Université du Maine

TRANSCRIPTION AUTOMATIQUE DE LA PAROLE SPONTANÉE

THÈSE

présentée et soutenue publiquement le 1 décembre 2010

pour l'obtention du

Doctorat de l'Université du Maine

(spécialité informatique)

par

RICHARD DUFOUR

Composition du jury

Présidente :	Mme Martine Adda-Decker	Directrice de Recherche	LPP, CNRS, Paris 3
Rapporteurs :	M. Guillaume Gravier	Chargé de Recherche	IRISA, CNRS, Rennes 1
	M. Denis Jouvet	Directeur de Recherche	LORIA, INRIA, Nancy
Examinateurs :	M. Paul Deléglise	Professeur	LIUM, Université du Maine
	M. Yannick Estève	Professeur	LIUM, Université du Maine

Laboratoire d'Informatique de l'Université du Maine

Remerciements

Cette page met un point final à la rédaction de ce manuscrit. Conscient de l'attente qu'elle suscite, et afin d'éviter tout oubli, je vais tout d'abord commencer par remercier toute personne ayant participé de près ou de loin à ce travail de thèse. Sachez que sans ces interactions, ce manuscrit n'aurait pu voir le jour. Ceci étant fait, je vais maintenant pouvoir sereinement détailler ces remerciements.

Je voudrais remercier en premier lieu Yannick Estève et Paul Deléglise pour m'avoir encadré pendant ces trois années. Sans votre participation active, votre patience, vos nombreux conseils ainsi que vos remarques éclairées, ce travail n'aurait pu voir le jour. Vous avez toujours su trouver les mots pour me remotiver après (quelques) "échecs", ponctués quand même de "succès" qui m'ont amenés à présenter mes travaux de recherche aux quatre coins du globe. Un encadrement de cette qualité est une chance, et je souhaite à tout doctorant de pouvoir travailler dans ces conditions.

Je tiens également à exprimer mes remerciements aux membres du jury pour le temps passé à relire et annoter ce manuscrit. Je remercie Guillaume Gravier, Chargé de Recherche à l'Irisa, Université de Rennes 1, ainsi que Denis Jouvet, Directeur de Recherche au Loria, Université de Nancy, pour avoir accepté d'être les rapporteurs de cette thèse. Je remercie également Martine Adda-Decker, Directrice de Recherche au LPP, Université de Paris 3, pour m'avoir fait l'honneur de présider ce jury de thèse. Vos différentes remarques ainsi que vos conseils judicieux m'ont permis d'améliorer ce manuscrit. Les discussions que nous avons eues au cours de la soutenance de thèse me permettront d'approfondir mes travaux de recherche.

Je remercie tout particulièrement Monika Woszczyna pour m'avoir permis de travailler quelques mois au sein de l'entreprise M*Modal à Pittsburgh. Ce séjour de recherche restera pour moi une des expériences les plus enrichissantes, tant au niveau professionnel qu'au niveau humain. Je souhaite également remercier mes collègues de bureau Elisa Vettier, Rebeca Dosal, Melody Hauber, Ehsan Variani et Dominic Telaar, pour cette agréable ambiance de travail et ce plaisir à travailler au quotidien. Je tiens à remercier Thomas Schaaf, Werakul Laoworakiat, Shahid Durrani, Matthew Flint et Mark Fuhs pour leur accueil dans l'équipe, ainsi que pour l'aide apportée. Enfin, je souhaite terminer cette parenthèse "made in USA" en remerciant la famille Vettier, dont la générosité dépasse largement le cadre de l'entraide americano-française. Merci de m'avoir consacré autant de votre temps lors des diverses sorties que nous avons eues, de m'avoir aidé à maintes et maintes reprises, et de m'avoir accueilli comme un membre de votre famille.

Je tiens à exprimer mes remerciements envers l'ensemble du personnel du LIUM. Pour leur sympathie et leur aide précieuse, je remercie Sylvain Meignier, Teva Merlin, Bruno Jacob ainsi qu'Étienne Micoulaut. Je n'oublie pas non plus Martine Turmeau, secrétaire du laboratoire, qui m'a toujours aiguillé dans les méandres de l'administration. Je remercie également Benoît Favre, Mickaël Rouvier, Frédéric Béchet et Georges Linarès pour avoir eu le privilège de travailler avec eux et pour avoir passé d'excellents moments pendant les diverses conférences auxquelles nous avons assistées.

Parmi les nombreux points positifs apportés par cette thèse, je ne pourrais pas ne pas citer mes collègues de bureau qui m'ont accompagné depuis le début de cette aventure : Thierry Bazillon, Vincent Jousse et Antoine Laurent. Merci de m'avoir supporté au quotidien, ce qui, je le conçois, n'est pas une mince affaire. J'ai, pour ma part, vécu de très grands moments avec vous, et ai pris plaisir à venir travailler tous les jours. Merci d'avoir constitué le meilleur remède anti-déprime. Même si nos chemins se séparent, cette amitié perdurera au delà de cette thèse. Je remercie également Fethi Bougares pour avoir complété l'équipe après le départ "Plein Sud" de Thierry. Je terminerai cette partie en remerciant Gaël Salaün, ancien Ingénieur d'Étude du laboratoire, pour tous les excellents moments partagés ensemble (et qui, d'ailleurs, devraient continuer).

Je remercie ma famille de m'avoir toujours apporté son soutien et ses encouragements. Merci d'avoir fait le déplacement pour assister à ma soutenance de thèse, qui, je l'espère, vous aura donné un aperçu du travail réalisé. Je mesure donc la chance que j'ai, et vous remercie pour tout ce que vous m'avez apporté, m'apportez, et m'apporterez au quotidien.

Enfin, je conclurai en remerciant Élodie pour avoir été présente ces dernières années et avoir participé très activement à la relecture de ce manuscrit. Ce travail a parfois demandé quelques sacrifices, que tu as tout le temps accepté.

Merci à tous et bonne lecture.

Table des matières

Table des figures **ix**

Liste des tableaux **xi**

Acronymes **1**

Introduction **3**

1 Le projet ANR EPAC 5
2 Problématique 6
3 Structure du document 7

Partie I Contexte de travail et état de l'art **9**

Chapitre 1
Reconnaissance de la parole **11**

1.1 Principe de base 13
1.2 Extraction de paramètres 15
1.3 Modèles acoustiques 15
1.3.1 Modèles de Markov Cachés 15
1.3.2 Apprentissage 17
1.3.2.1 Techniques 17

1.3.2.2 Dictionnaire de phonétisation 17
1.3.2.3 Alignement phonème/signal 18
1.3.3 Adaptation 19
1.3.3.1 Méthode MLLR 19
1.3.3.2 Adaptation SAT-CMLLR 20
1.3.3.3 Méthode MAP 20
1.4 Modèle de langage 21
1.4.1 Modèle n-gramme 21
1.4.2 Estimation des probabilités 22
1.4.3 Lissage 22
1.4.4 Évaluation du modèle de langage 23
1.4.5 Mesures de confiance 24
1.4.5.1 Théorie 24
1.4.5.2 Évaluation des mesures de confiance 24
1.4.6 Évaluation des systèmes de RAP 25
1.5 Système du LIUM 25
1.5.1 Apprentissage 26
1.5.1.1 Données d'apprentissage 26
1.5.1.2 Vocabulaire 28
1.5.1.3 Modèles acoustiques 29
1.5.1.4 Modèles de langage 30
1.5.2 Transcription 30
1.5.2.1 Système de segmentation et de regroupement en locuteurs 31
1.5.2.2 Système de transcription multi-passes 31
1.6 Campagnes d'évaluation ESTER 1 et 2 32

Chapitre 2
Traitement de la parole spontanée 35

2.1 Spécificités de la parole spontanée 36
2.1.1 Les disfluences 37
2.1.1.1 Les pauses 37
2.1.1.2 Les troncations, répétitions et faux-départs 38
2.1.1.3 L'élision 39
2.1.1.4 Les hésitations 40

2.1.2 Autres phénomènes 40
2.1.2.1 Agrammaticalité 40
2.1.2.2 L'intonation 41
2.1.2.3 Le débit de parole et l'état émotionnel du locuteur 41
2.2 Gestion des disfluences 42
2.2.1 Objectifs 42
2.2.2 Détection automatique 44
2.2.3 Correction automatique 48
2.3 Impacts et solutions pour la reconnaissance de la parole 51
2.3.1 Modélisation acoustique 51
2.3.2 Modélisation linguistique 53
2.3.3 Dictionnaire de prononciations 56
2.3.3.1 Approche guidée par les données 57
2.3.3.2 Approche à base de connaissances 59
2.4 Conclusion 60

Chapitre 3
Homophonie **63**

3.1 Description générale 64
3.1.1 Mots homophones sémantiquement différents 65
3.1.2 Mots homophones sémantiquement identiques 66
3.2 Systèmes de RAP et homophonie 67
3.2.1 Quelques particularités du français 67
3.2.2 Analyse des erreurs d'homophonie 68
3.3 Méthodes automatiques appliquées aux erreurs de reconnaissance des systèmes de RAP 70
3.3.1 Approches statistiques globales 70
3.3.1.1 Détection automatique des erreurs 70
3.3.1.2 Correction automatique des erreurs 72
3.3.2 Approches ciblées sur les homophones 74
3.3.2.1 Approches par règles linguistiques 74
3.3.2.2 Approches statistiques 77
3.3.3 Combinaison des approches 80
3.4 Conclusion 80

Partie II Contributions **83**

Chapitre 4
Étude comparative de la parole préparée et spontanée en français **85**

4.1 Caractérisation de la parole spontanée 87
4.1.1 Étiquettes et classes de spontanéité 87
4.1.2 Impact du degré de spontanéité 89
4.1.3 Extraction de caractéristiques de la parole spontanée 89
4.1.3.1 Caractéristiques prosodiques 90
4.1.3.2 Caractéristiques linguistiques 91
4.1.3.3 Mesures de confiance 92
4.2 Apprentissage automatique : le *Boosting* 93
4.2.1 Principe général 93
4.2.2 L'algorithme *AdaBoost* 93
4.3 Approche proposée 95
4.4 Détection automatique des segments de parole spontanée 95
4.4.1 Classification au niveau du segment 96
4.4.2 Décision globale au moyen d'un modèle probabiliste 98
4.4.2.1 Présentation du modèle 98
4.4.2.2 Résolution de l'équation 98
4.5 Expériences 101
4.5.1 Données expérimentales 101
4.5.1.1 Corpus 101
4.5.1.2 Performances du système de RAP 102
4.5.1.3 Détection et catégorisation automatiques des segments de parole 103
4.5.2 Conclusion 108

Chapitre 5
Modélisation spécifique de la parole spontanée pour la reconnaissance de la parole **111**

5.1 Dictionnaire et variantes de prononciation 114
5.1.1 Analyse de variantes de prononciation spécifiques à la parole spontanée 114
5.1.2 Construction du nouveau dictionnaire de prononciations 116
5.1.3 Expériences 116
5.1.4 Résultats 117
5.1.5 Analyse des erreurs 118
5.1.5.1 Au niveau de variantes de prononciation 118
5.1.5.2 Au niveau du type de parole 119
5.1.5.3 Au niveau du segment 120
5.2 Adaptation des systèmes de RAP 122
5.2.1 Principe général 122
5.2.1.1 Adaptation non-supervisée des modèles acoustiques et de langage 122
5.2.1.2 Combinaison des systèmes 123
5.2.2 Adaptation automatique des modèles 125
5.2.2.1 Modélisation acoustique 125
5.2.2.2 Modélisation linguistique 125
5.2.3 Corpus 126
5.2.4 Expériences 127
5.2.4.1 Analyse du système adapté 128
5.2.4.2 Combinaison des systèmes 129
5.2.5 Conclusion 130
5.3 Approches spécifiques : le cas de l'homophonie en français 131
5.3.1 Approche proposée 131
5.3.1.1 Méthodologie générale 131
5.3.1.2 Règle grammaticale 133
5.3.1.3 Méthode statistique 134
5.3.2 Expériences réalisées 137
5.3.2.1 Mots et classes de mots étudiés 137

5.3.2.2 Outils 137
5.3.2.3 Données expérimentales 138
5.3.3 Résultats obtenus 139
5.3.3.1 Avec les règles grammaticales 139
5.3.3.2 Avec la méthode statistique 140
5.3.4 Conclusion 146
5.4 Résultats finaux des méthodes spécifiques 147
5.5 Perspectives 148

Conclusion et perspectives **151**

1 Détecteur de la parole spontanée 152
2 Modélisation spécifique des systèmes de RAP à la parole spontanée 153
2.1 Apprentissage non-supervisé des modèles acoustiques et linguistiques 153
2.2 Combinaison des systèmes 153
3 Correction d'erreurs spécifiques d'homophonie 153
4 Perspectives 154

Bibliographie personnelle **157**

Bibliographie **161**

Résumé **174**

Table des figures

1.1 Schématisation du fonctionnement d'un système de RAP. 14
1.2 Exemple d'un MMC à 5 états. 16
1.3 Architecture générale du système de transcription automatique du LIUM, extrait de [Estève 2009]. 27

3.1 Exemple de formes fléchies homophones 68

4.1 Approche générale pour la détection de la classe de spontanéité de chaque segment de parole. 96
4.2 Approche générale pour le processus de décision globale dans l'attribution de la classe de spontanéité des segments de parole. 97
4.3 Exemple général d'une machine à états-finis. 99
4.4 Transducteur modélisant toutes les probabilités contextuelles de $P(s_i|s_{i-1}, s_{i+1})$ 100
4.5 Performances sur la détection des segments de parole *fortement spontanée* en fonction du seuil choisi sur le score de classification. 107

5.1 Exemple de mauvaise reconnaissance du mot "*elle*" par le système de RAP dans un flux de parole spontanée. 114
5.2 Taux d'erreur-mot au voisinage d'un mot erroné (toutes les erreurs *Global* comparées aux erreurs de mots Hors-Vocabulaire *HV*) sur la parole préparée et spontanée. 121
5.3 Principe d'adaptation des modèles d'un système de RAP pour la parole spontanée en utilisant les données d'apprentissage existantes. 124
5.4 Adaptation automatique des modèles acoustiques à la parole spontanée. 126
5.5 Méthodologie générale proposée pour la correction de mots homophones en sortie des systèmes de RAP. 132
5.6 Exemple d'application de notre approche utilisant une règle grammaticale pour corriger le mot "*vingt*". 134
5.7 Exemple de formes fléchies homophones d'un participe passé. 135
5.8 Approche générale de la méthode statistique traitant les accords en genre et en nombre des participes passés et adjectifs homophones. 138

Liste des tableaux

1.1 Répartition des mots dans le corpus d'apprentissage en fonction de la source du sous-corpus. 28
1.2 Taux d'erreur-mot obtenus par le système de RAP du LIUM durant les campagnes d'évaluation ESTER 1 et ESTER 2, sur les corpus de développement et de test. 34

4.1 Classes de spontanéité définies à partir du protocole d'annotation des segments de parole. 88
4.2 Comparaison des valeurs moyennes (en secondes) des caractéristiques acoustiques pour chaque classe de spontanéité. 90
4.3 Comparaison des variances moyennes des caractéristiques acoustiques pour chaque classe de spontanéité. 91
4.4 Comparaison des valeurs moyennes (proportion par segment) des caractéristiques linguistiques pour chaque classe de spontanéité sur les transcriptions de référence. 92
4.5 Comparaison des valeurs moyennes (proportion par segment) des caractéristiques linguistiques pour chaque classe de spontanéité sur les transcriptions automatiques. 92
4.6 Comparaison des valeurs moyennes des moyennes et variances des mesures de confiance selon la classe de spontanéité. 93
4.7 Performances du système de RAP en fonction de la classe de parole en termes de WER et de NCE. Le nombre de segments et la durée liés à la classe de parole sont également inclus. 102
4.8 Précision et rappel de la classification des segments de parole en fonction des trois classes de spontanéité et des caractéristiques extraites. 103
4.9 Précision et rappel de la classification des segments de parole en fonction des trois classes de spontanéité en appliquant un modèle global sur les résultats déjà obtenus pour chaque segment. 105
4.10 Matrice de confusion sur la classification des segments de parole en classe de spontanéité avec *all(rap)*. 106
4.11 Matrice de confusion sur la classification des segments de parole en classe de spontanéité avec *global(rap)*. 106

5.1 Comparaison des prononciations en parole préparée et en parole spontanée de certains mots fréquents en français. 115

5.2 Nombre et taux d'erreurs (substitutions et suppressions) pour les mots concernés par l'ajout de variantes de prononciation. 117
5.3 Nombre et taux d'erreurs globaux (substitutions et suppressions) avant et après l'ajout de nouvelles variantes de prononciation. 117
5.4 Taux d'erreur-mot globaux (suppressions, substitutions et insertions) avant et après l'ajout de nouvelles variantes de prononciation. 118
5.5 Analyse (nombre et proportion) des occurrences de mots correctement transcrites au moyen du nouveau dictionnaire par rapport au dictionnaire de base. . . 119
5.6 Analyse (nombre et proportion) des occurrences de mots correctement transcrites au moyen du dictionnaire de base par rapport au dictionnaire modifié. . . 119
5.7 Nombre et taux des mots correctement reconnus sachant que la variante de prononciation associée pendant le décodage diffère de celle choisie pendant le processus d'alignement. 120
5.8 Durée (en heures) des données d'apprentissage pour les modèles acoustiques généraux (*Global*) et leurs adaptations pour la parole spontanée (*Extrait*) en fonction de la bande de fréquence et du sexe du locuteur. 127
5.9 Taux d'erreur-mot sur les segments de parole *fortement spontanée*, obtenus sur le corpus de développement, avec le système de base (*Base*) et le système adapté à la parole spontanée (*Sponta*) selon la paire bande de fréquence/sexe du locuteur. 128
5.10 Comparaison des taux d'erreur-mot du système de RAP avec le système de base (*Base*), la méthode ROVER (*Sponta* $\oplus$ *Base*), et en calculant le score oracle (*Oracle*). 129
5.11 Taille (en mots) des données d'apprentissage, de développement et de test utilisées par le classifieur statistique sur les participes passés et adjectifs. 139
5.12 Proportion des erreurs d'accord avant (Baseline) et après correction (Correction) sur les mots "cent" et "vingt" en utilisant la règle linguistique. 140
5.13 Rappel, précision, taux d'erreurs introduites et taux de correction sur les participes passés homophones, en comparant la version de base avec la version enrichie de la méthode statistique. 141
5.14 Rappel, précision, taux d'erreurs introduites et taux de correction sur les différentes étapes de la méthode statistique lors de la correction des erreurs d'accords des participes passés sur le corpus de test. 142
5.15 Rappel, précision, taux d'erreurs introduites et taux de correction sur les adjectifs au moyen de la méthode statistique. 143
5.16 Nombre et taux d'erreurs dues à des erreurs d'homophones, par rapport au nombre total d'erreurs, sur les participes passés et les adjectifs dans le corpus de développement et de test. 143
5.17 Comparaison des résultats de correction des adjectifs et participes passés sur le corpus de test du système du LIUM, en utilisant les informations acoustiques précises, des informations acoustiques approximatives, ou aucune information acoustique. 144
5.18 Comparaison des taux de correction (*n/a* si indisponible) sur les adjectifs et les participes passés, en utilisant la méthode statistique sur quatre transcriptions en sortie de systèmes de RAP. 145

5.19 Taux d'erreur-mot avant (*Baseline*) et après correction (*Correction*) sur quatre sorties de systèmes de RAP. 146
5.20 Comparaison des taux d'erreur-mot du système de RAP en utilisant le système de base (*Base*), la méthode spécifique de la parole spontanée (*Sponta*), puis sa combinaison avec la méthode spécifique de correction d'homophones (*Sponta+Homoph*). 148

Acronymes

ANR	Agence Nationale de la Recherche
CMLLR	Constrained Maximum Likelihood Linear Regression
EM	Expectation-Maximisation
EPAC	Exploration de masse de documents audio pour l'extraction et le traitement de la PArole Conversationnelle
ESTER	Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophonique
FSM	Machine à états-finis (*Finite-State Machine*)
LPC	Linear Predictive Coding
MAP	Maximum A Posteriori
MFCC	Mel-scale Frequency Cepstral Coefficients
MMC	Modèles de Markov Cachés
MMIE	Maximum Mutual Information Estimation
MPE	Minimum Phone Error
NIST	National Institute of Standards and Technology
NCE	Entropie Croisée Normalisée (*Normalized Cross Entropy*)
PLP	Perceptual Linear Prediction
RAP	Reconnaissance Automatique de la Parole
SAT	Speaker Adaptive Training
TAL	Traitement Automatique des Langues
WER	Taux d'Erreur-Mot (*Word Error Rate*)

Introduction

Les objectifs liés aux systèmes de Reconnaissance Automatique de la Parole (RAP) ont largement évolué au cours du temps. En effet, les premiers systèmes ne cherchaient qu'à transcrire des mots isolés d'un langage, dont le vocabulaire était strictement délimité. Les attentes, aujourd'hui, sont bien plus grandes, puisque les systèmes de RAP doivent fournir des transcriptions textuelles de fichiers audio provenant de sources hétérogènes, contenant un vocabulaire très large, non limité à une thématique précise, et dont les locuteurs ne sont pas connus à l'avance. La transcription d'émissions radiophoniques est notamment le type de tâche difficile auquel les systèmes de RAP doivent faire face (différentes conditions d'enregistrement, locuteurs variés et inconnus, grande variabilité des thèmes abordés...). Ces différentes problématiques supposent donc des réponses adaptées à chaque problème, et de nombreux efforts de recherche ont été réalisés au cours de ces dernières années pour proposer des solutions. Grâce à ces avancées, la précision actuelle des systèmes de RAP est telle que ceux-ci peuvent être intégrés dans différentes applications (transcriptions manuelles assistées, dialogue homme-machine, sous-titrage automatique...). Notons que ces applications n'existent que si les systèmes de RAP sont suffisamment performants.

Les transcriptions automatiques fournies par ces systèmes ne sont cependant pas parfaites, et certains problèmes n'ont toujours pas de solution satisfaisante. Selon le type d'émission transcrit, les performances des systèmes de RAP sont très variables. En effet, les systèmes arrivent à fournir des transcriptions avec un haut niveau de précision lorsque la parole est préparée (très proche d'une parole lue, comme dans des journaux d'information), mais voient leurs performances chuter fortement lorsque la parole est dite spontanée, apparaissant lors de conversations non préparées (dialogues, débats...). Des travaux de recherche se sont focalisés sur cette difficulté. Le projet "Spontaneous Speech : Corpus and Processing Technology", conduit de 1999 à 2004, a permis d'obtenir un corpus japonais, appelé *Corpus of Spontaneous Japanese* (CSJ), contenant des enregistrements audio (650 heures) et leurs transcriptions manuelles (7 millions de mots). Le corpus est composé principalement de présentations orales spontanées. Il a aidé à réaliser des études à grande échelle et à fournir des solutions spécifiques sur ce type de parole. La récente campagne d'évaluation *Rich Transcription Fall 2004*[1], organisée par le NIST (National Institute of Standards and Technology), a notamment mis en lumière le fait qu'une importante chute au niveau des résultats des systèmes de RAP est visible lorsque ceux-ci devaient transcrire de la parole spontanée.

Ces différentes campagnes et projets permettent de proposer de nouvelles approches pour traiter des problèmes spécifiques à la reconnaissance de la parole. En France, le projet EPAC[2]

[1] Cette campagne intégrait de nombreuses tâches de transcription de la parole dans le contexte d'émissions d'information ainsi que des conversations téléphoniques dans différentes langues.

[2] `http://epac.univ-lemans.fr/`

[Estève 2010] (Exploration de masse de documents audio pour l'extraction et le traitement de la PArole Conversationnelle), de janvier 2007 à août 2010, poursuit ce travail sur la parole spontanée. L'objectif principal de ce projet est d'améliorer les systèmes de RAP sur ce type de parole, et de fournir des informations supplémentaires au niveau des transcriptions automatiques en sortie de ces systèmes (nommer les locuteurs, définir le genre de l'émission, réaliser un découpage syntaxique des segments...).

Le travail de cette thèse s'incrit dans le cadre du projet EPAC. Les travaux présentés dans ce mémoire sont directement liés à la problématique de ce projet. De plus, le projet EPAC a fourni les ressources nécessaires (corpus spécifique à la parole spontanée, études linguistiques...) à la mise en place et à la validation des solutions proposées. La partie suivante permet de mieux comprendre les objectifs et les attentes de ce projet.

1 Le projet ANR EPAC

Le projet EPAC, financé par l'ANR[3] (Agence Nationale de la Recherche), concerne le traitement de données audio non structurées. Il met en scène quatre laboratoires académiques durant 44 mois :

- l'Institut de Recherche en Informatique de Toulouse (IRIT),
- le Laboratoire d'Informatique de Tours (LI),
- le Laboratoire d'Informatique d'Avignon (LIA),
- le Laboratoire d'Informatique de l'Université du Maine (LIUM).

Le projet EPAC propose des méthodes d'extraction d'information et de structuration de documents spécifiques aux données audio, en prenant en compte l'ensemble des canaux d'information : segmentation du signal (parole / musique / jingle...), identification et suivi du locuteur, transcription de parole, détection et suivi de thèmes, détection d'émotions, analyse du discours, interactions conversationnelles... Ces tâches de traitement du signal et de la parole sont en grande partie maîtrisées par les différents partenaires du projet dont la plupart ont participé aux campagnes d'évaluation ESTER (voir section 1.6).

En particulier, ce projet met l'accent sur le traitement de la parole spontanée. Parmi les émissions radiophoniques ou télévisuelles d'information, la parole spontanée est souvent marginale : des techniques de détection et d'extraction de ce type de parole particulier doivent être proposées et développées. Le projet propose des méthodes de traitement de la parole spontanée en proposant des descripteurs pertinents et en développant les outils nécessaires à leur exploitation. Cette partie fut l'objet d'une collaboration entre chercheurs en linguistique (LI, LIUM) et chercheurs en traitement automatique de la parole (IRIT, LIA, LIUM). Enfin, un cadre

[3] `http://www.agence-nationale-recherche.fr`

d'évaluation commun aux différents acteurs du projet a été mis en place pour chacune des tâches étudiées. Cette évaluation portait sur une partie des 2 000 heures d'émissions radiophoniques disponibles, contribuant ainsi à la valorisation de ces données.

Le corpus, qui est réalisé dans le cadre du projet EPAC, se compose de transcriptions manuelles de 100 heures d'enregistrement audio. Ces transcriptions ont été annotées en partie grâce à une transcription assistée[4], le reste ayant été fait entièrement manuellement. Les enregistrements audio proviennent des 1 500 heures d'audio brut diffusées aux participants de la campagne ESTER 1. Il s'agit d'émissions de France Info, France Culture et RFI diffusées entre 2003 et 2004. Finalement, les sorties automatiques produites par les différents outils des partenaires du projet EPAC pour l'ensemble des 1 500 heures d'audio brut de ESTER 1 viennent s'ajouter à ces transcriptions manuelles.

2 Problématique

Dans le cadre du projet EPAC, l'un des objectifs finaux a été d'améliorer les systèmes de RAP sur la parole spontanée. Différents outils et solutions ont alors été apportés. La baisse des performances, pour ce type de parole, peut s'expliquer par ses multiples particularités que nous verrons dans la partie 2.1. De plus, lorsque l'on traite d'émissions radiophoniques, différents styles de parole peuvent apparaître. Il est ainsi possible rencontrer de la parole proche d'un texte lu (type présentation d'un journal), ou, au contraire, de la parole plus spontanée (lors de débats ou d'interviews). L'objectif est de proposer des méthodes améliorant la reconnaissance de la parole sur la parole spontanée, sans dégrader les performances sur la parole préparée.

Différents objectifs sont alors attendus pour gérer la parole spontanée. Le premier objectif est de pouvoir fournir un détecteur automatique fiable de la parole spontanée. Le deuxième objectif est d'améliorer les systèmes de RAP sur ce type de parole. Le détecteur peut alors s'avérer très utile pour proposer des solutions spécifiques. Le dernier objectif de ce travail de thèse est de proposer une solution pour gérer le problème de l'homophonie, erreur récurrente dans les transcriptions automatiques fournies par les systèmes de RAP. Ce travail est en fait une extension de celui réalisé sur la parole spontanée, visant à fournir des solutions spécifiques pour traiter des problèmes particuliers de la parole.

[4]Les transcriptions dites *assistées* sont obtenues en deux étapes : une première transcription est obtenue automatiquement au moyen d'un système de RAP, puis une correction est opérée manuellement.

3 Structure du document

Dans la première partie, une présentation générale d'un système de reconnaissance automatique de la parole sera tout d'abord exposée, afin d'en comprendre les concepts principaux. Nous nous intéresserons également au système de transcription du LIUM, sur lequel s'appuient les différentes expériences menées durant ce travail de thèse. Nous ferons un point sur un état de l'art de ce qui a pu se faire en parole spontanée. Ses particularités et son impact sur la reconnaissance de la parole seront présentés. Nous exposerons ensuite les différentes solutions mises en œuvre d'une part pour détecter et corriger ces spécificités, et d'autre part pour améliorer les systèmes de RAP. Après cette présentation de la parole spontanée suivra une description du phénomène de l'homophonie, qui nous amènera à survoler les solutions déjà proposées pour corriger ces mots particuliers, notamment au niveau de la reconnaissance automatique de la parole.

Nous nous focaliserons ensuite sur les apports de cette thèse concernant la parole spontanée dans le cadre de la reconnaissance de la parole. Dans un premier temps, nous présenterons une étude générale que nous avons effectué sur ce type de parole, en comparaison avec la parole préparée. Puis nous verrons plus particulièrement le détecteur automatique de parole spontanée que nous avons réalisé en nous basant sur les particularités de la parole spontanée.

Enfin, nous exposerons les travaux spécifiques que nous avons menés afin d'amélioration les performances des systèmes de RAP. Nous verrons tout d'abord nos approches pour traiter de manière plus efficace la parole spontanée, que ce soit au niveau du dictionnaire de prononciations (ajout de prononciations spécifiques), ou au moyen d'une adaptation non-supervisée des modèles acoustiques et modèles de langage. Cette adaptation utilise notamment le détecteur de parole spontanée que nous proposons. Nous nous consacrerons, dans cette même partie, à la correction de certains mots homophones, avec l'apport d'une nouvelle méthode permettant de corriger des erreurs spécifiques en post-traitement des transcriptions obtenues au moyen d'un système de RAP.

Première partie

Contexte de travail et état de l’art

Chapitre 1

Reconnaissance de la parole

Sommaire

1.1 Principe de base 13
1.2 Extraction de paramètres 15
1.3 Modèles acoustiques 15
1.3.1 Modèles de Markov Cachés 15
1.3.2 Apprentissage 17
1.3.2.1 Techniques 17
1.3.2.2 Dictionnaire de phonétisation 17
1.3.2.3 Alignement phonème/signal 18
1.3.3 Adaptation 19
1.3.3.1 Méthode MLLR 19
1.3.3.2 Adaptation SAT-CMLLR 20
1.3.3.3 Méthode MAP 20
1.4 Modèle de langage 21
1.4.1 Modèle n-gramme 21
1.4.2 Estimation des probabilités 22
1.4.3 Lissage 22
1.4.4 Évaluation du modèle de langage 23
1.4.5 Mesures de confiance 24
1.4.5.1 Théorie 24
1.4.5.2 Évaluation des mesures de confiance 24
1.4.6 Évaluation des systèmes de RAP 25
1.5 Système du LIUM 25
1.5.1 Apprentissage 26

1.5.1.1 Données d'apprentissage 26
1.5.1.2 Vocabulaire 28
1.5.1.3 Modèles acoustiques 29
1.5.1.4 Modèles de langage 30
1.5.2 Transcription 30
1.5.2.1 Système de segmentation et de regroupement en locuteurs 31
1.5.2.2 Système de transcription multi-passes 31
1.6 Campagnes d'évaluation ESTER 1 et 2 32

Les avancées technologiques réalisées dans le domaine de la reconnaissance automatique de la parole permettent aujourd'hui de fournir des systèmes de reconnaissance de la parole (RAP) performants, le but étant d'obtenir une transcription parfaite de la parole. Les enjeux actuels sont multiples pour ce domaine, puisque les systèmes de RAP sont de plus en plus utilisés dans de nombreuses applications (dialogue homme-machine, identification du locuteur, résumé automatique, recherche d'information, indexation d'archives audio...).

Afin d'obtenir une transcription correcte, il faut gérer les spécificités de la parole. Une difficulté réside dans le fait que n'importe quelle personne peut intervenir dans le document audio à transcrire : le signal audio peut varier selon le sexe de la personne (homme/femme), son âge, son niveau d'élocution ou de stress, ou encore son accent. Selon le contexte ou le locuteur, la prononciation des mots peut considérablement changer. À ces difficultés peuvent s'ajouter les conditions d'enregistrement du signal audio (par exemple la qualité du microphone utilisé), et les bruits extérieurs à la parole (inspirations, musique, éternuements...) qui parasitent le signal. Enfin, le découpage de la parole en mots et en phrases n'est pas simple pour un système de RAP, car ce flux est continu et aucune frontière n'est clairement définie.

Dans ce chapitre, nous allons premièrement nous intéresser aux principes généraux d'un système de RAP, en décrivant toutes les étapes indispensables au système pour transformer un signal audio en une transcription, mais aussi la façon dont les systèmes statistiques gèrent les différentes contraintes énoncées précédemment. Dans cette optique, nous présenterons la théorie principale de ces systèmes statistiques, puis nous nous intéresserons à la manière dont peuvent être extraits les paramètres acoustiques. Nous nous pencherons ensuite sur les modèles acoustiques et linguistiques utilisés dans les systèmes de RAP. Une présentation du système de RAP développé au sein du LIUM sera finalement entreprise.

1.1 Principe de base

Les systèmes de reconnaissance de la parole (RAP) utilisent actuellement le formalisme décrit dans [Jelinek 1976], présentant une approche statistique se fondant sur la théorie de l'information. Un système de RAP a pour but d'associer une séquence de mots à une séquence d'observations acoustiques. Ainsi, à partir de la séquence d'observations acoustiques $X = x_1 x_2 \cdots x_n$, un système de RAP recherche la séquence de mots $\hat{W} = w_1 w_2 \cdots w_k$ qui maximise la probabilité $P(W|X)$, qui est la probabilité d'émission de W sachant X. La séquence de mots $\hat{W}$ se doit alors de maximiser l'équation

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

En appliquant la règle de Bayes, on obtient la formule

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

Comme la séquence d'observations acoustiques X est fixée, $P(X)$ peut être considéré comme une valeur constante inutile dans l'équation 1.2. On obtient alors

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \tag{1.3}$$

Deux types de modèles probabilistes sont utilisés pour la recherche de la séquence de mots la plus probable : un modèle acoustique qui fournit la valeur de $P(X|W)$, et un modèle de langage qui fournit la valeur de $P(W)$. $P(X|W)$ peut se concevoir comme la probabilité d'observer X lorsque W est prononcé, alors que $P(W)$ se réfère à la probabilité que W soit prononcé dans un langage donné. Pour obtenir un système de RAP performant, il est essentiel de définir les modèles les plus pertinents possibles pour le calcul de $P(W)$ et $P(X|W)$. La figure 1.1 présente une schématisation générale du fonctionnement d'un système de RAP.

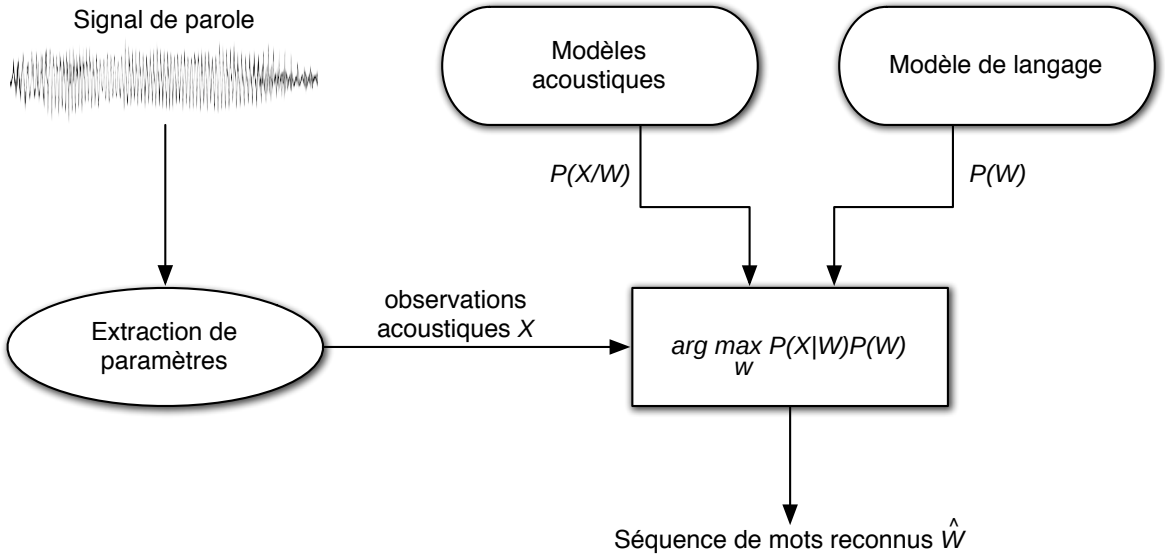

FIG. 1.1 – Schématisation du fonctionnement d'un système de RAP.

1.2 Extraction de paramètres

Comme nous avons pu le remarquer dans la figure 1.1, le signal de parole ne peut directement être transformé en hypothèses de séquences de mots. L'extraction de ses paramètres est une étape importante puisqu'elle doit déterminer les caractéristiques pertinentes du signal. Il est nécessaire de découper le signal audio par trame, en prenant généralement une taille fixe définie aux alentours de 25 ms, afin de rendre le signal quasi-stationnaire. Ce découpage est réalisé toutes les 10 ms. Un vecteur de paramètres est ensuite extrait pour chaque trame. Cette extraction peut se faire au moyen de multiples techniques, dont les plus connues sont l'analyse paramétrique, avec l'utilisation de la méthode LPC (*Linear Predictive Coding* [Atal 1971]), l'analyse cepstrale, avec par exemple la méthode MFCC (*Mel-scale Frequency Cepstral Coefficients* [Davis 1980]), ou encore la technique PLP (*Perceptual Linear Prediction* [Hermansky 1990]). Ces différentes méthodes permettent d'extraire des coefficients caractéristiques pour chaque trame. Cette extraction permet alors d'obtenir la séquence d'observations acoustiques X, où $X = x_1 x_2 \cdots x_n$ (x_i réprésentant une observation acoustique), *i.e.* un vecteur de paramètres associé à une trame.

1.3 Modèles acoustiques

1.3.1 Modèles de Markov Cachés

Les modèles acoustiques utilisés pour la reconnaissance de la parole sont, depuis des années, principalement basés sur les modèles de Markov cachés (MMC, connus en anglais sous le nom de HMM : *Hidden Markov Models*) [Jelinek 1976, Rabiner 1989]. Les MMC sont des automates probabilistes à états finis qui permettent de calculer la probabilité d'émission d'une séquence d'observations. Pour un système de RAP, les séquences d'observations sont les vecteurs de caractéristiques du signal de parole composés de coefficients MFCC ou PLP par exemple (voir section 1.2). Les MMC respectent l'hypothèse markovienne d'ordre 1 : la connaissance du passé se résume à celle du dernier état occupé. Pour capter certains comportements et évolutions du signal dans le temps, on intègre dans les vecteurs de caractéristiques du signal les dérivées premières et secondes des vecteurs de paramètres.

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

1. La parole est pondérée par une suite d'états stationnaires, représentés par des gaussiennes émettant des vecteurs (MFCC par exemple) et leurs dérivées premières et secondes.
2. L'émission d'une séquence de ces vecteurs est générée par un MMC respectant l'hypothèse markovienne d'ordre 1.

La figure 1.2 présente un exemple de MMC gauche-droit, avec saut d'état possible.

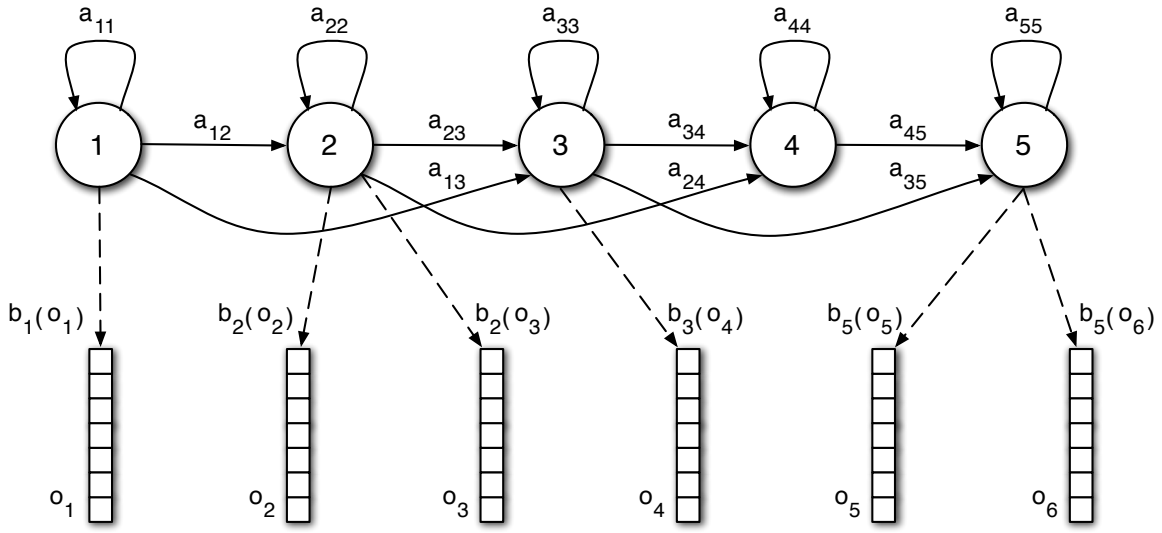

FIG. 1.2 – Exemple d'un MMC à 5 états.

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

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

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

De manière générale, l'unité de modélisation utilisée est le phonème. Ainsi la modélisation d'un mot s'effectue à partir de la concaténation des modèles de phonèmes qui composent ce mot. Pour tenir compte de la variabilité de prononciation d'un phonème, un MMC est construit pour un phonème donné, associé à un contexte gauche et un contexte droit particuliers. Un contexte gauche (resp. droit) d'un phonème est un phonème qui précède (resp. succède à) ce phonème. Ce triplet (contexte gauche, phonème, contexte droit) est appelé triphone, ou phonème en contexte. Pour affiner la modélisation d'un phonème en contexte, la position de ce phonème dans un mot (début, milieu, fin ou phonème isolé) est parfois prise en compte. Afin de

réduire la taille du modèle, une factorisation d'états similaires est effectuée : les états sont dits *partagés*.

1.3.2 Apprentissage

L'apprentissage des modèles acoustiques consiste à estimer les vecteurs de moyennes et les matrices de covariances d'un ensemble de gaussiennes, ainsi que les pondérations permettant d'établir des mélanges à partir de ces gaussiennes. Ces paramètres permettront de calculer des densités de probabilités qui constitueront des valeurs de vraisemblance associées à l'émission d'une observation par un état d'un MMC. À cela s'ajoute l'estimation des probabilités discrètes associées aux transitions entre les différents états des MMC. Nous nous intéresserons dans cette partie à deux techniques d'apprentissage des modèles acoustiques, puis nous verrons la manière dont peut être construit le lexique pour cette phase d'apprentissage. Dans la dernière sous-partie, nous décrirons la phase d'alignement des phonèmes sur le signal de parole.

