répétitions, faux départs, troncations ou autres disfluences sont autant d' « anomalies » langagières qu'ils n'ont pas l'habitude de rencontrer. Pour les premières citées, il est intéressant de constater que LIUM RT s ?est même, en de rares occasions, refusé à proposer deux occurrences consécutives du même mot, bien que la prononciation ne laissait planer aucune ambiguïté : « faut faut faut faut » : faut fois font fois Au sujet des troncations ou des faux départs, ils génèrent inévitablement de nouvelles alternatives homonymiques. Et à nouveau, le système de reconnaissance automatique se retrouve à traiter des suites de sons qu'il va chercher à associer à des mots qui lui sont connus, et jamais à des amorces ou fins de mots, particularités qu'on ne retrouve (presque) que dans la parole spontanée. Ce qui ne manque pas, à nouveau, de créer de nouvelles confusions : « bah s() » : basses « on a des r() » : on adhère « (en)fin » : fin

Autres

Enfin, nous mentionnerons pour terminer quelques problèmes généraux, que nous avons rencontrés dans une majorité de fichiers. Tout d'abord, et cela concerne surtout la parole spontanée, la parole superposée n'est pas correctement traitée. Naturellement, les enregistrements utilisés étant monophoniques, cette tâche est d'autant plus difficile, voire impossible à réaliser. Il n'en reste pas moins que la superposition de locuteurs fait partie intégrante de la parole spontanée, et que réussir à la traiter serait une avancée considérable dans le domaine de la reconnaissance automatique de la parole. Des mots relativement brefs comme « et » ou « ou » échappent assez régulièrement à la vigilance de LIUM RT, ce qui s'explique précisément par leur brièveté, et par le fait qu'ils soient souvent « aspirés » par les mots qui les précèdent ou les suivent. Enfin, il arrive fréquemment qu'une inspiration soit interprétée par le système de reconnaissance automatique comme une occurrence de la conjonction de subordination « que ».

Chapitre 6

Traduction automatique

Chapitre 7

Campagnes d'évaluation

Deuxième partie

Administration de la recherche et encadrement

Chapitre 8

Projets de recherche

8.1 Le projet Parole du LIUM

8.2 Coordination du projet ANR EPAC

8.3 Responsabilité scientifique au sein du LIUM pour le projet ANR PORT-MEDIA

8.4 Projets qui débutent

8.4.1 Coordination du projet ANR ASH

8.4.2 Le projet ANR COSMAT

8.4.3 Le projet européen EuroMatrixPlus

Chapitre 9

Encadrement de jeunes chercheurs

Chapitre 10

Conclusion et perspectives

Bibliographie

[1] Thierry Bazillon, Yannick Estève, and Daniel Luzzati, "Transcription manuelle vs assistée de la parole préparée et spontanée," in *JEP*, Avignon, France, 2008.

[2] Thierry Bazillon, Yannick Estève, and Daniel Luzzati, "Manual vs assisted transcription of prepared and spontaneous speech," in *LREC*, Marrakech, Maroc, 2008.

[3] Julie Mauclair, Yannick Estève, and Paul Deléglise, "Automatic detection of well recognized words in automatic speech transcription," in *LREC*, Gênes, Italie, 2006.

[4] Martine Garnier-Rizet, Gilles Adda, Frederik Cailliau, Sylvie Guillemin-Lanne, Claire Waast-Richard, Lori Lamel, Stephan Vanni, and Claire Waast-Richard, "Manual vs assisted transcription of prepared and spontaneous speech," in *LREC*, Marrakech, Maroc, 2008.

[5] Vincent Jousse, Yannick Estève, Frédéric Béchet, Thierry Bazillon, and Georges Linarès, "Caractérisation et détection de parole spontanée dans de larges collections de documents audio," in *JEP*, Avignon, France, 2008.

[6] Richard Dufour, Vincent Jousse, Yannick Estève, Frédéric Béchet, and Georges Linarès, "Spontaneous speech characterization and detection in large audio database," in *13-th International Conference on Speech and Computer - SPECOM*, Saint-Pétersbourg, Russie, 2009.

[7] Yannick Estève Paul Deléglise and Bruno Jacob, "Systèmes de transcription automatique de la parole et logiciels libres," *Traitement Automatique des Langues*, vol. 45, no. 2, 2004.