1.3.2.1 Techniques

Algorithme d'Espérance-Maximisation (EM)

L'algorithme d'Espérance-Maximisation (EM, en anglais *Expectation-Maximization*), proposé par [Dempster 1977], a pour objectif de trouver les valeurs des paramètres qui maximisent la vraisemblance de modèles probabilistes *a posteriori* et ce, lorsque le modèle dépend de variables latentes non observables. Cet algorithme se décompose en deux étapes. La première étape consiste à évaluer l'espérance (E) de la vraisemblance, qui sera calculée par rapport aux variables observées. La seconde étape consiste à maximiser (M) la vraisemblance des paramètres en utilisant la vraisemblance déjà calculée dans l'étape précédente. Ces deux étapes sont vues comme une itération de l'algorithme EM : les paramètres obtenus en sortie d'une itération sont réutilisés en entrée de l'itération suivante, jusqu'à convergence.

Méthode MMIE

Alors que l'algorithme EM présenté précédemment vise à maximiser la vraisemblance des données, l'apprentissage discriminant MMIE [Normandin 1994] (*Maximum Mutual Information Estimation*) cherche à maximiser la probabilité *a priori* du modèle correspondant aux données. L'idée est d'extraire les caractéristiques essentielles de chaque modèle, afin de pouvoir le différencier par rapport aux autres modèles.

1.3.2.2 Dictionnaire de phonétisation

Le dictionnaire de phonétisation, qui peut également être appelé *dictionnaire de prononciations* dans la littérature, est utilisé lors de l'apprentissage des modèles acoustiques. En effet,

puisque le système acoustique est basé sur les phonèmes, il est nécessaire d'associer à chaque entrée du lexique (que l'on peut communément appeler *mot*) une modélisation phonétique qui lui est propre. Cette modélisation est obtenue par concaténation des MMC de phonèmes (voir section 1.3.1). Comme un graphème [5] peut être associé à plusieurs phonèmes, il est nécessaire d'avoir toutes les séquences de phonèmes correspondantes à un mot dans le lexique.

Pour obtenir le dictionnaire de phonétisation, plusieurs approches sont possibles. La première approche envisagée est de créer le lexique manuellement. L'avantage de cette approche réside dans le fait que les prononciations possibles de chaque mot sont fiables, car vérifiées par un expert humain. Cependant, générer un lexique complet est très coûteux en ressources, et il est très difficile de couvrir la totalité des mots d'une langue. Des projets se sont notamment penchés sur la vérification manuelle de ces données, avec par exemple le projet *BDLex* [Pérennou 1987]. Une autre approche possible consiste à phonétiser les mots de manière automatique. Le système proposé par [Béchet 2001], propose l'utilisation d'une base de règles de phonétisation pour transcrire automatiquement les graphèmes en phonèmes. Généralement, ces deux approches sont utilisées conjointement, en utilisant les dictionnaires créés manuellement, et en les enrichissant des mots manquants au moyen d'une technique de phonétisation automatique.

Le choix de l'application conditionne la taille du lexique. Le lexique choisi doit couvrir tous les mots utilisés pendant la phase d'apprentissage. Notons également qu'un dictionnaire de prononciations, pouvant être différent de celui utilisé durant la phrase d'apprentissage, sera constitué pour la phase de décodage du système. Les données audio à transcrire peuvent alors contenir des mots hors-vocabulaire[6]

1.3.2.3 Alignement phonème/signal

Une fois la phonétisation des mots du vocabulaire terminée, il est possible de procéder à la phase d'alignement des phonèmes sur le signal. Généralement, cet alignement phonème/signal peut être obtenu en utilisant l'algorithme Viterbi [Viterbi 1967], ou en utilisant l'algorithme forward-backward [Baum 1972]. Pour obtenir des modèles acoustiques performants, il est nécessaire que la phonétisation de chaque transcription soit la plus proche possible de la prononciation effective de la phrase correspondante. Un problème se pose quand plusieurs prononciations sont envisageables pour une même entrée du dictionnaire : il faut pouvoir choisir la bonne prononciation, et il est impossible de vérifier humainement ce choix en écoutant le signal de

[5] Le graphème est défini comme l'écriture associée à un phonème. Le graphème peut être constitué d'une ou plusieurs lettres. Plusieurs graphèmes sont possibles pour un même phonème, et inversement.

[6] Mots devant être transcrits par le système de RAP, mais qui ne sont pas dans le lexique. Le système générera automatiquement une erreur, qui pourra se propager aux mots voisins.

parole lorsqu'il se mesure en dizaines d'heures de parole. Pour automatiser ce choix, une solution consiste à estimer grossièrement des premiers modèles acoustiques, puis de les utiliser pour définir la phonétisation utilisée. Cette méthode permet de forcer un alignement phonème/signal au moyen d'informations acoustiques et textuelles. Différentes techniques peuvent être mises en place pour en extraire les meilleures prononciations, comme par exemple [Estève 2004], qui choisit la phonétisation la plus courte lorsque plusieurs phonétisations pour un même mot sont possibles lors de l'estimation des premiers modèles acoustiques.

1.3.3 Adaptation

Les systèmes de RAP doivent faire face à de nombreuses contraintes liées au signal audio, comme par exemple devoir traiter un grand nombre de locuteurs différents, dans des conditions variables d'enregistrement. Le volume des données d'apprentissage ne peut couvrir tous les problèmes liés à ces conditions. Pour remédier à ces difficultés, des techniques d'adaptation ont été mises en place. L'objectif de ces techniques est de pouvoir gérer aux mieux les contraintes liées aux conditions, avec un volume de données beaucoup plus faible que ce qui aurait été nécessaire pour les méthodes d'apprentissage. L'idée est d'adapter les modèles déjà appris pour en créer des nouveaux beaucoup plus proches des conditions de test, en utilisant les modèles initiaux et un nombre restreint de nouvelles données. Ces techniques d'adaptation sont efficaces, par exemple, lorsque le système de RAP doit traiter un locuteur inconnu.

Des données extraites des données d'apprentissage permettront d'adapter les modèles existants en modèles spécialisés. Afin de pouvoir réaliser ce processus d'adaptation, une transcription des données audio doit être fournie. Ces méthodes d'adaptation sont alors dites *supervisées*. Il peut cependant arriver que la transcription de référence ne soit pas disponible : un décodage au moyen du système de RAP non-adapté permet alors d'obtenir une transcription. De nombreuses techniques existent, nous verrons succinctement les techniques les plus utilisées pour bien comprendre les enjeux de l'adaptation.

1.3.3.1 Méthode MLLR

La méthode MLLR [Leggetter 1995, Gales 1998] (*Maximum Likelihood Linear Regression*) est une technique d'adaptation des modèles acoustiques par régression linéaire, très efficace lorsque peu de données sont disponibles. Les transformations linéaires sont utilisées dans la procédure d'adaptation de modèles indépendants du locuteur [Anastasakos 1996]. Logiquement, les modèles dépendants d'un locuteur obtiennent de meilleurs résultats qu'un modèle indépendant du locuteur. La technique consiste alors à adapter le modèle indépendant du locuteur (appris sur un grand volume de données) à un locuteur précis, en essayant d'obtenir des

performances les plus proches possibles de celles obtenues au moyen d'un modèle dépendant du locuteur. Des transformations linéaires [Leggetter 1995] sont utilisées pour l'adaptation à un locuteur. Elles permettent d'estimer les transformations propres à chacun des locuteurs du corpus d'apprentissage ainsi que les paramètres des MMC. La grande force de cette méthode réside dans le fait que l'adaptation peut facilement être réalisée sur un nouveau locuteur (principalement dans le cas où le durée de parole d'un locuteur est courte). En effet, une ou plusieurs transformations sont réalisées au moyen de cette méthode, qui seront ensuite appliquées sur tous les modèles. Avec la méthode MLLR, les transformations sur les moyennes et variances des gaussiennes sont décorrelées les unes des autres.

1.3.3.2 Adaptation SAT-CMLLR

Au contraire de [Leggetter 1995], la méthode CMLLR [Digalakis 1995] (*Constrained Maximum Likelihood Linear Regression*) lie les transformations de la variance et de la moyenne.

$$v^{'} = Av - b \tag{1.4}$$

et

$$\Sigma^{'} = A\Sigma A^{'T} \tag{1.5}$$

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

Si des transformations linéaires identiques sont utilisées pour apprendre les modèles indépendants du locuteur, il est possible d'estimer conjointement les transformations propres à chacun des locuteurs de ce corpus et les paramètres de modèles markoviens. Les modèles qui en résultent sont ensuite plus facilement adaptables à un nouveau locuteur.

1.3.3.3 Méthode MAP

La méthode bayésienne d'estimation du Maximum *a posteriori* (MAP), introduite pour la reconnaissance de la parole dans [Gauvain 1994c], permet d'établir, dans l'apprentissage, des contraintes probabilistes sur les paramètres de modèles. Le critère MAP est appliqué aux modèles ayant fait l'objet d'un apprentissage préalable et pour lesquels on dispose de données *a priori*. Les modèles markoviens sont toujours estimés avec l'algorithme EM mais en maximisant la vraisemblance *a posteriori* (MAP) au lieu de la vraisemblance des données. Elle permet d'obtenir de nouveaux modèles en réduisant la variance des modèles initiaux, en

fournissant alors des modèles plus spécifiques grâce à l'utilisation d'un nombre restreint de données d'adaptation. La technique permet de modifier les paramètres acoustiques d'un modèle générique afin de le rapprocher des données de test, dans l'idée de créer, par exemple, des modèles spécifiques au sexe du locuteur, ou encore à des conditions acoustiques particulières. Au contraire de la méthode MLLR, la méthode MAP cherche à adapter chaque modèle au moyen de données spécifiques, la contrainte étant que la taille de ces données doit être assez importante pour permettre l'adaptation.

1.4 Modèle de langage

Le modèle de langage doit permettre, dans le système de RAP, de capturer les contraintes linguistiques, afin de guider le décodage acoustique en choisissant la suite de mots la plus adéquate. Il modélise ainsi les contraintes liées à une langue, et doit notamment pouvoir gérer les homophones hétérographes[7]. Les modèles de langage probabilistes attribuent une probabilité à une séquence de mots, qui s'exprime généralement par

$$P(W_1^k) = P(w_1)\prod_{i=2}^{k} P(w_i|w_1,\cdots,w_{i-1}) = P(w_1)\prod_{i=2}^{k} P(w_i|h_i) \tag{1.6}$$

où h_i est l'historique du mot w_i. On a : $h_i = w_1,\cdots,w_{i-1}$

1.4.1 Modèle n-gramme

Le modèle de type *n-gramme* est le modèle de langage probabiliste actuellement le plus utilisé en reconnaissance automatique de la parole. Pour ce genre de modèle, l'historique d'un mot est représenté par les $n-1$ mots qui le précèdent. Dans la pratique, il est à ce jour impossible de prendre en compte un historique trop large (limites techniques et manque de données d'apprentissage) : la valeur de n généralement choisie est de 3 ou 4. On parlera alors respectivement de modèles *trigrammes* et *quadrigrammes* (modèles *unigrammes* pour $n = 1$, et *bigrammes* pour $n = 2$). L'équation 1.6 devient alors

$$P(W_1^k) = P(w_1)\prod_{i=2}^{n-1} P(w_i|w_1,\cdots,w_{i-1})\prod_{i=n}^{k} P(w_i|w_{i-n+1},\cdots,w_{i-1}) \tag{1.7}$$

[7]Mots prononcés de la même manière mais ayant une graphie différente.

Même si ce genre de modèle semble particulièrement réducteur, en ne prenant en compte que des contraintes lexicales courtes, il contient suffisamment d'informations pour guider efficacement un système de RAP. Sa simplicité rend son emploi aisé lors de la phase de reconnaissance. De plus, l'utilisation d'un modèle de langage trigramme n'engendre pas un coût élevé en terme de calculs[8]. Enfin, une qualité fondamentale des modèles *n-grammes* est la couverture totale des phrases pouvant être exprimées dans un langage. Cette qualité est intéressante pour le traitement de la parole spontanée : l'utilisation de modèles probabilistes de type *n-gramme* permet de modéliser certains aspects du langage oral spontané incorrects d'un point de vue grammatical. Un modèle de langage à base de règles de grammaires formelles serait plus facilement mis en défaut dans ce type de situation. Bien entendu, il est évident que ces phénomènes typiques de la parole spontanée doivent être observés dans le corpus d'apprentissage pour être modélisés par le modèle *n-gramme*. En contrepartie, la précision des modèles *n-grammes* est limitée puisque ce type de modèle ne rejette aucune phrase, y compris celles n'appartenant pas au langage visé. Cependant, les scores affectés aux phrases composées de séquences de mots peu fréquentes (voire inexistantes) sont souvent pénalisés par rapport aux scores des phrases plus correctes. En effet, dans le corpus d'apprentissage du modèle de langage, il est plus probable de rencontrer les séquences de mots d'une phrase valide.

1.4.2 Estimation des probabilités

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

$$P_{MV}(w_i|h_i) = \frac{n(h_i, w_i)}{n(h_i)} \tag{1.8}$$

où $n(x)$ indique la fréquence d'apparition de x.

1.4.3 Lissage

Nous avons pu remarquer que les modèles de langage *n-grammes* dépendent des données textuelles collectées pour former le corpus d'apprentissage. Malheureusement, la quantité de

[8] Bien que l'utilisation d'un modèle de langage quadrigramme soit réalisable, ce temps de calcul est nettement plus élevé qu'un modèle trigramme.

données est en général insuffisante et certains *n-grammes* n'apparaissent jamais dans le corpus d'apprentissage. Il peut même arriver que certains mots du lexique soient absents du corpus d'apprentissage lorsque la construction de ce lexique n'exige pas leur présence.

Les techniques de lissage tentent de compenser cette carence : elles peuvent être vues comme une sorte de généralisation permettant d'attribuer une probabilité non nulle à un événement non vu dans le corpus d'apprentissage. Les principales techniques de lissage sont décrites dans [Chen 1996], où est également présentée une discussion sur leurs performances respectives.

Le lissage par repli (ou *back-off*), que l'on retrouve notamment dans les méthodes proposées par [Katz 1987] ou [Jelinek 1987], est un mécanisme permettant de pallier le manque de données d'apprentissage pour certains *n-grammes*. L'idée est alors d'utiliser une probabilité issue d'un ordre inférieur ($n-1$, $n-2\ldots$) lorsqu'aucune probabilité n'est disponible à l'ordre n pour un mot et un historique donnés. Pour chaque repli vers un ordre inférieur, la taille de l'historique est diminuée et les chances d'obtenir une probabilité estimée sur le corpus d'apprentissage augmente. En contrepartie, un coefficient de repli est habituellement associé à cette probabilité qui modifie la valeur finale de la probabilité proposée par le modèle pour le mot et l'historique donné. D'autres méthodes ont été proposées, comme le lissage *Witten Bell* [Witten 1991] ou encore le lissage *Kneser-Ney* [Kneser 1995].

1.4.4 Évaluation du modèle de langage

Pour être certain de la performance optimale d'un modèle de langage, la meilleure solution serait de le tester dans un système de RAP en conditions réelles d'utilisation. Évidemment, cette solution n'est pas toujours réalisable. Cette évaluation se fait alors le plus souvent en calculant la perplexité [Jelinek 1977] du modèle de langage, sur un texte non inclus dans le processus d'apprentissage du modèle. Plus la valeur obtenue est basse (l'idée étant de se rapprocher de 0), plus le modèle de langage possède des capacités de prédiction. La perplexité est calculée au moyen de la formule :

$$\log PP = -\frac{1}{n}\sum_{i=1}^{n} \log P(w_i|w_1, \cdots, w_{i-1}) \tag{1.9}$$

où $w_1, \cdots, w_{i-1}$ correspond à l'historique du mot w_i.

1.4.5 Mesures de confiance

1.4.5.1 Théorie

Les mesures de confiance sont des scores exprimant une estimation de la fiabilité des décisions prises par un système de RAP. Elles peuvent se retrouver dans de nombreuses applications, comme les systèmes de dialogue [Hazen 2000], en développant une stratégie différente selon les scores des mesures de confiance. Elles apparaissent également dans la reconnaissance automatique de la parole [Jiang 2005], présentant différentes utilisations et avancées des mesures de confiance en RAP, ou encore dans l'identification des langues [Metze 2000]. Dans le cadre de la reconnaissance automatique de la parole, une mesure de confiance CM associée à un mot w possède généralement un score compris dans l'intervalle $[0, 1]$. Plus la valeur de $CM(w)$ est proche de 0, plus le mot est potentiellement considéré comme erroné, et inversement, plus le score se rapproche de 1, plus le mot est considéré comme correct. Dans le cas idéal, $CM(w)$ vaut 0 ou 1. De manière théorique, le calcul de la moyenne des mesures de confiance de K mots ($K = w_1, \cdots, w_K$) respecte la formule

$$\mu(CM) = \frac{1}{K} \sum_{i=1}^{K} CM(w_i) \tag{1.10}$$

où $\mu(CM)$ doit être une approximation du taux de mots émis bien reconnus. Plus ce score se rapproche du taux d'erreur-mot du système de RAP, plus les mesures de confiance sont précises.

Les mesures de confiance sont estimées de multiples manières, les techniques les plus courantes pouvant être retrouvées dans [Mauclair 2006]. Les différentes méthodes proposées prennent en compte de nombreuses informations issues du système de RAP, aux niveaux acoustique et linguistique.

1.4.5.2 Évaluation des mesures de confiance

Dans la littérature, un grand nombre de métriques existe pour évaluer les mesures de confiance [Mauclair 2006]. La métrique de l'entropie croisée normalisée [Siu 1999] (*Normalized Cross Entropy* (NCE)) est parmi les plus utilisées. La métrique NCE a été employée lors des évaluations NIST afin d'évaluer la qualité des mesures de confiance. Elle permet de réaliser une estimation sur l'information supplémentaire apportée par la mesure de confiance à l'hypothèse fournie par le système de RAP. Plus la NCE se rapproche de 1, plus la mesure de confiance peut prédire si le mot est correct ou non. La métrique est définie par :

$$NCE = \frac{H_{max} + \sum\limits_{M corrects} \log_2(c(M)) + \sum\limits_{M incorrects} \log_2(1 - c(M))}{H_{max}} \tag{1.11}$$

où $Hmax = -n\log_2(p_c) - (N-n)\log_2(1-p_c)$,
n le nombre de mots correctement reconnus,
N le nombre total de mots reconnus,
p_c la probabilité moyenne qu'un mot reconnu soit correct (=n/N),
et enfin $c(M)$ la mesure de confiance associée au mot M.

1.4.6 Évaluation des systèmes de RAP

Afin de pouvoir évaluer plusieurs systèmes de reconnaissance de la parole (RAP), il convient de les comparer sur les mêmes données de test (par exemple les campagnes d'évaluation ESTER [Galliano 2005, Galliano 2009], voir partie 1.6). Classiquement, les systèmes de RAP sont évalués en termes de taux d'erreur-mot (*Word Error Rate*, WER). Le WER prend en compte les erreurs de :

- *Substitution* : mot reconnu à la place d'un mot de la transcription manuelle.
- *Insertion* : mot reconnu inséré par rapport à la transcription de référence.
- *Suppression* : mot de la référence oublié dans l'hypothèse fournie par le système de RAP.

Le WER s'exprime par la formule :

$$WER = \frac{nombre\ de\ substitutions + nombre\ d'insertions + nombre\ de\ suppressions}{nombre\ de\ mots\ de\ la\ référence} \tag{1.12}$$

Après cette présentation générale du fonctionnement d'un système de reconnaissance de la parole, la partie suivante présentera le système de RAP utilisé pour mener à bien les expériences présentées dans ce manuscrit.

1.5 Système du LIUM

Dans ce paragraphe, nous allons nous intéresser au système de RAP développé au sein du LIUM [Deléglise 2005, Deléglise 2009]. Ce système utilise comme base le décodeur *CMU*[9] *Sphinx* [Lee 1990], diffusé sous licence libre depuis 2001. Nous verrons que différentes modifications ont été apportées au décodeur *CMU Sphinx*, afin de rendre le système développé au sein du LIUM le plus performant possible. Deux versions du décodeur *Sphinx* sont utilisées pour réaliser le système du LIUM :

- *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 *CMU Sphinx*

[9] Carnegie Mellon University

[Ravishankar 2000]. Il utilise des modèles de Markov continus. Cette version a été développée en langage C.

– *Sphinx 4* : la version 4 de *Sphinx* est une réécriture complète d'un décodeur en Java décrit dans [Walker 2004]. L'objectif était de développer un décodeur au moins aussi performant que le décodeur *Sphinx 3*. Néanmoins, *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.

La figure 1.3 résume l'architecture générale du système de RAP du LIUM. La construction des ressources (l'apprentissage) et le processus de décodage y sont représentés, pendant que les bases de connaissance, créées pendant la phase d'apprentissage et utilisées lors de la transcription, y sont mises en évidence. Le système de RAP, présenté dans cette figure, 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 [Galliano 2009]. Nous reviendrons plus longuement sur les campagnes d'évaluation ESTER 1 et ESTER 2 dans la section 1.6.

Le système du LIUM 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.

1.5.1 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.

1.5.1.1 Données d'apprentissage

Les corpus, fournis par les organisateurs de la campagne ESTER 2, sont utilisés comme corpus d'apprentissage pour le système du LIUM. L'ensemble des participants a également eu accès à 40 heures d'enregistrements radiophoniques transcrits manuellement, et issus du projet EPAC (contenant principalement de la parole conversationnelle, voir section 1). De plus, le système intègre les données du corpus *French Gigaword Corpus*[10], qui regroupe un très grand nombre de dépêches AFP (*Agence France Presse*[11]) et APW (service français d'*Associated*

[10] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T17`
[11] `http://www.afp.com`

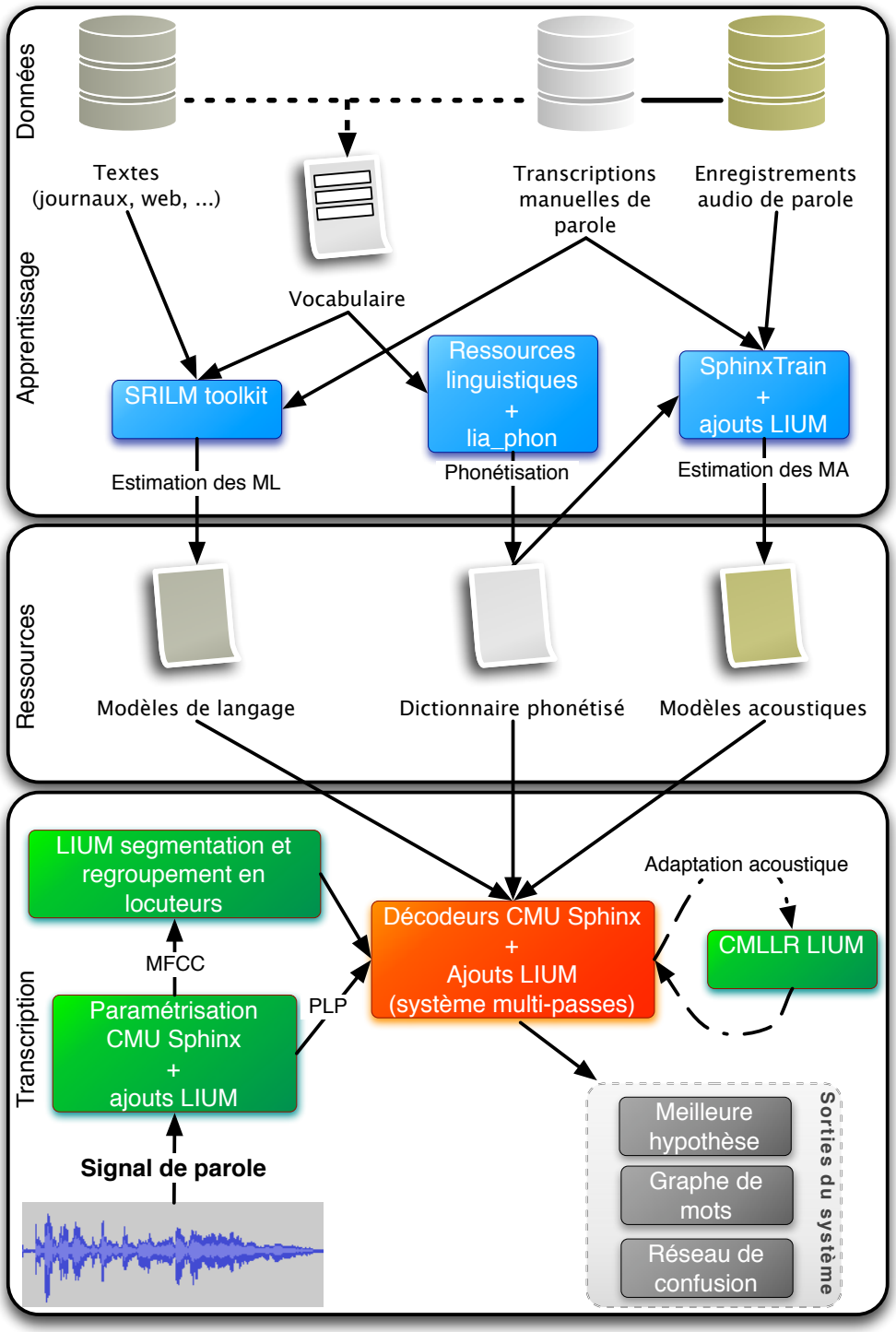

FIG. 1.3 – Architecture générale du système de transcription automatique du LIUM, extrait de [Estève 2009].

Press[12]) pour les années 1990 et 2000. Enfin, des données sur le web enrichissent les corpus d'apprentissage. Il s'agit 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, le système dispose de 240 heures d'enregistrements audio transcrits manuellement pour l'apprentissage des modèles acoustiques, ce qui correspond à environ 3,3 millions de mots pour estimer les modèles de langage. Le tableau 1.1 présente la répartition des mots du corpus d'apprentissage en fonction de leur origine. Nous constatons que la quantité de données spécifiques à la tâche visée, utilisée pour l'apprentissage des modèles de langage, représente moins de 0,4% de l'ensemble des données textuelles.

	Transcriptions manuelles d'émissions radiophoniques	**Presse écrite et dépêches**	**Archives de presse écrite (web)**
Nombre de mots	3,3M	1,0G	80M

TAB. 1.1 – Répartition des mots dans le corpus d'apprentissage en fonction de la source du sous-corpus.

1.5.1.2 Vocabulaire

La constitution du vocabulaire est une étape très importante et très délicate du développement d'un système de RAP (voir section 1.3.2.2). L'approche choisie pour construire le vocabulaire du système du LIUM suit celle proposée par [Allauzen 2004], consistant à :

1. estimer autant de modèles unigrammes que nous avons de sources d'apprentissage (tableau 1.1),
2. calculer les coefficients d'interpolation entre ces unigrammes, permettant de construire un modèle unigramme 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 (voir partie 1.3.2.1),
3. construire le modèle de langage unigramme avec les coefficients d'interpolation calculés lors de l'étape précédente,

[12]Principalement des informations financières : `www.ap.org/francais/`

4. extraire les N mots les plus probables du modèle unigramme, N étant la taille visée du vocabulaire.

La taille du vocabulaire a été fixée à 120 000 mots. Les taux de mots hors-vocabulaire sont de 0,66 % et 0,74 %, respectivement sur le corpus de développement et le corpus de test de la campagne ESTER 2.

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 [Strik 1999]. Le système développé pour le français utilise un jeu de 35 phonèmes. Le vocabulaire phonétisé suit ces étapes :

1. si le mot existait déjà dans le vocabulaire du système de RAP précédent, nous conservons la ou les phonétisations déjà utilisées,
2. sinon, si le mot existe dans BDLex, nous employons la ou les phonétisations proposées par BDLex,
3. si le mot n'apparaît dans aucune des étapes précédentes, une phonétisation automatique obtenue grâce à LIA_PHON [Béchet 2001] est utilisée.

1.5.1.3 Modèles acoustiques

Les modèles acoustiques employés par le système de RAP du LIUM, à base de modèles de Markov cachés, utilisent 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, exception faite des *fillers*, sont définis 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 et d'un descripteur de l'énergie, ainsi que des dérivées premières et secondes de ces descripteurs.

Le système dispose de différents ensembles de modèles acoustiques. Chacun de ces modèles est spécialisé en fonction du sexe du locuteur (*homme/femme*) et de la bande passante (*téléphone/studio*). Ces spécialisations ont été obtenues à partir d'une adaptation de type MAP (voir partie 1.3.3.3) au niveau des moyennes, covariances, et poids.

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

1. En première passe, les modèles sont composés de 6 500 états partagés. Chaque état étant modélisé par un mélange de 22 gaussiennes.

2. En seconde passe et par la suite, les modèles acoustiques sont composés de 7 500 états, toujours modélisés par une mixture de 22 gaussiennes. Ces modèles ont été estimés à partir d'un apprentissage de type SAT [Anastasakos 1997] (*Speaker Adaptive Training*) combiné à un apprentissage discriminant de type MPE [Povey 2002] (*Minimum Phone Error*). De plus, une matrice de transformation CMLLR [Digalakis 1995] a été calculée pour chaque locuteur et appliquée sur les paramètres acoustiques de chacun des locuteurs respectifs.

1.5.1.4 Modèles de langage

Les modèles de langage du système de RAP du LIUM sont, comme pour la très grande majorité des systèmes de RAP, des modèles de langage n-grammes. Le système utilise des modèles trigrammes dans les premières passes de décodage, et des modèles quadrigrammes pour les dernières passes.

L'estimation des modèles de langage n-grammes est faite au moyen de la technique de *discounting* dite de Kneser-Ney modifié [Kneser 1995] avec interpolation des n-grammes d'ordres inférieurs. Une autre des caractéristiques de ces modèles n-grammes est qu'aucun *cut-off* n'a été appliqué : tous les n-grammes 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. Il est aussi utilisé pour éliminer des coquilles ou des séquences de mots erronées (à cause d'une faute d'orthographe par exemple).

Les données d'apprentissage sont les données textuelles décrites dans la partie 1.5.1.1. Chaque corpus d'apprentissage a été utilisé pour estimer un modèle n-gramme. Ensuite, sur le corpus de développement approprié, les coefficients d'interpolation ont été optimisés avec l'algorithme EM afin de minimiser la mesure de perplexité du modèle interpolé sur ce corpus. Ces manipulations sont réalisées à l'aide du *SRILM toolkit*[13] [Stolcke 2002].

1.5.2 Transcription

La constitution des bases de connaissance (modèles acoustiques, dictionnaire phonétisé, modèles de langage) est primordiale pour obtenir de bonnes performances de reconnaissance de la parole. Leur construction est alors élaborée de façon minutieuse. Dans cette section, nous allons décrire la façon dont sont exploitées ces bases de connaissance dans le système de RAP du LIUM.

[13]The SRI Language Modeling Toolkit

1.5.2.1 Système de segmentation et de regroupement en locuteurs

Les transcriptions automatiques ont besoin de frontières précises au niveau des segments. Ainsi, les segments ne contenant pas, par exemple, de parole, doivent être supprimés afin de minimiser les insertions de mots. L'idée est de découper les segments, qui seront transcrits par un système de RAP, en fonction de zones ne contenant pas de parole (comme par exemple en détectant les pauses, voir section 2.1.1.1). Il apparaît que découper un segment au milieu d'un mot a pour conséquence d'augmenter le taux d'erreur-mot des systèmes de RAP [Deléglise 2005, Meignier 2010].

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

- Le signal est décomposé en petits segments homogènes.
- Les segments sont ensuite regroupés par locuteur sans changer les frontières.
- Dans la phase finale, les frontières sont ajustées.

En détail, le système de segmentation et de regroupement en locuteurs, initialement développé pour la campagne d'évaluation ESTER 1 [Deléglise 2005], est composé d'une segmentation acoustique, suivie d'une classification hiérarchique, toutes deux basées sur BIC, permettant d'obtenir différentes classes. Chaque classe représente un locuteur, et est modélisée avec une gaussienne à covariance complète. Un décodage Viterbi [Viterbi 1967] est réalisé pour ajuster les frontières des segments au moyen de mélanges de gaussiennes (*GMM*) de chaque classe. La musique et les interludes musicaux sont supprimés en décodant avec 8 GMM. Le sexe des locuteurs ainsi que les conditions d'enregistrement sont détectés avec 4 GMM spécialisées. Enfin, les segments de parole sont limités à 20 secondes en coupant les segments trop longs au moyen d'un détecteur de silences, utilisant des GMM.

Le système de segmentation et de regroupement en locuteurs développé au sein du LIUM est présenté en détails dans [Meignier 2010].

1.5.2.2 Système de transcription multi-passes

Sans inclure la phase de segmentation et de regroupement en locuteurs, le système de RAP du LIUM est un système multi-passes. La notion de passe est assez subjective et consiste, ici, en l'utilisation d'un algorithme de recherche manipulant les données de la passe précédente.

Chaque passe peut proposer une nouvelle hypothèse de reconnaissance. Le système se déroule en cinq passes :

1. La première passe consiste en un traitement utilisant la version 3.7 du décodeur rapide de *Sphinx 3* appliqué sur des paramètres acoustiques PLP ; cette passe applique un modèle de langage trigramme et des modèles acoustiques adaptés au sexe du locuteur (homme/femme) et aux conditions acoustiques (studio/téléphone). Elle sert également à obtenir des modèles acoustiques généraux, qui seront ensuite adaptés au moyen de la méthode CMLLR [Digalakis 1995].
2. La seconde passe utilise de nouveau la version 3.7 du décodeur rapide de *Sphinx 3* appliquant les mêmes paramètres acoustiques PLP. Cependant, une matrice de transformation CMLLR a été calculée de façon à adapter les paramètres acoustiques aux modèles acoustiques. Ces modèles ont été estimés au moyen des méthodes SAT [Anastasakos 1997] et MPE [Povey 2002]. Le modèle de langage trigramme est toujours appliqué.
3. La troisième passe permet de remédier aux approximations inter-mots faites par le décodeur lors du calcul des scores acoustiques des phonèmes[14]. En utilisant le graphe de mots généré lors de la seconde passe comme espace de recherche, il est possible de corriger ces imprécisions inter-mots grâce au vrai contexte droit des phonèmes en fin de mot. Ce sont les mêmes modèles acoustiques et linguistiques (modèle de langage trigramme) que lors de la seconde passe qui sont utilisés, avec l'application de la même matrice de transformation CMLLR sur les paramètres acoustiques.
4. La quatrième passe consiste à recalculer, à l'aide d'un modèle de langage quadrigramme, les scores linguistiques des mots du graphe de mots généré en passe 3.
5. Enfin, la dernière passe transforme le graphe de mots issu de la quatrième passe en un réseau de confusion. Une variante de la méthode de consensus [Mangu 2000] est alors appliquée, permettant d'obtenir l'hypothèse de reconnaissance finale, avec pour chaque mot des probabilités *a posteriori* utilisables comme mesures de confiance.

1.6 Campagnes d'évaluation ESTER 1 et 2

Les campagnes d'évaluation ESTER 1 [Galliano 2005] et 2 [Galliano 2009] (Évaluation des Systèmes de Transcription d'Émissions Radiophoniques) visent l'évaluation des performances des systèmes de transcription d'émissions d'information radiophoniques. Les transcriptions sont enrichies par un ensemble d'informations annexes comme le découpage automatique en

[14] L'approximation inter-mots permet un gain de temps, mais engendre une baisse des performances du système de RAP.

tours de parole, le marquage des entités nommées... La transcription enrichie vise à obtenir une transcription lisible et une représentation structurée du document à des fins d'extraction d'informations. Les émissions comportent des phases de parole lue et des phases de parole spontanée (interviews, débats, conversations téléphoniques).

Pour la campagne d'évaluation ESTER 1 [Galliano 2005], les données mises à la disposition des participants pour l'apprentissage et la mise au point des systèmes proviennent d'enregistrements sonores d'émissions radiophoniques francophones (90 heures d'émissions provenant de France Inter, France Info, Radio France Internationale (RFI), Radio Télévision Marocaine (RTM)), de transcriptions enrichies issues de ces 90 heures d'enregistrements, ainsi que des dictionnaires de phonétisation de mots en français. De plus, un corpus textuel correspondant aux années 1987 à 2003 du journal "Le Monde", augmenté du corpus MLCC (corpus multilingue et parallèle[15]) contenant des transcriptions des débats du Conseil Européen, est fourni. Enfin, cette campagne donne un corpus audio non transcrit d'environ 2 000 heures datant du dernier trimestre 2003 à septembre 2004. Ce corpus contient les mêmes radios que celles comprises dans les données transcrites, plus des enregistrements de France Culture. La phase de test de la campagne ESTER 1, qui a eu lieu début 2005, porte sur de nouvelles données : 10 heures d'émissions provenant de RFI, RTM, France Info, France Inter, France Culture et Radio Classique. Les différents corpus ainsi que le package d'évaluation contenant les protocoles d'évaluation et les outils de mesure de performance, sont diffusés par ELDA (*Evaluations and Language resources Distribution Agency*).

La seconde campagne d'évaluation (ESTER 2 [Galliano 2009]), démarrée fin janvier 2008, a pour but de mesurer les progrès effectués depuis la première campagne et de lancer de nouveaux axes de recherche. Elle est organisée conjointement par la DGA (Direction Générale à l'Armement) et l'AFCP (Association Francophone de la Communication Parlée), avec le concours de ELDA. Cette nouvelle campagne réutilise les données de ESTER 1, et en ajoute de nouvelles. Ces données incluent notamment des programmes étrangers mais francophones (nouvelle difficulté des accents étrangers en français), tout comme un nombre plus important de programmes contenant de la parole spontanée. Il est alors possible de trouver dans les données d'ESTER 2, en complément des émissions d'information radiophoniques, des débats, ainsi que des programmes africains issus de la radio Africa N° 1. Le nouveau corpus contient alors 100 heures d'émissions radiophoniques transcrites, enregistrées entre 1998 et 2004, auxquelles s'ajoutent 6 heures d'émissions radiophoniques pour le développement, et 6 heures pour le test. Des transcriptions, réalisées rapidement, de 40 heures d'émissions radiophoniques africaines étaient également disponibles. Enfin, les ressources textuelles ont été étendues avec des articles provenant du journal "Le Monde" de 2004 à 2006.

[15] http ://catalog.elra.info/product_info.php ?products_id=764

Le système de RAP, développé au LIUM pour les campagnes ESTER 1 [Deléglise 2005] et ESTER 2 [Deléglise 2009], a atteint de très bonnes performances : il est le meilleur système open-source dans les deux campagnes, et se trouve second en prenant en compte tous les participants. Le tableau 1.2 récapitule les taux d'erreur-mot obtenus par le système du LIUM durant les deux campagnes d'évaluation, sur les corpus de développement et de test.

Campagne	**Développement**	**Test**
ESTER 1	17,8 %	23,6 %
ESTER 2	24,2 %	17,8 %

TAB. 1.2 – Taux d'erreur-mot obtenus par le système de RAP du LIUM durant les campagnes d'évaluation ESTER 1 et ESTER 2, sur les corpus de développement et de test.

Chapitre 2

Traitement de la parole spontanée

Sommaire

2.1 Spécificités de la parole spontanée 36
2.1.1 Les disfluences 37
2.1.1.1 Les pauses 37
2.1.1.2 Les troncations, répétitions et faux-départs 38
2.1.1.3 L'élision 39
2.1.1.4 Les hésitations 40
2.1.2 Autres phénomènes 40
2.1.2.1 Agrammaticalité 40
2.1.2.2 L'intonation 41
2.1.2.3 Le débit de parole et l'état émotionnel du locuteur . . . 41
2.2 Gestion des disfluences 42
2.2.1 Objectifs 42
2.2.2 Détection automatique 44
2.2.3 Correction automatique 48
2.3 Impacts et solutions pour la reconnaissance de la parole 51
2.3.1 Modélisation acoustique 51
2.3.2 Modélisation linguistique 53
2.3.3 Dictionnaire de prononciations 56
2.3.3.1 Approche guidée par les données 57
2.3.3.2 Approche à base de connaissances 59
2.4 Conclusion 60

Les systèmes de transcription automatique de la parole progressent grâce à l'augmentation de la puissance de calcul, la quantité de données disponibles, et les progrès de la recherche dans le domaine. Après avoir longtemps proposé des solutions pour transcrire des corpus de phrases lues et enregistrées dans des conditions de studio (vocabulaire ciblé aux conditions de reconnaissance, conditions acoustiques claires, parole lue...), les travaux, en reconnaissance automatique de la parole, se sont récemment orientés vers le traitement des enregistrements d'actualités radiophoniques, beaucoup plus difficiles à transcrire (différents types de parole, conditions acoustiques hétérogènes, locuteurs inconnus et nombreux...) [Forsberg 2003]. Dans le cadre d'émissions radiophoniques, les systèmes de RAP obtiennent des performances raisonnables lorsqu'ils doivent transcrire de la parole préparée (parole proche d'un texte lu), mais ont de réelles difficultés à traiter de la parole spontanée [Furui 2003]. Dans notre cas d'étude, nous considérons la parole spontanée dans le sens où "*un énoncé est perçu et conçu au fil de son énonciation*" [Luzzati 2004] (débats, interviews...). Ce type de parole se place alors en opposition à la parole préparée.

La communauté scientifique du domaine poursuit ses efforts en orientant ses travaux vers le traitement de la parole spontanée. Ce type de parole est beaucoup plus difficile à traiter, car, entre autres, ses spécificités ne suivent pas toujours les règles linguistiques précises, et créent des ruptures au niveau de la parole (agrammaticalité, disfluences[16], état émotionnel du locuteur...). Les systèmes de RAP ne sont pas, à la base, conçus pour faire face à ces spécificités [Adda-Decker 2004]. Nous verrons donc dans une première partie les différentes études, notamment linguistiques, faisant ressortir les spécificités de la parole spontanée. Cette première partie permettra de prendre conscience de ces différences, et des difficultés rencontrées lors de sa transcription par les systèmes de RAP. Puis, nous ferons un état des lieux des travaux déjà réalisés pour détecter et corriger ces spécificités, et leur impact sur la reconnaissance de la parole. Pour clore ce chapitre, nous nous intéresserons aux différentes solutions existantes pour améliorer la transcription de ce type de parole dans les différentes composantes d'un système de RAP.

2.1 Spécificités de la parole spontanée

La parole spontanée a été décrite dans diverses études linguistiques, l'objectif étant de comprendre les singularités de ce type de parole. Elle apparaît comme un objet d'étude à part entière, en comparaison de la parole préparée [Gendner 2002]. Bien que la parole spontanée soit une particularité du langage finalement assez peu étudiée, certaines études décrivent les disfluences

[16]Les disfluences peuvent être définies comme des problèmes, des dysfonctionnements, apparaissant au cours de la parole.

comme étant la plus grande caractéristique de la parole spontanée en langue française. Par exemple, nous pouvons citer les travaux sur le français de [Adda-Decker 2004, Bazillon 2008b, Henry 2004] (de nombreux autres travaux existent, une présentation plus détaillée se trouvant dans les parties suivantes). La présence de disfluences n'est cependant pas un phénomène propre au français : cette particularité se retrouve dans de nombreuses autres langues [Vasilescu 2004, Shriberg 1999, Ward 1989]. Ainsi, comme défini dans la littérature, les disfluences sont visibles sous différentes formes (morphèmes spécifiques, répétitions, hésitations ...). Bien que ces disfluences soient généralement étudiées séparément, des études montrent que certaines d'entre elles peuvent être liées lors de la construction du discours [Campione 2004, Henry 2004], en démontrant par exemple que les répétitions étaient fortement associées aux pauses. À ces disfluences s'ajoutent l'agrammaticalité et un registre de langage différent de celui retrouvé dans les textes écrits [Boula de Mareüil 2005]. En fonction du locuteur, de l'état émotionnel et du contexte, le langage utilisé peut varier énormément. L'intonation, utilisée lors d'échanges verbaux, peut également être considérée comme discriminante pour caractériser la parole spontanée [Martin 2006].

Nous allons donc détailler, dans les sous-parties suivantes, les différentes particularités de la parole spontanée que l'on peut retrouver actuellement dans la littérature. Tout d'abord, nous présenterons les disfluences les plus caractéristiques de la parole spontanée. Puis nous nous intéresserons plus précisément à certains phénomènes prosodiques de la parole spontanée, notamment l'intonation ainsi que le débit phonémique.

2.1.1 Les disfluences

2.1.1.1 Les pauses

Dans le cadre de la parole spontanée, les pauses apparaissent comme des marqueurs très utiles, car très différents de ce que l'on peut trouver dans de la parole lue. Ces pauses peuvent être classées en deux grandes catégories [Duez 1982] :

Les pauses silencieuses, parfois appelées pauses non-sonores, pouvant être plus ou moins longues. Elles marquent souvent, dans le cadre de la parole spontanée, la rupture au niveau d'une idée. Ces pauses permettent de structurer le discours [Campione 2004]. De plus, dans [Bazillon 2008b], les auteurs ont affirmé que les pauses de respiration dans ce type de parole étaient beaucoup plus nombreuses et plus longues que celles que l'on pouvait retrouver en parole préparée. Cette différence est due au fait que ce type de parole est conçu à l'instant où le locuteur parle, il lui arrive donc de devoir s'arrêter pour continuer à construire son discours.

Les pauses sonores, plus communément appelées pauses remplies, sont des phénomènes typiques de l'oral. En effet, les pauses remplies regroupent les morphèmes tels que "*euh*", "*hum*"

ou encore "*ben*"[17]. Dans [Bazillon 2008b], la fonction de ces pauses remplies, appelées aussi morphèmes spécifiques, reste difficile à définir. Le morphème "*euh*" est alors catégorisé en tant qu'hésitation, mais pour les autres morphèmes, une catégorisation reste plus délicate. Les auteurs donnent alors l'exemple des emplois de "*ben*", qui peut être adverbe, conjonction de coordination...

2.1.1.2 Les troncations, répétitions et faux-départs

Bien que les études sur les disfluences relatent de manière générale les phénomènes de troncations, répétitions et faux-départs, les auteurs dans [Bazillon 2008b] les étudient un peu plus en détail. Ainsi, ils présentent la troncation comme une spécificité de la parole spontanée, qui se manifeste à l'intérieur d'un mot, possédant plusieurs départs : ses premiers phonèmes sont répétés un certain nombre de fois, sans que le mot ne soit réellement terminé. Le mot est ensuite finalement bien prononcé[18]. Parfois, la troncation rejoint le phénomène de bégaiement [Pallaud 2007].

Les faux-départs sont des phénomènes désignant une "*interruption à l'intérieur d'un énoncé, et non à l'intérieur d'un mot*" [Bazillon 2008b]. En d'autres termes, un faux-départ peut être vu comme le commencement d'une idée (les mots sont réellement prononcés, à la différence du phénomène de troncation) qui n'est pas achevée : le locuteur enchaîne directement sur un énoncé complètement différent du premier (souvent une rupture au niveau du sens et de la syntaxe)[19]. Il est également possible qu'un faux-départ apparaisse dans la continuité de la première idée du locuteur ; le terme de semi faux-départ peut alors être employé[20].

Enfin, les répétitions sont fréquentes en parole spontanée, et doivent être traitées avec attention [Henry 2004]. En effet, tout comme les autres disfluences présentées dans cette partie, elles constituent une rupture lors de l'énoncé d'un locuteur. Les répétitions peuvent ici avoir lieu au niveau du mot[21], mais aussi au niveau d'un groupe de mots[22]. Tout comme les troncations, le bégaiement peut les expliquer. Afin de mieux comprendre ce phénomène, une étude linguistique a été entreprise dans [Henry 2002]. L'auteure laisse entrevoir la possibilité d'utiliser cette étude linguistique pour améliorer les systèmes de reconnaissance automatique de la parole.

[17] Exemple d'utilisation de pauses remplies : "*ben* là *hum* je fais *euh* ce que je veux tu vois".
[18] Un exemple de troncation : "mais cette personne ne pré() pré() présentera que la première partie".
[19] Un exemple de faux-départ : "non mais l'homme dans [*rupture*] ne cherche pas tes clés".
[20] Par exemple : "Il ne faut cependant pas chercher () vas voir par là".
[21] Exemple de répétitions au niveau du mot : "non mais je je suis très très très attentif à ce que tu dis".
[22] Exemple d'une répétition au niveau d'un groupe de mots : "tu es tu es responsable de cet échec".

2.1.1.3 L'élision

L'élision du schwa est une élision très connue et largement étudiée dans le milieu linguistique. Par définition, en linguistique, le schwa (quelque fois appelé *e muet* ou *e caduc*) désigne une voyelle neutre, centrale, notée [ə] en Alphabet Phonétique International (API). L'élision du schwa est, par conséquent, l'assimilation de cette voyelle. Cette "perte" peut apparaître au milieu d'un mot ou entre deux mots [Fougeron 1999]. Ainsi, par exemple, ce phénomène peut apparaître :

- **À l'intérieur d'un mot** Prenons l'exemple du groupe de mots "un *petit* chat" où l'élision du schwa peut se réaliser sur le mot "petit" (prononciation standard [pəti] et après élision [pti] — *p'tit*).
- **Entre deux mots** Le mot "je", dont la prononciation standard est [ʒə], est un exemple récurrent du phénomène du schwa dans la littérature. En effet, lorsqu'il est associé à un autre mot commençant par une consonne sourde, le phonème [ə] est amené à disparaître. Par exemple les mots "je prends" peuvent se prononcer [ʒə pʁɑ̃] (parole lue), ou [ʒ pʁɑ̃] (*j' prends* en parole spontanée, avec le phénomène d'élision du schwa).
- **En changeant de phonème** Dans certains cas, les phonèmes utilisés peuvent également être amenés à changer. En reprenant l'exemple de "je pense", le son [ʒ] peut se transformer en [ʃ] (*ch' pense*). Cette modification se fait en deux étapes : la première étape est une assimilation du son [ə] (*j' pense*), puis, la seconde étape transforme le son [ʒ] (consonne sonore), au contact du son [p] (consonne sourde), en [ʃ] (consonne sourde). Ce phénomène d'assimilation progressive des consonnes sonores en consonnes sourdes est appelé *dévoisement*.

Une étude quantitative, dans [Adda-Decker 1999], permet de connaître de manière plus détaillée la proportion des phonèmes d'élision du schwa dans le cadre d'un corpus français de parole spontanée (*MASK*[23]) et d'un corpus de parole préparée (*BREF*[24]), en analysant les différentes variantes de prononciation. Les résultats montrent que le phénomène d'élision du schwa est beaucoup plus prononcé en parole spontanée.

D'autres élisions existent en parole spontanée. Les auteurs, dans [Bazillon 2008b], se sont intéressés au problème particulier de la vibrante [r], que l'on retrouve en parole spontanée en français. Il existe alors une règle concernant les mots se terminant par "-bre", "-cre", "-dre", "-tre" ou "-vre". Dans le cas d'une prononciation correcte de ces mots, le phonème / r / est obligatoire. Par exemple, le verbe conjugué "ouvre" doit être prononcé [uvr]. Cependant en parole spontanée, le phonème / r / peut parfois être omis à la fin d'un mot, la prononciation [uv] pouvant alors apparaître.

[23] 35 heures de parole spontanée via un système de dialogue de demandes d'information de voyages.
[24] 120 heures de parole lue.

Il est très difficile d'énumérer tous les phénomènes d'élision de la parole spontanée pour tous les mots, car peu de règles bien définies existent. Nous pouvons néanmoins noter que de nombreux mots voient leur variante de prononciation modifiée en fonction du registre de langue, ou encore de la vitesse d'élocution, phénomènes que nous présenterons dans les parties suivantes.

2.1.1.4 Les hésitations

Dans de nombreuses études, comme dans [Campione 2004, Pallaud 2004, Vasilescu 2004], les hésitations reviennent comme un phénomène important de la parole spontanée. Ce phénomène ne fonctionne pas seul, puisqu'il reprend quelques unes des disfluences caractéristiques de la parole spontanée. Finalement, le terme d'*hésitation* peut plutôt être vu comme le regroupement de particularités du langage. L'hésitation comprend ainsi le phénomène de pauses remplies [Campione 2004], et est susceptible d'être liée aux phénomènes dits d'allongements *d'hésitations*, que l'on peut retrouver sous le nom d'allongements vocaliques [Candéa 2000]. Un allongement vocalique, est, comme son nom l'indique, l'étirement d'un son, souvent en fin de mot[25] [Bazillon 2008b]. Dans la littérature, certains allongements sont appelés *mélismes* [Caelen-Haumont 2002a], désignant un allongement syllabique en fin de mot. De plus, les troncations, les répétitions, et les faux-départs, peuvent parfois être englobés dans le terme plus générique d'*hésitations* [Pallaud 2004].

2.1.2 Autres phénomènes

2.1.2.1 Agrammaticalité

L'agrammaticalité est également très présente en parole spontanée, à la différence des textes lus. En effet, lorsque les textes sont préparés à l'avance (par exemple le cas d'un journaliste qui lit son texte), les phrases respectent les règles de syntaxe de la langue utilisée, puisque l'auteur de ce texte a pris le temps, en amont, de le construire. Dans le cas de la parole spontanée, ce texte est construit au fil de l'énonciation. La construction de la phrase ne suit alors pas de manière précise les règles grammaticales conventionnelles, puisque l'intérêt premier est de se faire comprendre : les règles "structurelles" ne sont pas indispensables dans le cadre de l'oral spontané. Cette idée de l'agrammaticalité se retrouve dans la plupart des études linguistiques traitant de la parole spontanée, et notamment celles que nous avons déjà étudiées avec les disfluences. Il est possible de retrouver ce concept dans [Luzzati 2004], sans cependant avoir d'études détaillées sur l'agrammaticalité dans le cadre de la parole spontanée.

[25] Exemple : "Je ne cherche pas **à** répondre objectivement" où le phonème [a] du mot **à** dure quelques secondes.

2.1.2.2 L'intonation

L'intonation, ou *intonation prosodique*, qui est ici étudiée dans un cadre linguistique, est, par définition, le *fonctionnement signifiant des variations de la fréquence fondamentale dans l'énoncé*[26]. Plus simplement, l'intonation peut être vue comme la façon d'attaquer un son lorsque l'on parle (*i.e.* le ton employé). Ainsi, bien que le sens grammatical de la phrase ne change pas, l'intonation permet de structurer le discours (elle permet par exemple de fournir l'information d'une phrase interrogative ou exclamative). Des études se sont penchées sur les variations au niveau de l'intonation pouvant exister entre une parole lue et une parole plus spontanée [Martin 2006]. Les auteurs de cet article utilisent alors un corpus comprenant ces deux types de parole afin d'en extraire des caractéristiques comparables au niveau acoustique, et de chercher à savoir si des différences existent. Au moyen de diverses expériences, les auteurs montrent que des différences sont visibles.

Ces mêmes particularités ont été constatées par [Adda-Decker 2004], qui, dans le cadre de la reconnaissance automatique de la parole, montre que l'intonation peut avoir un impact sur les performances des applications visées. Dans le cadre de l'annotation manuelle de la parole spontanée, les auteurs, dans [Bazillon 2008b], ont ainsi remarqué qu'il était beaucoup plus difficile pour les annotateurs humains d'établir la ponctuation d'une phrase, alors que ce même travail pour la parole préparée apparaissait plus simple. Cette difficulté a pu être expliquée par l'agrammaticalité et les disfluences (voir sections précédentes), mais aussi par les intonations des locuteurs, permettant aux annotateurs d'obtenir des informations structurelles sur les phrases énoncées. En effet, les auteurs prennent en exemple la fin d'un segment de parole : le locuteur ne sait pas toujours, dans le cadre de la parole spontanée, lorsque sa phrase se terminera. Il ne fournira donc pas de marque intonative de fin de phrase, ce qui rend le travail des transcripteurs beaucoup plus difficile. Tout comme l'article présenté précédemment, les auteurs, dans [Adda-Decker 2004], évoquent le problème de la ponctuation des phrases dans le cadre de la parole spontanée. Les travaux se positionnent dans le domaine de l'indexation des documents et de l'étiquetage syntaxique, où la ponctuation peut avoir une influence importante au niveau des résultats. L'intonation est, pour les auteurs, un des phénomènes pouvant aider à la recherche de ponctuation.

2.1.2.3 Le débit de parole et l'état émotionnel du locuteur

Le débit de parole, que l'on peut retrouver sous la dénomination de *débit phonémique*, peut se définir comme la variation de la vitesse de production des sons par un locuteur. Dans certaines études, cette vitesse est caractéristique de la parole spontanée. Des analyses menées sur de la

[26]Définition extraite du dictionnaire encyclopédique **Larousse**. `http://www.larousse.fr`

parole lue ont permis de constater que ce débit phonémique change peu, alors que dans le cadre de la parole dite spontanée, ce débit a tendance à varier au cours de l'énonciation. Selon l'élocution et les ruptures au niveau de la parole (parfois dues aux disfluences), le débit phonémique peut se mettre à tout moment à ralentir, si par exemple le locuteur ne sait pas exactement ce qu'il va dire, ou au contraire accélérer brusquement [Bazillon 2008b]. Cette variation est visible dans les différentes expériences qui ont été menées dans cet article de revue. Dans [Rouas 2004], les auteurs cherchent à mesurer automatiquement les débits de parole dans le cadre de la parole spontanée, et ce, dans plusieurs langues. Ils ont alors expliqué que la durée des phonèmes et la durée des syllabes étaient de bons indicateurs du débit de parole. En conclusion, les auteurs affirment qu'en plus de la langue parlée, le débit de parole est dépendant du locuteur. Cependant, ces explications ne sont pas suffisantes : dans [Rouas 2004], ils considèrent que la variation du débit de parole peut avoir plusieurs sources, les disfluences (pauses, hésitations...) pouvant être une première cause.

L'état émotionnel du locuteur peut également être un problème de la parole spontanée. Les émotions, dans la parole spontanée, semblent jouer un rôle beaucoup plus important que dans la parole lue. Dans [Caelen-Haumont 2002a], les expériences menées semblent montrer que l'état émotionnel (un bon exemple pourrait être le stress) du locuteur est parfois beaucoup plus marqué en parole spontanée, ce qui a pour conséquence d'influer sur la manière dont la phrase va s'articuler. En effet, en parole préparée, si le locuteur est en état de stress, il pourra toujours s'appuyer sur son texte préparé. Son état émotionnel aura donc une influence assez faible sur les idées et les mots associés. Or, dans un contexte spontané, cet état émotionnel peut rendre complexe la construction et l'organisation des idées du locuteur ; il peut avoir plus de mal à parler, la construction de ses phrases peut être beaucoup plus difficile et moins compréhensible (nombreuses disfluences, avec des hésitations, des pauses, des élisions...). Finalement, comme nous l'avons déjà vu dans [Rouas 2004], les différentes particularités de la parole spontanée sont souvent liées entre elles. Nous pouvons citer les travaux de [Caelen-Haumont 2002a] qui développent cette même idée de lien entre le débit phonémique et l'état émotionnel du locuteur.

2.2 Gestion des disfluences

2.2.1 Objectifs

Les études présentées précédemment sur les spécificités de la parole spontanée ont permis de mieux comprendre ce type de parole. Ces particularités peuvent avoir une incidence dans différents domaines, notamment ceux du traitement automatique des langues (TAL) : parmi eux le dialogue homme-machine [Goto 1999], le résumé automatique [Honal 2003],

ou encore la reconnaissance automatique de la parole [Goldwater 2010, Adda-Decker 2003]. Pouvoir endiguer ce problème des disfluences permettrait, par exemple, d'aider la transcription manuelle d'émissions radiophoniques (une présentation de ces problèmes est visible dans [Adda-Decker 2004, Bazillon 2007]). Globalement, ces spécificités rendent la reconnaissance automatique de la parole (au moyen de systèmes de RAP) beaucoup plus difficile. Des faiblesses apparaissent dans plusieurs modules du système de RAP, que ce soit au niveau du dictionnaire de prononciations, de la modélisation acoustique (avec principalement le problème des phénomènes prosodiques), ou encore de la modélisation linguistique (l'agrammaticalité, les troncations, ou encore les répétitions qui posent problèmes au niveau des modèles de langage, souvent appris sur des textes grammaticalement corrects). Les disfluences semblent avoir une influence sur les systèmes de RAP et leurs taux d'erreur-mot [Adda-Decker 2003].

Dans la continuité de ces recherches, des études se sont focalisées sur la détection et la correction automatique de ces particularités, afin d'améliorer les systèmes ayant des difficultés à gérer ces phénomènes (principalement pour le domaine du TAL). Ces travaux ont commencé il y a une quinzaine d'années, avec notamment la thèse de [Lickley 1994]. Dans cette thèse, l'auteur cherche à connaître les caractéristiques acoustiques et linguistiques permettant à un humain de détecter la présence de discontinuités (disfluences) dans les segments de parole. Les résultats de ses expériences permettent de conclure qu'une détection des disfluences est possible très tôt dans le segment[27].

La récente campagne d'évaluation *NIST Rich Transcription Fall 2004*[28] a permis de mettre en lumière le fait qu'une importante chute des performances des systèmes de RAP est visible lorsque ceux-ci devaient transcrire de la parole spontanée. Cette baisse au niveau des résultats peut s'expliquer par le bruit généré par les phénomènes de la parole spontanée dans les systèmes de RAP, non conçus pour gérer ces spécificités. Certaines tâches de cette campagne se sont alors intéressées à la détection et la correction automatiques des disfluences dans le contexte de la parole spontanée.

Dans ce contexte, nous chercherons à décrire, dans les parties suivantes, les études réalisées pour détecter automatiquement les disfluences contenues dans un flux de parole spontanée, et ensuite les approches envisagées pour corriger ces phénomènes particuliers. L'idée finale est d'aider les systèmes à mieux gérer ces spécificités, et d'améliorer leurs performances.

[27] L'auteur parle, dans sa thèse, d'une détection automatique possible d'une zone d'interruption (disfluence) dès l'apparition du premier mot suivant cette interruption.

[28] La campagne d'évaluation intégrait de nombreuses tâches de transcription de la parole, dans le contexte d'émissions d'information, ainsi que des conversations téléphoniques dans différentes langues. Plus d'informations sur le site : `http://itl.nist.gov/iad/mig/tests/rt/2004-fall/index.html`

2.2.2 Détection automatique

De manière générale, la détection automatique de certaines particularités de la parole spontanée a été traitée dans de nombreux travaux. Des études linguistiques, permettant de comprendre ces spécificités, sont indispensables pour mettre au point des outils de détection automatique. Récemment, des travaux se sont penchés sur le problème des disfluences[29], en essayant de les détecter dans un texte. Afin de rendre les textes plus lisibles pour les humains, et d'améliorer les systèmes issus du TAL (dialogue, reconnaissance de la parole...), les auteurs dans [Liu 2005] comparent plusieurs méthodes automatiques permettant de détecter les disfluences. Ils prennent alors en compte des indices linguistiques (transcriptions automatiques ou manuelles), comme la répétition de mots ou encore les informations morpho-syntaxiques. Pour pallier les erreurs introduites par les systèmes de RAP dans les transcriptions automatiques, les auteurs extraient également des informations prosodiques, comme la durée, la fréquence fondamentale (F0), ou encore les pauses. Ces caractéristiques sont ensuite utilisées dans trois approches statistiques différentes : avec des MMC (voir section 1.3.1), avec une méthode de classification CRF [Lafferty 2001] (*Conditional Random Field*), et avec la méthode du maximum d'entropie (*MaxEnt*). Les méthodes statistiques utilisant les CRF et le MaxEnt obtiennent les meilleurs résultats. Ils permettent aux auteurs de conclure avec des résultats encourageants sur la détection des disfluences (ce que les auteurs appellent la détection des *points d'interruption*[30]).

Dans la même lignée que [Liu 2005], les auteurs dans [Johnson 2004, Lease 2006] cherchent à détecter les pauses remplies (*euh*, *ben*...), les éditions des mots[31], et enfin les points d'interruption (voir paragraphe précédent). Afin de détecter les disfluences, plusieurs méthodes sont alors utilisées :

- Une approche statistique pour détecter les réparations, avec l'utilisation d'informations linguistiques au moyen d'un modèle de langage syntactique.
- Des règles manuelles pour détecter les pauses remplies.
- La combinaison de la méthode statistique et des règles manuelles pour détecter les éditions de mots.

À travers leurs différentes approches, les auteurs ont obtenu des résultats positifs dans la détection de ces trois disfluences, que ce soit au niveau des transcriptions manuelles ou des transcriptions automatiques, concluant même avec des performances obtenues au moyen de

[29] Les auteurs cherchent ici à détecter certaines disfluences, en particulier celles interrompant la phrase (pauses, hésitations, répétitions...).

[30] La détection des *points d'interruption*, est le fait de prédire, pour chaque pause phonétique inter-mots, si la parole devient disfluente à partir de ce point, le mot précédant terminant une réparation, ou le mot suivant étant une pause remplie.

[31] Les éditions de mots font référence aux problèmes de répétitions et de troncations, et notamment le problème des mots "non terminés". Par exemple, si on prend la phrase "je vais man() manger", "man()" fera partie des problèmes d'édition.

leurs méthodes automatiques relativement proches de celles obtenues par un annotateur humain. Ces travaux ont été réalisés dans l'idée de participer à trois tâches (correspondant aux trois disfluences détectées) de la campagne d'évaluation *NIST Rich Transcription Fall 2004*. Cette méthode a permis aux auteurs de se classer premier dans chacune de ces tâches [Lease 2006].

Pour [Stolcke 1998], la détection des disfluences est très importante. Elle peut avoir un intérêt lors de la segmentation de la parole, en définissant les frontières entre les phrases. Ce problème est important en RAP, où les systèmes doivent traiter un flux audio sans cependant connaître les délimitations des phrases (voir partie 1.5.2.1). Les auteurs cherchent à étudier ces deux sous-domaines de recherche, à savoir la détection des disfluences et la segmentation en phrases. Ces travaux, qui se veulent comme une base au niveau des résultats de détection automatique de disfluences (pauses, hésitations, faux-départs...), permettent de résumer des caractéristiques intéressantes pour ce type de tâche :

- **Des caractéristiques prosodiques** : les durées des pauses et des voyelles finales, la fréquence fondamentale (F0), l'énergie et le ratio signal/bruit.
- **Des caractéristiques linguistiques** : utiliser un modèle de langage spécifique pour les disfluences ; utiliser des informations morpho-syntaxiques (comme les classes grammaticales, mais qui ne sont pas étudiées dans cet article).

Plus récemment, les auteurs dans [Liu 2003], ont eux aussi cherché à détecter les disfluences (répétitions, faux-départs et troncations) de la parole spontanée en combinant différentes sources d'information, proches de celles définies dans [Stolcke 1998]. Ces travaux utilisent des caractéristiques prosodiques (durée des phonèmes et des pauses, fréquences fondamentales...) ainsi que des informations obtenues avec trois différents modèles de langage :

- Un modèle de langage n-gramme de mots, basé sur des transcriptions manuelles. Les transcriptions manuelles fournies ne contiennent pas simplement les mots, mais aussi les disfluences contenues dans la phrase (les interruptions par exemple). Ces marqueurs sur les disfluences sont considérés comme des mots lors de l'apprentissage du modèle.
- Un modèle de langage n-classe, appris sur les catégories syntaxiques des mots (nom, adjectif, adverbe...) obtenues au moyen d'un analyseur syntaxique statistique. Tout comme le modèle de langage précédent, les disfluences sont présentes dans le corpus d'apprentissage.
- Un modèle de langage n-gramme appris sur un regroupement des problèmes de répétitions en classes : au lieu de se focaliser simplement sur les mots, ceux-ci sont regroupés au sein d'un même modèle plus général. Par exemple, une répétition peut se caractériser par un point d'origine, une rupture, puis une répétition : ce schéma particulier constitue une classe. Les mots sont alors catégorisés à l'intérieur de ces classes, devant remédier au problème des mots peu fréquents.

Ces trois modèles sont utilisés séparément, puis leurs probabilités postérieures sont interpolées ensemble, les auteurs remarquant que la complémentarité de ces trois modèles permettait d'obtenir les meilleurs résultats sur la détection de disfluences.

Nous venons de voir quelques travaux sur la détection de nombreuses disfluences de manière générale. Cependant, certains phénomènes de la parole spontanée ont été étudiés individuellement. Les pauses remplies ont notamment fait l'objet de travaux particuliers. Dans le cadre de la reconnaissance de la parole, les auteurs dans [O'Shaughnessy 1992] ont obtenu des résultats très intéressants sur la détection de pauses, la plupart d'entre elles ayant pu être identifiées. Ils ont alors suggéré que les pauses et les hésitations, si elles étaient isolées, permettraient d'enlever des hypothèses des systèmes de RAP, ce qui pourrait conduire à l'amélioration de leurs performances. Dans ce travail, une distinction est réalisée entre les pauses grammaticales, et les pauses non-grammaticales, respectivement des pauses utilisées pour marquer la fin d'une phrase (avant d'en développer une nouvelle), et des pauses utilisées à l'intérieur d'une phrase (la phrase n'est pas terminée). En effet, elles ne présentent pas exactement les mêmes caractéristiques, particulièrement au niveau de la fréquence fondamentale.

Ces premières recherches ont permis de fournir des premiers résultats encourageants sur la détection des pauses. Dans [Goto 1999], les auteurs partent de ces précédentes conclusions, et cherchent à améliorer la détection de ce type précis de disfluences[32] dans la langue japonaise, afin d'aider les systèmes de dialogue homme-machine à gérer de manière plus efficace, par exemple, les tours de parole lors d'une conversation. L'idée de cet article est de trouver des caractéristiques acoustiques, en utilisant une analyse de fréquence, afin de détecter les pauses remplies : une faible transition au niveau de la fréquence fondamentale (F0), et enfin une faible déformation de l'enveloppe spectrale, sont alors retenues comme de bons indicateurs de pauses remplies. Finalement, leur détecteur automatique de pauses remplies a été utilisé (de manière expérimentale) dans le processus d'alignement automatique des phonèmes sur le signal de parole (voir partie 1.3.2). L'alignement, réalisé ici au moyen de l'algorithme Viterbi, peut avoir des difficultés à gérer les pauses remplies : le détecteur est utilisé pour contrôler la durée des phonèmes.

Le phénomène de l'élision peut aussi avoir un effet sur les outils développés dans le cadre de la reconnaissance automatique de la parole. Ainsi, dans [Auran 2004], les auteurs s'intéressent aux corpus annotés et alignés automatiquement au moyen d'un système de RAP, et soulignent les difficultés rencontrées dans ce type de tâche, particulièrement lorsque les corpus contiennent de la parole spontanée. Le problème réside dans le fait que le dictionnaire de prononciations peut

[32]Notons que l'allongement des mots est également étudié, car il présente des caractéristiques acoustiques communes avec les pauses remplies. Par soucis de simplicité, les auteurs appellent *pauses remplies* ces deux phénomènes.

ne pas être adapté à ce contexte de spontanéité s'il ne possède pas, entre autres, des phonétisations pour le problème de l'élision (disparition de phonèmes). Afin d'améliorer la qualité de l'alignement, les auteurs mettent au point une méthode semi-automatique[33], qui consiste, dans un premier temps, à obtenir un dictionnaire de prononciations intégrant le phénomène d'élision (obtenu au moyen de règles manuelles), pour ensuite l'utiliser lors du processus d'alignement. Cet alignement sera ensuite amélioré au moyen d'un algorithme de prédiction d'élision de phonèmes (entraîné sur un petit corpus étiqueté manuellement). Finalement, les auteurs ont trouvé que cette approche semi-automatique permettait d'obtenir une meilleure fiabilité de l'alignement (en comparaison d'une méthode totalement automatique). Des travaux sur l'alignement automatique se sont également intéressés à l'analyse du schwa, notamment dans le cadre de l'analyse phonétique [Bürki 2008].

Fort de ses résultats positifs sur la détection des pauses, l'auteur dans [O'Shaughnessy 1992] se penche ensuite sur la détection du phénomène prosodique des faux-départs [O'Shaughnessy 1993]. La compréhension de ce phénomène particulier des faux-départs est toujours étudiée, afin d'extraire les caractéristiques acoustiques utiles à leurs détections. Les indices acoustiques retenus par les auteurs, et qui ont été repris dans [Goto 1999], sont la durée des phonèmes, la fréquence fondamentale (F0), et enfin l'enveloppe spectrale aux alentours d'une pause. Leur utilisation a permis de détecter automatiquement les reprises simples (reprises sans changement de mots), avec une précision de plus de 80 %. Ce détecteur de faux-départs a été réalisé dans le contexte des systèmes de RAP (extraction des données acoustiques).

La gestion des émotions des locuteurs est également un phénomène qui a fait l'objet d'études au niveau de sa détection automatique. Nous pouvons citer les travaux réalisés dans [Devillers 2004], qui cherchent à détecter les émotions dans le cadre du dialogue oral, domaine du TAL devant faire face aux spécificités de la parole spontanée. Cette détection peut alors, dans le cadre du dialogue, permettre de connaître l'état émotionnel du locuteur, et donc d'interagir différemment selon ses réactions. Cette détection est réalisée automatiquement en extrayant des indices prosodiques corrélés aux états émotionnels (par exemple le débit de parole, ou encore l'énergie). Les expériences menées montrent que ces indices permettent d'obtenir des résultats intéressants au niveau de cette première étape de détection des émotions. Ils ouvrent alors des perspectives aux Interactions Homme-Machine dans le cadre du dialogue oral.

[33] Gain de ressources par rapport aux approches entièrement manuelles, et gain en précision par rapport aux méthodes complètement automatiques.

2.2.3 Correction automatique

Comme nous venons de le voir, des études se sont penchées sur la détection, de manière automatique, des disfluences. Très souvent, cette détection s'est faite en ayant comme objectif final d'aider les outils informatiques, et particulièrement les outils du TAL, à mieux gérer les problèmes de la parole spontanée. Bien que la détection de certaines disfluences puisse avoir des effets bénéfiques pour la reconnaissance de la parole[34], une simple détection ne semble pas suffisante pour d'autres phénomènes de la parole spontanée : leur correction doit être envisagée. Cette volonté de correction automatique des spécificités de la parole spontanée a tout d'abord été motivée par le domaine du dialogue homme-machine [Bear 1992, Heeman 1996]. Dans [Bear 1992], les auteurs veulent combiner la détection et la correction pour certaines disfluences de la parole spontanée, comme les faux-départs, les hésitations, ou encore les répétitions. Leur outil, développé pour traiter des phrases provenant d'un dialogue, utilise des informations syntaxiques, sémantiques et acoustiques. Ces trois sources d'information sont combinées, car peu efficaces séparément, et utilisées au moyen de règles manuelles et d'un filtrage par motif (*pattern-matching*). Ces études préliminaires ont permis d'amener de nouveaux travaux sur la correction automatique, toujours sur ces problèmes de "réparations" (hésitations, répétitions...). Dans l'article proposé par [Heeman 1996], l'idée est non pas de traiter séparément détection et correction, mais d'utiliser conjointement ces deux modules. Ainsi, le module de correction aide à savoir si un problème de "réparation" est présent (phase de détection). Ils parlent alors "d'auto-correction". Les auteurs se sont rendus compte, durant leurs expériences, que cette approche (correction aidant la détection) permettait d'améliorer les résultats de détection.

La correction automatique des disfluences, dans le cadre de la parole spontanée, est toujours d'actualité, puisque des études essayent toujours de résoudre ce problème. Les travaux de [Spilker 2000] se sont orientés sur la correction de ce phénomène de "réparation", très présent dans les dialogues oraux. Tout comme les autres auteurs, ces travaux se positionnent dans le cadre du TAL. L'objectif est de corriger en amont ces problèmes de la parole spontanée, en les détectant le plus tôt possible. La méthode proposée cherche à insérer des nouvelles hypothèses corrigées (sans les disfluences étudiées) à l'intérieur d'un treillis contenant déjà des hypothèses. La méthode se déroule en différentes étapes, et fait intervenir plusieurs sources de connaissances :

– Premièrement, un module acoustique (avec extraction d'informations acoustiques et prosodiques) génère des hypothèses sur la présence ou non de "réparations". Le but de ce module est de détecter les points d'interruption dans la séquence de parole.

[34] Nous pouvons notamment citer les travaux de [Goto 1999, O'Shaughnessy 1992] sur la détection des pauses.

– Ensuite, un modèle statistique essaie de trouver une correction appropriée de chaque point d'interruption détecté. L'idée est d'envisager le processus de correction comme un problème de traduction automatique statistique, où l'on doit remplacer le mot précédant le point d'interruption par le mot corrigé (se trouvant après ce point). Un modèle de traduction (avec informations syntaxiques, en cherchant les ruptures de syntaxe dans la phrase, et sémantiques, en manipulant les classes grammaticales des mots) est alors utilisé, et les hypothèses qui en résultent sont insérées dans le treillis.
– La dernière étape consiste simplement à sélectionner le meilleur chemin dans le treillis.

Les expériences ont été menées sur la langue allemande, sur le corpus *Verbmobil*[35], et ont permis d'obtenir des résultats positifs pour la correction de ce type d'erreur, tout en proposant un modèle simple et rapide.

Les "réparations" sont au centre des recherches en correction des disfluences. Ainsi, dans [Honal 2003], les auteurs s'intéressent encore à ce phénomène du langage, ainsi qu'aux faux-départs, qu'ils définissent comme une spécificité à part. Poursuivant l'approche proposée dans [Spilker 2000], un système de traduction automatique statistique sera utilisé. Pour décrire cette approche, les auteurs se servent de la terminologie issue de la traduction :

– Le langage source (*i.e.* le langage que l'on doit traduire) est dans cette étude le texte contenant des disfluences.
– Le langage cible (*i.e.* le langage vers lequel le langage source sera traduit) est le texte sans disfluence.

La méthode cherche alors la phrase cible la plus probable (parmi un ensemble de phrases cibles possibles) étant donné une phrase source. Dans l'approche proposée, en plus de la probabilité d'émission d'une phrase cible étant donné une phrase source, viennent s'ajouter des probabilités obtenues au moyen de modèles statistiques entraînés sur certaines propriétés des disfluences (par exemple le nombre de mots à supprimer de la phrase source pour obtenir la phrase cible). De nombreuses corrections de disfluences sont étudiées : les faux-départs, les répétitions/corrections, les pauses remplies, et enfin les interjections. Au contraire d'autres approches [Bear 1992], aucune information linguistique n'est nécessaire pour cette méthode (l'annotation des disfluences est suffisante), aucune règle manuelle n'est utilisée (seulement des modèles statistiques), et la méthode est potentiellement extensible à la correction d'autres disfluences. Les résultats présentés dans cette étude dépassent largement les autres méthodes présentées jusqu'à maintenant, avec un rappel atteignant 77,2 % et une précision de 90,2 % sur la correction des disfluences étudiées sur un corpus en langue anglaise.

[35]Corpus de dialogues pour la prise de rendez-vous. Les données ont été collectées pour l'anglais, l'allemand et le japonais.

Les résultats obtenus dans [Honal 2003] ont poussé les auteurs à continuer dans cette voie, en cherchant dans [Honal 2005] à améliorer les performances de leur méthode grâce aux paramètres du système, en sélectionnant intelligemment les données d'apprentissage et de développement. Le principal apport de cet article est l'introduction du locuteur en tant qu'élément important dans la production des disfluences. En effet, nous avons pu voir dans la partie 2.1 traitant des spécificités de la parole spontanée, que le locuteur pouvait avoir un rôle important en parole spontanée (qualité d'élocution, état émotionnel. . .). Des expériences ont été menées pour connaître l'impact de leur système en séparant les locuteurs en deux groupes : ceux prononçant de nombreuses disfluences, et ceux en produisant peu. Le but est de voir si un système, entraîné avec des locuteurs organisés en groupes, permet de mieux corriger les disfluences d'un locuteur. Bien que les résultats de ces expériences ne permettent pas de conclure que les performances du système, prenant en compte un groupe de locuteurs, soient meilleures pour tous les locuteurs, ils ont tout de même pu obtenir des gains sur certains rappels/précisions par rapport à un apprentissage global. Une des limites de cette approche par groupe de locuteurs réside dans le fait qu'il soit nécessaire de connaître le groupe de chaque locuteur *a priori* pour que cette approche fonctionne.

Le problème des "réparations" ne concerne pas simplement le domaine du dialogue homme-machine, mais également le domaine de la reconnaissance automatique de la parole, avec, entre autres, des travaux motivés par l'amélioration des alignements forcés (voir partie 1.3.2.3) entre le signal acoustique et la transcription manuelle qui lui est associée (la performance des modèles acoustiques dépend en partie de cet alignement). Le problème se situe parfois au niveau des transcriptions manuelles fournies : celles-ci ne contiennent que les mots, les autres phénomènes n'étant pas annotés. Or, lorsque l'on se trouve face à des disfluences, il est nécessaire d'avoir une annotation précise de ces phénomènes : dans le cas contraire, l'alignement forcé peut ne pas fonctionner de manière optimale. La méthode proposée par [Stolcke 2004] cherche à corriger ces transcriptions manuelles de faible qualité, dans l'optique de réaliser des alignements plus précis. Il est à noter que cette approche, au contraire de celles présentées précédemment, cherche plutôt à ajouter des disfluences au lieu de les supprimer. L'idée centrale de cette méthode n'est pas d'aligner les mots sur le signal acoustique, mais de créer un graphe de mots. L'alignement est appelé "flexible" car le graphe de mots est assez souple (possibilité d'ignorer un mot du graphe par exemple). Un décodage au moyen d'un système de RAP devrait permettre d'obtenir une meilleure transcription, qui sera ensuite utilisée pour apprendre de nouveaux modèles acoustiques. Finalement, le taux d'erreur-mot du système a pu être diminué grâce à cet alignement flexible, ajoutant des disfluences dans les transcriptions.

2.3 Impacts et solutions pour la reconnaissance de la parole

Comme nous avons pu le constater, les spécificités multiples de la parole spontanée rendent sa transcription plus difficile. Bien que les études précédentes tendent à surmonter ces particularités, des taux d'erreur-mot (WER) plus élevés sont obtenus par les meilleurs systèmes de RAP lorsqu'ils doivent faire face à de la parole spontanée [Furui 2003]. Les spécificités de la parole spontanée sont directement liées à cette baisse de performance. En effet, une analyse quantitative menée par [Kawahara 2003] sur le corpus japonais de parole spontanée "Corpus of Spontaneous Japanese" (CSJ) montre que les disfluences, la perplexité des mots (augmentation des mots hors-vocabulaire par exemple) ainsi que le débit de parole ont un impact négatif sur le WER des systèmes de RAP. Pour améliorer les transcriptions fournies en sortie de ce type de parole, une solution envisagée par la communauté scientifique est d'adapter les systèmes à ce type de parole. L'adaptation au niveau des systèmes de RAP peut intervenir à différents niveaux, puisque la modélisation de la "parole" est intégrée dans différents composants. De plus, les auteurs dans [Adda-Decker 2004], en étudiant des émissions de débats (style d'émission favorisant les disfluences), cherchent à connaître l'impact des disfluences sur les transcriptions de parole. Ils concluent en affirmant que les prononciations de taille réduite (peu de phonèmes) sont une source d'erreurs pour les systèmes de RAP. Il convient alors de mieux modéliser, dans les différents composants des systèmes de RAP, les spécificités de ce type de parole.

Nous nous intéressons dans cette partie aux travaux réalisés au niveau de la modélisation acoustique, de la modélisation du langage, et du dictionnaire de prononciations des systèmes de RAP pour améliorer la transcription de la parole spontanée.

2.3.1 Modélisation acoustique

Au regard des phénomènes spécifiques à la parole spontanée présentés dans la partie 2.1, l'acoustique est particulièrement touchée par ces différences. Au niveau des systèmes de RAP, et plus précisément lors de la reconnaissance d'émissions de journaux d'information (thèmes variés, grand vocabulaire, locuteurs inconnus à l'avance...), les modèles généralement utilisés ne sont pas assez robustes pour traiter tous les types de parole [Furui 2003, Shriberg 2005]. Dans [Nakamura 2008], les auteurs cherchent les différences (au niveau acoustique) entre la parole lue et la parole spontanée, afin de comprendre la raison d'une telle baisse au niveau des performances entre ces types de parole. Ils analysent alors de manière quantitative les caractéristiques acoustiques entre deux corpus japonais, "Corpus of Spontaneous Japanese (CSJ)" et "Japanese Newspaper Article Sentences (JNAS)", le premier contenant principalement de la parole spontanée, et le second des textes lus. Les travaux et expériences menés leur ont permis de conclure que ces différences sont visibles au niveau de l'espace spectral, où un espace

spectral plus réduit est constaté pour la parole spontanée. Finalement, cette réduction de l'espace spectral est, pour les auteurs, la raison principale de cet écart de performance.

Le "Corpus of Spontaneous Japanese (CSJ)", introduit dans le paragraphe précédent, fait partie d'un grand projet japonais débuté en 1999 : "Spontaneous Speech : Corpus and Processing Technology". Ce projet, orienté sur la parole spontanée, avait pour objectif de créer manuellement un large corpus de parole spontanée (le corpus CSJ, avec principalement des monologues, type présentations orales), mais aussi d'améliorer les performances des systèmes de RAP sur ce type de parole (principalement au niveau des modèles acoustiques et linguistiques), et de créer un système de résumé automatique de parole spontanée. Ce projet ambitieux a permis de poser certains problèmes de la parole spontanée dans les systèmes de RAP, et de tenter de fournir des solutions pour remédier à ces difficultés. Dans [Furui 2000], le corpus CSJ a permis de construire des modèles acoustiques spécifiques pour la parole spontanée. Partant du constat que les modèles actuels utilisés en reconnaissance de la parole ne permettaient pas de traiter tous les types de parole, les auteurs ont cherché à simplement entraîner de nouveaux modèles acoustiques sur des données entièrement "spontanées". Différentes expériences de transcription de la parole ont été menées sur un corpus de test contenant des présentations spontanées de dix locuteurs "hommes" (environ 5 heures de parole), en créant, au niveau acoustique, deux modèles : un modèle appris sur de la parole lue (environ 40 heures), et un modèle appris sur des présentations spontanées (environ 59 heures). Les résultats de ces expériences sont sans appel ; le modèle appris sur la parole spontanée est, comme attendu, bien meilleur que le modèle appris sur de la parole lue, lorsque l'on transcrit de la parole spontanée. Nous verrons dans la partie suivante que cette même spécialisation peut être réalisée au niveau linguistique. Les travaux entrepris dans [Furui 2000] peuvent être retrouvés dans [Shinozaki 2001, Furui 2003]. L'accent est plutôt mis dans ces travaux sur le résumé automatique de présentations orales.