[8] Kai-Fu Lee, Hsiao-Wuen Hon, and Raj Reddy, "An Overwiew of the SPHINX Speech Recognition System," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, Janvier 1990.

[9] Xuedong Huang, Fileno Alleva, Hsiao wuen Hon, Mei yuh Hwang, and Ronald Rosenfeld, "The SPHINX-II Speech Recognition System : An Overview," *Computer, Speech and Language*, vol. 7, pp. 137–148, 1992.

[10] Mosur Ravishankar, Rita Singh, Bhiksha Raj, and Richard M. Stern, "The 1999 CMU 10x real time broadcast news transcription system," in *Proc. DARPA workshop on Automatic Transcription of Broadcast News*, 2000.

[11] Arthur Chan, Mosur Ravishankar, and Alex Rudnicky, "On improvements of CI-based GMM selection," in *Interspeech*, Lisbonne, Portugal, 2005.

[12] Wille Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel, "Sphinx-4 : A flexible open source framework for speech recognition," Tech. Rep. TR-2004-139l, Sun Microsystems Laboratories, Novembre 2004.

[13] David Huggins-daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alex I. Rudnicky, "Pocketsphinx : A free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of ICASSP*, 2006.

[14] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the cmu-cambridge toolkit," in *Eurospeech*, Rhodes, Grèce, 1997, vol. 1, pp. 2707–2710.

[15] Andreas Stolcke, "SRILM - An extensible language modeling toolkit," in *Proceedings of ICASSP*, Denver, Colorado, USA, 2002.

[16] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Interspeech*, Brighton, Royaume-Uni, Septembre 2009.

[17] Alexandre Allauzen and Jean-Luc Gauvain, "Construction automatique du vocabulaire d un système de transcription," in *Journée d'Étude sur la Parole*, Fès, Maroc, 2004.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, " Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B,*, vol. 39, no. 1, pp. 1–38, Novembre 1977.

[19] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR : A survey of the literature," *Speech Communication*, vol. 29, pp. 225–246, 2001.

[20] G. Pérennou and M. D. Calmès, "BDLEX lexical data and knowledge base of spoken and written French," in *European Conference on Speech Technology*, Edinbourg, Ecosse, 1987.

[21] Frédéric Béchet, "LIA_PHON, un système complet de phonétisation de texte," *Traitement automatique des langues*, vol. 42, pp. 47–68, 2001.

[22] Antoine Laurent, Teva Merlin, Sylvain Meignier, Yannick Estève, and Paul Deléglise, "Combined systems for automatic phonetic transcription of proper nouns," in *LREC*, Marrakech, Maroc, 2008.

[23] Antoine Laurent, Teva Merlin, Sylvain Meignier, Yannick Estève, and Paul Deléglise, "Iterative filtering of phonetic transcriptions of proper nouns," in *ICASSP*, Taïpei, Taiwan, 2009.

[24] Antoine Laurent, Paul Deléglise, and Sylvain Meignier, "Grapheme to phoneme conversion using an SMT system," in *Interspeech*, Brighton, United Kingdom, Septembre 2009.

[25] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[26] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[27] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker Adaptive Training : A Maximum Likelihood Approach to Speaker Normalization," in *ICASSP '97 : Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 2*, Washington, DC, USA, 1997, p. 1043, IEEE Computer Society.

[28] D. Povey and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP '02 : Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Floride, USA, 2002, vol. 1, pp. 105–108.

[29] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep., Cambridge University Engineering Department, Mai 1997.

[30] R. Kneser and H. Ney, "Improved backing-off for n-gram language modeling," in *ICASSP'95 : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995, pp. 181–184.

[31] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Tech. Rep. TR-10-98, Center for Research in Computing Technology (Harvard University), août 1998.

[32] H. Mangu, E. Brill, and Stolcke A., "Finding consensus in speech recognition : Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[33] Paul Deléglise, Yannick Estève, Sylvain Meignier, and Teva Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ?," in *Interspeech*, Brighton, Royaume-Uni, 2009.

[34] Paul Deléglise, Yannick Estève, Sylvain Meignier, and Teva Merlin, "The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news," in *Interspeech*, Lisbonne, Portugal, 2005.

[35] Julie Mauclair, *Mesures de confiance en traitement automatique de la parole et applications*, Ph.D. thesis, Université du Maine, 2006.