Continuant le travail entrepris dans [Furui 2000], les auteurs dans [Furui 2005], résument leurs avancées. Ainsi, en démontrant qu'ils pouvaient obtenir de meilleures performances en utilisant des données spécifiques d'apprentissage pour traiter la parole spontanée (environ 59 heures d'enregistrements audio), ils ont cherché à augmenter de manière notable ce corpus d'apprentissage (utilisation de 509 heures d'enregistrements audio). Les résultats obtenus sont très intéressants, puisqu'en augmentant le corpus d'apprentissage pour les modèles acoustiques, mais aussi pour les modèles de langage (voir partie 2.3.2), une amélioration du taux d'erreur-mot (WER) de 2,9 points en absolu a été constatée (passage d'un taux d'erreur-mot de 27,2 % à 25,3 %). Les auteurs concluent dans leur article que pour pouvoir gérer la parole spontanée au niveau des systèmes de RAP, il est obligatoire que les modèles utilisés soient adaptés à la parole spontanée. La taille du corpus d'apprentissage joue également un rôle déterminant au niveau des performances : plus le corpus est grand, meilleurs sont les résultats. Les auteurs

expliquent ce phénomène par la variabilité de la parole spontanée : il est nécessaire d'avoir un corpus d'apprentissage assez large pour traiter toutes les variations de ce type de parole.

D'autres techniques ont également été étudiées pour adapter les modèles acoustiques à la parole spontanée. Dans [Bertoldi 2001], les auteurs cherchent à adapter un système de RAP devant traiter des journaux radiophoniques d'information (domaines variés/nombreux locuteurs) vers un système spécifique devant gérer seulement des dialogues spontanés. Les auteurs conservent l'idée d'utiliser des modèles acoustiques adaptés à la parole spontanée, mais les modèles ne seront pas ici totalement entraînés sur ce type de parole : les modèles acoustiques généraux vont être adaptés grâce aux nouvelles données proches du domaine et de la tâche. L'adaptation réalisée dans ce travail est totalement supervisée, avec une forte intervention humaine : les données sont collectées *a priori*, et traitent directement du sujet/type de parole abordé dans les fichiers audio de test. De manière incrémentale, une faible quantité de données permet d'adapter n fois[36] le modèle acoustique général (environ 36 heures d'apprentissage) au moyen de la technique MLLR (voir partie 1.3.3.1). Les résultats avancés par les auteurs prouvent qu'adapter de manière supervisée les modèles acoustiques à la parole spontanée permet d'améliorer grandement les résultats. Les taux d'erreur-mot, sur leurs données de test[37], ont été diminués de 26,0 % et 28,4 % en relatif (en ne prenant simplement en compte que l'adaptation au niveau acoustique, un modèle de langage général étant utilisé). De plus, les auteurs ont constaté qu'il n'était pas obligatoire de posséder des données d'adaptation annotées au niveau des disfluences (bien que les résultats semblent légèrement meilleurs en possédant ces informations), le modèle acoustique générique étant assez robuste pour s'en occuper.

2.3.2 Modélisation linguistique

Prendre en compte simplement les phénomènes acoustiques de la parole spontanée ne suffit pas à résoudre tous les problèmes des systèmes de RAP. En effet, bien que nous ayons vu dans la partie précédente que de nombreux efforts aient été réalisés pour rendre les modèles acoustiques plus robustes aux particularités de la parole spontanée, il convient également de gérer les problèmes "linguistiques". Dans la partie 2.1, nous avons constaté, par exemple, que l'agrammaticalité était largement plus répandue en parole spontanée qu'en parole lue : entraîner des modèles de langages sur des textes bien formés grammaticalement ne semble donc pas, a priori, la meilleure approche pour ce type de parole. Parallèlement au travail réalisé au niveau acoustique, [Furui 2000] a également cherché à entraîner des modèles linguistiques pour la parole spontanée. Une comparaison a été entreprise sur trois modèles de langage, appris à partir

[36] Ici, un corpus de deux heures de parole audio annotée a été découpé en 4 sous-corpus de 30 minutes.

[37] Deux adaptations avec des données différentes ont été réalisées, le corpus de test traitant de deux domaines différents.

de trois sources de données différentes : un premier modèle appris à partir des données textuelles provenant d'Internet (environ 2M de mots), un second interpolant (de manière équiprobable) les données textuelles du premier modèle avec des textes provenant de livres traitant de parole spontanée (environ 63k mots), et enfin un dernier modèle entraîné à partir de transcriptions manuelles de présentations scientifiques (issues du corpus CSJ et contenant environ 1,5M de mots). Comme pour les modèles acoustiques, les meilleurs résultats sont obtenus avec le modèle linguistique appris sur des données spécifiquement spontanées (et qui sont très proches des données de test). Il est cependant intéressant de noter que les auteurs ont trouvé des disparités au niveau des résultats entre les différents locuteurs des données de test : les spécificités de la parole spontanée rendent la transcription de ce type de parole plus difficile selon le locuteur (débit de parole, quantité de disfluences...). Tout comme pour les modèles acoustiques, les auteurs ont cherché à analyser l'influence de la taille du corpus d'apprentissage spécifique à la parole spontanée sur les résultats des systèmes de RAP. Dans [Furui 2005], les données d'apprentissage sont passées de 63k mots à 6,84M de mots. Comme évoqué dans la partie précédente, le WER a globalement diminué de 2,9 points en absolu (en utilisant les modèles acoustiques et linguistiques spécifiques, avec le corpus d'apprentissage augmenté), ainsi qu'une baisse de 62 % du taux de mots hors-vocabulaire.

Les modèles de langage peuvent également, comme l'ont été les modèles acoustiques, être adaptés de manière supervisée. La seconde partie de [Bertoldi 2001] concernait l'adaptation supervisée des modèles de langage, toujours de manière incrémentale[38] grâce à l'application d'un schéma d'interpolation. Les auteurs voulaient montrer qu'une adaptation des modèles à un domaine particulier, même avec peu de données, améliorait les performances des systèmes de RAP. Des expériences ont été menées avec un corpus, spécialisé au domaine traité, d'environ 15k mots (2 heures d'enregistrements audio), découpé en 4 sous-corpus égaux. Chacun de ces sous-corpus, ajouté de manière incrémentale (30 minutes, puis 1 heure, 1h30, et enfin 2 heures) a ensuite été interpolé avec le modèle de langage général (environ 215M de mots) afin de créer des modèles adaptés. Les auteurs ont constaté une baisse du WER de 44,3 % sur le corpus de test grâce à l'utilisation d'un modèle de langage adapté au domaine traité. Ces résultats peuvent s'expliquer par des modèles de langage qui semblent mieux appropriés pour traiter un domaine particulier (amélioration de la perplexité et baisse des mots hors-vocabulaire).

Bien que les études présentées soient relativement récentes, le problème de l'adéquation entre les modèles de langage et la parole spontanée a débuté quelques années auparavant. Déjà, dans [Suhm 1994], améliorer les modèles de langage était une nécessité pour obtenir

[38] Plusieurs sous-corpus spécialisés au niveau du domaine et du type de parole ont été choisi pour spécialiser les modèles.

de meilleures performances dans les systèmes de RAP, et ce dans de nombreuses langues[39]. La motivation initiale de cette étude, se déroulant dans le projet JANUS[40], est la traduction automatique de "*parole à parole*"[41] (connue sous le terme *speech-to-speech translation*) de dialogues de parole spontanée. Dans ce processus de traduction automatique se trouve la reconnaissance automatique de la parole : le dialogue est initié par des humains, qui, si l'on veut traduire ce qu'ils viennent de produire, doit être transcrit au moyen d'un système de RAP. Au niveau du modèle de langage, qui n'est, selon les auteurs, pas adapté à la parole spontanée, un travail de prise en compte des spécificités de ce type de parole a été réalisé. Trois pistes sont étudiées :

- En utilisant un modèle de langage à base de classes (appelé *cluster based language model*), qui présente l'avantage de nécessiter une quantité de données d'apprentissage plus faible que les modèles de langage classiques à base de mots. En se basant sur la méthode proposée par [Kneser 1993], l'approche proposée est d'apprendre automatiquement des classes de mots en minimisant la perplexité (sur des données non incluses dans les données d'apprentissage) du modèle de langage à base de classes entraîné.
- En utilisant un modèle de langage à base de groupes de mots (appelé *word phrase based language model*). L'idée est de ne pas considérer les mots de manière indépendante, mais plutôt de les considérer comme des groupes de mots s'ils sont utilisés fréquemment[42]. La sélection de ces groupes de mots est réalisée de manière automatique, et, comme l'approche précédente, en cherchant à minimiser la perplexité du modèle de langage.
- Enfin, en utilisant *a priori* un analyseur sémantique afin de catégoriser les différentes parties de chaque phrase (les catégories peuvent être des "interjections", des "informations temporelles"...), pour ensuite créer des modèles de langage (ici, bigrammes) à l'intérieur des catégories (si n catégories, n modèles de langage bigrammes seront entraînés), et entre les catégories (probabilité de voir apparaître une catégorie sachant la catégorie précédente). La baisse de la perplexité est toujours la mesure utilisée.

Finalement, bien que ces approches aient à chaque fois permis de réduire la perplexité des modèles de langage, les résultats des expériences réalisées au moyen d'un système de RAP et évaluées en taux d'erreur-mot, sont plus mitigés. En effet, en prenant en tant que système de référence l'utilisation d'un modèle de langage bigramme classique, seul le modèle de langage à base de groupes de mots a permis d'améliorer les résultats, les autres approches augmentant le WER.

[39] L'étude menée dans [Suhm 1994] traite de l'anglais, de l'allemand et de l'espagnol.

[40] `http://www.is.cs.cmu.edu/mie/janus.html`

[41] L'objectif est de partir d'un signal audio et de fournir de manière automatique et à travers divers processus, une traduction en sortie des mots prononcés dans le signal en entrée. La traduction en sortie est parlée, c'est-à-dire qu'elle est fournie par un système de synthèse de la parole.

[42] Les auteurs dans [Suhm 1994] donnent l'exemple du groupe de mots "I'll be out of town" où "out of town", très fréquent, est considéré comme un seul mot.

2.3.3 Dictionnaire de prononciations

Dans le cadre de la reconnaissance de la parole, le dictionnaire de prononciations est très important puisqu'il permet de faire le lien entre le mot et sa représentation phonémique (voir partie 1.3.2.2). Il convient d'adapter ce dictionnaire à la tâche de reconnaissance visée, qui sera par exemple différent si l'on se trouve dans le cadre de la reconnaissance à grand vocabulaire, comme dans le cas des journaux d'information (modélisation des liaisons inter-mots, changements de phonèmes...), ou au contraire dans le cas de mots isolés. Dans la partie 2.1, nous avons pu constater que les phénomènes propres à la parole spontanée amènent les prononciations "standardisées" (régies par de règles de prononciation strictes) à changer. Un même mot peut alors être prononcé différemment, si l'on est dans un contexte de parole préparée, lue, ou si l'on est plutôt dans un contexte de parole spontanée, construite au fil de la pensée. Ces différences peuvent s'expliquer à plusieurs niveaux. Ainsi, Fosler-Lussier dans [Fosler-Lussier 1999] a analysé les variations des prononciations utilisées par un locuteur lorsque son débit de parole varie, dans le cadre d'un contexte conversationnel. Il a montré que le débit de parole influe sur les variantes de prononciation utilisées, ce qui rend la transcription de ce type de parole très difficile par les systèmes de RAP, puisqu'ils ne prennent pas en compte toutes ces variations phonétiques, surtout au niveau de la modélisation et de l'apprentissage de ces phénomènes. De plus, l'état émotionnel du locuteur peut également être un facteur influant sur la manière dont il va parler. Dans [Polzin 1998], les auteurs montrent que la précision des systèmes de RAP varie significativement en fonction de ce facteur émotionnel. Les auteurs ont constaté que cette baisse au niveau des performances était due aux prononciations des mots employées par les locuteurs, qui pouvaient changer énormément selon leur état émotionnel. Ces différences sont, de plus, visibles dans différentes langues, comme le montre les recherches effectuées par exemple en allemand [Sloboda 1996], ou encore en mandarin [Byrne 2001]. La langue française n'échappe pas à ces variations au niveau des prononciations [Adda-Decker 2008].

Les dictionnaires de prononciations doivent donc prendre en compte ces variations au niveau de la prononciation [Strik 1999] et se doivent d'être adaptés, tout comme doivent l'être les modèles acoustiques et modèles de langage. La modélisation des différentes prononciations possibles d'un mot est très importante, puisqu'en l'absence de sa prononciation effective, il apparaît très difficile pour un système de reconnaissance de la parole de transcrire correctement le mot. Les variantes de prononciation ont également un impact au niveau de la segmentation automatique des systèmes de RAP (voir partie 1.5.2.1), et se doivent d'être modélisées en fonction du type de parole. En effet, dans [Kipp 1997], les auteurs démontrent que l'utilisation de variantes de prononciation adaptées à la parole spontanée permettent d'obtenir de meilleures performances lors de la tâche de segmentation automatique, en opposition à l'utilisation de variantes de prononciation génériques. Durant leurs expériences, le dictionnaire de prononciations

créé au moyen de règles linguistiques classiques, obtient des résultats inférieurs au dictionnaire de prononciations créé statistiquement en utilisant un corpus annoté manuellement de parole spontanée. Des études se sont intéressées à la modélisation des variantes de prononciation dans la cadre de la parole spontanée. La modélisation des prononciations des différents mots a fait l'objet de multiples recherches, et peut s'effectuer de plusieurs façons : en les créant manuellement (par exemple [Gauvain 1994b]), en les obtenant de manière automatique ([Byrne 1997]), ou encore en utilisant conjointement ces deux approches ([Riley 1999]).

Nous allons nous intéresser, dans cette partie, à différentes recherches effectuées pour créer des dictionnaires de prononciations adaptés à la parole spontanée, soit en utilisant des approches guidées par les données, en utilisant généralement des données d'apprentissage et en essayant d'inférer des prononciations à partir de celles-ci, soit en utilisant des approches à base de connaissances, c'est-à-dire des approches nécessitant des expertises linguistiques.

2.3.3.1 Approche guidée par les données

Appliquer des règles phonétiques nécessite préalablement une expertise humaine coûteuse. Des recherches se sont penchées sur l'obtention automatique des variantes de prononciation, en proposant des approches guidées par les données. Les auteurs, dans [Sloboda 1996], ont ainsi cherché à modéliser les phénomènes propres à la parole spontanée, en essayant d'étendre et d'adapter les dictionnaires phonétiques à la reconnaissance de la parole spontanée. Ils partent du postulat qu'il vaut mieux modéliser, dans le dictionnaire, les prononciations les plus fréquentes et les mieux adaptées à un type de parole, plutôt que d'utiliser la prononciation "correcte" d'un mot. En effet, les meilleures prononciations ne sont pas forcément les plus usitées. De plus, plus la taille du dictionnaire de prononciations augmente, plus la difficulté de reconnaissance pour les systèmes de RAP augmente. Il ne suffit donc pas de modéliser toutes les prononciations possibles d'un mot mais de les contrôler pour ne conserver que les prononciations qui seront effectivement prononcées dans le document audio à transcrire. Pour ce faire, les auteurs proposent une approche guidée par les données (*Data-Driven Approach*) afin d'ajouter dans le dictionnaire des nouvelles variantes de prononciation, qui seront le reflet des occurrences de mots observées dans les données d'apprentissage. L'idée générale de cet algorithme est de récupérer toutes les variantes de prononciation possibles d'un même mot à partir d'un corpus d'apprentissage contenant les documents audio ainsi que leur transcription manuelle. Un alignement des mots sur le signal acoustique est réalisé pour récupérer les différentes prononciations. Finalement, leur technique permet de n'ajouter dans le dictionnaire que les prononciations les plus performantes pour la tâche de reconnaissance de la parole.

Plusieurs autres études se sont penchées sur l'obtention de variantes de prononciation spécifiques à la parole spontanée, en se servant notamment d'une base d'apprentissage pour créer

automatiquement de nouvelles variantes de prononciation. Dans le travail exploratoire proposé par [Byrne 1997, Byrne 1998], l'idée est d'utiliser des transcriptions phonétiques annotées manuellement (mots + variantes de prononciation) afin de pouvoir s'en servir comme base pour inférer automatiquement de nouvelles variantes de prononciation. Ce travail est totalement orienté sur la parole spontanée et motivé par le fait que ce type de parole présente une plus grande variabilité au niveau des variantes de prononciation. Les auteurs explorent les arbres de décision de deux façons différentes pour définir de nouvelles variantes de prononciation :

- En utilisant des arbres de décision entraînés sur une portion de données d'apprentissage manuellement annotées (variantes de prononciation), afin de définir statistiquement des règles de phonétisation (au moyen des arbres de décision), qui permettront de fournir de nouvelles prononciations au dictionnaire. L'intérêt de cette approche est de permettre la généralisation des données d'apprentissage, et de pouvoir traiter les mots inconnus (non vus dans le corpus).
- En ne conservant que les variantes de prononciation les plus fréquemment rencontrées dans le corpus d'apprentissage. La méthode précédente est appliquée dans un premier temps pour obtenir de nouvelles variantes de prononciation. Puis, au moyen d'un alignement forcé sur le corpus d'apprentissage, seules les prononciations les plus fréquentes sont conservées. Cette méthode est potentiellement plus robuste que la précédente, puisque les données ont été "observées" et non simplement inférées à partir d'une portion de données manuellement annotées. Il faut également noter que cette approche propose un nombre de variantes de prononciation moindre que la précédente. Cependant, cette proposition ne peut traiter que les mots observés dans le corpus d'apprentissage.

Afin de réduire la confusion entre les différentes variantes de prononciation homophones créées (prononciations identiques mais graphies différentes, voir partie 3.1), un poids pour chaque variante est fourni (en fonction de sa fréquence d'apparition), ainsi qu'une contextualisation des prononciations (certaines prononciations ne peuvent se réaliser que si un mot, précédent ou suivant, est présent). Les performances obtenues sont bonnes, les auteurs ayant constaté un gain de 0,9 points en absolu sur leur taux d'erreur-mot (WER) de référence (utilisant un dictionnaire de base) pour les deux méthodes. Il faut cependant noter, pour la première méthode, que pour obtenir ces résultats, des phases de ré-apprentissage ont été nécessaires[43]. Enfin, les modèles acoustiques ont été ré-entrainés en utilisant le dictionnaire de prononciations obtenu au moyen

[43]Le premier dictionnaire de prononciations, créé au moyen de données manuellement annotées, est ensuite utilisé pour décoder les données d'apprentissage (au moyen de modèles acoustiques de base) et pour obtenir une transcription automatique. Ces données, acquises automatiquement, servent ensuite de ré-entrainer les arbres de décision, fournissant en résultat un nouveau dictionnaire. Cette méthode a permis de ré-entrainer le dictionnaire deux fois. Le dernier dictionnaire obtenu est celui présentant les meilleurs résultats.

de la première technique présentée (arbres de décision + données manuellement annotées). Une réduction du WER de 2,2 % en absolu a été observée.

Cette étude sur les variantes de prononciation en parole spontanée peut également se retrouver dans l'article de [Riley 1999]. Les auteurs y décrivent, de manière plus précise, plusieurs approches statistiques pour la construction d'un dictionnaire adapté à la parole spontanée, toujours en inférant automatiquement les prononciations à partir du corpus d'apprentissage disponible. La méthode statistique choisie dans cet article consiste à appliquer un arbre de décision basé sur des modèles de prononciation spécifiques à la parole spontanée (obtenus à partir de données d'apprentissage manuellement annotées), proposant des prononciations alternatives. Dans cet article, les expériences y sont plus détaillées et d'autres expérimentations ont été entreprises, notamment sur le modèle de langage. De nouveaux résultats positifs sont présentés, allant jusqu'à 2,7 % de réduction du WER.

Bien que ces approches soient intéressantes car peu coûteuses au niveau "expertise humaine" (un faible corpus d'apprentissage est souvent nécessaire), il convient de noter qu'il est très difficile de contrôler les variantes de prononciation créées, et que leur correction nécessiterait une vérification par un expert. Une autre perspective, que nous présenterons dans la partie suivante, vise à créer des dictionnaires de prononciations en s'appuyant sur des connaissances linguistiques.

2.3.3.2 Approche à base de connaissances

L'utilisation de techniques statistiques, à partir de données d'apprentissage, pour inférer de nouvelles prononciations permet d'obtenir des résultats intéressants, en minimisant l'expertise humaine. Cependant, ces approches fournissent des prononciations difficilement vérifiables et corrigibles[44]. La prononciation, en tant qu'objet d'étude linguistique, est étudiée depuis des dizaines d'années. Les dictionnaires des systèmes de RAP ont profité de cette expertise linguistique pour définir des variantes de prononciation, afin d'obtenir les meilleurs performances de transcription automatique.

Il est assez difficile de trouver des travaux traitant de cette approche. En effet, construire des dictionnaires de prononciations au moyen de connaissances linguistiques ne nécessite pas de techniques innovantes (des publications scientifiques n'étant pas d'un grand intérêt), mais plutôt l'utilisation d'expertises sur la langue. Pour pouvoir trouver des travaux sur cette approche, il faut regarder au niveau des articles comparatifs entre l'approche à base de connaissances et l'approche guidée par les données (présentée dans la partie 2.3.3.1). Ainsi, dans [Wester 2000], les auteurs cherchent à comparer les performances de ces deux approches sur un corpus en

[44]Leur performance est souvent évaluée en terme de taux d'erreur-mot, et leur correction nécessiterait une expertise humaine.

langue danoise sur des dialogues de renseignements d'horaires de train entre un homme et une machine. Bien que ce corpus ne puisse être directement considéré comme de la parole complètement spontanée [45], certains de ses problèmes peuvent s'y retrouver, notamment une variation au niveau des prononciations employées. L'approche à base de connaissances, qui nous intéresse dans cette partie, est décrite dans cet article. L'idée de cette approche est de manipuler conjointement des dictionnaires de prononciations (réalisés manuellement) et des études linguistiques (au moyen de règles automatiques). Notons que dans les expériences menées dans cet article, une amélioration très faible a été constatée par rapport au système de référence utilisant l'approche à base de connaissances, alors que l'approche guidée par les données permettait d'obtenir une amélioration significative du WER. Cependant, un gain est possible, car l'article de [Kessens 1999], traitant du même corpus, a vu les performances de reconnaissance s'améliorer avec une approche à base de connaissances.

Au niveau des articles spécifiques à la parole spontanée, il faut s'intéresser aux particularités de ce type de parole, avec par exemple [Adda-Decker 1999] qui se penche sur les phénomènes du schwa et de l'élision. Le dictionnaire de prononciations peut donc être une solution pour gérer ces spécificités au niveau des systèmes de RAP.

En conclusion, bien que l'approche à base de connaissances soit peu détaillée en reconnaissance automatique de la parole, et particulièrement pour la parole spontanée, il semble que les travaux effectués se dirigent vers une combinaison des deux approches présentées. En effet, lorsque les systèmes de RAP doivent traiter des journaux radiophoniques d'information (large vocabulaire, sujets variés, nombreux types de parole…), il est nécessaire d'avoir les dictionnaires de prononciations les plus efficaces possibles. Pour pouvoir construire de tels dictionnaires (au minimum supérieurs à 60k mots), il convient d'utiliser des méthodes automatiques pour les construire. Mais l'expertise humaine reste indispensable pour contrôler les variantes de prononciation [Gauvain 2005].

2.4 Conclusion

Dans ce chapitre, nous nous sommes intéressés à la parole spontanée, en opposition à la parole préparée (proche d'un texte lu). La parole spontanée possède de multiples spécificités qui la rendent unique, mais qui font que, dans le cadre de la reconnaissance automatique de la parole, sa transcription est beaucoup plus difficile. De nombreuses études se sont appliquées à mieux comprendre ce type de parole, en essayant tout d'abord de détecter automatiquement ses particularités, et/ou de les corriger. Les objectifs de ces travaux étaient multiples :

[45] L'interaction entre un homme et une machine n'est pas naturelle, le dialogue est donc différent d'un dialogue homme-homme.

– Supprimer certaines particularités de la parole spontanée (disfluences, répétitions, faux-départs…) des transcriptions, afin d'améliorer les traitements ultérieurs sur ces mêmes transcriptions (par exemple pour le dialogue homme-machine).
– Détecter les disfluences pour améliorer les performances des systèmes de RAP.

La détection et correction de ces phénomènes ont eu un impact positif sur les applications en traitement automatique des langues.

Dans le cadre de la reconnaissance automatique de la parole, les études se sont principalement portées sur l'amélioration de trois composants principaux des systèmes pour traiter la parole spontanée : les modèles acoustiques, les modèles de langage, et enfin les dictionnaires de prononciations. En prenant en compte les particularités de la parole spontanée, et donc en adaptant ces modules à ce type de parole, une amélioration des performances des systèmes de RAP a pu être constatée. Cependant, les différents travaux présentés concluent fréquemment que les performances des systèmes de RAP sur la parole spontanée sont encore loin des performances obtenues en parole préparée.

Chapitre 3

Homophonie

Sommaire

3.1 Description générale 64
3.1.1 Mots homophones sémantiquement différents 65
3.1.2 Mots homophones sémantiquement identiques 66
3.2 Systèmes de RAP et homophonie 67
3.2.1 Quelques particularités du français 67
3.2.2 Analyse des erreurs d'homophonie 68
3.3 Méthodes automatiques appliquées aux erreurs de reconnaissance des systèmes de RAP 70
3.3.1 Approches statistiques globales 70
3.3.1.1 Détection automatique des erreurs 70
3.3.1.2 Correction automatique des erreurs 72
3.3.2 Approches ciblées sur les homophones 74
3.3.2.1 Approches par règles linguistiques 74
3.3.2.2 Approches statistiques 77
3.3.3 Combinaison des approches 80
3.4 Conclusion 80

L'homophonie est une particularité du langage ; des mots différents peuvent se prononcer exactement de la même manière. Dans certains cas, ces mots homophones possèdent une écriture différente. Nous pouvons alors parler de mots *homophones hétérographes*. La seule façon de choisir le mot correct est de connaître le contexte dans lequel ce mot est utilisé. Pour les êtres humains, qui doivent faire face fréquemment à ce phénomène, cette désambiguïsation se fait naturellement. Or, si pour un être humain ce problème est relativement bien traité, il est difficile de connaître de manière précise tout ce qui entre en jeu pour choisir le mot correct, et particulièrement en extraire des règles. De plus, ce problème semble relativement fréquent et touche plusieurs langues. Les auteurs, dans [Gauvain 2005], rapportent un taux de mots homophones sur des textes écrits de journaux d'information, de 20 % en anglais, et même de 75 % en français.

Le problème des homophones hétérographes, que l'on retrouve dans le domaine de la reconnaissance de la parole, est difficile car les systèmes de reconnaissance automatique de la parole (RAP), de part leur conception initiale, ont souvent des difficultés à lever une ambiguïté entre des mots homophones [Forsberg 2003]. Le fait de ne pas posséder de règles précises pour résoudre ce problème le rend encore plus complexe pour un traitement automatique. Pour pouvoir gérer cette particularité, il semble inévitable de mettre au point d'autres techniques, en prenant notamment en compte de nouvelles sources d'information (classes grammaticales, contexte élargi…).

Dans ce chapitre, nous chercherons dans un premier temps à comprendre, de manière générale, l'homophonie et ses particularités. Nous nous focaliserons essentiellement, dans cette présentation, sur les mots homophones hétérographes, qui, comme nous le verrons ensuite, posent souvent problème au niveau du traitement automatique des langues. Puis, nous entrerons plus en détail sur le problème de l'homophonie dans les systèmes de RAP en nous intéressant à l'homophonie dans le cadre de la langue française. Nous verrons que des études ont été réalisées pour comprendre les erreurs faites par les systèmes de RAP, et que des solutions ont été envisagées pour résoudre ce problème, que ce soient des solutions globales pour traiter toutes les erreurs, ou plus spécifiques, pour ne s'intéresser qu'à l'homophonie.

3.1 Description générale

De manière générale, plusieurs types d'erreurs, au niveau des mots, peuvent apparaître dans les textes. Cependant, selon le type d'erreur, l'impact auprès des utilisateurs n'est pas le même. En effet, certaines erreurs modifient le sens de la phrase et peuvent alors être très pénibles pour la compréhension globale. Ces erreurs empêchent des retours corrects de la part des utilisateurs,

le sens de la phrase pouvant être altéré. En revanche, d'autres erreurs, ne gênant pas la compréhension, sont souvent négligées car elles ne sont pas indispensables pour comprendre une phrase.

Dans le cas de l'homophonie, ces deux types d'erreurs sont possibles. Par définition, deux mots sont considérés comme homophones[46] lorsqu'ils se prononcent exactement de la même manière, mais possèdent une graphie différente (la manière dont ils sont écrits). En fait, ces mots homophones ne peuvent se différencier que dans le contexte d'une phrase ou d'un groupe de mots, leur simple énonciation orale n'étant pas suffisante. De plus, dans certaines langues, des mots, qui ne sont pourtant pas, à la base, homophones selon leur définition linguistique, le deviennent dans un contexte précis car ils sont très proches phonétiquement (par exemple en contexte de parole spontanée [Vasilescu 2009]). Afin de bien comprendre la manière dont fonctionnent les mots homophones, leur contexte d'apparition et leur difficulté, nous verrons, dans un premier temps, les mots homophones ne partageant pas le même sens. Dans une seconde partie, nous nous intéresserons aux mots homophones possédant le même sens.

3.1.1 Mots homophones sémantiquement différents

Ce phénomène d'homophonie où deux (voire plusieurs) mots peuvent avoir la même prononciation mais des graphies et des sens différents[47], est présent dans de très nombreuses langues. Pour pouvoir différencier ce type de mots homophones, il est nécessaire de connaître leur contexte d'utilisation (place dans la phrase, catégorie grammaticale, sujet abordé au cours de la conversation...), mais également, dans certains cas, les règles grammaticales associées. Afin d'illustrer notre propos, nous allons voir deux cas d'étude de l'homophonie, en français. Le premier cas se base sur le contexte de la phrase pour différencier les mots homophones, et le second cas, bien qu'utilisant également le contexte, nécessite la prise en compte d'une règle grammaticale :

- *"Le roi a apposé son **sceau**"* : le mot *sceau* a pour mots homophones possibles les mots *sot*, *saut* ou encore *seau*. Tous ces mots ont les mêmes propriétés grammaticales : nom masculin/singulier. Ici, les mots *roi* et *apposé* permettent de choisir le mot *sceau*, puisque, dans ce contexte, c'est le mot ayant le sens le plus approprié.
- *"L'enfant aime **ce** jeu"* : le mot *ce* est homophone avec le mot *se*. Les deux mots ne remplissent cependant pas la même fonction grammaticale dans une phrase : *ce* est un pronom démonstratif, et *se* un pronom personnel. Une règle grammaticale précise que le pronom personnel *se* ne peut être placé que devant un verbe ou un pronom. Or, dans ce

[46] En étant plus précis, nous pouvons parler de mots *homophones hétérographes*.

[47] Ce type d'homophonie ne doit pas être confondu avec les mots *homophones homographes*, qui peuvent avoir un sens différent tout en possédant la même graphie.

cas précis, si une analyse grammaticale de la phrase est réalisée, le mot suivant est un nom commun : le mot utilisé ne peut donc être que le pronom démonstratif *ce*.

3.1.2 Mots homophones sémantiquement identiques

Les mots sémantiquement identiques sont sûrement les mots homophones les plus difficiles à différencier. En effet, outre le fait qu'ils soient phonétiquement identiques, il faut généralement étudier de manière précise et détaillée la phrase, pour en extraire les informations permettant de choisir le bon mot homophone. Ces mots sémantiquement identiques peuvent se retrouver au niveau :

- Des déclinaisons des verbes : selon le temps (présent, passé simple, futur...), la personne (1ère personne du singulier, 2ème personne du pluriel...) ou encore son groupe (1er groupe, 2ème groupe ou 3ème groupe). Par exemple, si nous prenons le cas du français, et du verbe "manger". Nous trouvons que des déclinaisons de ce verbe sont homophones et hétérographes ([mãʒɛ]) à la 1ère personne du singulier de l'imparfait ("mangeais"), à la 3ème personne du pluriel de l'imparfait ("mangeaient"), ou encore à la 1ère personne du singulier du passé simple ("mangeai").
- Sur les accords en genre et en nombre, que l'on peut retrouver sur les noms, les adjectifs, les participes passés...Sans entrer dans les détails (nous y reviendrons dans la partie 3.2.1), si nous prenons une nouvelle fois la langue français, les possibilités d'écriture de ces catégories grammaticales de mots varient selon le genre (féminin/masculin) et/ou le nombre (singulier/pluriel). Il est alors possible d'avoir, dans certains cas, deux, et jusqu'à quatre, combinaisons différentes pour un même mot. Prenons l'exemple du nom commun "enfant" : il est homophone au singulier ("enfant") et au pluriel ("enfants") avec la prononciation associée [ãfã]. Si nous nous intéressons au participe passé du verbe "manger" ([mãʒe]), pouvant s'accorder en genre et nombre, ses quatre formes sont homophones "mangé" (masc./sing.), "mangés" (masc./plu.), "mangée" (fem./sing.) et "mangées" (fem./plu.).

Ces types de mots homophones sont notamment sources de nombreuses fautes d'accord. En effet, pour pouvoir choisir la bonne flexion, il est nécessaire de connaître à la fois les règles de grammaire et les informations contenues dans les mots au voisinage de ce mot. Dans certaines langues, comme le français, ce problème de flexion "homophone" touche un grand nombre de mots : dans [Béchet 1999b], les auteurs ont constaté que sur un corpus de textes journalistiques, en prenant les mots pouvant apparaître au singulier et au pluriel, 72 % d'entre eux ont une flexion homophone hétérographe.

3.2 Systèmes de RAP et homophonie

Les systèmes de reconnaissance automatique de la parole (RAP) sont de plus en plus efficaces. Leurs performances actuelles sont suffisantes pour qu'ils soient utilisés dans de nombreuses applications (dialogue homme-machine, indexation, recherche d'information...). Comme nous l'avons vu dans la partie 3.1, les erreurs ne modifiant pas le sens de la phrase sont souvent négligées car elles ne sont pas cruciales pour mener à bien certaines opérations, particulièrement celles cherchant à obtenir le sens d'une phrase et non une transcription précise (dialogue, résumé automatique...). Cependant pour d'autres applications, comme le sous-titrage ou la transcription assistée [Bazillon 2008a], ces erreurs sont plus importantes : leur répétition, même si elles ne modifient pas le sens, est très fatigante pour l'utilisateur. De plus, elles peuvent décrédibiliser le système de RAP aux yeux des utilisateurs, pour qui ces erreurs semblent facilement corrigibles. Enfin, ces erreurs peuvent avoir un impact négatif sur le reste de la phrase à transcrire : en effet, comme nous le constaterons à travers diverses expériences dans la partie 5.1.5.2 et présenté dans [Dufour 2008a], un mot mal transcrit a un impact sur les mots se trouvant à son voisinage. Il convient donc de traiter ces erreurs dues à l'homophonie, qui sont, comme nous le verrons dans la partie suivante, une source importante d'erreurs en français au niveau des systèmes de RAP. Nous nous intéresserons ensuite à des analyses quantitatives des erreurs d'homophonie produites par les systèmes de RAP.

3.2.1 Quelques particularités du français

Comme nous avons pu l'entrevoir dans l'introduction de ce chapitre, l'homophonie est un phénomène rendant difficile la transcription automatique en français [Gauvain 2005]. Dans la partie 3.1, nous avons vu que l'homophonie revêt principalement deux formes principales : des mots homophones possédant le même sens (ils ne diffèrent généralement que par leur flexion) ou au contraire ayant un sens totalement différent. Dans le cadre de la Reconnaissance Automatique de la parole (RAP), la complexité de l'homophonie représente un réel problème qui n'est pas simple à gérer. En effet, si nous nous intéressons au cas de la flexion en genre et en nombre en français, celui-ci représente l'un des aspects les plus difficiles pour ces systèmes statistiques. Cette difficulté, pour les systèmes de RAP, peut s'expliquer par le fait que les différentes formes fléchies homophones d'un mot ne peuvent être correctement distinguées que par le modèle de langage (les modèles acoustiques sont inefficaces dans ce cas précis).

L'accord en genre et en nombre n'est pas toujours bien modélisé par les systèmes de RAP car, d'une part la longueur des contraintes modélisées par les modèles de langage de type n-gramme (voir partie 1.4.1) peut ne pas être suffisante, et d'autre part la taille des données d'apprentissage peut être trop faible. Ainsi, les auteurs, dans [Gauvain 1994a], montrent qu'il

faut deux fois plus de mots différents dans le vocabulaire d'un système de RAP en français qu'il n'en faudrait pour un système de RAP en anglais pour obtenir la même couverture de mots[48]. De plus, de nombreuses règles grammaticales complexes ne peuvent être modélisées avec un modèle de langage n-gramme. La figure 3.2.1 présente un exemple du verbe "adopter", qui doit, dans cette phrase, être accordé au féminin/singulier, et donc apparaître sous la forme "adoptée". La première difficulté est que le participe passé du verbe "adopter" possède 4 formes fléchies homophones (toutes prononcées [adɔpte]), en fonction du genre (masculin ou féminin) et du nombre (singulier ou pluriel). La modélisation acoustique de ce participe passé, issue du système de RAP, ne peut pas aider à trouver sa bonne flexion. La seconde difficulté, pour trouver le bon accord de ce verbe, réside dans le fait que l'information sur son genre et son nombre se situe, dans cet exemple, sur le mot féminin/singulier "réforme", distant de 6 mots : le modèle de langage n-gramme utilisé par les systèmes de RAP dépasse rarement le quadrigramme, et ne peut donc pas capturer cette contrainte trop éloignée dans l'historique. Au final, le système de RAP ne peut pas choisir précisément la bonne forme fléchie de ce mot.

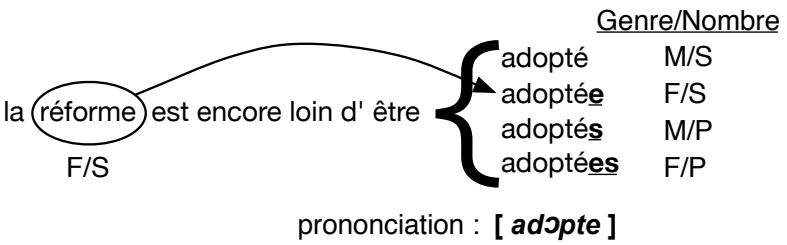

FIG. 3.1 – Exemple de formes fléchies homophones

Dans la partie suivante, nous nous intéressons à des analyses quantitatives réalisées sur les erreurs d'homophonie dans le cadre du traitement automatique des langues.

3.2.2 Analyse des erreurs d'homophonie

Les mots homophones, comme nous venons de le voir pour le français, ont besoin de leur contexte d'utilisation pour être différenciés. Le vaste domaine du traitement automatique des langues (TAL) s'est notamment penché sur ce problème d'homophonie, en essayant de l'analyser plus finement, et de comprendre comment ces mots homophones pourraient être plus efficacement reconnus. Ainsi, dans [Nemoto 2008a, Nemoto 2008b], les auteurs partent du postulat que des mots considérés comme homophones ne le sont en réalité pas complètement. De légères

[48] La couverture de mots s'exprime, dans ce cas, par le nombre de graphies différentes.

différences au niveau du signal acoustique existeraient, et des outils informatiques pourraient être capables de les capter. Les auteurs s'intéressent aux mots *et* (conjonction de coordination) et *est* (verbe *être* conjugué à la 3ème personne du singulier), ainsi qu'aux mots *à* (préposition) et *a* (verbe *avoir* conjugué à la 3ème personne du singulier). Dans l'optique de les distinguer, différentes caractéristiques acoustiques sont extraites au moyen d'un système de RAP pour chacun des mots homophones ciblés : les trois premiers formants[49], la fréquence fondamentale (F0), l'intensité, leur durée, ou encore la co-occurrence gauche/droite de pauses. Ces caractéristiques ont ensuite été regroupées et testées au sein de diverses méthodes de classification automatique. Les transcriptions manuelles de deux corpus différents ont été utilisées : le corpus de la campagne d'évaluation ESTER, contenant principalement de la parole préparée (voir partie 1.6), et le corpus Phonologie du Français Contemporain (PFC), contenant principalement de la parole spontanée. Il est intéressant de noter, que, bien que cet article ne soit pas orienté sur la parole spontanée, des analyses ont montré des résultats allant dans le sens des études présentées dans la partie 2.1. En effet, des caractéristiques extraites dans ces travaux obtiennent des valeurs moyennes différentes selon le type de parole (par exemple, la durée des mots homophones étudiés est plus courte en parole spontanée). Les résultats des expériences montrent qu'il semble possible de discriminer certains mots homophones à travers différentes caractéristiques acoustiques. Cependant, des limites ont été énoncées par les auteurs : les mots *et* et *est* ne sont pas complètement homophones, puisque des sons différents devraient normalement être utilisés. Ces différences s'atténuent lors de la parole spontanée, où les phonèmes deviennent quasiment identiques. Enfin, des transcriptions manuelles sont utilisées, ce qui est le cas le plus favorable possible (aucun mot n'est erroné, au contraire de transcriptions automatiques).

Comme dans l'étude précédente, les auteurs, dans [Vasilescu 2009], cherchent à comprendre, au moyen d'une analyse d'erreurs, les raisons qui font qu'un humain arrive à désambiguïser les mots homophones alors que les systèmes de RAP ont beaucoup plus de difficultés. L'analyse se porte, comme dans les travaux de [Nemoto 2008b], sur des mots quasi-homophones[50] courts dans deux langues, leurs confusions étant parmi les erreurs les plus fréquentes des systèmes de RAP. Ils étudient ainsi les mots homophones *et*/*est* en français et *and*/*in* en anglais. L'idée générale de cette expérience perceptive est d'extraire, dans les transcriptions automatiques des systèmes de RAP, les passages où ces mots ont été mal transcrits[51], à l'instar des passages où ces mots ont été correctement reconnus. Les trois mots précédents et suivants sont également extraits, ce qui permet d'avoir une portion de texte entourant le mot ciblé. Il est important de

[49] Un formant représente un maxima d'énergie au niveau du spectre sonore d'un son de parole. Les trois premiers formants sont particulièrement intéressants pour obtenir des informations sur les voyelles.

[50] Les prononciations varient légèrement, mais sont assez proche pour que l'on parle de quasi-homophonie.

[51] Les erreurs peuvent être multiples : suppression, insertion, ou encore substitution du mot ciblé (par un mot homophone ou non).

noter que les mots voisins peuvent être erronés car ils sont aussi extraits d'une transcription automatique. Le but est ensuite de fournir ces portions de textes à des humains pour correction et de vérifier si leur taille (7 mots) est suffisante pour corriger les mots homophones. Les auteurs ont constaté que les humains arrivent à corriger ces mots 5 à 6 fois mieux que les systèmes de RAP, et ce pour les deux langues. La question de la faiblesse du modèle de langage des systèmes de RAP a alors été soulevée, sachant que seuls des informations linguistiques ont été fournies (portion de texte). En comparant les transcriptions automatiques et manuelles sur les mots homophones étudiés au moyen d'un modèle de langage quadrigramme, il apparaît possible de mesurer l'ambiguïté de certains mots. Cependant, il faut noter que, à la différence des humains, les systèmes de RAP ne prennent pas en compte les informations sémantiques contenues dans chaque portion de texte, pouvant aider les humains dans le processus de correction.

3.3 Méthodes automatiques appliquées aux erreurs de reconnaissance des systèmes de RAP

Les erreurs produites par les systèmes de RAP peuvent avoir un impact négatif dans d'autres domaines, si ceux-ci utilisent directement les transcriptions fournies. Comme nous l'avons vu dans la partie 1.1, présentant un système classique de RAP, les modèles acoustiques, les modèles de langage, et les dictionnaires de prononciations, sont les composants principaux permettant la transcription de parole. Afin de corriger les erreurs laissées par ces composants, différentes techniques ont été envisagées. Parmi ces erreurs se trouvent, en quantité plus ou moins élevée selon la langue transcrite, les homophones. Nous nous intéresserons, dans cette partie, aux techniques envisagées pour améliorer la robustesse des systèmes de RAP, en nous focalisant sur deux approches : chercher à corriger de manière générale les erreurs produites par les systèmes de RAP, ou fournir des solutions spécifiques pour corriger les erreurs dues aux mots homophones.

3.3.1 Approches statistiques globales

3.3.1.1 Détection automatique des erreurs

En compréhension du langage naturel notamment, les erreurs de reconnaissance peuvent avoir un impact négatif important. En général, pour les corriger, les propositions cherchent à réparer globalement toutes les sortes d'erreurs. Certains travaux ont cherché à corriger ces erreurs au niveau du dialogue, après la phase de correction [Sagawa 2004]. Dans cette approche, l'interaction avec l'utilisateur permet de réparer les erreurs d'incompréhension. Les auteurs,

dans [Walker 2000], définissent, quant à eux, une méthode pour détecter automatiquement les erreurs dans le cadre d'un dialogue homme-machine en langage naturel. Pour pouvoir fonctionner correctement, un système de dialogue doit analyser ce que l'utilisateur vient de dire afin de déclencher les actions appropriées, la difficulté étant que le système doit quasiment réagir en temps réel. Or, si cette transcription est incomplète ou difficilement compréhensible, le système peut inutilement chercher une réponse, voire même se bloquer. L'approche propose d'extraire automatiquement plusieurs caractéristiques, pouvant être obtenues dans un laps de temps très court (pour respecter le temps réel du dialogue). L'objectif est de les combiner dans une tâche de classification afin de détecter les erreurs potentielles dans le dialogue. Il est intéressant de noter que, dans les caractéristiques extraites, se trouvent des informations fournies par les systèmes de RAP. Bien que cet article ne cherche pas à corriger directement les erreurs des transcriptions, il cherche néanmoins à contourner les erreurs produites par les systèmes de RAP, corrélées aux performances des systèmes de compréhension de la langue naturelle. Les mots homophones peuvent alors être vus non plus comme un simple problème de reconnaissance, mais comme un problème de compréhension.

Les erreurs en reconnaissance automatique de la parole peuvent être détectées au moyen de techniques en post-traitement du système de RAP. Les homophones peuvent faire partie des informations utiles servant à la détection de ces erreurs. La méthode employée dans [Allauzen 2007] cherche à détecter automatiquement, dans le réseau de confusion[52], les erreurs produites par le système de RAP. L'idée est de fournir, pour chaque ensemble de confusion, une mesure de confiance, qui permettra de prendre une décision sur le caractère erroné ou non de la meilleure hypothèse proposée pour chaque ensemble d'hypothèses. Cette mesure de confiance sera estimée au moyen de la méthode de régression logistique en prenant en compte trois caractéristiques différentes, provenant :

- Dans un premier temps, seulement de l'ensemble de confusion (meilleur score postérieur de probabilité, durée du mot hypothèse, nombre de mots hypothèses en concurrence...).
- Par la suite, d'un modèle de langage n-gramme, en l'appliquant sur le réseau de confusion (probabilité n-gramme du mot hypothèse, meilleur score de l'ensemble des n-grammes obtenus en prenant en compte l'ensemble de confusion courant et celui se trouvant à $n-1$...).
- Pour finir, des propriétés lexicales des mots (probabilité unigramme, nombre d'homophones, découpage syntaxique en catégorie...).

Les différentes expériences, à travers ces trois sources d'information, ont été réalisées sur des émissions radiophoniques d'information, en utilisation les 10 heures de corpus de test de la campagne d'évaluation ESTER 1 (voir la section 1.6). Les auteurs définissent, pour les mesures

[52]Le réseau de confusion contient, de manière séquentielle, les hypothèses de reconnaissance.

de confiance (CM) calculées, un seuil à 0,5 (une mesure de confiance inférieure à ce seuil considère que le mot hypothèse est erroné). Les résultats sont fournis en terme de taux d'erreurs de classification (*Classification Error Rate* (CER)), selon la formule :

$$CER = \frac{\#\ mesures\ de\ confiance\ incorrectement\ classifiées}{\#\ total\ des\ mesures\ de\ confiance} \tag{3.1}$$

Les auteurs obtiennent des résultats positifs pour chacune des trois sources de caractéristiques utilisées séparément, bien que celles directement extraites du réseau de confusion obtiennent les meilleurs résultats. La combinaison de ces trois sources permet également d'améliorer légèrement les résultats, pour au final, passer d'un CER de 17,2 % à 12,3 %.

3.3.1.2 Correction automatique des erreurs

Contrairement aux méthodes cherchant à détecter les erreurs faites par les systèmes de RAP, d'autres approches se penchent sur leur correction. Si l'on se positionne dans le domaine de la compréhension automatique de la parole, un des objectifs est d'extraire des concepts ; l'idée n'étant pas simplement d'obtenir les mots prononcés mais de comprendre ce qui a été dit[53]. Les systèmes de RAP permettent alors d'extraire, dans un premier temps, les mots prononcés (l'analyse des concepts est un module supplémentaire). Dans la littérature, nous trouvons des méthodes alternatives pour pallier les difficultés des systèmes de RAP, dont les mots homophones peuvent être problématiques. Ainsi, si un mot homophone est choisi à la place d'un autre, un concept erroné pourrait y être associé. Des solutions ont permis d'améliorer les performances des systèmes de RAP pour l'extraction de concepts [Servan 2006]. Dans ce travail, les auteurs montrent que l'approche classique, qui consiste à partir de la meilleure hypothèse en sortie des systèmes de RAP, pour ensuite chercher les concepts qui y sont associés, n'est pas la meilleure solution à envisager. En effet, les systèmes de compréhension sont directement fragilisés par les sorties erronées des systèmes de RAP. Ils proposent d'intégrer directement dans les systèmes de RAP la recherche des concepts. Pour ce faire, l'approche consiste à utiliser les graphes de mots produits par un système de RAP, d'associer dans ce graphe les mots avec des concepts, et de ré-estimer les poids de chaque entrée du graphe avec un modèle de langage de concepts. Les expériences menées par les auteurs sur un corpus de dialogue montrent qu'il est possible d'améliorer les systèmes de RAP sur des tâches devant utiliser les transcriptions automatiques. L'intégration d'informations supplémentaires à celles contenues dans les systèmes de RAP est cependant nécessaire.

[53]Cette compréhension permet notamment, dans les systèmes de dialogue, de réaliser certaines actions spécifiques.

Dans [Stolcke 1997], les auteurs partent du constat qu'en reconnaissance automatique de la parole, le principe général est de minimiser le taux d'erreurs d'une phrase (voir partie 1.1). Bien entendu, le fait de minimiser le taux d'erreurs d'une phrase conduit à minimiser le taux d'erreur-mot (évaluant classiquement les systèmes de RAP). Ils cherchent à savoir, à travers ce travail, si la minimisation du taux d'erreurs au niveau du mot entraîne les mêmes résultats que la minimisation du taux d'erreurs au niveau de la phrase, et si un algorithme était possible pour minimiser le taux d'erreur-mot (WER). Pour ce faire, un algorithme, utilisé en post-traitement des systèmes de RAP, est développé. Il doit ré-estimer les scores des n meilleures hypothèses produites, afin de prendre une décision alternative à celle proposée par le système de RAP, en espérant diminuer le WER. Le principe de cet algorithme est de calculer un taux d'erreur-mot[54] pour chaque phrase de la liste des meilleures hypothèses : une hypothèse est prise pour la phrase de référence et un WER sera calculé avec chacune des autres hypothèses restantes dans la liste. Une moyenne de ces WER sera ensuite calculée, et sera combinée avec la probabilité postérieure de la phrase prise comme référence. Ce processus sera répété tant que toutes les n meilleures hypothèses n'auront pas été évaluées en tant qu'hypothèse de référence. Au final, le choix de la meilleure hypothèse se fera sur ce WER estimé. Grâce à leurs expériences, les auteurs ont montré que le choix de la meilleure hypothèse proposée par un système de RAP n'était pas forcément la meilleure, puisqu'en utilisant leur méthode, une baisse du WER de 0,5 point en absolu est obtenue. Ces expériences confirment bien le besoin de traiter différemment les hypothèses des systèmes de RAP.

En continuant l'approche amorcée par [Stolcke 1997], les auteurs dans [Huet 2007, Huet 2010] cherchent à trouver la meilleure hypothèse dans une liste de n meilleures hypothèses fournies par un système de RAP. La méthode qu'ils proposent est réalisée en post-traitement d'un système de RAP, introduisant des informations morpho-syntaxiques dans le choix de la meilleure hypothèse. Dans une première étape, un découpage morpho-syntaxique est obtenu, au moyen d'un étiqueteur automatique, pour chaque hypothèse de la liste. L'idée est que ce découpage morpho-syntaxique contient des informations supplémentaires à celles utilisées par un système de RAP et qu'elles peuvent être utiles pour trouver la meilleure hypothèse. Au moyen d'un modèle de langage entraîné sur les catégories grammaticales, la méthode doit permettre de trouver la meilleure séquence de catégories (un modèle de langage classique utilisant une séquence de mots). Les modèles de langage n-grammes de mots, de par les contraintes actuelles, dépassent rarement les modèles quadrigrammes. Cet historique étant jugé trop faible, un modèle de langage 7-grammes est utilisé. Ce nouveau score, obtenu grâce au découpage syntaxique, sera ensuite ajouté, pour chacune des n meilleures hypothèses, aux

[54] Ce taux d'erreur-mot ne reflète pas complètement la réalité, la transcription de référence n'étant, bien entendu, pas fournie.

scores acoustiques et linguistiques (modèle de langage de mots) fournis par le système de RAP. Cette méthode a été testée sur le corpus de test de la campagne d'évaluation ESTER (voir partie 1.6). Les résultats atteints grâce à cette méthode sont positifs, puisque son utilisation en post-traitement des systèmes de RAP a permis de réduire significativement le taux d'erreur-mot. Cette désambiguïsation morpho-syntaxique permet notamment de corriger les erreurs dues à l'homophonie. L'utilisation d'informations morpho-syntaxiques semble donc importante et manque aux systèmes de RAP.

3.3.2 Approches ciblées sur les homophones

Les approches précédentes ont cherché à détecter et corriger de manière générale les erreurs des systèmes de RAP. Cependant, les erreurs dues à des mots homophones peuvent être considérées comme des erreurs particulières, et à ce titre, elles nécessitent des traitements spécifiques. Nous verrons dans les prochaines parties, différentes études et travaux réalisés spécifiquement pour tenter de corriger ces problèmes d'homophonie. Nous nous intéresserons ainsi à différentes approches tentant de corriger ces erreurs : en utilisant des règles linguistiques et des méthodes statistiques.

3.3.2.1 Approches par règles linguistiques

Des études se sont penchées sur la détection et la correction des erreurs, notamment orthographiques et grammaticales, au moyen de règles linguistiques. Les études proposent principalement des techniques pour corriger des textes. Elles proviennent principalement du domaine du traitement automatique des langues (TAL). Ces travaux ne se sont pas directement focalisés sur les problèmes spécifiques à la reconnaissance automatique de la parole, mais sont cependant intéressants pour le traitement des mots homophones. En effet, un des objectifs de ces règles est de corriger les erreurs grammaticales concernant les accords (singulier et pluriel), très souvent homophones en français. Ces règles grammaticales seraient donc potentiellement bénéfiques aux transcriptions en sortie des systèmes de RAP, particulièrement dans l'optique d'un post-traitement. Ces travaux étant assez vastes, nous nous focaliserons, dans cette partie, aux travaux les plus connus. Nous présenterons alors, de manière générale, l'approche par règles linguistiques. Nous n'évoquerons pas volontairement les erreurs autres que les erreurs grammaticales (fautes de frappes par exemple). Elles n'entrent pas dans le cadre des mots homophones et de la reconnaissance automatique de la parole.

De nombreux outils, développés dans le cadre du TAL, cherchent principalement à détecter des erreurs grammaticales et des problèmes au niveau du style (comme la ponctuation). Dans

[Bustamante 1996], les auteurs présentent un outil, appelé *GramCheck*, devant détecter et fournir un diagnostic des erreurs grammaticales rencontrées dans un texte. La langue espagnole, qui comme le français, est une langue fléchie (voir partie 3.2.1), est étudiée dans ce travail. Bien que les flexions de l'espagnol ne soient pas homophones, la difficulté peut être considérée comme équivalente au français car la correction ne se fait qu'à partir de données textuelles et non audio. Afin de pouvoir détecter les erreurs grammaticales, l'outil développé par les auteurs utilise des règles grammaticales, implémentées au niveau informatique grâce au langage de programmation logique *Prolog*.

D'autres correcteurs, plus récents, ont été mis au point. Ils sont opérationnels et largement intégrés dans des systèmes de traitement de texte (*Word*, *OpenOffice*, *Pages*...), ou en tant qu'outil complémentaire (*Cordial*[55], *Antidote*[56]...). Cependant, beaucoup de ces programmes étant destinés à la commercialisation, il est impossible de connaître précisément la manière dont ceux-ci ont été réalisés, notamment au niveau de la correction grammaticale. Cependant, il est possible, en parcourant leur site Internet de présentation, de se faire une idée globale de certaines approches. En effet, le logiciel *Cordial* intègre un module de correction automatique des fautes les plus fréquentes[57], que les auteurs présentent comme "*le fait de corriger les fautes dans le texte sans aucune intervention de l'utilisateur (...) il est également possible de corriger beaucoup de fautes de grammaire automatiquement*". Ce correcteur semble intégrer des règles grammaticales avec des approches statistiques entraînées sur un corpus d'apprentissage.

Même si ces logiciels commerciaux ne donnent que peu d'informations sur les méthodes développées, des correcteurs grammaticaux gratuits et libres présentent leurs approches. Ainsi, le travail réalisé dans [Naber 2003], ayant abouti à la création du logiciel *LanguageTool*, a pour objectif de fournir des outils permettant de modéliser des règles grammaticales, indépendamment de la langue. Ces règles grammaticales sont déclenchées, si, dans le texte, les mots correspondent exactement au schéma défini (principe du *pattern-matching*). Il est donc indispensable de définir manuellement ses propres règles. De plus amples informations sur les correcteurs grammaticaux peuvent se retrouver dans [Souque 2007].

Bien que des solutions aient été proposées, la correction des erreurs grammaticales demeure un domaine d'étude ouvert, dont les méthodes proposées ne permettent toujours pas une correction complète. Dans [Souque 2008], l'auteur part du constat que, pour le français, il est difficile de trouver des outils permettant de corriger les erreurs grammaticales. Cette étude leur a permis de tirer la conclusion que les approches utilisées actuellement dans le domaine du TAL, pour la correction grammaticale, ne sont pas optimales. Leur idée suit celle de certains outils présentés

[55] `http://www.synapse-fr.com/`

[56] `http://www.druide.com/antidote.html`

[57] Une description succincte est fournie sur le site `http://www.synapse-fr.com/descr_technique/Correction_automatique.htm`

précédemment, à savoir fournir un outil libre, transposable dans d'autres langues, et facilement modifiable par les utilisateurs (ajout de nouvelles règles grammaticales). L'intérêt de la méthode réside dans la recherche d'un niveau supérieur dans la modélisation de règles linguistiques, à la différence d'autres approches où chaque règle pour chaque erreur doit être modélisée (exemple de l'outil *LanguageTool*). La proposition faite dans ce travail consiste à utiliser deux concepts pour corriger les erreurs grammaticales :

- L'utilisation d'un découpage syntaxique en catégories (appelé *chunks* ou *syntagmes*), réunissant un groupe de mots (groupe nominal, groupe verbal…) au lieu de prendre en compte uniquement un mot isolé. L'intérêt est que le découpage en syntagmes est réalisé selon une structure précise : comparativement aux mots, une généralisation est, dans ce cas, plus simple. La ponctuation[58] est, par exemple dans cette étude, un bon indicateur des syntagmes.
- L'unification des structures de traits, pouvant simplement se définir comme la non-compatibilité entre deux catégories grammaticales (nom, adjectif, pronom…)[59], permet encore de généraliser la correction.

Leur combinaison devrait permettre de vérifier grammaticalement les accords : les syntagmes définissant les frontières de recherche, et la vérification de l'unification des traits à l'intérieur de ces syntagmes trouvant les erreurs potentielles. Dans les expériences menées par les auteurs, les gains sont pour l'instant minimes, et surtout portés sur les erreurs d'accord. Mais ces expériences montrent qu'une généralisation est intéressante en termes de coûts (informatique et humain). Cependant, la technique présentée, comme les précédentes, nécessite une ponctuation précise pour fonctionner. De plus, la correction se limite à chaque syntagme ; la taille d'un syntagme semble trop courte pour prendre en compte les erreurs nécessitant des informations très éloignées dans l'historique.

En suivant l'affirmation de [Souque 2008], les auteurs dans [Clément 2009] partent du constat que la correction grammaticale reste peu étudiée et que les solutions proposées sont difficilement généralisables. Des solutions pour des langues précises, ainsi que des études linguistiques poussées sont souvent nécessaires. Les auteurs exposent, dans cet article, une approche cherchant à analyser de manière globale la phrase, pour en extraire des erreurs et des propositions de corrections possibles. Dans un premier temps, la méthode transforme une phrase en un graphe acyclique orienté[60] en ayant, pour chaque état, le lemme correspondant à

[58] L'étude porte sur des textes dont la ponctuation permet d'organiser les phrases supposées grammaticalement bien structurées. Ces textes ne reflètent pas un discours oralisé, où une ponctuation est plus difficile, et fatalement moins juste (agrammaticalité, disfluences…).

[59] Par exemple, "le chats" ne peut pas être unifié car "le" (masculin/singulier) n'est pas compatible avec "chats" (masculin/pluriel). Si "chats" était au singulier, cette unification aurait fonctionné (déterminant avec adjectif, même genre, même nombre).

[60] Le parcours de ce graphe se fait dans un seul sens, sans possibilité de se retrouver dans le même état.

chaque mot. Si, dans la phrase, des mots sont potentiellement homophones, ceux-ci sont ajoutés au graphe. Cependant, dans cette méthode, les mots homophones ne sont qu'une particularité, des lemmes correspondants à d'autres sortes de confusions peuvent être ajoutés. L'analyse syntaxique se fera donc en omettant, tout d'abord, les erreurs d'accord. Les possibilités d'accord sont ensuite insérées dans le graphe : son parcours permettra, pour chaque phrase possible, d'estimer un coût pour ce chemin. Au final, le coût minimum est retenu (le chemin de la phrase hypothèse) : si ce chemin possède un coût nul, la phrase est supposée correcte. Dans le cas contraire, un coût est associé à chaque "assignation de traits" (par exemple un groupe nominal, contenant plusieurs "traits" comme un déterminant et un nom), et donne une idée sur la validité de l'hypothèse de mot de la phrase. Un ensemble d'alternatives se trouvant dans le graphe, il est alors envisageable de fournir une proposition de correction, en choisissant la plus plausible (coût minimum). Ce travail tend vers des travaux plus statistiques que ceux présentés précédemment. En effet, bien que des règles linguistiques soient à la base de ces travaux (la notion de "traits" par exemple), ceux-ci se différencient des autres en présentant une méthode non plus guidée par des règles grammaticales strictes, mais par une vision plus générale de la phrase.

Les approches par règles grammaticales qui, bien qu'elles aient été très largement traitées dans le domaine du TAL, n'ont vu aucune réelle implémentation orientée pour les systèmes de RAP. Les auteurs, dans leurs articles, donnent un début de réponse : les règles grammaticales sont complexes à mettre en œuvre et nécessitent une expertise linguistique poussée. De plus, elles ont généralement besoin de phrases bien structurées, d'une ponctuation correcte, et d'une segmentation en phrases[61] pour fournir une correction.

3.3.2.2 Approches statistiques

Les méthodes basées sur des règles grammaticales générales semblent difficilement applicables à des sorties de systèmes de RAP (phrases grammaticalement correctes, ponctuation présente, intervention humaine parfois nécessaire...). Des approches statistiques automatiques, fournissant des solutions plus ciblées, ont été entreprises. En effet, la correction des mots homophones peut nécessiter des traitements particuliers, qui ne pourraient fonctionner au moyen de règles linguistiques générales.

Dans [Béchet 1999a, Béchet 1999b], les auteurs s'intéressent à un cas particulier d'homophones hétérographes en français, le problème du singulier/pluriel. En effet, comme nous l'avons déjà introduit dans la section 3.2.1, les formes du singulier et du pluriel sont très souvent homophones en français, et les modèles de langage ne sont pas assez larges et robustes pour traiter efficacement cette forme. Afin de remédier à cette faiblesse, les auteurs proposent

[61] Ce qui n'est pas le cas pour la reconnaissance de la parole, où les segments sont généralement découpés lorsqu'un silence ou une interruption sont rencontrés, comme une pause.

deux méthodes différentes. Ces deux méthodes utilisent un graphe d'homophones[62] créé spécialement pour y effectuer différents traitements. L'idée est de fournir des probabilités pour chaque hypothèse dans le graphe, afin de pouvoir extraire la meilleure phrase hypothèse. Trois modèles de langage sont entraînés, leur combinaison permettant d'obtenir une probabilité pour chaque hypothèse du graphe. Le premier modèle de langage utilisé est un modèle de langage n-gramme classique sur les mots, estimant la probabilité d'une hypothèse de mot selon ses mots précédents (ici, un modèle de langage trigramme). Le second modèle est un modèle n-classe (dans cette étude un modèle triclasse), qui, au lieu de prendre en considération les mots, prend en compte une séquence de classes représentant la catégorie syntaxique des mots (genre, nombre et classe grammaticale[63]). Enfin, un modèle de langage sur les syntagmes est utilisé (un modèle trigramme), qui est, en fait, un regroupement de plusieurs mots au sein d'une unité supérieure commune[64]. Ce dernier modèle possède l'avantage de pouvoir prendre en compte un historique plus grand qu'un modèle n-gramme classique.

L'originalité du travail réalisé dans [Béchet 1999b] est l'utilisation d'un nouveau modèle appelé *modèle Homophone-Cache*, dont l'objectif est de différencier les mots homophones au singulier et au pluriel. En effet, bien que les trois modèles de langage présentés précédemment aident à la correction de cette particularité homophonique, ils sont parfois inopérants dans ces cas spécifiques. Par exemple, ces modèles ne sont pas robustes aux erreurs de transcription des systèmes de RAP. La méthodologie générale de ce modèle peut se décomposer en deux étapes :

- À partir d'un corpus d'apprentissage, une mémoire cache de 10 mots précédant chaque mot homophone singulier/pluriel est conservée. L'idée est de mettre à jour la mémoire cache (composée d'un jeu de classes de 105 catégories syntaxiques) de chacune des formes (singulier et pluriel) des mots homophones en parcourant le corpus d'apprentissage. Chaque mot homophone possédera au final un vecteur de 105 composantes (représentant les catégories syntaxiques) pour sa forme au singulier et sa forme au pluriel.
- La seconde étape calcule deux distances durant la phase de décodage[65] :
  - premièrement, la distance entre le cache courant (calculé lors du décodage) et le cache appris à partir du corpus d'apprentissage pour le mot au singulier,
  - dans un second temps, la distance, en suivant ce même calcul, pour le mot au pluriel.

  Enfin, si la différence entre ces deux distances dépasse un certain seuil (optimisé sur un corpus de développement), le mot homophone contenant le score le plus faible est choisi.

[62] Ce graphe d'homophones est construit à partir d'une phrase, à laquelle on rajoute, pour chaque mot, l'ensemble de ses homophones possibles.

[63] Par exemple : déterminant masculin/singulier, nom féminin/pluriel...

[64] Par exemple, le syntagme *"les enfants"* est un groupe nominal masculin/pluriel.

[65] Seulement si les formes homophones d'un mot, dans le cadre de l'homophonie singulier/pluriel, apparaissent comme hypothèses possibles durant le décodage.

Les expériences, menées par les auteurs sur un corpus de test d'émissions journalistiques (texte grammaticalement correct car "préparé"), montrent que si l'on compare ces quatre approches séparément (les trois modèles de langage et le modèle Homophone-Cache), les meilleurs résultats sont obtenus avec le modèle n-gramme de mots et le modèle de classes. Il faut noter que la seule utilisation d'un modèle Homophone-Cache ne permet pas d'obtenir des performances intéressantes, ce que les auteurs attribuent à la trop grande permissivité au niveau des contraintes textuelles. Cependant, la combinaison de ces quatre modèles permet d'améliorer fortement les résultats, les modèles étant complémentaires dans leurs approches.

Comme dans l'étude précédente, les auteurs, dans [Lavecchia 2006], s'intéressent au problème du "nombre" homophone en français, c'est-à-dire des formes "singulier/pluriel" homophones, ainsi qu'au problème du genre ("féminin/singulier") en français. En effet, comme nous l'avons vu dans la partie 3.2.1, la combinaison de ces formes est très souvent homophone en français. L'idée, développée dans ce travail, repose sur l'utilisation d'un modèle Cache [Kuhn 1990]. Ce modèle doit capturer les informations utiles afin de déterminer le genre et le nombre du mot considéré. Les auteurs partent ainsi du constat que les informations, permettant de définir le nombre et le genre d'un mot, se retrouvent dans un contexte proche de ce mot. Le modèle Cache doit permettre aux auteurs de prendre en considération cette observation, en augmentant la probabilité d'apparition d'un mot lorsque celui-ci a été observé dans un passé proche. L'hypothèse est que si un mot est déjà apparu, il a de fortes chances de réapparaître dans un laps de temps relativement court. L'approche propose alors de prendre en compte des caractéristiques utiles pour définir le genre et le nombre d'un mot, au lieu de prendre en compte les mots, et de les utiliser au moyen d'un modèle Cache (appelé modèle *Features-Cache* dans ce travail). Chaque mot est donc transformé selon les caractéristiques qui lui sont associées, à savoir son genre (féminin, masculin, ou invariant) et son nombre (singulier, pluriel, ou invariant). Par exemple, le mot "chaise" est féminin/singulier et le mot "tapis" est masculin/invariant. Le nouveau modèle *Features-Cache* créé est interpolé avec un modèle de langage n-gramme classique afin d'obtenir une probabilité du mot et de son genre/nombre. Le modèle *Features-Cache* intègre également un second traitement, qui prend en compte un regroupement des mots selon leur catégorie syntaxique. Ainsi, "le petit chat" fait partie d'un groupe nominal masculin/singulier : seule cette information sera maintenant prise en compte, et non plus les caractéristiques des trois mots. Les expériences, menées par les auteurs, ont permis d'obtenir des résultats encourageants.En effet, partant d'un taux d'erreur-mot de 40,2 % avec un modèle de langage trigramme classique, le taux d'erreur-mot a baissé de 1,2 % en absolu en utilisant le modèle *Features-Cache*.

3.3.3 Combinaison des approches

Deux approches existent pour corriger les mots homophones : les approches par règles grammaticales, et les approches automatiques. Cependant, il semble possible de combiner ces deux approches. Au niveau des correcteurs grammaticaux, que nous avons étudiés dans la partie 3.3.2.1, le logiciel *Grac* (*GRAmmar Checker*)[66] propose une combinaison de ces approches. Son auteur, dans [Biais 2005], présente les concepts du logiciel, et aborde le problème de la correction grammaticale sous les deux approches. Ainsi, l'approche probabiliste permet de corriger certaines erreurs (ou tout du moins de les détecter), en complément d'une approche basée sur des règles linguistiques. Les informations morpho-syntaxiques constituent la principale base de connaissance du potentiel outil de correction. Bien que l'approche reste trop vague pour être exploitable, et qu'aucune réelle solution nouvelle ne soit apportée pour régler le problème de la correction, il faut retenir que l'idée d'une combinaison est intéressante, ces approches pouvant être complémentaires.

Cette combinaison a été mentionnée plus récemment dans le travail réalisé par [Mudge 2009]. Les auteurs y utilisent des règles linguistiques pour trouver les erreurs grammaticales dans un texte et fournir des propositions de correction possibles, dans l'optique de filtrer les propositions de correction au moyen d'approches statistiques. Dans un premier temps, l'outil développé va segmenter un texte brut (sans information sur sa structure *a priori*), grâce à des règles grammaticales, pour définir le début et la fin d'une phrase. L'étape suivante de détection, couplée à la proposition de correction, utilise des règles grammaticales, définies manuellement. Ces règles entrent automatiquement en application au moyen d'expressions régulières selon les informations contenues dans les mots de la phrase (informations morphologiques, classes grammaticales des mots, début et fin de phrase). Les différentes propositions de correction seront ensuite filtrées au moyen d'un modèle de langage trigramme classique sur les mots, afin de ne garder que les propositions les plus pertinentes. L'utilisateur choisira enfin s'il désire, ou non, corriger l'erreur détectée, à partir des propositions de correction (aucune correction automatique n'est réalisée). Un outil gratuit et libre a été créé suite à ce travail[67].

3.4 Conclusion

Dans ce chapitre, nous nous sommes intéressés au problème de l'homophonie, apparaissant dans de nombreuses langues. Nous avons pu constater qu'en français, ce phénomène est beaucoup plus présent que dans les autres langues telles que l'anglais ou l'allemand. Dans le cadre de

[66] `http://grac.sourceforge.net/`
[67] `http://open.afterthedeadline.com/about/technology-overview/`

la reconnaissance automatique de la parole, des études se sont penchées sur cette particularité du langage, afin de comprendre pourquoi les systèmes de RAP sont limités pour gérer les différents cas de l'homophonie.

Concernant les travaux réalisés dans le domaine, deux points de vue différents s'affrontent. D'un côté, des études présentent des solutions pour corriger toutes les erreurs que ne peut produire un système de RAP classique (le problème de l'homophonie étant inclus). D'un autre côté, des travaux s'intéressent à ce phénomène de façon plus ciblée, fournissant des solutions spécifiques pour corriger les mots homophones. Les travaux spécifiques sur l'homophonie en reconnaissance automatique de la parole sont peu nombreux et il faut regarder plus particulièrement du côté du traitement automatique des langues (TAL) pour voir ce phénomène traité plus largement. Les études semblent unanimes sur le fait que l'homophonie reste un problème difficile, et qu'aucune solution proposée ne permet de corriger toutes les confusions qui en résultent.

L'application des solutions fournies en TAL ne semble pas transposable directement à la reconnaissance automatique de la parole. En effet, le plupart des études du TAL nécessitent une organisation grammaticalement correcte, avec notamment un découpage en phrases possédant une ponctuation précise, ce que les systèmes de RAP ne peuvent fournir précisément. Bien que d'autres approches s'orientent vers l'utilisation conjointe des approches par règles grammaticales manuelles et approches statistiques, les solutions proposées restent souvent des propositions de correction potentielles, et non une correction automatique par un système informatique.

Deuxième partie

Contributions

Chapitre 4

Étude comparative de la parole préparée et spontanée en français

Sommaire

4.1 Caractérisation de la parole spontanée 87

4.1.1 Étiquettes et classes de spontanéité 87

4.1.2 Impact du degré de spontanéité 89

4.1.3 Extraction de caractéristiques de la parole spontanée 89

4.1.3.1 Caractéristiques prosodiques 90

4.1.3.2 Caractéristiques linguistiques 91

4.1.3.3 Mesures de confiance 92

4.2 Apprentissage automatique : le *Boosting* 93

4.2.1 Principe général 93

4.2.2 L'algorithme *AdaBoost* 93

4.3 Approche proposée 95

4.4 Détection automatique des segments de parole spontanée 95

4.4.1 Classification au niveau du segment 96

4.4.2 Décision globale au moyen d'un modèle probabiliste 98

4.4.2.1 Présentation du modèle 98

4.4.2.2 Résolution de l'équation 98

4.5 Expériences 101

4.5.1 Données expérimentales 101

4.5.1.1 Corpus 101

4.5.1.2 Performances du système de RAP 102

4.5.1.3 Détection et catégorisation automatiques des segments de parole 103
4.5.2 Conclusion 108

Les taux d'erreur-mot obtenus sur les transcriptions d'un système de reconnaissance de la parole montrent très clairement la difficulté de transcrire de la parole conçue au cours de l'énonciation, et donc dite spontanée (par exemple les débats, interviews…), par rapport à la parole lue et préparée (comme les émissions de journaux d'information). La définition de [Luzzati 2004] illustre bien la subjectivité de la classification de la parole en parole spontanée ou préparée, la frontière étant relativement floue. De nombreuses particularités existent en parole spontanée, comme par exemple les disfluences ou les répétitions, ou encore une "agrammaticalité" dans la construction des phrases. Il apparaît donc utile de pouvoir automatiquement annoter un corpus au moyen d'étiquettes représentant la spontanéité de chacun des segments de parole. En effet, une détection automatique des zones de spontanéité permettrait d'y réaliser différentes études sans avoir besoin d'écouter l'ensemble des documents audio. De plus, pouvoir détecter au plus tôt le type de parole des segments de parole donnerait la possibilité aux systèmes de RAP d'adapter leurs modèles, et de pouvoir, à terme, améliorer la transcription de la parole dite spontanée.

Posséder un corpus de référence annoté en niveaux de spontanéité est une étape nécessaire pour définir les différences entre les types de parole. Ce corpus, ainsi que les différentes caractéristiques intéressantes de la parole spontanée, seront détaillés dans le paragraphe 4.1.

La détection de la parole spontanée de manière automatique est nécessaire si l'on veut l'intégrer aux systèmes de RAP sans intervention humaine. La partie 4.3 présente la méthode méthode que nous proposons pour détecter automatiquement les classes de spontanéité de chaque segment.

4.1 Caractérisation de la parole spontanée

Afin de pouvoir définir le type de spontanéité d'un segment de parole, une connaissance des différences fondamentales entre les types de parole est indispensable. L'état de l'art sur la parole spontanée, décrit dans la partie 2.1, met en lumière ses particularités permettant de la caractériser de manière précise. Pour rendre automatique l'étape de détection du type de spontanéité, un corpus annoté manuellement doit être fourni, sur lequel nous pourrons extraire des caractéristiques utiles.

4.1.1 Étiquettes et classes de spontanéité

Idéalement, il faudrait demander à chaque locuteur d'annoter ses propres phrases, lui seul pouvant définir avec exactitude le degré de spontanéité de son intervention (à condition que lui-même sache différencier la fluidité de ses interventions). Cette hypothèse n'est, bien entendu,

pas réalisable. Nous avons opté pour l'utilisation du corpus présenté dans [Bazillon 2008b], où les auteurs ont choisi de définir un protocole d'annotation des segments de parole en degrés de spontanéité. Chaque locuteur possédant des aptitudes différentes en contexte de parole spontanée[68], le degré de spontanéité représente un degré de fluidité dans le discours, jugeant les locuteurs sur leur niveau d'élocution, les ruptures dans le discours, les hésitations... L'idée générale est de fournir le protocole à un juge humain, qui l'utilisera pour définir le degré de spontanéité d'un segment. L'approche propose 10 étiquettes (correspondant chacune à un degré de spontanéité) pour annoter manuellement les segments de parole d'un corpus. Les étiquettes s'étendent de l'échelon 1, qui sera choisi pour définir des segments de parole préparée comparables à de la parole lue, jusqu'à l'échelon 10, pour caractériser des segments de parole très spontanée, à la limite de l'incompréhension par un humain.

Les différentes étiquettes associées aux segments de parole ont ensuite été regroupées en trois classes de spontanéité, choisies pour caractériser les limites entre la parole spontanée et préparée. Le tableau 4.1 présente les classes de spontanéité issues des étiquettes associées aux segments de parole. Ces trois classes de spontanéité ont l'avantage d'être plus facilement exploitables que l'étiquetage précis défini précédemment. En effet, il est déjà très difficile de trouver une frontière entre la parole préparée et spontanée (la fluidité du discours reste subjective), fournir automatiquement un degré de spontanéité (étiquette de 1 à 10) du segment de parole apparaît impossible.

Classe	**Préparée**	**Légèrement spontanée**	**Fortement spontanée**
Etiquette	1	2 à 4	5 à 10

TAB. 4.1 – Classes de spontanéité définies à partir du protocole d'annotation des segments de parole.

Deux annotateurs ont séparément étiqueté en degrés de spontanéité le même corpus. Ce corpus a, au préalable, été segmenté automatiquement au moyen du système de segmentation du LIUM, détaillé dans la partie 1.5.2.1. Aucune transcription n'a été fournie aux annotateurs. Pour pouvoir évaluer l'accord inter-annotateurs sur cette tâche, le coefficient *Kappa* [Cohen 1960] de cet accord a été calculé sur une heure d'émission radiophonique. Le score obtenu pour les trois classes de spontanéité était très haut (0,852), un score supérieur à 0,8 étant généralement considéré comme excellent [Di Eugenio 2004].

Après avoir calculé ce coefficient *Kappa*, les deux annotateurs humains, toujours séparément, ont fini d'annoter en degrés de spontanéité le corpus étudié. L'étiquetage en segments

[68]Certains locuteurs peuvent, par exemple, posséder un niveau d'élocution et de fluidité proche de la parole lue, même s'ils se trouvent dans un contexte spontané.

de parole a été choisi car il permet de trouver, dans un même flux de parole, des segments spontanés entourés de segments préparés, et inversement. Cet étiquetage permet ainsi d'être plus précis qu'un étiquetage sur un paragraphe. Mais surtout, les systèmes de RAP utilisant un découpage en segments (par exemple pour l'apprentissage des modèles acoustiques), il est nécessaire de posséder l'information sur le type de parole segment par segment si l'on veut utiliser cette information au niveau du système de RAP.

4.1.2 Impact du degré de spontanéité

Définir différents niveaux de spontanéité est une tâche difficile, puisqu'en partie subjective. Cette tâche n'est cependant pas impossible, comme l'a prouvé le coefficient *Kappa*. De plus, ce travail est essentiel. En effet, dans [Jousse 2008], nous constatons très clairement que le taux d'erreur-mot (WER) est fortement corrélé au niveau de spontanéité choisi par l'annotateur humain. Les auteurs ont utilisé les transcriptions annotées en niveaux de spontanéité, et ont, pour chacun, calculé le WER correspondant. Ces taux d'erreurs ont été regroupés selon les classes de spontanéité énoncées dans la section 4.1.1 (une description plus détaillée du corpus se trouve dans la partie 4.5.1.1) :

- **Parole préparée** : les auteurs ont obtenu un WER se situant approximativement autour de 20 %.
- **Légèrement spontanée** : le taux d'erreurs se situe quasiment à 40 %.
- **Fortement spontanée** : le taux d'erreurs, beaucoup plus important, oscille entre 45 et 60 %

4.1.3 Extraction de caractéristiques de la parole spontanée

Les variations des taux d'erreurs entre ces classes de spontanéité prouvent, comme nous l'avons déjà vu dans le chapitre 2, que des différences existent et que les systèmes de RAP semblent moins bien conçus pour traiter de la parole spontanée. Afin de différencier ces types de parole, il est indispensable d'extraire les caractéristiques propres à la parole spontanée. Ce problème a récemment été étudié comme une tâche particulière de la campagne d'évaluation *NIST Rich Transcription Fall 2004*, destinée à détecter les disfluences de la parole. Les systèmes de RAP utilisant des informations de natures diverses (acoustiques et linguistiques notamment), nous verrons dans les parties suivantes les différentes caractéristiques qui apparaissent utiles pour caractériser le type de parole, en réalisant une analyse quantitative de certaines d'entre-elles au niveau des trois classes de spontanéité présentées précédemment.

4.1.3.1 Caractéristiques prosodiques

Dans un premier temps, nous nous sommes intéressés aux caractéristiques prosodiques des différents types de parole. Différentes spécificités ont été extraites dans [Jousse 2008], que nous avons enrichies par la suite dans [Dufour 2009b, Dufour 2009a].

Durée D'après les travaux de [Shriberg 1999], la durée des voyelles et leur allongement en fin de mot semblent être des critères discriminants. Cet allongement, étudié dans [Caelen-Haumont 2002a] et associé au concept de mélisme, a également été choisi pour sa pertinence. Le *mélisme*, issu du domaine musical, désigne le groupement de plusieurs notes en une seule syllabe, où le nombre de sons perçus pour un mot est supérieur au nombre de syllabes de ce mot. Ce concept, adapté au domaine du traitement automatique de la parole dans [Caelen-Haumont 2002b], tend à montrer que la prononciation d'un même mot peut varier selon le locuteur et son état émotionnel (ici selon le type de parole utilisé).

Débit phonémique Des études précédentes [Caelen-Haumont 2002b] ont mis en lumière une corrélation entre les variations du débit de parole et l'état émotionnel du locuteur. De ces constats a émergé l'idée d'estimer le débit de parole mot par mot et segment par segment, qui s'est révélée être une caractéristique intéressante dans [Jousse 2008] pour caractériser la spontanéité d'un segment. Le débit phonémique est extrait au moyen de deux caractéristiques, d'une part sa variance pour chaque mot, et d'autre part sa moyenne pour le segment entier, pauses et morphèmes spécifiques inclus ("*ben*", "*hum*", "*euh*"...) puis exclus. Nous avons ajouté la variance du *pitch* [Dufour 2009b] comme caractéristique potentielle.

Afin de se rendre compte de l'impact possible de ces différentes caractéristiques acoustiques, nous présentons dans le tableau 4.2 les valeurs moyennes de ces caractéristiques (en secondes), et dans le tableau 4.3 leurs variances moyennes pour chaque classe de spontanéité.

Caractéristiques (en secondes)	**Préparée**	**Légèrement spontanée**	**Fortement spontanée**
Durée voyelles	0,075	0,081	0,091
Mélismes	0,082	0,094	0,110
Durée phonémique	0,078	0,081	0,087

TAB. 4.2 – Comparaison des valeurs moyennes (en secondes) des caractéristiques acoustiques pour chaque classe de spontanéité.

Les résultats montrent qu'une corrélation existe entre ces caractéristiques et la classe de spontanéité. Les valeurs sont ainsi plus élevées dans les segments étiquetés en parole *fortement spontanée*, que ce soit au niveau des durées ou des variances.

	Préparée	**Légèrement spontanée**	**Fortement spontanée**
Durée voyelles	0,0018	0,0033	0,0071
Mélismes	0,0025	0,0051	0,01130
Pitch	1068,19	1085,58	1106,07
Durée phonémique	0,0017	0,0026	0,0046

TAB. 4.3 – Comparaison des variances moyennes des caractéristiques acoustiques pour chaque classe de spontanéité.