[36] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit word error minimization in N-best list rescoring," in *Eurospeech*, Rhodes, Grèce, septembre 1997.

[37] F. Wessel, K. Macherey, and H. Ney, "A comparison of word graph and N-best list based confidence measures," in *Eurospeech*, Budapest, Hongrie, septembre 1999, pp. 315–318.

[38] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Eurospeech*, Rhodes, Grèce, septembre 1997.

[39] F. Wessel, K. Macherey, and R. Schlüter, "Using word probabilities as confidence measures," in *ICASSP*, Seattle, USA, mai 1998, pp. 225–228.

[40] G. Evermann and P.C. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination," in *Proc. Speech Transcription Workshop*, College Park, 2000.

[41] D. Falavigna, R. Gretter, and G. Riccardi, "Acoustic and word-lattice based algorithm for confidence scores," in *ICSLP*, Denver, USA, septembre 2002.

[42] S. Young, "Detecting misrecognitions and out-of-vocabulary words," in *ICASSP*, Adélaïde, Australie, avril 1994.

[43] Christian Raymond, Frédéric Béchet, Renato De Mori, Géraldine Damnati, and Yannick Estève, "Automatic learning of interpretation strategies for spoken dialogue systems," in *ICASSP*, Montréal, Canada, 2004, pp. 929–932.

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

[45] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Morgan Kaufmann Publishers*, 1996.

[46] C. Uhrik and W. Ward, "Confidence metrics based on n-gram language model backoff behaviors," in *Eurospeech*, Rhodes, Grèce, septembre 1997.

[47] M. Siu and H. Gish, "Evaluation of word confidence for speech recognition systems," *Computer Speech and Language*, vol. 13, no. 4, pp. 299–319, octobre 1999.

[48] Julie Mauclair, Yannick Estève, and Paul Deléglise, "Probabilité a posteriori : amélioration d'une mesure de confiance en reconnaissance de la parole," in *JEP*, Dinard, France, 2006.

[49] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[50] J. Simonin, L. Delphin-Poulat, and G. Damnati, "Gaussian Density Tree Structure in a Multi-Gaussian HMM-Based Speech Recognition System," in *ICSLP*, 1998.

[51] R. Prasad, S. Matsoukas, C.-L. Kao, J.Z. Ma, D.-X. Xu, T. Colthurst, O. Kimball, R. Schwartz, J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, , and F. Lefevre, "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System," in *Interspeech*, Lisbonne, Protugal, 2005.

[52] B. Hoffmeister, T. Klein, R. Schluter, and H. Ney, "Frame based system combination and a comparison with weigthed rover and cnc," in *Interspeech*, Pittsburgh, PA, USA, 2006, pp. 537–540.

[53] J.M Fiscus, "A post processing system to yield reduced word error rates : Recognizer Output Voting Error Reduction (ROVER)," in *IEEE ASRU Workshop*, 1997, pp. 347–352.

[54] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, "The LIA ?s French broadcast news transcription system," in *SWIM : Lecture Masters in Speech Processing*, Maui, Hawaii, USA, 2004.

[55] Benjamin Lecouteux, Georges Linares, J.F. Bonastre, and Pascal Nocera, "Imperfect transcript driven speech recognition," in *Interspeech*, Pittburgh, PA, USA, 2006.

[56] Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Julie Mauclair, "System combination by driven decoding," in *ICASSP*, Honolulu, Hawaii, USA, 2007.

[57] Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Guillaume Gravier, "Generalized driven decoding for speech recognition system combination," in *ICASSP*, Las Vegas, Nevada, USA, 2008.

[58] Guillaume Gravier, Stéphane Huet, and Pascale Sébillot, "Morphosyntactic processing of N-best lists for improved recognition and con ?dence measure computation," in *Interspeech*, Anvers, Belgique, 2007.

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

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

[61] S. E. Tranter, "Who really spoke when ? Finding speaker turns and identities in broadcast news audio," in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, vol. 1, pp. 1013–1016.

[62] J. Mauclair, S. Meignier, and Y. Estève, "Speaker diarization : about whom the speaker is talking ?," in *IEEE Odyssey 2006*, San Juan, Puerto Rico, USA, June 2006.

[63] M. Chengyuan, Patrick Nguyen, and Milind Mahajan, "Finding speaker identities with a conditional maximum entropy model," in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, April 2007.