4.1.3.2 Caractéristiques linguistiques

Après cela, des caractéristiques linguistiques ont été étudiées dans [Dufour 2009b], en nous focalisant plus particulièrement sur le contenu lexical et syntaxique du segment. Le concept des *disfluences* est le principal concept que l'on retrouve dans la parole spontanée. De nombreuses études existent sur le sujet, et se retrouvent au niveau acoustique [Shriberg 1999] et lexical [Siu 1996] (voir partie 2.1.1). En analysant les différentes disfluences existantes, les auteurs se sont concentrés sur les morphèmes spécifiques déjà étudiés au niveau acoustique. Ainsi, le nombre d'occurrences de ces morphèmes par rapport au nombre de mots dans un segment constitue une première caractéristique linguistique. En continuant leur analyse, les auteurs ont noté que le nombre de répétitions d'unigrammes et de bigrammes constituait une information pertinente pour caractériser le niveau de spontanéité. Nous utiliserons par conséquent le nombre de répétitions par rapport au nombre de mots du segment comme caractéristique. Enfin, utiliser des paquets de n-grammes de mots et fournir un découpage syntaxique en catégories (groupe nominal, groupe prépositionnel) permet de fournir des informations supplémentaires au niveau linguistique.

Dans l'optique d'améliorer la détection des niveaux de spontanéité, nous avons enrichi les caractéristiques fournies par [Jousse 2008] avec de nouvelles informations. Nous avons pris en compte une expérience réalisée dans [Bazillon 2008b] qui tendait à prouver que plus un discours était préparé, plus l'utilisation des noms propres était fréquente. Nous avons alors ajouté, comme présenté dans [Dufour 2009b] la fréquence d'apparition d'un nom propre en fonction du nombre de mots dans un segment pour caractériser son niveau de spontanéité.

Ainsi, nous cherchons à comprendre l'impact des différentes caractéristiques linguistiques que nous avons choisies pour chaque classe de parole. Dans le tableau 4.4, les valeurs moyennes sont analysées pour tous les segments étiquetés pour chaque type de parole sur les transcriptions de référence. Le tableau 4.5 présente les mêmes expériences, mais dans ce cas sur les transcriptions automatiques fournies par le système de RAP du LIUM.

Caractéristiques (proportion par segment)	**Préparée**	**Légèrement spontanée**	**Fortement spontanée**
# morphèmes spécifiques	0,0030	0,0213	0,0451
Taille découpage syntaxique	0,5751	0,5974	0,6340
# répétitions	0,0014	0,0065	0,0315
# noms propres	0,0833	0,0561	0,0299

TAB. 4.4 – Comparaison des valeurs moyennes (proportion par segment) des caractéristiques linguistiques pour chaque classe de spontanéité sur les transcriptions de référence.

Caractéristiques (proportion par segment)	**Préparée**	**Légèrement spontanée**	**Fortement spontanée**
# morphèmes spécifiques	0,0021	0,0163	0,0381
Taille découpage syntaxique	0,5841	0,6021	0,6403
# répétitions	0,0014	0,0051	0,0224
# noms propres	0,0739	0,0533	0,0345

TAB. 4.5 – Comparaison des valeurs moyennes (proportion par segment) des caractéristiques linguistiques pour chaque classe de spontanéité sur les transcriptions automatiques.

Les tableaux 4.4 et 4.5 montrent la corrélation existante entre la classe de spontanéité assignée aux segments de parole de notre corpus et les caractéristiques linguistiques présentées. Nous constatons que l'utilisation de ces caractéristiques sur les transcriptions automatiques fournies par le système de RAP est toujours pertinente, malgré un fort taux d'erreur-mot sur les segments de parole spontanée. De plus, nous pouvons voir que ces critères résistent aux erreurs de reconnaissance du système de RAP. En effet, les mêmes tendances, entre les différents types de parole, ont pu être observées sur les transcriptions de référence et les transcriptions automatiques.

4.1.3.3 Mesures de confiance

Pour terminer, nous avons ajouté les mesures de confiance fournies par le système de RAP du LIUM comme caractéristique pour définir le type de parole. Ces mesures de confiance sont des scores exprimant la fiabilité des décisions de reconnaissance prises par un système de RAP (voir partie 1.4.5). Ces scores sont utilisés pour caractériser la spontanéité des segments de parole. Effectivement, comme cela est mentionné dans [Dufour 2009b], les systèmes de RAP ont beaucoup plus de difficultés à transcrire des segments de parole spontanée que des segments de parole préparée. Le tableau 4.6 présente les valeurs moyennes des moyennes et des variances

des mesures de confiance obtenues pour chaque classe de spontanéité sur tous les segments annotés.

	Préparée	**Légèrement spontanée**	**Fortement spontanée**
Moyenne	0,91	0,88	0,82
Variance	0,021	0,026	0,036

TAB. 4.6 – Comparaison des valeurs moyennes des moyennes et variances des mesures de confiance selon la classe de spontanéité.

Les résultats montrent que les mesures de confiance fournies par le système de RAP semblent constituer un bon indicateur du degré de spontanéité des segments de parole.

4.2 Apprentissage automatique : le *Boosting*

4.2.1 Principe général

Le *Boosting* est un principe issu du domaine de l'apprentissage automatique [Schapire 2003]. Son objectif principal est d'améliorer (de "booster") la précision de n'importe quel algorithme d'apprentissage permettant d'associer, à une série d'exemples, leur classe correspondante[69]. Le principe général du *Boosting* est assez simple : la combinaison pondérée d'un ensemble de classifieurs binaires (appelés classifieurs *faibles*), chacun associé à une règle différente de classification très simple et peu efficace[70] (que l'on retrouve parfois sous le nom de *rule of thumb*), permettant au final d'obtenir une classification robuste et très précise (classifieur *fort*). L'algorithme permettant la construction de classifieurs faibles est appelé *apprenant faible*.

4.2.2 L'algorithme *AdaBoost*

AdaBoost est l'algorithme le plus utilisé en *Boosting* [Freund 1995] car il présente de nombreux avantages. En effet, l'algorithme est très rapide et simple à programmer, applicable à de nombreux domaines, adaptable aux problèmes multi-classes... Il fonctionne sur le principe du *Boosting*, où un classifieur *faible* est obtenu à chaque itération de l'algorithme *AdaBoost*. Chaque exemple d'apprentissage possède un poids, chaque tour de classification permettant de les re-pondérer selon le classifieur *faible* utilisé. Le poids d'un exemple bien catégorisé

[69] Le jeu de "*pile ou face*" peut, par exemple, être un problème de classification.

[70] La seule contrainte de ces classifieurs *faibles* est d'obtenir des performances meilleures que le hasard. Dans le cas des classifieurs binaires, l'objectif est de classifier plus de 50 % des données correctement.

est diminué, au contraire d'un exemple mal catégorisé, où son poids est augmenté. L'itération suivante se focalisera ainsi sur les exemples les plus "difficiles" (poids les plus élevés).

De manière formelle :

- Soit un échantillon d'apprentissage $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, où $x_i \in X$ et $y_i \in \{-1, 1\}$
- Pour $t = 1, \ldots, T$:
  - On initialise la distribution D_t des exemples par :

$$D_1(i) = \frac{1}{m}, i = 1 \ldots m$$

  - Trouver un classifieur *faible* :

$$h_t : X \rightarrow \{-1, 1\}$$

    tel que l'erreur de classification ϵ_t sur D_t soit minimisée :

$$\epsilon_t = Pr_{i \sim D_t}(h_t(x_i) \neq y_i)$$

- Choisir :

$$\alpha_t = \frac{1}{2}\, ln\, \left(\frac{1 - \epsilon_t}{\epsilon_t}\right)$$

- La pondération des exemples d'apprentissage est ensuite actualisée :

$$D_{t+1}(i) = \frac{D_t(i) e^{\alpha_t y_i h_t(x_i)}}{Z_t}$$

  où Z_t est un facteur de normalisation (choisi pour que D_{t+1} soit une distribution de probabilité)
- Combiner les h_t de manière à obtenir le classifieur *fort* final :

$$H(x) = sign\, \left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

L'algorithme a été implémenté dans l'outil de classification à large-marge *BoosTexter* [Schapire 2000], permettant de fournir comme caractéristiques au classifieur des données numériques mais également des données textuelles. Les classifieurs *faibles* sur les données textuelles

peuvent prendre en compte des n-grammes[71], s-grammes [72], ou f-grammes[73] de mots. Une version *open-source* de cet outil a été réalisée à l'*International Computer Science Institute (ICSI)* sous le nom d'*Icsiboost* [Favre 2007].

4.3 Approche proposée

Notre approche, exposée dans [Dufour 2009b], consiste à combiner les différentes caractéristiques présentées dans la partie 4.1.3 afin de prendre une décision sur la classe de parole du segment. Cette combinaison sera faite pour chaque segment, au moyen d'un *classifieur*, qui nous fournira la classe de parole la plus probable associée à ce segment, en fonction des valeurs des caractéristiques fournies. La figure 4.1 présente l'approche générale suivie pour prendre une décision sur la classe de spontanéité au niveau de chaque segment de parole.

Bien que la classification, effectuée pour chaque segment, permette une précision intéressante pour les systèmes de RAP nous avons l'impression intuitivement qu'il est rare d'observer un segment de parole *fortement spontanée* entouré de segments de parole *préparée*. Nous avons alors cherché à modéliser ce phénomène afin de ne plus considérer simplement la classe de spontanéité au niveau du segment, mais à prendre en compte les informations sur la classe potentielle de spontanéité fournies par les segments voisins. Ces informations seront résumées au moyen du score de confiance attribué par le *classifieur* à chacune des trois classes de spontanéité, et ce, pour chaque segment. Ainsi, nous proposons dans [Dufour 2009a] de prendre en considération la nature des segments de parole contigus. Cette décision implique que la catégorisation de chaque segment de parole issu d'un fichier audio ait un impact sur la catégorisation des autres segments de parole. Le processus de décision devient alors un processus global et ne se limite plus au segment courant. Le figure 4.2 montre le processus général de décision globale, utilisant les résultats obtenus sur les classes de spontanéité pour chaque segment.

4.4 Détection automatique des segments de parole spontanée

Afin d'extraire automatiquement les caractéristiques acoustiques et linguistiques nécessaires à la catégorisation des segments de parole en fonction de leur classe de spontanéité, nous avons utilisé le système de RAP du LIUM. Les différentes phases de ce décodeur ont été présentées dans la partie 1.5.

[71] 1 à n mots consécutifs. Exemple pour $n = 2$: "le le_petit petit_chat chat".

[72] Bigrammes séparés par n mots. Exemple pour $n = 1$: "je_vais * la_plage" où * représente n'importe quel mot.

[73] n-grammes de taille n uniquement : pas de repli en dessous de n. Exemple pour $n = 3$: "le_petit_chat".

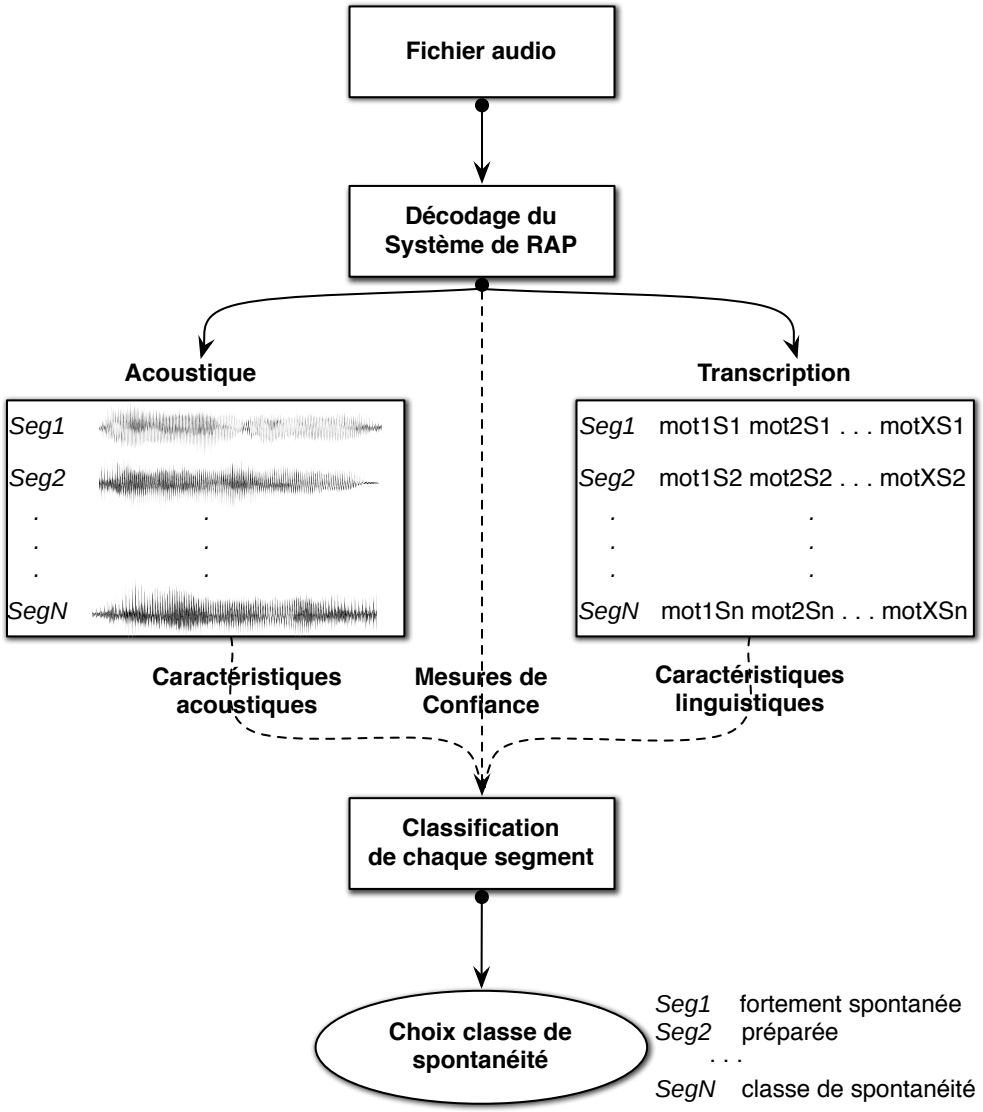

FIG. 4.1 – Approche générale pour la détection de la classe de spontanéité de chaque segment de parole.

4.4.1 Classification au niveau du segment

Les caractéristiques présentées dans la partie 4.1.3, issues des différentes études sur la parole spontanée, sont évaluées sur le corpus étiqueté (présenté dans la partie 4.5.1.1) au cours d'une tâche de classification. Ce processus de classification doit permettre de prendre une décision parmi les trois classes de spontanéité que nous avons choisi pour les segments de parole : préparée, légèrement spontanée ou fortement spontanée.

L'outil de classification choisi est *Icsiboost*, un outil open-source s'appuyant sur l'algorithme *AdaBoost*, qui a initialement été implémenté dans l'outil *BoosTexter*. De plus amples

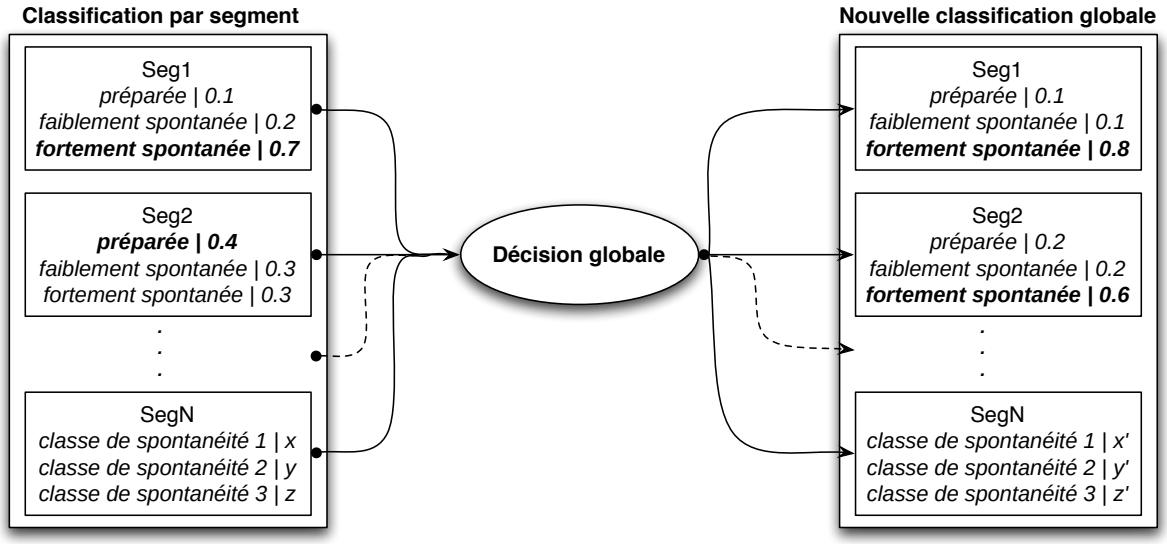

FIG. 4.2 – Approche générale pour le processus de décision globale dans l'attribution de la classe de spontanéité des segments de parole.

précisions sur le fonctionnement de cet algorithme se trouve dans la partie 4.2. Pour notre utilisation, les différentes caractéristiques que nous avons retenues sont fournies en entrée pour l'apprentissage du classifieur. Des valeurs numériques discrètes ou continues (valeurs des caractéristiques acoustiques, linguistiques et des mesures de confiance) et des valeurs textuelles (pour les caractéristiques linguistiques, comme les mots et l'étiquetage syntaxique) sont alors données. Nous avons choisi d'ajouter d'autres informations propres au processus de décodage du système de RAP, comme la durée de chaque segment ainsi que le nombre de mots reconnus dans chaque segment [Dufour 2009b]. À la fin de ce processus d'apprentissage, la liste des classifieurs sélectionnés est obtenue tout comme le poids de chacun d'entre eux, afin d'utiliser les exemples les plus discriminants pour chaque classe de spontanéité.

Ensuite, ces mêmes caractéristiques (acoustiques, linguistiques et mesures de confiance) seront fournies en entrée pour chaque segment dont nous voulons trouver la classe de spontanéité. Ces segments seront alors catégorisés au moyen de nos données d'apprentissage, cette classification assignant en sortie la classe la plus probable pour chaque segment. Chaque segment est catégorisé individuellement.

4.4.2 Décision globale au moyen d'un modèle probabiliste

4.4.2.1 Présentation du modèle

La classification en parole spontanée est, pour l'instant, réalisée en utilisant simplement les informations locales du segment de parole. Nous avons alors choisi d'utiliser une approche statistique classique en utilisant une méthode du maximum de vraisemblance : la classification des segments devient un processus de décision globale.

Soit s_i une classe de spontanéité du segment de parole i, avec $s_i \in$ {"*fortement spontanée*", "*légèrement spontanée*", "*préparée*"}. Nous définissons $P(s_i|s_{i-1}, s_{i+1})$ comme la probabilité d'observer un segment de parole i associé à la classe s_i, sachant que le segment précédent est associé à la classe s_{i-1} et que le segment suivant est associé à la classe s_{i+1}. Soit $c(s_i)$ la mesure de confiance fournie par la classifieur *AdaBoost* par rapport au choix de la classe s_i pour le segment de parole i, en prenant en compte les valeurs des caractéristiques extraites de ce segment. S est une séquence de classes s_i associée à la séquence de tous les segments de parole i (simplement une classe par segment). Le processus de décision globale consiste à choisir la séquence hypothèse de classes $\bar{S}$ qui maximise le score global obtenu en combinant $c(s_i)$ et $P(s_i|s_{i-1}, s_{i+1})$ pour chaque segment de parole i détecté sur le fichier audio. La séquence $\bar{S}$ est calculée en utilisant la formule suivante :

$$\bar{S} = \underset{S}{\operatorname{argmax}}\ c(s_1) \times c(s_n) \times \prod_{i=2}^{n-1} \times c(s_i) \times P(s_i|s_{i-1}, s_{i+1}) \tag{4.1}$$

où n est le nombre de segments de parole détectés automatiquement dans le fichier audio.

4.4.2.2 Résolution de l'équation

En pratique, pour résoudre l'équation 4.1 présentée précédemment, plusieurs solutions sont envisageables : utilisation d'un décodeur, création d'un outil spécifique... Pour des raisons pratiques (facilité de mise en œuvre, outils existants...), nous avons projeté le problème au moyen du paradigme des machines à états-finis (connues également sous le nom de *Finite-State Machine*, FSM). Par définition, les FSM sont constituées d'états et de transitions orientées entre ces états. Ces états existent en un nombre fini. Un symbole (contenu dans un alphabet constitué d'un nombre fini de symboles) est fourni en entrée et dirigera le comportement général de la machine. Celle-ci se déplacera d'états en états (en fonction des symboles déjà reconnus) jusqu'à se trouver dans un état final. Une probabilité peut être assignée à chaque transition d'état. La figure 4.3 présente un exemple général d'une machine à états-finis avec l'alphabet {0, 1}.

L'exemple de ce FSM contient trois états (*État 1* est l'état d'entrée et *État 3* est l'état de sortie), et quatre transitions. L'idée est de fournir en entrée un symbole à la machine (dans notre

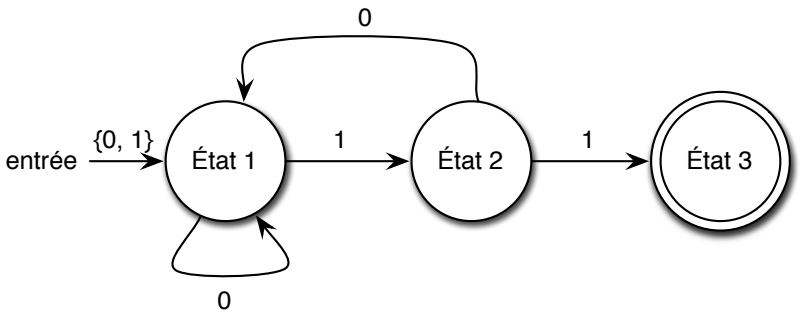

FIG. 4.3 – Exemple général d'une machine à états-finis.

exemple *0* ou *1*), ce qui permettra d'arriver dans l'*État 1*. Au niveau de l'*État 1*, deux transitions sont possibles, permettant soit d'aller dans l'*État 2*, soit de rester dans l'*État 1*. Si l'on arrive dans l'*État 2*, deux transitions sont possibles : aller dans l'*État 3*, qui est notre état de sortie (fin du FSM), ou retourner dans l'*État 1*. Bien sûr, cet exemple est simplifié, les états, les transitions, et les états de sortie sont généralement plus nombreux.

En pratique, nous avons choisi de nous appuyer sur l'article de [Mohri 2002] en utilisant des transducteurs d'états-finis pondérés. Pour utiliser les transducteurs à états-finis et mettre en application notre équation, l'outil *AT&T FSM toolkit*[74] a été choisi.

Pour résoudre notre problème, nous devons représenter le modèle contenant les probabilités $P(s_i|s_{i-1}, s_{i+1})$ pour le 3-tuple (s_{i-1}, s_i, s_{i+1}). Ce 3-tuple sera représenté dans un transducteur.

La figure 4.4 montre la topologie proposée pour représenter toutes les probabilités en utilisant le formalisme FSM. Les différentes classes de spontanéité sont représentées dans cette figure par :

- *prepa* correspond à la classe de parole *préparée*,
- *low* correspond à la classe de parole *faiblement spontanée*,
- *high* correspond à la classe de parole *fortement spontanée*.

Par soucis de clarté dans la figure 4.4, le FSM étant très volumineux, les termes génériques suivants y apparaissent :

- *tag* désigne la classe de spontanéité,
- *next tag* désigne la classe de spontanéité du segment suivant,
- *previous tag* désigne la classe de spontanéité du segment précédent.

En analysant la modélisation représentée dans la figure 4.4, nous constatons que nous n'avons pas pris en compte la classe de spontanéité du premier segment du fichier audio analysé, qui

[74] `http://www.research.att.com/~fsmtools/fsm/`

constituera notre condition d'entrée, et que la classe de spontanéité du dernier segment constituera notre condition de sortie.

Les probabilités $P(tag|previous, next)$ sont estimées au moyen du corpus d'apprentissage étiqueté en classes de spontanéité (voir partie 4.5.1.1), selon le rapport entre le nombre d'occurrences de $tag \in \{prepa, low, high\}$ dans le 3-tuple $previous, tag, next$ et le nombre total d'occurrences de $previous, *, next$ où $*$ représente n'importe quelle valeur dans tag.

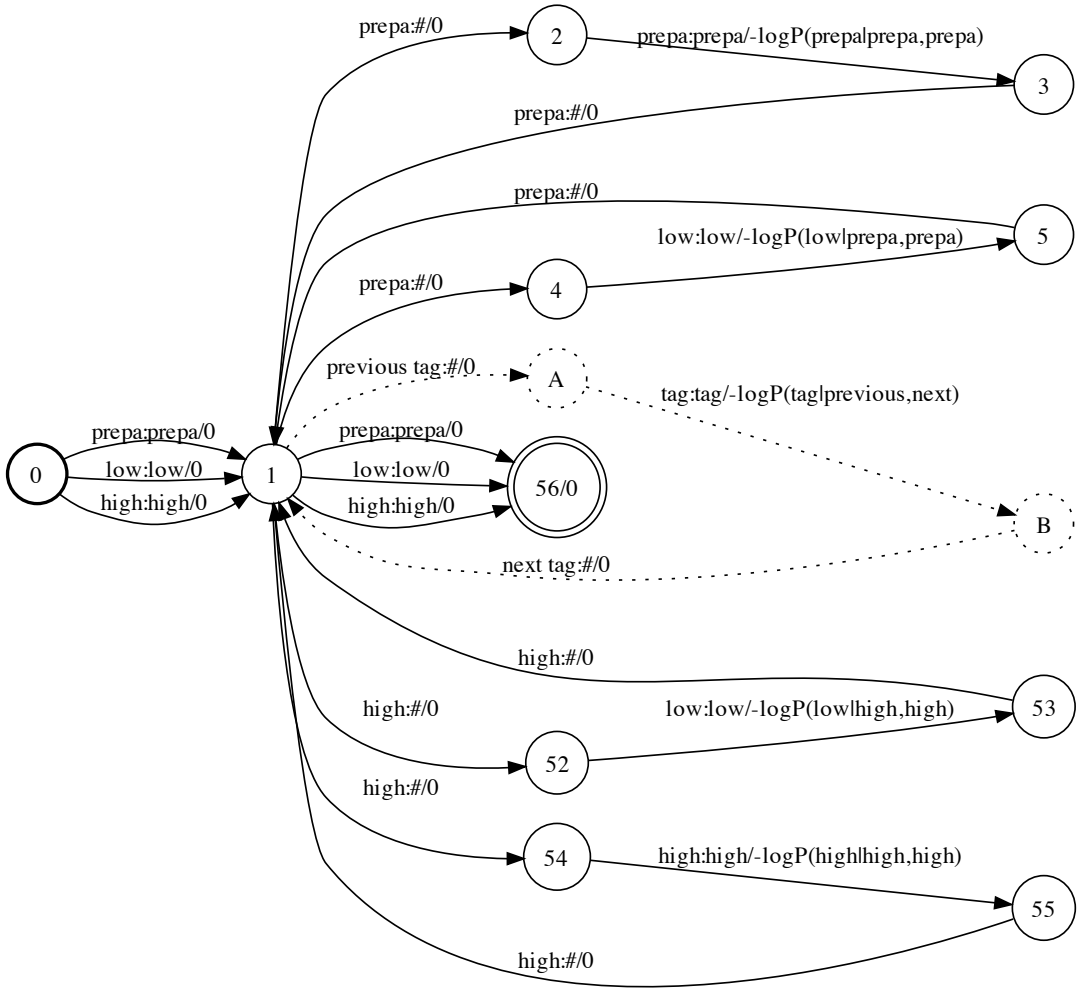

FIG. 4.4 – Transducteur modélisant toutes les probabilités contextuelles de $P(s_i|s_{i-1}, s_{i+1})$

En terme de rapidité d'exécution, nous avons noté que l'utilisation du paradigme FSM est très rapide (moins de 6 secondes de temps de calcul pour traiter les 11 fichiers audio sur un portable MacBook Pro Apple, soit 11 heures de parole et 3 814 segments). Cependant, cette

rapidité est due au nombre limité de classes (trois classes de spontanéité dans notre étude) et au nombre moyen de segments détectés dans un fichier audio (en moyenne 346 segments par fichier).

Bien sûr, un outil optimisé en terme de temps de calcul pourrait être implémenté pour calculer la meilleure séquence de classes en respectant la formule 4.1 sans le formalisme FSM. Cependant, ce dernier nous a permis de réaliser nos expériences sans avoir à développer un nouvel outil spécifique, pour un temps de calcul raisonnable.

4.5 Expériences

Afin de mettre en place le système de détection de la classe de spontanéité de chaque segment de parole, nous avons mené différentes expériences au moyen du corpus annoté par les juges humains, avec les différentes caractéristiques proposées précédemment. La méthode proposée évaluera alors l'intérêt de l'utilisation de ces caractéristiques au niveau du segment, puis au niveau d'un flux de parole (modèle global).

4.5.1 Données expérimentales

4.5.1.1 Corpus

Le corpus expérimental, composé de 11 fichiers audio issus d'émissions radiophoniques françaises, a été étiqueté par deux annotateurs humains, comme décrit dans la partie 4.1.1. En détail, ces fichiers proviennent :

- 6 fichiers du corpus de test ESTER 2 (voir partie 1.6) :
  - Radio Classique : 20041006_0700_0800.
  - France Culture : 20041006_0800_0900.
  - France Inter : 20041007_0800_0900 et 20041011_1300_1400.
  - France Info : 20041008_1800_1830 et 20041012_1800_1830.
- 5 fichiers du corpus du projet EPAC (voir partie 1) :
  - RFI : 20040107_1655_1810 (divers journaux et émission "Culture Vive").
  - France Info : 20040409_0455_0705 (interviews, débat et conférence).
  - France Culture : 20040901_2155_2217 (début de l'émission "Double Culture"), 20040920_2155_2217 (extrait de l'émission "Les chemins de la connaissance") et 20040921_0655_0905 (émission "Les matins de France Culture").

La durée totale est de 11 heures, pour un total de 3 814 segments (obtenus au moyen d'une segmentation automatique). Parmi ces segments, 1 228 ont été annotés avec la classe *parole préparée*, 1 339 avec la classe *légèrement spontanée*, et 1 247 avec la classe *fortement*

spontanée. Durant ces expériences, nous avons utilisé la méthode du *Leave one out* : 10 fichiers sont choisis pour l'apprentissage, pendant que le fichier restant sert pour l'évaluation. Puis, ce processus est répété jusqu'à ce que tous les fichiers aient été évalués. Cette méthode nous a également permis d'estimer les probabilités contextuelles $P(s_i|s_{i-1}, s_{i+1})$ de la méthode globale (voir partie 4.4.2.2).

4.5.1.2 Performances du système de RAP

Les caractéristiques acoustiques et linguistiques utilisées pour caractériser le niveau de spontanéité proviennent du système de RAP du LIUM décrit dans la section 1.5. Le tableau 4.7 présente les résultats en terme de taux d'erreur-mot (WER) et d'entropie croisée normalisée (NCE) de ce système de RAP sur les données expérimentales. Nous n'avons pas inclus ces données durant la phase d'apprentissage et de développement des modèles utilisés par le système de RAP. Le WER est la métrique classique d'évaluation des performances des systèmes de RAP, alors que le NCE est habituellement choisi pour évaluer les mesures de confiance proposées par un système de RAP.

Le tableau 4.7 montre que les performances globales du système de RAP, avec un WER de 15,0 % et un NCE de 0,383, sont très bonnes pour un système traitant des journaux d'information en français. Comme attendu, plus la parole est fluide, plus le WER est bas ; de 10,1 % pour les segments annotés manuellement en parole *préparée*, jusqu'à 28,5 % pour les segments annotés en parole *fortement spontanée*. Il est intéressant de noter qu'il existe une corrélation entre l'annotation subjective en classe de spontanéité et le WER obtenu pour un système de RAP.

Classe	**Durée**	**# Segments**	**WER**	**NCE**
Préparée	3h40	1 228	10,1 %	0,395
Légèrement spontanée	3h50	1 339	18,4 %	0,376
Fortement spontanée	3h30	1 247	28,5 %	0,348
Total	11h	3 814	15,0 %	0,383

TAB. 4.7 – Performances du système de RAP en fonction de la classe de parole en termes de WER et de NCE. Le nombre de segments et la durée liés à la classe de parole sont également inclus.

Notons que le temps de calcul du système de RAP du LIUM pour cette tâche est de 10 fois le temps réel, incluant l'alignement phonétique et le calcul des mesures de confiance. De plus, un tel calcul est facilement fractionnable. Ainsi, nous pouvons, par exemple, utiliser de manière répartie un processeur par fichier à traiter, pour réduire le temps d'attente de traitement, le décodage de chaque fichier fonctionnant en parallèle.

4.5.1.3 Détection et catégorisation automatiques des segments de parole

Afin de mesurer le gain fourni pour chaque famille de descripteurs, nous nous intéressons dans un premier temps à la détection et la catégorisation de chaque segment de manière indépendante. Quatre conditions sont évaluées :

- *ling(ref)* : les caractéristiques linguistiques seules sur la transcription de référence,
- *ling(rap)* : les caractéristiques linguistiques seules sur la transcription automatique (fournie par notre système de RAP),
- *acous(rap)* : les caractéristiques acoustiques seules sur la transcription automatique,
- *all(rap)* : toutes les caractéristiques (acoustiques et linguistiques) sur la transcription automatique.

Nous pouvons alors comparer les résultats sur les transcriptions de référence avec les transcriptions obtenues automatiquement par le système de RAP. Notons que nous n'avons pas de résultats sur les caractéristiques acoustiques de référence, puisque l'obtention des valeurs de ces informations peut difficilement se faire manuellement. De plus, bien que ces caractéristiques puissent être calculées à partir des transcriptions manuelles au moyen d'un système de RAP, il reste le biais du dictionnaire de phonétisation. En effet, celui-ci n'intègre pas forcément la prononciation exacte de chaque occurrence de mots de la référence.

Le tableau 4.8 présente les résultats sur la détection (précision et rappel) pour chaque classe de spontanéité. Nous constatons que les performances sur la détection des segments *légèrement spontanés* sont moins bonnes que sur les autres classes. Ce constat n'est pas surprenant, les segments pouvant facilement être faussement catégorisés en parole *préparée* d'un côté, et en parole *fortement spontanée* de l'autre.

Préparée				
Caractéristique	*ling(ref)*	*ling(rap)*	*acous(rap)*	*all(rap)*
Rappel	64,1	61,8	59,0	65,1
Précision	56,0	53,0	56,3	59,8
Légèrement spontanée				
Caractéristique	*ling(ref)*	*ling(rap)*	*acous(rap)*	*all(rap)*
Rappel	37,7	31,7	41,2	43,5
Précision	43,8	40,7	44,3	47,0
Fortement spontanée				
Caractéristique	*ling(ref)*	*ling(rap)*	*acous(rap)*	*all(rap)*
Rappel	65,9	60,4	62,2	66,3
Précision	65,2	58,0	60,5	66,7

TAB. 4.8 – Précision et rappel de la classification des segments de parole en fonction des trois classes de spontanéité et des caractéristiques extraites.

Dans ce tableau, nous remarquons une baisse des performances lorsque nous utilisons les caractéristiques linguistiques issues des transcriptions automatiques (*ling(rap)*) par rapport aux caractéristiques linguistiques issues des transcriptions de référence (*ling(ref)*). Cette différence est due aux erreurs du système de transcription. Cependant, nous pouvons voir que cet écart est compensé, et même amélioré, grâce aux caractéristiques acoustiques, plus robustes aux erreurs des systèmes de RAP. Ainsi, au moyen d'un classifieur s'appuyant sur toutes les caractéristiques (acoustiques, linguistiques et mesures de confiance) extraites automatiquement (*all(rap)*), une amélioration des performances est visible. En effet, peu importe la classe de spontanéité ou la métrique utilisée, nous obtenons de meilleurs résultats en utilisant la combinaison des caractéristiques issues du système de RAP par rapport à celles extraites des transcriptions de référence.

Nous avons ensuite pris en compte les résultats obtenus sur l'attribution des classes de spontanéité de chaque segment afin de réaliser une nouvelle décision pour chaque segment, mais cette fois-ci en prenant une décision globale et non plus locale. Pour mesurer le gain obtenu grâce à cette méthode statistique globale, nous avons réalisé deux expériences :

- *global(ref)* : le modèle global utilisant les résultats d'étiquetage des segments obtenus avec les caractéristiques des transcriptions de référence *ling(ref)*,
- *global(rap)* : le modèle global utilisant les résultats d'étiquetage des segments obtenus avec les caractéristiques des transcriptions automatiques *all(rap)*.

Le tableau 4.9 présente les résultats sur la détection (précision et rappel) pour chaque classe de spontanéité, en appliquant un modèle global sur les résultats déjà obtenus pour chaque segment. Les gains relatifs, par rapport à la détection au niveau du segment (*ling(ref)* et *all(rap)*), sont fournis pour pouvoir plus facilement comparer les résultats. Nous constatons que l'utilisation du modèle contextuel probabiliste améliore fortement les performances sur l'étiquetage des classes de spontanéité, et ce quelle que soit la classe de spontanéité et la métrique choisie (rappel/précision).

Nous remarquons aussi que le gain obtenu grâce au modèle global probabiliste est important, tant au niveau de l'utilisation des transcriptions de référence (*global(ref)*), que de l'utilisation de l'ensemble des caractéristiques obtenues automatiquement au moyen du système de RAP (*global(rap)*). Notons également que les performances sont globalement meilleures avec *global(rap)* que *global(ref)*. Ces constats montrent bien la robustesse de notre système pour la détection automatique de la classe de spontanéité des segments de parole, notamment avec l'utilisation de caractéristiques acoustiques issues du système de RAP.

Nous constatons, dans le tableau 4.9, que la proportion des nouveaux gains obtenus, en comparaison à notre méthode se focalisant sur chaque segment indépendamment, n'est pas la même selon la classe de spontanéité. En effet, la classe de spontanéité *légèrement spontanée* obtient la meilleure progression au niveau des résultats, très visible surtout sur *global(rap)*

Préparée		
Caractéristique	*global(ref)*	*global(rap)*
Rappel	66,5 (+3,6%)	69,6 (+6,5%)
Précision	61,6 (+9,1%)	66,8 (+10,5%)
Légèrement spontanée		
Caractéristique	*global(ref)*	*global(rap)*
Rappel	42,8 (+11,9%)	51,8 (+16,0%)
Précision	46,9 (+6,6%)	54,3 (+13,4%)
Fortement spontanée		
Caractéristique	*global(ref)*	*global(rap)*
Rappel	71,5 (+7,8%)	73,5 (+9,8%)
Précision	70,3 (+7,3%)	73,0 (+8,6%)

TAB. 4.9 – Précision et rappel de la classification des segments de parole en fonction des trois classes de spontanéité en appliquant un modèle global sur les résultats déjà obtenus pour chaque segment.

(gain relatif de 16,0 % pour le rappel et de 13,4 % pour la précision). Le rappel pour cette classe (nombre de segments bien étiquetés par rapport au nombre total à étiqueter) est ainsi beaucoup plus élevé, tout en permettant de nettement améliorer la précision de la détection (le système réalise moins d'erreurs de classification).

Pour bien comprendre la façon dont ont été catégorisés les segments et l'impact précis de la méthode statistique globale, le tableau 4.10 présente la matrice de confusion obtenue pour la détection au niveau des segments en utilisant toutes les caractéristiques (acoustiques et linguistiques) obtenues automatiquement (*all(rap)*). Puis le tableau 4.11 présente la matrice de confusion obtenue en utilisant le modèle probabiliste global (*global(rap)*). La matrice de confusion représente le nombre global de segments de parole catégorisés automatiquement par classe de spontanéité. Ces résultats permettent d'estimer le rappel et la précision pour chaque classe. Exemple de lecture pour la matrice de confusion 4.10, si nous prenons le cas de la parole *préparée* :

- En lisant le tableau de gauche à droite, nous avons 1 228 segments à classifier en tant que parole *préparée*. Sur ces 1 228 segments, 799 segments ont effectivement été bien catégorisés en tant que parole *préparée*, 343 segments ont été faussement catégorisés en parole *faiblement spontanée*, et 86 segments ont été faussement catégorisés en parole *fortement spontanée*.
- En lisant le tableau de haut en bas, 1 337 segments ont été estimés en tant que parole *préparée* par le classifieur. Sur ces 1 337 segments, 799 segments ont effectivement été bien catégorisés en tant que parole *préparée*, 430 segments auraient dû être catégorisés

en tant que parole *faiblement spontanée*, et 108 segments auraient dû être catégorisés en tant que parole *fortement spontanée*.

		Classe estimée			
		Préparée	**Faiblement spon.**	**Fortement spon.**	**Total**
Classe réelle	*Préparée*	799	343	86	1 228
	Faiblement spon.	430	582	327	1 339
	Fortement spon.	108	312	827	1 247
	Total	1 337	1 237	1 240	

TAB. 4.10 – Matrice de confusion sur la classification des segments de parole en classe de spontanéité avec *all(rap)*.

		Classe estimée			
		Préparée	**Faiblement spon.**	**Fortement spon.**	**Total**
Classe réelle	*Préparée*	855	315	58	1 228
	Faiblement spon.	363	694	282	1 339
	Fortement spon.	62	268	917	1 247
	Total	1 280	1 277	1 257	

TAB. 4.11 – Matrice de confusion sur la classification des segments de parole en classe de spontanéité avec *global(rap)*.

En analysant la matrice de confusion 4.10 pour *all(rap)*, nous constatons que la classe de spontanéité *faiblement spontanée* est un réel point faible pour notre système. Cependant, cela ne semble pas surprenant. En effet, l'objectif initial de ce détecteur automatique de type de parole est de retrouver le plus précisément possible les segments contenant de la parole *fortement spontanée*. Or, nous constatons ici que le nombre de segments faussement étiquetés entre la parole *préparée* et *fortement spontanée* est assez élevé. C'est ce type d'erreur que nous voulions corriger avec notre méthode probabiliste globale. Et en analysant le tableau 4.11, nous nous apercevons que le fait de prendre en compte les segments voisins lors du processus de décision permet de grandement diminuer cette erreur, tout en améliorant les résultats sur la classe *faiblement spontanée*. En comparant les matrices, nous nous apercevons que nous pouvons réduire de 38,1 % en relatif les erreurs de classification entre les classes de parole *préparée* et *fortement spontanée*.

Comme nous l'évoquions précédemment dans la formule 4.1, nous utilisons les mesures de confiance $c(s_i)$ fournies par le classifieur pour chaque classe de spontanéité de chaque segment de parole. De plus, en combinant les scores $c(s_i)$ avec les probabilités $P(s_i|s_{i-1}, s_{i+1})$ fournies par le modèle contextuel, il est possible de filtrer la proposition de classe en appliquant un seuil

sur la valeur de $c(s_i) \times P(s_i|s_{i-1}, s_{i+1})$. La figure 4.5 présente les performances de détection de la classe de parole *fortement spontanée* obtenues en faisant varier ce seuil. La figure montre d'une part les résultats obtenus par le classifieur en faisant varier $c(s_i)$ (*all(rap)* et *ling(rap)*), et d'autre part les résultats obtenus au moyen de la méthode globale (*global(ref)* et *global(rap)*). Nous choisissons, ici, de nous focaliser sur la classe *fortement spontanée* car elle représente la classe de spontanéité sur laquelle nous souhaitons améliorer les performances des systèmes de RAP en réalisant des traitements spécifiques.

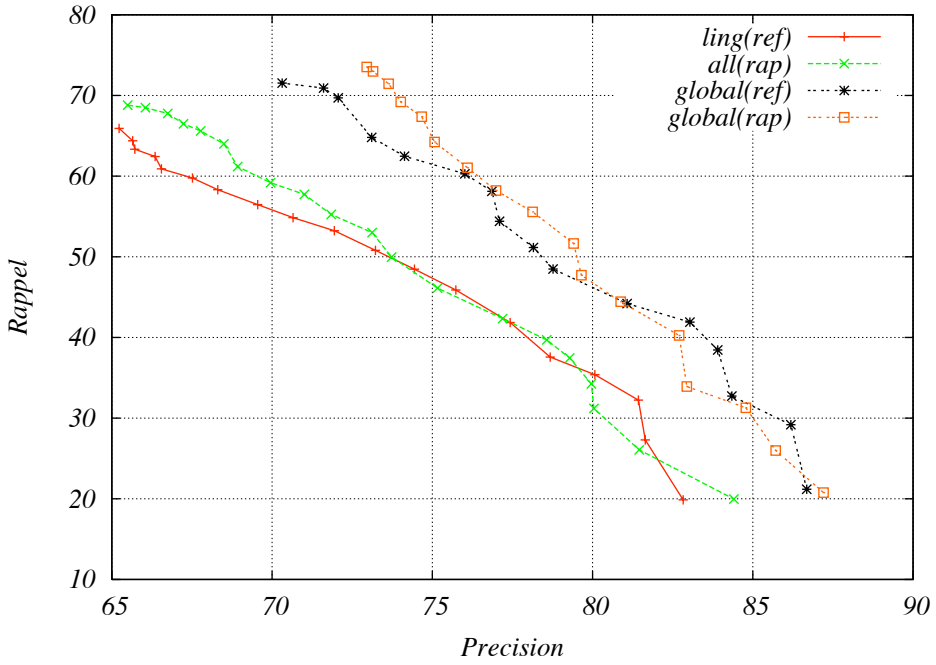

FIG. 4.5 – Performances sur la détection des segments de parole *fortement spontanée* en fonction du seuil choisi sur le score de classification.

Dans la figure 4.5, nous constatons que notre système peut être plus précis quand moins de décisions sont prises (baisse du rappel et augmentation de la précision). Cette possibilité de seuillage permet d'adapter l'utilisation de la méthode de classification en cherchant le meilleur compromis entre le rappel et la précision pour l'application souhaitée.

De plus, nous pouvons remarquer que *ling(ref)* et *all(rap)* suivent la même tendance, même lorsque le seuil de décision varie. Nous pouvons alors en conclure que les caractéristiques acoustiques, linguistiques, et les mesures de confiance (*all(rap)*) extraites dans un système de RAP, permettent d'atteindre le même niveau de performance que l'utilisation seule des caractéristiques linguistiques issues de la transcription de référence *ling(ref)*. La perte d'informations due aux erreurs de transcription peut donc être compensée au moyen d'autres caractéristiques extraites d'un système de RAP.

4.5.2 Conclusion

Dans ce chapitre, nous avons proposé une méthode permettant de caractériser la classe de spontanéité d'un segment de parole, plus particulièrement adapté à la détection de la parole très spontanée. Un corpus a été préalablement étiqueté manuellement par deux annotateurs humains en niveaux de spontanéité pour chaque segment de parole. À partir de cet étiquetage, trois classes de parole ont été définies : parole *préparée*, *légèrement spontanée* et *fortement spontanée*. Ce corpus annoté constitue la base de nos expérimentations. Nous nous en servons pour tester notre méthode de classification des segments de parole. Différentes caractéristiques acoustiques et linguistiques ont été définies pour pouvoir caractériser la classe de spontanéité de chaque segment. La méthode se déroule en deux étapes. Tout d'abord, les différentes caractéristiques (acoustiques, linguistiques, et mesures de confiance) sont extraites au moyen d'un système de RAP pour chaque segment de parole, puis elles sont utilisées lors d'un processus de classification, au moyen d'un classifieur intégrant l'algorithme *AdaBoost*. Pour chaque segment, une classe de spontanéité lui est alors attribuée. À noter que dans cette étape, la classification est indépendante pour chaque segment de parole. La seconde étape consiste à ne plus considérer le segment de parole comme isolé, mais plutôt lié à un flux de parole. La méthode prend alors en compte les classes potentielles de spontanéité des segments voisins pour prendre une décision au niveau du segment courant. Le processus d'attribution d'une classe de spontanéité devient alors global. Au moyen d'un modèle probabiliste, une réestimation des scores de classification, obtenus à l'étape précédente, a été réalisée, permettant d'affiner le processus de décision sur la classe de spontanéité.

La méthode proposée permet d'atteindre des niveaux de classification en classes de spontanéité très intéressants pour chaque segment. Elle peut notamment permettre une réalisation de traitements spécifiques selon la classe de spontanéité. De plus, il est possible de définir un seuil sur les scores fournis par la méthode. Ainsi, selon l'application visée, il est possible de rendre la détection des classes de spontanéité encore plus précise (en réduisant bien sûr le nombre de choix pris par le système). Ces résultats de classification, notamment pour la parole très spontanée, peuvent être utilisés pour améliorer les systèmes de RAP [Dufour 2010]. Nous en verrons une application détaillée dans la partie suivante. Cette méthode peut facilement être combinée avec d'autres approches. Une combinaison avec des approches issues de l'identification automatique des langues est présentée dans [Rouvier 2010].

Dans le chapitre suivant, des expériences sur le dictionnaire de prononciations seront réalisées. Ce dictionnaire de prononciations sera modifié, en y ajoutant des variantes de prononciation spécifiques à la parole spontanée. L'idée est d'estimer l'impact de ces variantes sur

le décodage de segments de parole spontanée dans un système de RAP. Nous nous intéresserons également aux modèles acoustiques et linguistiques. Nous proposons une méthode non-supervisée pour adapter ces modèles à la parole spontanée, sans ajout de données d'apprentissage supplémentaires.

Chapitre 5

Modélisation spécifique de la parole spontanée pour la reconnaissance de la parole

Sommaire

5.1 Dictionnaire et variantes de prononciation 114

5.1.1 Analyse de variantes de prononciation spécifiques à la parole spontanée 114

5.1.2 Construction du nouveau dictionnaire de prononciations 116

5.1.3 Expériences 116

5.1.4 Résultats 117

5.1.5 Analyse des erreurs 118

5.1.5.1 Au niveau de variantes de prononciation 118

5.1.5.2 Au niveau du type de parole 119

5.1.5.3 Au niveau du segment 120

5.2 Adaptation des systèmes de RAP 122

5.2.1 Principe général 122

5.2.1.1 Adaptation non-supervisée des modèles acoustiques et de langage 122

5.2.1.2 Combinaison des systèmes 123

5.2.2 Adaptation automatique des modèles 125

5.2.2.1 Modélisation acoustique 125

5.2.2.2 Modélisation linguistique 125

5.2.3 Corpus 126

5.2.4 Expériences 127

5.2.4.1 Analyse du système adapté 128
5.2.4.2 Combinaison des systèmes 129
5.2.5 Conclusion 130
5.3 Approches spécifiques : le cas de l'homophonie en français 131
5.3.1 Approche proposée 131
5.3.1.1 Méthodologie générale 131
5.3.1.2 Règle grammaticale 133
5.3.1.3 Méthode statistique 134
5.3.2 Expériences réalisées 137
5.3.2.1 Mots et classes de mots étudiés 137
5.3.2.2 Outils 137
5.3.2.3 Données expérimentales 138
5.3.3 Résultats obtenus 139
5.3.3.1 Avec les règles grammaticales 139
5.3.3.2 Avec la méthode statistique 140
5.3.4 Conclusion 146
5.4 Résultats finaux des méthodes spécifiques 147
5.5 Perspectives 148

Les chapitres précédents ont permis de comprendre les difficultés des systèmes de RAP pour traiter la parole spontanée. Les différences entre ce type de parole et la parole préparée sont importantes. Cela conduit inévitablement à rendre les systèmes de RAP moins performants lorsqu'ils doivent transcrire de la parole spontanée, puisqu'ils ne gèrent pas correctement les spécificités de ce type de parole. Les données souvent disponibles pour entraîner les modèles (au niveau acoustique et linguistique) des systèmes de RAP sont principalement composées de parole préparée (journaux d'information, textes lus...) et de textes écrits (articles de journaux, Internet...). Comme nous l'avons développé dans le chapitre 2.3, il semble indispensable d'adapter les différents modèles d'un système de RAP (modèle de langage, modèle acoustique, dictionnaire de prononciations...) sur la parole spontanée pour améliorer les performances générales.

Les dictionnaires de prononciations, utilisés dans les systèmes de RAP, sont composés de variantes de prononciation obtenues par des experts linguistes, mais aussi par des approches statistiques automatiques (voir section 1.3.2.2). Les variantes ainsi obtenues sont souvent définies dans un cadre général de prononciation et ne prennent pas en compte les spécificités de prononciation de la parole spontanée. Au niveau des modèles acoustiques et modèles de langage, il semble relativement difficile de construire des modèles spécifiques à la parole spontanée, puisqu'un grand nombre de données est nécessaire.

Certaines propositions traitent la parole spontanée comme un type de parole spécifique, qui requiert des traitements particuliers (voir partie 2.3). Nous cherchons, à travers nos travaux, à poursuivre cette approche en proposant une méthode spécifique pour l'amélioration des systèmes de RAP sur la parole spontanée. Cette idée de spécificité nous a également amenés à nous intéresser à une particularité de la langue, l'homophonie, que l'on retrouve très fréquemment en français. Au même titre que la parole spontanée, l'homophonie peut nécessiter des approches spécifiques, comme nous l'avons vu dans la partie 3.3.2.

La section 5.1 s'intéressera au travail que nous avons réalisé au niveau du dictionnaire de prononciations. Nous avons choisi d'y ajouter, manuellement, des variantes de prononciation adaptées à la parole spontanée. Ensuite, la section 5.2 présentera la technique d'adaptation automatique des données d'apprentissage que nous proposons, afin de rendre les modèles de langage et modèles acoustiques plus adaptés à la parole spontanée. Enfin, nous nous intéresserons au cas de l'homophonie en français. Nous décrirons notre proposition de méthodes spécifiques de correction d'homophones hétérographes, en post-traitement des sorties de systèmes de RAP.

5.1 Dictionnaire et variantes de prononciation

5.1.1 Analyse de variantes de prononciation spécifiques à la parole spontanée

Le dictionnaire de prononciations, du système de RAP du LIUM, a été développé pour couvrir des journaux radiophoniques d'information, et notamment pour une reconnaissance de la parole continue à vocabulaire étendu. En d'autres termes, le système de RAP ne doit pas être contraint par un domaine particulier, le vocabulaire devant couvrir des thèmes variés. Ces variantes de prononciation, contenues dans le dictionnaire, ne conviennent pas nécessairement à la parole spontanée.

Le fait de ne pas posséder la bonne prononciation d'un mot dans le dictionnaire, a pour conséquence de voir ce mot supprimé ou substitué par un autre mot par le système de RAP. Certaines prononciations peuvent changer selon le type de parole. La figure 5.1 présente un exemple de prononciation pour le mot "*elle*" dont la prononciation peut évoluer dans le cas d'un flux de parole spontanée. Lorsqu'il est prononcé de manière lue (en prononçant parfaitement tous les phonèmes), le mot "*elle*" devrait contenir les deux phonèmes [ɛl]. Cette seule entrée est présente dans le dictionnaire de prononciations. Or, dans ce cas, le mot "*elle*" peut très bien se prononcer [ɛ] dans le cadre de la parole spontanée, en omettant le phonème / l /, élidé dans le flux de parole. Cette variante de prononciation, inexistante dans le dictionnaire, conduit alors, dans notre exemple, le système de RAP à transcrire le mot "*est*" à la place du mot attendu "*elle*".

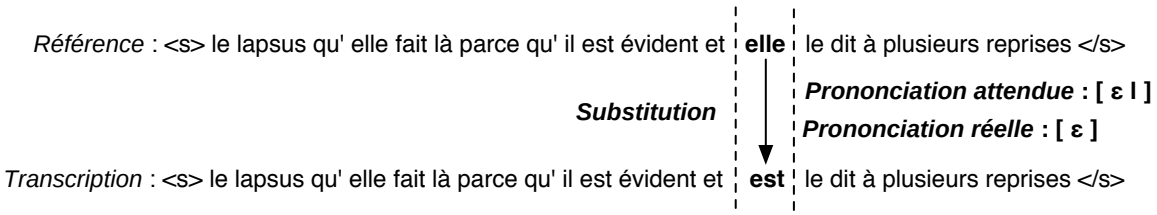

FIG. 5.1 – Exemple de mauvaise reconnaissance du mot "*elle*" par le système de RAP dans un flux de parole spontanée.

Au travers du constat réalisé dans cet exemple, nous avons voulu évaluer l'impact de l'ajout de variantes de prononciation spécifiques à la parole spontanée dans le dictionnaire de prononciations du système de RAP, comme présenté dans [Dufour 2008a]. Afin de pouvoir ajouter de nouvelles variantes, un travail préliminaire d'analyse de variantes de prononciation pour la parole spontanée a été entrepris. Ce premier travail a été réalisé avec un linguiste, possédant

les connaissances théoriques sur le changement de prononciation des mots. En effet, différentes règles de prononciation existent lorsque la parole est dite spontanée. Nous la retrouvons généralement dans les interventions non préparées à l'avance. Nous nous sommes alors basés sur l'analyse réalisée par [Bazillon 2008b], qui entreprend de décrire la parole spontanée en opposition à la parole préparée. Cette analyse a mis en lumière des différences au niveau de la prononciation de certains mots.

Nous avons ainsi choisi différentes règles et prononciations spécifiques pour notre étude, en nous basant sur les études linguistiques existantes à ce sujet (voir partie 2.1). Nous nous sommes tout d'abord focalisés sur le problème particulier de la vibrante / ʁ /, que l'on retrouve en parole spontanée en français, comme développé dans la partie 2.1.1.3. Cette règle s'applique spécialement sur les mots de terminant par "*-bre*", "*-cre*", "*-dre*", "*-tre*" ou "*-vre*", sachant que le phonème / ʁ / est obligatoire lorsque l'on prononce ces mots dans un registre de langage standard. Ce phonème / ʁ / a tendance à disparaître plus fréquemment en parole spontanée. Nous avons également observé des différences de prononciation, lors de parole spontanée, sur certains mots très courants de la langue française. Le tableau 5.1 récapitule les mots étudiés, avec leur variante de prononciation "standard", comparable à de la parole lue, préparée, et leur variante de prononciation dans le cadre de la parole spontanée, conçue au fil de l'énonciation.

	Phonèmes "parole préparée"	**Phonèmes "parole spontanée"**
il	[il]	[i]
elle	[ɛl]	[ɛ]
enfin	[ãfɛ̃]	[fɛ̃]
parce-que	[paʁs(e)kə]	[pas(e)kə]
je	[ʒ(ə)]	[ʃ]
au-dessus	[od(ə)sy]	[ot(ə)sy]
de	[də]	[t]
puis	[pɥi]	[pi]

TAB. 5.1 – Comparaison des prononciations en parole préparée et en parole spontanée de certains mots fréquents en français.

Dans ce tableau, nous constatons très souvent que les mots subissent un phénomène d'élision en parole spontanée. En effet, les mots tels que "*elle*" ou encore "*enfin*" perdent des phonèmes (respectivement / l / et / ʁ /). Des phénomènes d'assimilation sont également visibles au niveau du mot "*je*" [ʒ(e)], qui en parole spontanée, devient [ʃ] devant une consonne sourde, comme développé dans la partie 2.1.1.3 (exemple "*j' pense*" [ʃpãs(e)]). La fréquence de ces mots peut aussi expliquer ce changement de prononciation, avec le phénomène d'assimilation : l'effort de prononciation est minime, leur compréhension étant aisée.

5.1.2 Construction du nouveau dictionnaire de prononciations

Nous avons opté pour l'ajout des variantes de prononciation spécifiques à la parole spontanée pour mesurer l'impact du dictionnaire de prononciations lors d'un décodage au moyen d'un système de RAP [Dufour 2008a]. Dans cette optique, nous nous sommes focalisés sur certaines spécificités décrites dans la partie 5.1.1.

La règle de la vibrante / r / a été très simplement modélisée. Cela nous a permis d'ajouter aux mots se terminant par "*-bre*", "*-cre*", "*-dre*", "*-tre*" ou "*-vre*" ces nouvelles prononciations. De plus, nous avons inséré les variantes de prononciation aux mots décrits dans le tableau 5.1.

Initialement, le dictionnaire de prononciations contenait environ 165 000 variantes de prononciation pour un total d'environ 62 000 mots. Suite à la modification du dictionnaire, 1 761 nouvelles variantes de prononciation ont été ajoutées, 638 mots ont été modifiés, soit au total 1,02 % des mots du dictionnaire, et un ajout de 1,06 % au niveau des variantes de prononciation.

5.1.3 Expériences

Un corpus de test est nécessaire pour évaluer l'impact de l'ajout des variantes de prononciation dans le dictionnaire (voir 5.1.2) lors de la transcription au moyen d'un système de RAP. Pour mener à bien nos expériences, nous avons utilisé une partie des fichiers audio issus du projet EPAC, présenté dans la partie 1. Ces fichiers audio proviennent d'enregistrements radiophoniques de RFI, France Inter, et France Culture, auxquels nous avons ajouté des enregistrements audio provenant de RMC. L'ensemble de ces fichiers représente un total de 14 heures, une très forte proportion de parole spontanée justifiant ce choix.

Les transcriptions automatiques initiales obtenues à partir des fichiers audio (sans modification du dictionnaire de prononciations) contiennent 13 355 mots distincts. En comparant ces transcriptions avec les variantes de prononciation ajoutées dans le dictionnaire, 203 mots distincts (ce qui représente 1,52 % des mots) sont susceptibles d'être influencés cette modification (au moins une nouvelle variante de prononciation pour chacun de ces mots).

L'analyse des occurrences de mots est cependant légèrement différente de l'analyse des mots distincts. En effet, 178 960 occurrences de mots sont présentes dans les transcriptions des fichiers audio. De plus, 14 943 occurrences de mots ont au moins une nouvelle prononciation si l'on utilise le nouveau dictionnaire. La proportion des occurrences de mots éventuellement influencées par le nouveau dictionnaire est alors de 8,35 %. Cette large différence peut s'expliquer par le fait que les prononciations ajoutées à certains mots ont été faites sur des mots usuels très courants en français parlé, d'où un nombre d'occurrences élevé.

5.1.4 Résultats

Pour analyser l'impact des nouvelles variantes de prononciation, les premières expériences ont été évaluées en simplifiant le taux d'erreur-mot (WER) en ne comptant que les erreurs de substitutions et de suppressions. En effet, nous voulons savoir si les nouvelles variantes de prononciation permettent de mieux reconnaître les mots concernés par ces ajouts. Cette meilleure reconnaissance peut être visible à deux niveaux :

– soit en remplaçant un mot mal transcrit par le bon mot (phénomène de substitution),
– soit en faisant apparaître un mot qui n'était pas transcrit auparavant et qui l'est grâce à l'ajout d'une nouvelle prononciation (limitation du phénomène de suppression).

Le tableau 5.2 se focalise sur les mots affectés par la modification du dictionnaire de prononciations en présentant le nombre ainsi que le taux d'erreurs (pour les suppressions et substitutions) de ces mots avant (*Base*) et après modification du dictionnaire (*Base+n^{elles} pron.*). Le gain entre les deux expériences est également présenté. Comme attendu, les nouvelles variantes de prononciation ont un impact positif sur les erreurs de substitution et de suppression des mots affectés.

	Base	**Base + n^{elles} pron.**	**Gain**
Nombre d'erreurs	3 369	3 111	**258**
Taux d'erreurs	22,54 %	20,82 %	**1,72 %**

TAB. 5.2 – Nombre et taux d'erreurs (substitutions et suppressions) pour les mots concernés par l'ajout de variantes de prononciation.

Cette expérience ne s'est intéressée qu'aux mots ayant subi une modification au niveau du dictionnaire de prononciations. Nous voulons maintenant évaluer l'impact au niveau de la transcription générale. Dans ce contexte, le tableau 5.3 présente le nombre et le taux d'erreurs sur l'ensemble de la transcription des fichiers audio avant (*Base*), et après modification du dictionnaire de prononciations (*Base + n^{elles} pron.*), en se focalisant sur les erreurs de suppression et de substitution. L'ajout de ces prononciations permet d'obtenir une réduction, même si elle apparaît relativement faible, des erreurs de substitution et de suppression.

	Base	**Base + n^{elles} pron.**	**Gain**
Nombre	46 509	45 061	**1 448**
Taux	25,99 %	25,18 %	**0,81 %**

TAB. 5.3 – Nombre et taux d'erreurs globaux (substitutions et suppressions) avant et après l'ajout de nouvelles variantes de prononciation.

Finalement, il faut évaluer l'impact global de l'ajout de ces variantes de prononciation au niveau du taux d'erreur-mot global, c'est-à-dire en ajoutant les erreurs d'insertion aux erreurs de substitution et de suppression. Le tableau 5.4 présente les taux d'erreur-mot globaux avant et après l'ajout de variantes de prononciation.

	Base	**Base + n^{elles} pron.**	**Gain**
Taux	37,5 %	38,6 %	- 1,1 %

TAB. 5.4 – Taux d'erreur-mot globaux (suppressions, substitutions et insertions) avant et après l'ajout de nouvelles variantes de prononciation.

Globalement, l'ajout des variantes de prononciation a un impact négatif sur le WER. Nous pouvons donc en conclure que le taux d'insertions a nettement augmenté avec le nouveau dictionnaire. Cette augmentation s'explique par l'ajout de variantes de prononciation de mots usuels courts (par exemple "*de*", "*je*" ou encore "*il*"). Notons également que le reste du système n'a pas été modifié lors de l'utilisation du nouveau dictionnaire : l'optimisation de différents paramètres, comme la pénalité d'insertion des mots, pourrait améliorer le système. De plus, assigner des poids aux variantes de prononciation permettrait de privilégier certaines variantes de prononciation. En effet, toutes les prononciations fournies dans le dictionnaire ne possèdent pas la même probabilité d'apparition, et ce, selon le type de parole. Ainsi, dans le cadre de la parole spontanée, il pourrait être utile de proposer un poids supplémentaire aux variantes spécifiques à ce type de parole. Finalement, cette étude initiale que nous avons menée prouve qu'un dictionnaire de prononciations adapté est important, car un gain apparaît envisageable : les variantes de prononciation introduites ont permis de réduire les erreurs de substitution et de suppression sur les mots touchés par cet ajout.

5.1.5 Analyse des erreurs

5.1.5.1 Au niveau de variantes de prononciation

Une analyse sur les occurrences de mots bien transcrites avec le dictionnaire modifié, et dont une nouvelle prononciation a été choisie lors du décodage, a été entreprise afin de mesurer l'impact de ces prononciations sur le taux d'erreurs (par rapport au dictionnaire de base). Le nombre d'occurrences de mots correctement transcrites avec le dictionnaire modifié atteint le nombre de 500. Le tableau 5.5 présente l'analyse réalisée sur ces occurrences de mots correctement transcrites par le dictionnaire modifié, par rapport au dictionnaire de base.

Aucun effet n'a été constaté sur 56,8 % des occurrences de mots bien transcrites, c'est-à-dire que ces occurrences étaient déjà bien transcrites avec le dictionnaire de base. Par contre,

	Aucun effet	**Erreurs corrigées**
Nombre	284	**216**
Proportion	56,8 %	**43,2 %**

TAB. 5.5 – Analyse (nombre et proportion) des occurrences de mots correctement transcrites au moyen du nouveau dictionnaire par rapport au dictionnaire de base.

43,2 % d'entre elles sont maintenant correctement transcrites avec le dictionnaire modifié alors qu'elles ne l'étaient pas avec le dictionnaire de base.

En s'inspirant de l'analyse précédente, nous cherchons désormais à savoir si l'utilisation des nouvelles variantes ne dégrade pas la reconnaissance des mots initialement bien transcrits au moyen du dictionnaire de base, et dont une variante de prononciation (au moins) a été ajoutée. Le nombre d'occurrences de mots concernées par une nouvelle variante de prononciation (et bien transcrites au moyen du dictionnaire de base) est de 4 221. Le tableau 5.6 présente l'analyse réalisée sur ces occurrences correctement transcrites par le dictionnaire de base, par rapport au dictionnaire modifié.

	Erreurs ajoutées	**Aucun effet**
Nombre	144	**4 077**
Proportion	3,4 %	**96,6 %**

TAB. 5.6 – Analyse (nombre et proportion) des occurrences de mots correctement transcrites au moyen du dictionnaire de base par rapport au dictionnaire modifié.

Il apparaît que 3,4 % des occurrences de mots correctement transcrites au moyen du dictionnaire de base sont maintenant mal reconnues avec le nouveau dictionnaire. Aucun effet n'a cependant été constaté sur 96,6 % d'entre elles, dans le sens où ces occurrences étaient bien reconnues au moyen du dictionnaire de base, et le sont toujours avec le dictionnaire modifié. En conclusion de ces deux précédentes analyses, nous pouvons dire que, sur les mots touchés par le nouveau dictionnaire de prononciations, le gain obtenu est supérieur à l'ajout d'erreurs résultant de ces nouvelles variantes de prononciation.

5.1.5.2 Au niveau du type de parole

L'analyse précédente (partie 5.1.5.1) s'est intéressée à la reconnaissance des mots au moyen d'un dictionnaire de prononciations modifié (ajout de variantes de prononciation) sur de la parole spontanée. Nous voulons maintenant savoir si les dictionnaires manipulés (base et modifié) ont le même impact selon le type de parole rencontré dans les fichiers audio. Nous cherchons à comparer l'influence des prononciations sur la parole préparée et sur la parole spontanée.

Afin de réaliser cette analyse, des fichiers audio issus de la campagne ESTER ont été utilisés, contenant une majeure partie de parole préparée, ainsi que des fichiers audio issus du projet EPAC, contenant principalement de la parole spontanée. Les taux d'erreur-mot obtenus par le système de RAP du LIUM atteste bien des différences entre les deux corpus qui seront analysés : le taux d'erreur-mot est de 19,6 % sur les données ESTER, alors que le taux d'erreur-mot est de 29,7 % sur le corpus EPAC.

Pour les données ESTER et EPAC, les phonèmes ont été alignés sur le signal de parole au moyen du nouveau dictionnaire de prononciations que nous avons présenté dans les expériences précédentes (voir partie 5.1.2). Cette étape d'alignement constitue notre référence, la variante de prononciation associée à chaque mot étant supposée la meilleure possible. Durant la phase de décodage, le dictionnaire de base sera utilisé, afin de savoir si, lorsque la meilleure prononciation pour transcrire un mot est absente, le mot est malgré tout correctement transcrit.

Le tableau 5.7 présente les nombres et taux des mots bien reconnus par le système de RAP, bien que leur variante de prononciation, choisie durant le processus de décodage, ne corresponde pas exactement à la meilleure variante de prononciation (obtenue par le processus d'alignement). Cette analyse mesure la capacité des systèmes de RAP à corriger la distance entre la prononciation attendue et celle proposée par le dictionnaire.

	Parole préparée (ESTER)	**Parole spontanée (EPAC)**
Nombre	2 210	2 785
Taux de reconnaissance	71,85 %	49,51 %

TAB. 5.7 – Nombre et taux des mots correctement reconnus sachant que la variante de prononciation associée pendant le décodage diffère de celle choisie pendant le processus d'alignement.

Il apparaît très clairement que le taux de mots correctement reconnus est significativement meilleur pour la parole préparée que pour la parole spontanée. Selon notre analyse, la proportion des mots non associés avec la meilleure variante de prononciation est relativement similaire pour les deux corpus étudiés. Le dictionnaire de prononciations ne peut donc pas être l'unique problème. Il est possible que le modèle de langage ne soit pas capable de corriger la distorsion des prononciations lors du traitement de la parole spontanée, alors que cela semble être le cas pour la parole préparée.

5.1.5.3 Au niveau du segment

Dans cette partie, nous cherchons à connaître l'impact de la reconnaissance d'un mot sur un segment de parole, puisque les mots sont liés entre eux lors d'un décodage au moyen d'un système de RAP, notamment au niveau du modèle de langage n-gramme.

Ainsi, pour les deux corpus déjà analysés dans la partie 5.1.5.2 (corpus ESTER et EPAC), nous récupérons les mots aux alentours de chaque mot mal reconnu par le système, afin d'analyser la propagation des erreurs au niveau du voisinage de ce mot. Ce voisinage est caractérisé par une fenêtre d'observation n, correspondant aux n mots gauches et aux n mots droits du mot mal reconnu. Le WER est calculé pour les valeurs de n, en faisant varier la fenêtre d'observation de 2 à 5. Deux cas sont distingués :

– toutes les erreurs de la transcription sont prises en compte (*Global*),
– seuls les mots erronés hors-vocabulaire sont analysés (*HV*).

La figure 5.2 montre les taux d'erreur-mot des mots autour d'un mot erroné (de $n = 2$ à $n = 5$), en utilisant les deux cas précédents. Les taux d'erreur-mot généraux (*WER total*) obtenus pour chaque corpus sont également rappelés dans cette figure.

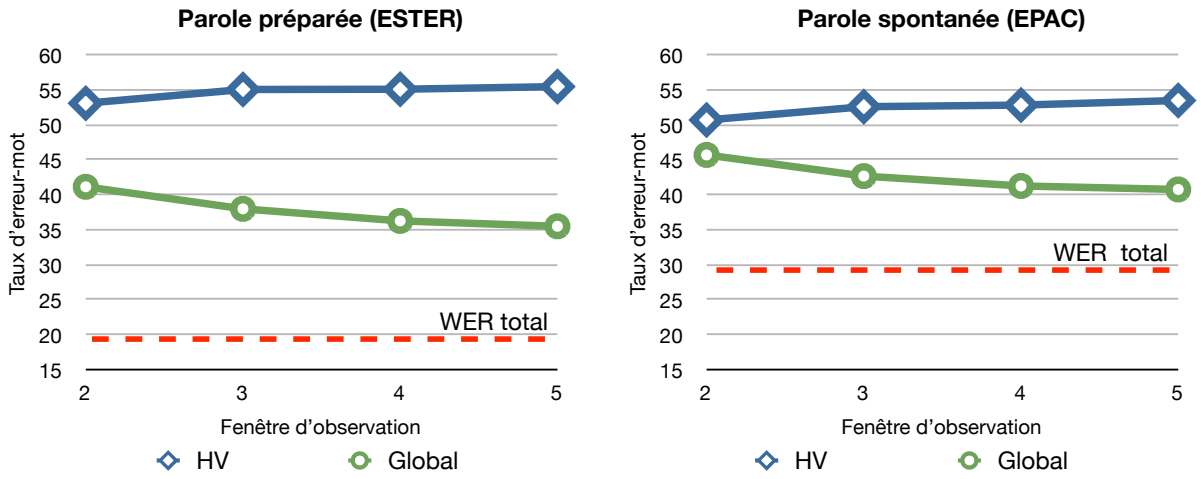

FIG. 5.2 – Taux d'erreur-mot au voisinage d'un mot erroné (toutes les erreurs *Global* comparées aux erreurs de mots Hors-Vocabulaire *HV*) sur la parole préparée et spontanée.

Les résultats montrent que le WER autour d'un mot mal reconnu a tendance, en général, à décliner lorsque la fenêtre d'observation augmente. D'un autre côté, le WER augmente lorsque nous nous focalisons sur les mots au voisinage des mots hors-vocabulaire. Ces mêmes évolutions peuvent être observées sur les deux corpus (ESTER et EPAC). Ces observations peuvent s'expliquer par le modèle de langage utilisé dans les systèmes de RAP. En effet, le modèle de langage n-gramme (voir partie 1.4.1) se fonde sur l'historique d'un segment pour pouvoir émettre des hypothèses : si cet historique est erroné, il semble logique que les hypothèses émises à partir de celui-ci soient fausses. Lorsqu'un mot à reconnaître est hors-vocabulaire, le taux d'erreurs autour de ce mot a tendance à augmenter car le mot peut être complètement

différent du mot à reconnaître. Ce constat est différent pour une simple erreur, le taux d'erreur-mot diminuant légèrement. Dans ce cas, le mot hypothèse peut être relativement proche du mot correct, le modèle de langage "compensant" l'erreur.

5.2 Adaptation des systèmes de RAP

5.2.1 Principe général

5.2.1.1 Adaptation non-supervisée des modèles acoustiques et de langage

Comme nous avons pu le constater dans la partie 2.3, des travaux insistent sur la nécessité d'obtenir une grande quantité de données pour pouvoir créer des modèles fiables [Furui 2005]. Collecter de telles données s'avère cependant difficile et coûteux. Généralement, la quantité de données d'apprentissage disponible n'est pas suffisante pour créer des modèles spécialisés, notamment pour la parole spontanée. Le principe que nous suivons est assez proche de celui adopté par [Ariki 2003], cherchant à adapter les modèles généraux en utilisant des données spécifiques (émissions radiophoniques sur le base-ball). Dans cet article, les auteurs proposent d'adapter les modèles des systèmes de RAP (acoustique et linguistique) au domaine des émissions sportives afin d'extraire des mots clés à partir des transcriptions finales (pour la recherche documentaire). Dans un premier temps, les auteurs ont adapté de manière supervisée [75] et non-supervisée[76] les modèles acoustiques au thème précis du sport. Ce deux types d'adaptation ont été réalisées au moyen des méthodes MLLR et MAP (voir partie 1.3.3).

Cependant, l'approche que nous proposons dans [Dufour 2010] se différencie par le fait qu'aucune donnée n'est ajoutée pour adapter les modèles acoustiques et de langage. En fait, nous voulons extraire une partie des données déjà disponibles dans le corpus servant à l'apprentissage des modèles acoustiques et de langage généraux. Cette extraction se veut la plus automatique possible. Cette idée se rapproche de celle développée dans [Bigi 2000], proposant de choisir dynamiquement les modèles de langage adaptés aux thématiques abordées, mais surtout de l'approche détaillée dans [Chen 2003]. En effet, ce dernier propose une méthode pour adapter de manière non-supervisée les modèles de langage pour les systèmes de RAP devant transcrire des émissions radiophoniques. Les émissions radiophoniques contenant des thématiques variées, l'utilisation d'un seul modèle de langage général ne semble pas la meilleure solution pour traiter tous ces sujets. Les auteurs ont choisi d'utiliser des techniques issues de la recherche d'information pour détecter le sujet abordé. Cette détection leur permet d'adapter le

[75] Avec des données audio annotées manuellement.
[76] Avec des données audio annotées au moyen d'un système de RAP.

modèle de langage au domaine qu'il est en train de traiter. Cette adaptation utilise des données spécifiques qui sont extraites automatiquement d'un corpus d'émissions radiophoniques.

Comme aucune donnée ne doit être fournie au système de RAP, la principale difficulté réside dans le fait de réussir à extraire les données "utiles" du corpus d'apprentissage pour adapter les modèles à la parole spontanée. L'idée est d'utiliser l'outil que nous avons développé dans [Dufour 2009a] pour la détection automatique de segments de parole spontanée (présenté dans la partie 4.4) afin de catégoriser notre corpus d'apprentissage selon le type de parole. Il sera ainsi possible d'adapter les modèles acoustiques et de langage généraux du système de RAP au moyen de ce sous-corpus spécialisé. La figure 5.3 résume le principe d'adaptation que nous proposons, en présentant les différentes étapes possibles permettant d'obtenir un système adapté de manière non-supervisée à la parole spontanée.

5.2.1.2 Combinaison des systèmes

Les nouveaux modèles adaptés à la parole spontanée, que nous obtiendrons au moyen de notre méthode, permettront au système de RAP de fournir de nouvelles hypothèses de reconnaissance. Nous espérons que le système de RAP adapté permettra d'améliorer les performances sur tous les segments de parole spontanée. Cependant, l'approche que nous proposons se voulant entièrement automatique, nous n'avons aucune certitude sur le fait que les performances seront améliorées sur ce type de parole. En effet, le fait d'utiliser un détecteur automatique de parole spontanée conduit inévitablement à classifier de façon erronée certains segments de parole. Cette mauvaise classification peut alors fausser les résultats obtenus. De plus, ces erreurs d'étiquetage peuvent se multiplier : tout d'abord, durant la phase d'adaptation des modèles, où l'étiqueteur possède une marge d'erreur, puis lors de la phase de décodage, si le détecteur de parole spontanée choisit les mauvais segments à décoder.

Au regard de ces difficultés, nous choisissons, dans notre approche, de ne pas simplement nous focaliser sur les segments de parole spontanée lors du décodage au moyen du système de RAP adapté. L'idée, ici, est de réaliser deux décodages. Le premier utilise les modèles généraux, et le second les modèles adaptés. Comme chaque système possède une base de connaissance différente, nous espérons qu'ils pourront proposer des hypothèses complémentaires, qui devraient potentiellement permettre d'améliorer les performances sur la reconnaissance de parole continue à grand vocabulaire [Ellis 2000].

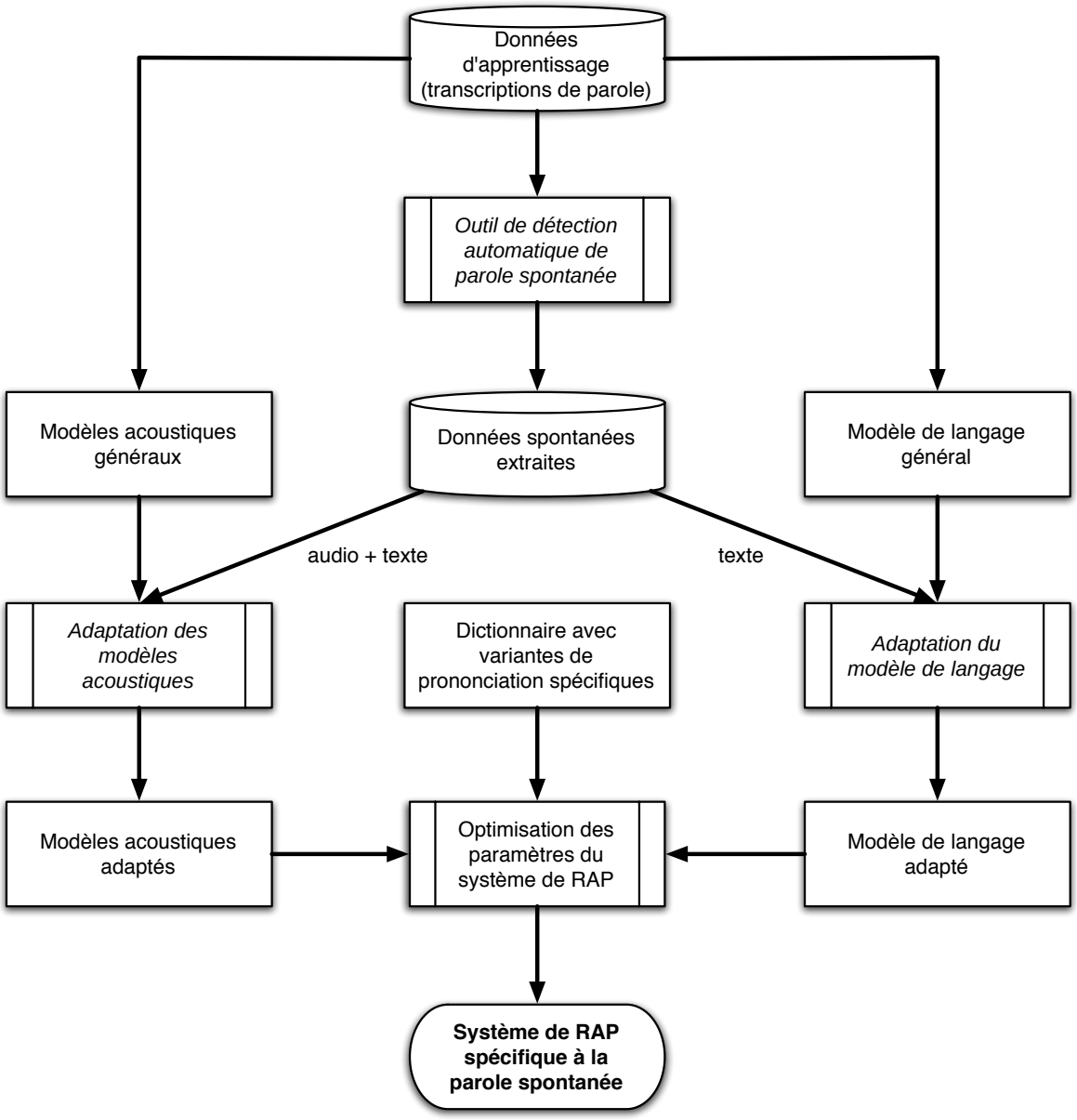

FIG. 5.3 – Principe d'adaptation des modèles d'un système de RAP pour la parole spontanée en utilisant les données d'apprentissage existantes.

5.2.2 Adaptation automatique des modèles

5.2.2.1 Modélisation acoustique

Pour réaliser nos expériences, les nouveaux modèles, spécialisés pour traiter la parole spontanée, seront créés à partir des modèles acoustiques généraux appris par le système du LIUM (voir partie 1.5.1.3). Pour rappel, pendant la phase d'apprentissage, plusieurs modèles généraux sont créés. Premièrement, des modèles sont entraînés selon les bandes de fréquence (bande large/bande étroite), composés de 6 500 états. Ensuite, ils sont adaptés, au moyen de la méthode MAP, pour générer des modèles dépendants du sexe du locuteur (homme/femme). Quatre modèles spécialisés sont alors obtenus : Bande large/Homme (BL_H), Bande étroite/Homme (BE_H), Bande large/Femme (BL_F), et Bande étroite/Femme (BE_F). Ils sont utilisés pour calculer les matrices de transformation CMLLR afin d'obtenir des modèles SAT-CMLLR composés de 7 500 états.

La figure 5.4 présente le processus d'adaptation des modèles acoustiques que nous proposons. Le corpus de parole spontanée n'étant pas suffisant grand pour entraîner des modèles acoustiques robustes à ce type de parole, les deux premières passes du système du LIUM utilisent les modèles acoustiques généraux de base. De plus, nous avons vu que les deux premières passes utilisent une approximation des triphones inter-mots (une manipulation précise de ces triphones n'intervient qu'à la passe 3). Ainsi, les modèles spécialisés n'apparaissent qu'à partir de la passe 3 jusqu'à la passe 5, où une adaptation MAP est appliquée aux modèles généraux appris avec une adaptation SAT. Enfin, l'utilisation du décodeur à partir de la passe 3 permet de réduire le temps de calcul du système de RAP, dont le traitement dure, dans ce cas trois, fois le temps réel (toutes les passes nécessitent dix fois le temps réel). Les modèles étant déjà adaptés au sexe et à la bande de fréquence, nous appliquons une adaptation MAP sur chacun d'entre eux. Nous obtenons, au final, quatre modèles spécialisés pour la parole spontanée, selon le sexe et la bande.