[64] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news," in *LREC, Language Evaluation and Resources Conference*, Genoa, Italy, May 2006.

[65] V. Jousse, C. Jacquin, S. Meignier, Y. Estève, and B. Daille, "étude pour l'amélioration d'un système d'identification nommée du locuteur," in *JEP*, Avignon, France, June 2008.

[66] L. Canseco-Rodriguez, *Speaker Diarization in Broadcast News*, Ph.D. thesis, Ecole Doctorale Sciences et Technologies de l'Information des Télécommunications et des Systèmes, Université Paris XI, July 2006.

[67] R. Kuhn and R. De Mori, "The application of semantic classification trees to natural language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 449–460, 1995.

[68] Yannick Estève, Syvlain Meignier, Paul Deléglise, and Julie Mauclair, "Extracting true speaker identities from transcriptions," in *Proc. of Interspeech, European Conference on Speech Communication and Technology*, Antwerp, Belgium, Sept. 2007.

[69] E. El Khoury, S. Meignier, and C. Sénac, "Segmentation et regroupement en locuteurs pour la parole conversationnelle," in *JEP*, Avignon, France, June 2008.

[70] C. Barras, X. Zhu, S. Meignier, and J.L. Gauvain, "Multi-stage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1505–1512, September 2006.

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

[72] N Fourour, *Identification et catégorisation automatiques des entités nommées dans les textes français*, Ph.D. thesis, Thèse en informatique de l'université de Nantes, 2004.

[73] T Grass, "Typologie et traductibilité des noms propres de l'allemand vers le français," in *Traitement automatique des langues*, 2000, vol. 41(3), pp. 643–670.

[74] F. Bechet, A. Nasr, and F. Genet, "Tagging unknown proper names using decision trees," in *ACL, 38th Annual Meeting of the Association for Computational Linguistics*, Hong-Kong, China, Oct. 2000, pp. 77–84.

[75] G. Gravier, J.-F. Bonastre, S. Galliano, and E. Geoffrois, "The ESTER evaluation campaign of rich transcription of french broadcast news," in *LREC, Language Evaluation and Resources Conference*, Lisbon, Portugal, May 2004.

[76] Jean-Luc Gauvain, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Veronique Gendner, Lori Lamel, and Holger Schwenk, "Where are we in transcribing french broadcast news ?," in *Proc. of Interspeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[77] D. Hakkani-Tur and G. Tur, "Statistical Sentence Extraction for Information Distillation," *ICASSP 2007*, vol. 4, 2007.

[78] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection," *InterSpeech 2005*, 2005.

[79] M. Lease, Johnson M., and E. Charniak, "Recognizing Disfluencies in Conversational Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1566–1573, 2006.

[80] P.B. de Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda, and P. Paroubek, "A quantitative study of disfluencies in French broadcast interviews," *Proceeding of the workshop Disfluency In Spontaneous Speech (DISS), Aix-en-Provence, France*, 2005.

[81] D. Luzzati, "Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané," in *MIDL*, Paris, France, 2004.

[82] Cohen J., "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[83] Barbara Di Eugenio and Michael Glass, "The Kappa statistic : A second look," *Computational Linguistics*, vol. 30, no. 1, pp. 95–101, 2004.

[84] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[85] J.-F. Yeh and C.-H. Wu, "Edit Disfluencies Detection and Correction Using a Cleanup Language Model and an Alignment Model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1574–1583, 2006.

[86] E. Shriberg, "Phonetic consequences of speech disfluency," *Proceedings of the International Congress of Phonetic Sciences (ICPhS-99)*, pp. 619–622, 1999.

[87] G. Caelen-Haumont, "Perlocutory Values and Functions of Melisms in Spontaneous Dialogue," *Proceedings of the 1st International Conference on Speech Prosody, SP*, pp. 195–198, 2002.

[88] M.H. Siu and M. Ostendorf, "Modeling disfluencies in conversational speech," *ICSLP 1996*, vol. 1, 1996.

[89] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.d. dissertation, Department of Engineering, University of Cambridge, United Kingdom, 2004.

[90] J. Durand, B. Laks, and C. Lyche, "La phonologie du français contemporain : usages, variétés et structure," *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*, pp. 93–106, 2002.

[91] Robert E. Schapire and Yoram Singer, "BoosTexter : A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000.

[92] Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 69–88, 2002.