5.2.2.2 Modélisation linguistique

Les modèles de langage trigrammes et quadrigrammes initiaux ont été estimés au moyen de sept corpus textuels d'apprentissage différents : cinq d'entre eux proviennent de différents journaux d'information, un est issu de données collectées sur le Web, et le dernier contient les transcriptions annotées manuellement des fichiers audio utilisés pour l'apprentissage des modèles acoustiques généraux. Sur ce dernier corpus ont été extraits, grâce à l'outil d'extraction automatique de parole spontanée, les segments contenant ce type de parole. Les transcriptions manuelles associées à ces segments de parole spontanée sont alors choisies pour fournir un

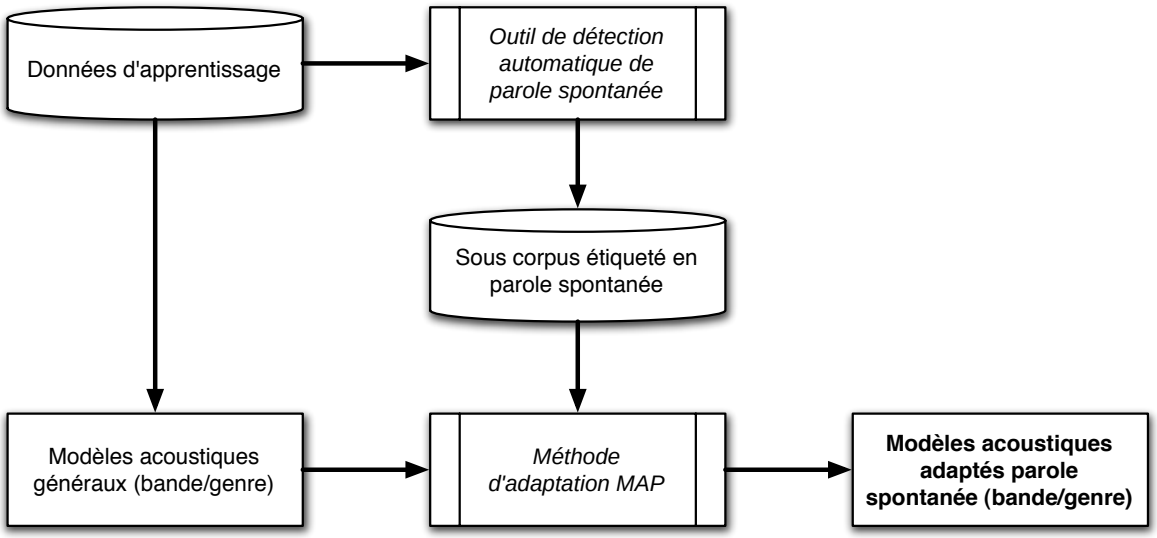

FIG. 5.4 – Adaptation automatique des modèles acoustiques à la parole spontanée.

nouveau corpus d'apprentissage (le huitième). Ces huit corpus sont utilisés pour estimer les modèles de langage trigrammes et quadrigrammes, comme expliqué dans la partie 1.5.1.4.

Les modèles de langage trigrammes ont été combinés linéairement afin de produire le nouveau modèle de langage trigramme fourni pour les passes 1 à 3 du système de RAP du LIUM. Le modèle de langage quadrigramme, utilisé dans la passe 4, est également estimé de la même manière. Les coefficients linéaires ont été optimisés sur les transcriptions manuelles associées aux segments de parole spontanée détectés dans le corpus de développement.

Ainsi, la différence principale entre les modèles de langage initiaux et ceux spécialisés pour traiter la parole spontanée, se trouve au niveau du corpus de développement choisi pour optimiser les poids linéaires : le corpus de développement complet pour les modèles initiaux et seulement les segments de parole spontanée pour les modèles spécialisés. De plus, un renforcement des observations survenant dans les segments de parole spontanée du corpus d'apprentissage a été réalisé au moyen d'un corpus spécifique contenant les transcriptions manuelles associées à ces segments. En effet, ces segments étaient déjà présents dans le corpus d'apprentissage initial utilisé pour entraîner les modèles acoustiques.

5.2.3 Corpus

Cette partie s'intéresse au corpus employé pour entraîner les modèles acoustiques, en incluant la quantité de données utilisée pour adapter les modèles à la parole spontanée (extraction d'une partie du corpus initial). Pour construire les modèles acoustiques et linguistiques, un

grand corpus audio est essentiel. Le corpus utilisé est composé de 240 heures fournies par les campagnes d'évaluation ESTER 1 et ESTER 2 (principalement de la parole préparée, voir partie 1.6), plus 80 heures d'émissions radiophoniques françaises transcrites manuellement et diffusées par le projet EPAC (principalement de la parole spontanée, voir partie 1).

Dans l'optique de construire les modèles acoustiques et de langage spécialisés, une partie du corpus d'apprentissage a été extraite grâce au détecteur automatique de parole spontanée. Le détecteur a permis d'extraire 133 heures de parole très spontanée, ce qui représente au niveau textuel environ 132K segments et 2,31M de mots. Globalement, le corpus contient 256 heures de parole, 260K segments, et 4,17M mots. Un seuil (défini à 0,8) a été appliqué sur les mesures de confiance issues de notre détecteur automatique de type de parole (voir partie 4.3) afin d'augmenter la précision de cette classification sur les segments de parole très spontanée. Notre choix s'est porté sur ce seuil car il apparaissait comme le meilleur compromis entre la proportion de données utilisées pour l'apprentissage[77] et entre l'amélioration de la précision du détecteur automatique. Les modèles acoustiques étant adaptés en fonction de la bande de fréquence et du sexe du locuteur, le tableau 5.8 présente les durées des données d'apprentissage (en heures) utilisées pour entraîner les modèles acoustiques initiaux (*Global*) et pour adapter les modèles acoustiques à la parole spontanée (*Extrait*).

Corpus	**Global**	**Extrait**
BL_M	161	84
BL_F	53	21
BE_M	33	23
BE_F	9	5

TAB. 5.8 – Durée (en heures) des données d'apprentissage pour les modèles acoustiques généraux (*Global*) et leurs adaptations pour la parole spontanée (*Extrait*) en fonction de la bande de fréquence et du sexe du locuteur.

5.2.4 Expériences

Les expériences ont été menées sur les corpus officiels de test et de développement de la campagne d'évaluation ESTER 2 (les documents audio provenant de la radio Africa N° 1 n'ont cependant pas été utilisés durant nos expériences), ainsi que ceux du projet EPAC. La concaténation de ces deux corpus de développement et de test, permet d'obtenir respectivement 13 heures et 16 heures d'enregistrements audio.

[77] La quantité de données extraites doit être assez conséquente pour permettre une adaptation correcte des modèles.

5.2.4.1 Analyse du système adapté

Pour être en mesure d'examiner l'impact de modèles adaptés à la parole spontanée, nous avons mené plusieurs expériences de transcription de la parole. L'idée de base est de réaliser deux décodages au moyen du système de RAP du LIUM, en utilisant séparément les modèles de base lors d'un premier décodage, et les modèles adaptés lors d'un second. Dans un premier temps, nous voulons voir l'impact des deux modèles sur les segments de parole fortement spontanée. Ces segments, étiquetés en parole *fortement spontanée*, ont été obtenus au moyen de notre outil de détection automatique de parole spontanée, suivant le même processus que celui pour l'adaptation des modèles acoustiques. Le tableau 5.9 présente les taux d'erreur-mot, obtenus sur le corpus de développement, sur les segments contenant ce type de parole, avec le système de base (*Base*) et le système adapté à la parole spontanée (*Sponta*).

	Base	**Sponta**
BL_M	28,0	**27,3**
BL_F	25,0	26,2
BE_M	31,2	**30,6**
BE_F	27,5	28,4
Global	27,1	**27,0**

TAB. 5.9 – Taux d'erreur-mot sur les segments de parole *fortement spontanée*, obtenus sur le corpus de développement, avec le système de base (*Base*) et le système adapté à la parole spontanée (*Sponta*) selon la paire bande de fréquence/sexe du locuteur.

Nous pouvons voir que le WER global est différent selon le système utilisé lorsque nous nous focalisons sur les segments de parole fortement spontanée, avec une très légère baisse du WER avec le système adapté à la parole spontanée. Nous notons également que les résultats sont différents selon la paire bande de fréquence / sexe du locuteur : les gains sont obtenus sur les locuteurs *Hommes*, alors que les pertes sont obtenues sur les locuteurs *Femmes*.

Le gain relativement faible atteint sur les segments de parole fortement spontanée peut s'expliquer par la différence entre les valeurs des caractéristiques extraites par le système de détection automatique de parole spontanée sur le corpus d'apprentissage et le corpus de développement. En effet, les descripteurs du corpus d'apprentissage sont extraits par le système de RAP sur des transcriptions manuelles, au contraire du corpus de développement, où ils ont été extraits des transcriptions automatiques. Cette différence peut conduire à la baisse des performances de notre détecteur de parole spontanée. De plus, les résultats disparates entre le sexe du locuteur et la bande de fréquence peuvent être dus à la quantité différente de données d'adaptation utilisée, comme nous pouvons le constater dans le tableau 5.8. Pour toutes ces

raisons, nous focalisons nos efforts sur la combinaison du système de base et celui adapté à la parole spontanée.

5.2.4.2 Combinaison des systèmes

Nous venons de montrer, dans le paragraphe précédent, que l'adaptation que nous avons réalisée ne semble pas être la meilleure solution pour réduire le taux d'erreur-mot d'un système de reconnaissance de parole continue grand vocabulaire. En se basant sur les résultats du tableau 5.9, il est raisonnable de penser que de meilleures hypothèses de mot sont introduites par le système adapté à la parole spontanée. Dans ce contexte, une combinaison des sorties du système de RAP peut être une solution pour extraire les meilleures hypothèses de chaque système.

Le système de RAP du LIUM peut fournir des mesures de confiance (voir partie 1.4.5) estimées à partir des scores acoustiques et linguistiques et associées à chaque mot transcrit. Les mesures de confiance sont un bon indicateur pour choisir la meilleure hypothèse parmi différents systèmes [Lecouteux 2008]. Pour toutes ces raisons, nous combinons les sorties du système de base et du système adapté à la parole spontanée au moyen de la méthode ROVER [Fiscus 1997]. Cette méthode cherche à réduire le WER en combinant différentes sorties de systèmes de RAP, en prenant en considération les mesures de confiances associées aux hypothèses de mot.

Premièrement, nous avons cherché à minimiser le WER lors de la combinaison des sorties (systèmes de base et adapté) en optimisant les paramètres de la méthode ROVER sur le corpus de développement. Notre optimisation utilise simplement les scores de confiance de chaque hypothèse. Le tableau 5.10 présente les résultats de combinaison de sorties obtenus au moyen de la méthode ROVER, ainsi que le WER minimum que nous pouvons espérer obtenir si la meilleure hypothèse de mot est toujours choisie entre les deux systèmes (score *oracle*).

	Développement	**Test**
Base	21,08	18,85
Sponta $\oplus$ *Base*	20,67 (-1,94 %)	18,53 (-1,70 %)
Oracle	18,52 (-12,1 %)	16,47 (-12,6 %)

TAB. 5.10 – Comparaison des taux d'erreur-mot du système de RAP avec le système de base (*Base*), la méthode ROVER (*Sponta* $\oplus$ *Base*), et en calculant le score oracle (*Oracle*).

La combinaison des transcriptions en sortie des systèmes montre leur complémentarité. Cette combinaison permet d'améliorer la précision de 1,7 points en relatif sur notre corpus de test. Même si ce gain semble minimal, le score oracle montre que, potentiellement, un très fort gain est possible si la combinaison est idéale (12,6 points en relatif). Afin de vérifier si les résultats de la combinaison des systèmes (Sponta $\oplus$ Base) par rapport au système de base sont

significatifs, nous avons réalisé, avec l'outil *sc_stats* fourni par l'institut *NIST*, le test "Matched Pairs Sentence-Segment Word Error" (MAPSSWE) [Pallet 1990]. Ce test indique que l'amélioration obtenue au moyen de la combinaison de systèmes est statistiquement significative à un niveau de **p=0,001**. Ce test se focalise sur les segments où les systèmes proposent des hypothèses différentes. Ces hypothèses doivent être entourées par au moins deux mots correctement reconnus par les systèmes de RAP.

5.2.5 Conclusion

Dans ce travail, nous cherchons à améliorer les systèmes de RAP sur le traitement de la parole spontanée. Dans un premier temps, nous nous sommes intéressés au dictionnaire de prononciations. Au moyen d'une étude préliminaire, nous avons constaté que ce dictionnaire, qui doit traiter tout type de parole, est beaucoup plus efficace lorsqu'il est appliqué sur des enregistrements audio de parole préparée, en comparaison avec les passages contenant de la parole spontanée. Nous cherchons ainsi, au moyen d'une analyse manuelle, à ajouter dans le dictionnaire certaines prononciations spécifiques à la parole spontanée. Les résultats obtenus laissent entrevoir un gain possible, puisque le système utilisant le nouveau dictionnaire réduit les erreurs de substitution et de suppression. Cependant, améliorer globalement les performances reste difficile, à cause notamment d'un plus fort taux d'insertion de mots.

Au niveau des modèles acoustiques et de langage, en partant du principe que la parole spontanée dégrade significativement les performances des systèmes de reconnaissance de la parole, de nombreux travaux ont concentré leurs efforts sur la collecte de données spécifiques (en employant des transcripteurs humains). Ces données permettent d'entraîner des modèles acoustiques et linguistiques sur la parole spontanée. Malheureusement, réunir une quantité de données suffisante pour pouvoir créer de tels modèles est coûteux. D'autres stratégies doivent ainsi être considérées pour traiter la parole spontanée. Nous avons proposé une méthode qui cherche à adapter les modèles acoustiques et linguistiques à la parole spontanée. L'idée directrice est qu'aucune donnée spécifique à la parole spontanée ne sera ajoutée aux modèles déjà existants. Les données d'apprentissage sont simplement utilisées différemment. La méthode d'adaptation proposée est complètement automatique, en se servant de notre outil de détection de la parole spontanée [Dufour 2009a]. Le sous-corpus extrait, au moyen de notre outil, est ensuite utilisé pour adapter les modèles acoustiques et le modèle de langage.

Les expériences montrent que la combinaison des transcriptions du système de base et du système adapté à la parole spontanée a un impact positif sur le taux d'erreur-mot. Cette combinaison a été réalisée au moyen de la méthode ROVER dans nos expériences. Une réduction statistiquement significative du taux d'erreur-mot a été observée, avec un gain relatif

de 1,7 % sur le corpus de test. De plus, cette méthode n'a pas besoin d'expertise humaine, ni de transcripteur humain pour collecter de données spécifiques, tout en restant raisonnable au niveau du décodage (un seul décodage supplémentaire parallélisable).

5.3 Approches spécifiques : le cas de l'homophonie en français

Les différentes travaux précédemment présentés se sont intéressés à la parole spontanée en tant que problème particulier, auquel il convient d'appliquer des traitements spécifiques. Or, au cours de nos travaux, nous avons été amenés à analyser les erreurs des systèmes de RAP, principalement pour nos travaux sur la modification du dictionnaire de prononciations (voir 5.1). Parmi les erreurs les plus fréquentes, beaucoup étaient dues à des phénomènes de l'homophonie. Ce phénomène particulier est difficilement traité dans les systèmes de RAP. Ainsi, par extension aux problèmes spécifiques que nous avons traités pour la parole spontanée, nous nous sommes intéressés à cet autre problème particulier : l'homophonie. Le traitement de l'homophonie a été réalisé sur la langue française, où l'homophonie y est très fréquente (voir partie 3.2.1). Certaines erreurs dues à l'homophonie sont négligées, principalement lorsqu'elles ne gênent pas la compréhension (nous pouvons penser aux erreurs grammaticales d'accord, voir partie 3.1). Cependant, pour certaines applications, comme le sous-titrage ou les transcriptions assistées, ces erreurs sont plus importantes : leurs répétitions, même si elles ne modifient pas le sens, sont très fatigantes pour l'utilisateur.

Les correcteurs orthographiques ou grammaticaux sont incapables de traiter les sorties d'un système de RAP. Cela est dû aux erreurs de reconnaissance des systèmes de RAP, à l'absence de ponctuation dans les transcriptions, ou encore au non respect des règles de grammaire (voir partie 3). Dans cette partie une méthode de correction d'homophones sera présentée, en décrivant tout d'abord l'approche générale choisie dans la partie 5.3.1. Ensuite, les expériences réalisées au moyen de cette méthode sur certains mots homophones de la langue française seront présentées dans la partie 5.3.2, pour exposer, enfin, les résultats obtenus dans la partie 5.3.3.

5.3.1 Approche proposée

5.3.1.1 Méthodologie générale

Nous cherchons à corriger les transcriptions fournies par les systèmes de RAP au moyen d'une méthode de post-traitement permettant de corriger certains mots homophones contenus dans ces transcriptions. L'approche que nous proposons a pour vocation d'être générique et

réutilisable. En effet, la méthode ne doit pas être dépendante d'un système de RAP. Elle doit pouvoir fonctionner quelle que soit la transcription fournie, et ce, peu importe le système de RAP. De plus, la méthode ne doit pas être spécifique à un domaine précis, et doit fonctionner dans le cadre de la reconnaissance de la parole continue à grand vocabulaire.

Notre approche consiste à construire une solution spécifique pour chaque type d'erreur, afin de pouvoir corriger certaines erreurs homophones spécifiques [Dufour 2008b]. La première étape de cette approche est très pragmatique. Elle consiste à analyser, les erreurs les plus fréquentes réalisées par les systèmes de RAP au niveau des paires de confusion, calculées au moyen de l'outil *NIST SCLITE*. À partir de cette analyse, les erreurs homophones les plus fréquentes qui semblent corrigibles sont choisies manuellement, soit en prenant l'erreur sur la paire de confusion[78], soit en prenant un groupe d'erreurs [79].

Nous proposons ensuite de détecter automatiquement les mots potentiellement erronés pour chaque type d'erreur choisi à l'étape précédente. L'idée est d'obtenir une liste d'erreurs potentielles, contenues dans les transcriptions en sortie des systèmes de RAP. Finalement, ces erreurs seront corrigées automatiquement. Le processus de correction automatique peut être réalisé de deux manières différentes, soit en utilisant des règles grammaticales (voir partie 5.3.1.2), soit en utilisant une méthode statistique (voir partie 5.3.1.3) si les règles linguistiques sont trop difficiles à mettre en œuvre. La figure 5.5 présente la méthode générale proposée pour corriger certains mots homophones dans les transcriptions issues de systèmes de RAP.

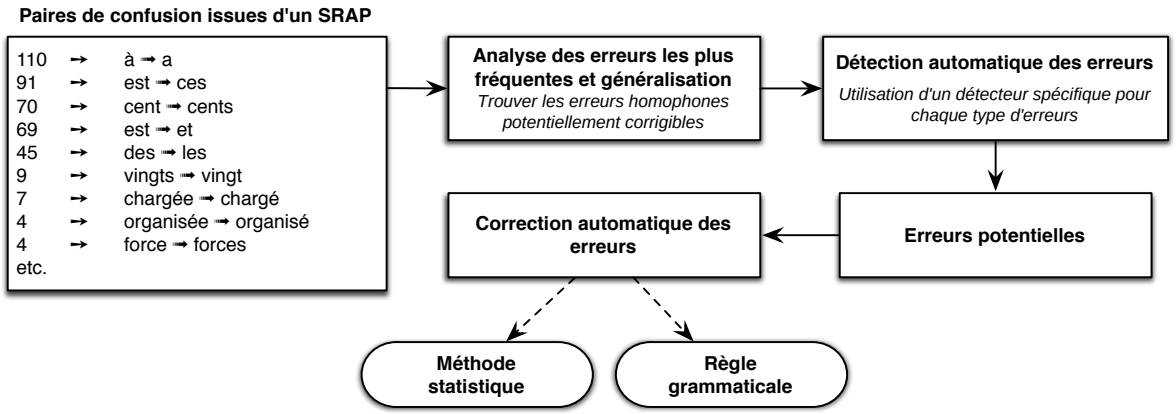

FIG. 5.5 – Méthodologie générale proposée pour la correction de mots homophones en sortie des systèmes de RAP.

[78]Par exemple, la confusion entre la préposition "*à*" et le verbe "*avoir*" conjugué à la troisième personne du singulier "*a*" est une erreur potentiellement corrigible.

[79]Par exemple, les erreurs sur les participes passés.

5.3.1.2 Règle grammaticale

Un langage est régi par des lois générales clairement établies, communes aux différents utilisateurs de la langue. À l'écrit, et notamment si nous l'appliquons dans notre cas aux homophones, il est possible de choisir le mot correct à utiliser par rapport à son sens, son positionnement dans une phrase, et aux règles grammaticales qu'il doit respecter (voir partie 3.1). Or, dans le cas d'un système de RAP, fondé sur des méthodes statistiques, ces règles ne sont modélisées qu'au moyen du modèle de langage n-gramme. Dans la majorité des cas, le modèle de langage est suffisant pour capter ces règles, mais certaines, plus complexes, ne peuvent être appliquées. Le modèle de langage est également limité au corpus d'apprentissage fourni, ce qui peut expliquer que certains cas précis d'une règle grammaticale ne peuvent être correctement appliqués. Notre méthode se voulant spécifique à certaines erreurs, la règle grammaticale s'appliquant au mot à corriger sera utilisée dans notre système si, et seulement si, celle-ci est simple à mettre en application.

Pour illustrer ce point, nous nous sommes intéressés à la correction des paires de mot "*cent/cents*" et "*vingt/vingts*", dont la règle d'accord en nombre est très simple. Cette règle est cependant mal gérée par les modèles statistiques, d'une part car ces mots sont homophones, et d'autre part, parce que la règle a besoin de la classe grammaticale du mot précédent et du mot suivant pour prendre une décision sur son nombre (singulier ou pluriel). Enfin, les modèles de langage n-gramme ne modélisent que les événements observés : si le contexte, dans lequel apparaît un mot, est manquant, le mot correct peut ne pas être choisi (voir partie 1.4.1).

La figure 5.6 présente un exemple de confusion entre les mots "*vingt/vingts*", ainsi que la méthode de correction au moyen de sa règle grammaticale sur un extrait de transcription du système de RAP du LIUM. Dans cet exemple, le mot "*vingt*" est au singulier, alors que la règle grammaticale appliquée à cette portion de texte donne le mot au pluriel ("*vingts*").

La première étape consiste à détecter, dans les transcriptions issues d'un système de RAP, les mots hypothétiquement erronés. Dans l'exemple de la figure 5.6, nous cherchons à savoir si l'hypothèse du mot "*vingt*" est bien au singulier dans le segment "*<s> de cent quatre vingt locomotives de fret </s>*". Pour pouvoir trouver l'erreur, un étiquetage grammatical, obtenu au moyen d'un étiqueteur grammatical automatique, est réalisé sur le mot précédent et le mot suivant du mot potentiellement corrigible. Ce processus d'étiquetage fournit des informations sur la catégorie grammaticale de ces mots. Un processus de détection permet ensuite d'appliquer la règle grammaticale (ici sur le mot "*vingt*"), afin de trouver si ce mot est erroné. Dans le cas d'une erreur, le mot bien corrigé est trouvé (dans cet exemple, le mot devait être au pluriel). Ce mot corrigé remplacera, dans la transcription finale, le mot erroné[80].

[80]Dans cet exemple, le mot "*vingts*" remplacera le mot "*vingt*".

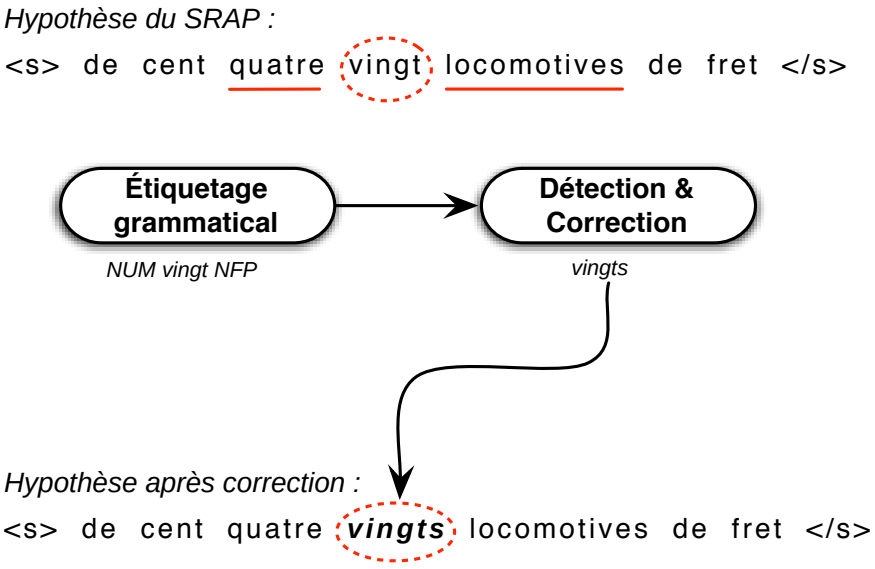

FIG. 5.6 – Exemple d'application de notre approche utilisant une règle grammaticale pour corriger le mot "*vingt*".

5.3.1.3 Méthode statistique

Lorsque les règles grammaticales, s'appliquant à un mot ou classe de mots, sont trop difficiles à mettre en œuvre au moyen d'un outil informatique, nous proposons d'utiliser une méthode statistique. En effet, les règles formelles ne sont pas, entre autres, très robustes aux erreurs se trouvant dans le contexte lexical du mot à corriger. Nous pouvons notamment penser à la règle générale d'accord des participes passés. L'encyclopédie Larousse[81] définit les règles d'accord du participe passé comme suit (*extrait*) :

Variabilité :

a. Ils s'accordent avec le sujet lorsqu'ils sont conjugués avec l'auxiliaire **être**.

b. Ils s'accordent avec le complément d'objet direct antéposé lorsqu'ils sont conjugués avec l'auxiliaire **avoir**

Invariabilité :

a. Lorsqu'ils sont conjugués avec l'auxiliaire **avoir**.

b. Les participes passés des verbes transitifs lorsqu'ils sont suivis d'un infinitif ou d'une proposition complément d'objet direct.

[81] `http://www.larousse.fr/encyclopedie/nom-commun-nom/participe/77217`

Comme nous pouvons le constater dans cet extrait, les quelques règles présentées ici reflètent bien la difficulté de l'accord du participe passé. L'accord est subordonné aux différents acteurs de la phrase (auxiliaire, sujet, compléments d'objet direct...) et à leur position. Dans un système de RAP classique, les moyens pour désambiguïser l'accord des participes passés se trouvent au niveau des modèles linguistiques et acoustiques. Cependant, les différentes formes fléchies (différents accords) que peut prendre un participe passé sont très souvent homophones. Les modèles acoustiques sont donc inefficaces pour capter ces accords. Le modèle de langage est le seul modèle pouvant aider à les différencier. Or, les modèles n-grammes, de par leurs limites, ne peuvent pas, dans tous les cas, satisfaire aux différentes règles linguistiques du participe passé : ils sont dépendants de leur corpus d'apprentissage et sont limités au niveau de la longueur du n-gramme (dans le système du LIUM par exemple, le modèle de langage est maximum quadrigramme). La figure 5.7 présente un exemple du mot "*adoptée*", qui ne peut être correctement orthographié qu'en prenant en compte le mot "*réforme*", se trouvant six mots avant lui. Le modèle de langage quadrigramme utilisé par le système de RAP ne peut capter cette contrainte de distance. De plus, le modèle acoustique ne peut précisément choisir la bonne forme fléchie de ce mot puisque les différents accords sont homophones (prononcés [adopte]).

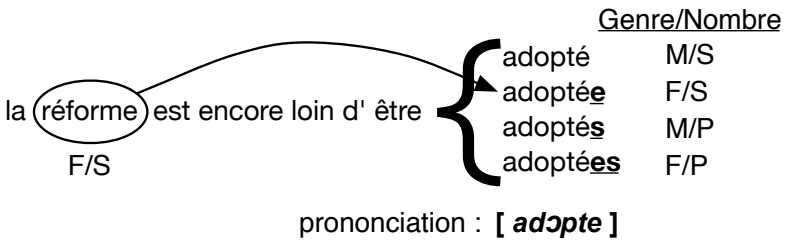

FIG. 5.7 – Exemple de formes fléchies homophones d'un participe passé.

Pour pouvoir traiter ces règles difficiles, l'utilisation d'une méthode statistique est envisageable. Nous voulons choisir, de manière automatique, la forme fléchie la plus probable du mot potentiellement erroné. Cette approche peut, par exemple, s'appuyer sur un classifieur statistique pour prendre une décision sur l'hypothèse fournie par un système de RAP, et éventuellement proposer une meilleure solution. La méthode que nous présentons se focalise sur les participes passés et les adjectifs qui sont, comme nous l'avons vu dans l'exemple précédent, difficiles à modéliser au moyen de règles linguistiques.

Afin de construire la méthode statistique, de nombreuses informations sont extraites. Ces informations permettent de fournir des données plus riches que celles actuellement utilisées dans les systèmes de RAP. Dans l'optique de pallier la faiblesse des modèles de langage, la

méthode statistique prendra en compte les huit mots précédents et huit mots suivants le mot potentiellement corrigible : cet élargissement des n-grammes permet de capter plus d'informations lors du processus de correction. De plus, comme déjà vu dans la partie 5.3.1.2, un étiquetage grammatical est réalisé au niveau de la portion de texte extraite, procurant ainsi de nouvelles informations, en plus des mots. Le classifieur qui sera construit aura alors la possibilité de choisir entre quatre classes : masculin/singulier, masculin/pluriel, féminin/singlier et féminin/pluriel.

Le classifieur aura pour tâche de tester le mot fourni par le système de RAP et de proposer un couple genre/nombre pour ce mot. La prise en compte des informations acoustiques fournies par le système de RAP constitue un apport important de cette méthode. En effet, chaque système de RAP possède un dictionnaire de prononciations, contenant une liste de mots (vocabulaire du système) où chaque mot est associé à une ou plusieurs prononciations (voir partie 1.3.2.2). Nous pouvons obtenir, en sortie d'un décodeur, la variante de prononciation du dictionnaire utilisée pour chaque mot transcrit. Cette information acoustique est très importante, car elle va permettre de filtrer les choix lors de la correction d'un mot. Prenons par exemple une correction proposée par l'outil statistique sur le mot masculin/singulier "*transcrit*" fourni en hypothèse par le système de RAP, associé aux phonèmes (également choisis par le système de RAP) [tʁãskʁi]. Admettons que le mot soit considéré comme erroné par notre classifieur, et que celui-ci propose la forme féminin/singulier. Cette forme donnerait [tʁãskʁi**t(e)**]. La nouvelle forme acoustique ne coïncide donc pas avec l'hypothèse acoustique donnée par le décodeur. Par conséquent, celle-ci ne sera pas retenue comme correction. Cette première proposition de méthodologie a été présentée dans l'article [Dufour 2008b].

À cette première proposition s'ajoute, au niveau du classifieur, un découpage syntaxique en catégories (groupe nominal, groupe prépositionnel) réalisé sur la portion de phrase, permettant de fournir une nouvelle information sur le groupement de mots. Enfin, le lemme du mot cible est la dernière information donnée au classifieur statistique. Cette étape de lemmatisation est importante car le mot cible, récupéré dans la transcription d'un système de RAP, est déjà accordé en genre et en nombre. Lui enlever ces informations permet de ne pas "pré-guider" la détection et correction d'erreurs.

D'autres informations sont également utilisées pour le filtrage des corrections, en sus du filtrage acoustique. Ainsi, en considérant que le classifieur statistique est multiclasse ($n > 2$ classes, où n est le nombre de classes possibles), n classifieurs binaires spécifiques seront créés. Dans le cas des participes passés et adjectifs, proposant quatre classes, quatre nouveaux classifieurs binaires seront entraînés : masculin-singulier/autre, féminin-singulier/autre, masculin-pluriel/autre et féminin-pluriel/autre. Nous obtenons alors cinq hypothèses de classes en sortie des classifieurs. Une décision ne sera prise que si les cinq classifieurs convergent vers la même proposition de classe. Par exemple, si le classifieur, devant prendre une décision

sur quatre classes, propose la classe masculin/singulier, il faut que le classifieur binaire du masculin/singulier fournisse également le même résultat (et non "autre"), et que les 3 autres classifieurs binaires fournissent en classe de sortie "autre".

Enfin, une correction n'est appliquée que si la mesure de confiance (fournie par le système de RAP) d'une hypothèse de mot potentiellement corrigible, est inférieure à un seuil défini empiriquement (optimisé sur un corpus de développement). La figure 5.8 résume les différentes étapes de notre méthode pour traiter les accords en genre et en nombre des participes passés et adjectifs homophones. L'exemple de cette figure est le même que celui présenté dans la figure 5.7. Deux propositions de la méthode statistique sont décrites : le système de base, comme présenté dans [Dufour 2008b], et le système enrichi, avec toutes les informations supplémentaires apportées par la suite.

5.3.2 Expériences réalisées

5.3.2.1 Mots et classes de mots étudiés

La méthode proposée cherchant à corriger des erreurs homophones particulières, nous avons ciblé nos expériences sur certains mots ou classes de mots. Dans un premier temps, pour notre approche utilisant des règles grammaticales, nous voulons corriger les erreurs de nombre (singulier/pluriel) sur les mots "*vingt*" et "*cent*". Ces erreurs homophones sont très fréquentes en français dans les sorties du système de RAP du LIUM. Ensuite, pour notre méthode statistique, nous avons choisi d'étudier les classes de mots concernant les participes passés et les adjectifs, présentant des similarités au niveau de leur accord en genre (masculin/féminin) et en nombre (singulier/pluriel).

5.3.2.2 Outils

Les expériences ont été menées en utilisant le système de RAP du LIUM, décrit dans la section 1.5. Le système nous fournit les transcriptions sur lesquelles seront appliquées les deux méthodes de correction en post-traitement (par règle grammaticale et par méthode statistique). L'étiquetage grammatical, la lemmatisation des mots, ainsi que le découpage syntaxique automatique en catégories, ont été réalisés au moyen de l'outil *lia_tagg*[82], distribué sous licence GPL. Nous utilisons également le classifieur à l'état de l'art *Icsiboost*, basé sur l'algorithme *AdaBoost*, permettant de construire un ensemble de classifieurs à partir de textes et de valeurs continues (voir partie 4.2).

[82] `http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html`

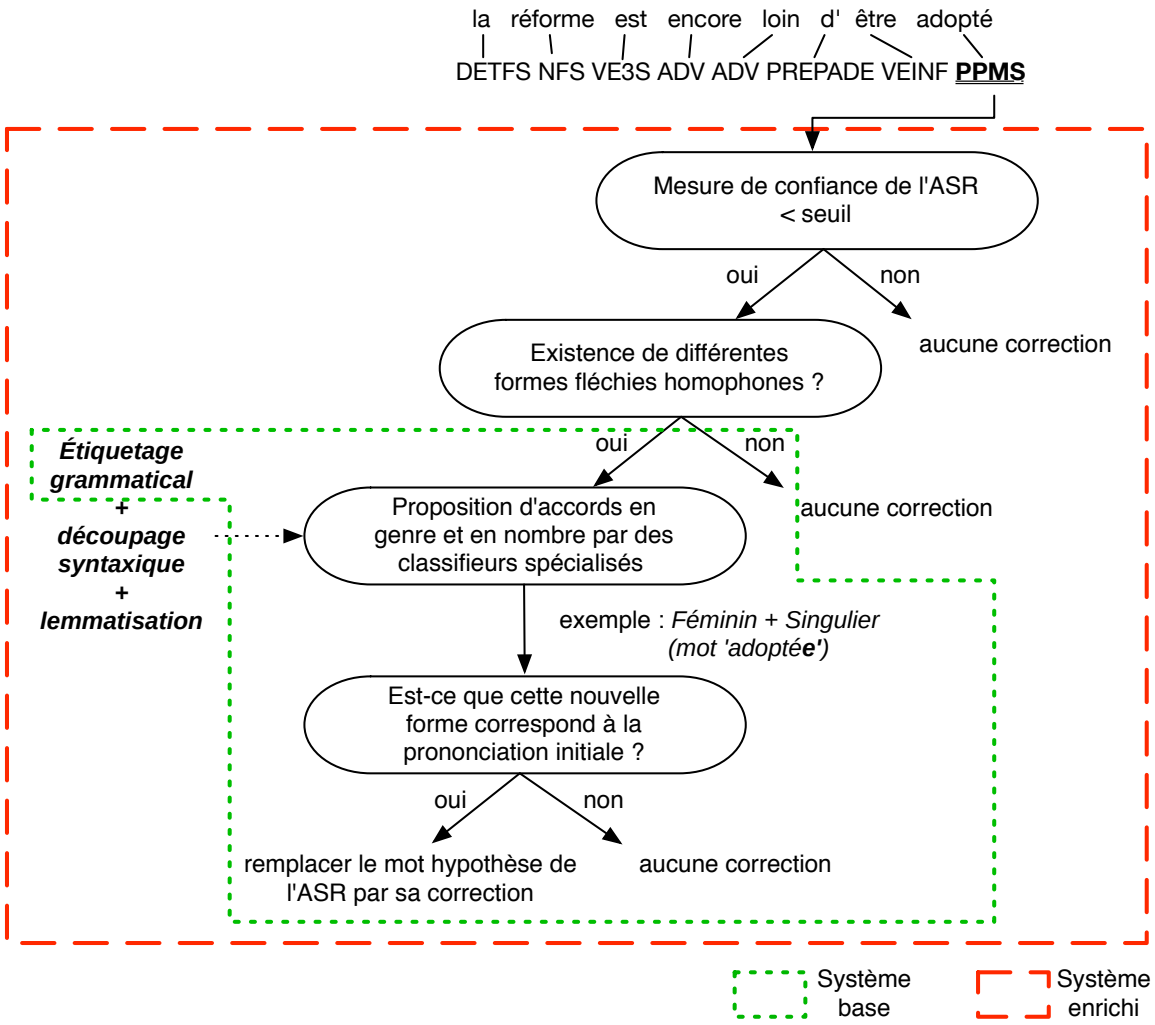

FIG. 5.8 – Approche générale de la méthode statistique traitant les accords en genre et en nombre des participes passés et adjectifs homophones.

5.3.2.3 Données expérimentales

Les données utilisées pour les expériences proviennent des campagnes d'évaluation ESTER 1 et ESTER 2. Les données d'ESTER 1 servent à présenter les résultats de la première version de notre méthodologie. Les données d'ESTER 2 (voir section 1.6) permettent de tester notre méthode enrichie de nouvelles informations. Sur les données de test, le système de RAP du LIUM possède un taux d'erreur-mot de 22,2 % sur ESTER 1, et de 19,3 % sur ESTER 2. Cependant, dans notre approche, nous cherchons à corriger les mots homophones : aucun impact ne sera observé sur les erreurs d'insertion et de suppression présentes dans les transcriptions

automatiques. Seul les taux d'erreurs sur les substitutions, qui atteignent 13,6 % sur ESTER 1 et 11,8 % sur ESTER 2, peuvent varier.

Les classifieurs statistiques ont été entraînés à partir de transcriptions provenant des données d'apprentissage des campagnes ESTER 1 et ESTER 2. Le tableau 5.11 récapitule la taille (en mots) des données d'apprentissage utilisées par la méthode statistique, ainsi que la taille des données testées (en développement et en test) sur les participes passés et les adjectifs.

Classe	Apprentissage	Développement	Test
participes passés	55 237	2 185	2 280
adjectifs	201 684	4 927	4 644

TAB. 5.11 – Taille (en mots) des données d'apprentissage, de développement et de test utilisées par le classifieur statistique sur les participes passés et adjectifs.

Les informations acoustiques (phonèmes associés aux mots) se trouvent dans le dictionnaire de prononciation du décodeur du LIUM. Ce dictionnaire contient environ 301K variantes de prononciation, pour environ 121K mots. Enfin, pour pouvoir corriger les mots, il est nécessaire d'avoir les différentes formes fléchies de chaque mot potentiellement corrigible avec notre méthode. Pour obtenir ces informations, nous avons utilisé la base de données *lexique3*[83], contenant environ 135K mots en français.

5.3.3 Résultats obtenus

5.3.3.1 Avec les règles grammaticales

Nous avons, dans un premier temps, comparé les taux d'erreur-mot sur les transcriptions en sortie du système de RAP du LIUM avant et après correction, afin de vérifier si l'approche modélisant directement les règles grammaticales aide à corriger certaines erreurs spécifiques. Le premier taux d'erreurs sera calculé sur les données de développement. Il permet de savoir si la méthode est fiable. Le second taux d'erreurs, calculé sur les données de test, permet de confirmer son efficacité. Nous avons choisi de nous focaliser sur les transcriptions de la campagne ESTER 1 car les erreurs d'accord des mots *vingt* et *cent* sont parmi les plus fréquentes sur ces sorties. L'impact sera donc beaucoup plus visible. Le tableau 5.12 présente la proportion des erreurs d'accord des mots *cent* et *vingt* par rapport au nombre total d'erreurs sur ces mots, avant (Baseline) et après correction au moyen de la règle linguistique (Correction). Le gain relatif obtenu est également présenté.

[83] `http://www.lexique.org`

	Développement	**Test**
Baseline	16,7	6,6
Correction	0,6	0,9
Gain relatif	**96,7**	**86,4**

TAB. 5.12 – Proportion des erreurs d'accord avant (Baseline) et après correction (Correction) sur les mots "cent" et "vingt" en utilisant la règle linguistique.

L'impact des règles formelles est positif. Pendant la phase de développement, nous avons obtenu un gain très élevé, avec une correction de 87 mots erronés sur un total de 90. Au vu de ces premiers résultats, nous avons appliqué la méthode sur les données de test et observé que seulement 6 mots étaient toujours mal accordés après correction. Cette impossibilité de correction peut s'expliquer par une mauvaise transcription de l'historique entourant le mot à corriger. Notons que durant cette phase de correction, aucune nouvelle erreur n'a été introduite dans la transcription. La méthode a permis de corriger 28 erreurs du mot *cent* et 13 du mot *vingt*. En utilisant les règles linguistiques en post-traitement des sorties d'un système de RAP, nous pouvons voir qu'il est possible de corriger des erreurs d'accord, avec un gain relatif de 86,4 % sur les mots ciblés.

5.3.3.2 Avec la méthode statistique

Comparaison du système de base et du système enrichi

La première expérience sur la correction d'homophones, au moyen de la méthode statistique, a été réalisée en comparant le système de base, présenté dans [Dufour 2008b], avec le système enrichi par toutes les informations présentées dans la section 5.3.1.3. Le tableau 5.13 présente les gains obtenus sur les deux versions de la méthode statistique en corrigeant les transcriptions fournies par le système de RAP du LIUM. Comme la première version de la méthode (base) ne s'intéressait qu'aux participes passés, la comparaison ne sera réalisée que sur cette classe d'erreurs. La comparaison sera effectuée sur les données de développement et de test de la campagne d'évaluation ESTER 2, le système de RAP étant plus performant[84]. Les performances globales du système seront évaluées en *taux de correction*, calculé selon la formule :

$$taux\ de\ correction = \frac{\#erreurs\ corrigées - \#erreurs\ introduites}{\#erreurs\ initiales} \tag{5.1}$$

[84]Notons cependant que la proportion d'erreurs sur les participes passés, due à une homophonie, est quasiment équivalente sur les deux campagnes d'évaluation.

où :

- une *erreur corrigée* est la correction d'une hypothèse de mot erronée dans la transcription,
- une *erreur introduite* est une hypothèse de mot correcte dans la transcription remplacée par un mot erroné,
- et les *erreurs initiales* sont toutes les erreurs dans la transcription dues à une homophonie (sur les mots ou classes de mots choisis) et qui sont potentiellement corrigibles par notre méthode statistique.

Notons que le taux de correction peut être négatif si le nombre d'erreurs introduites est supérieur au nombre d'erreurs corrigées.

Version	Données	**Rappel**	**Précision**	**Taux erreurs introduites**	**Taux correction**
Enrichie	dev.	37,7	51,0	15,6	26,1
Enrichie	test	**40,3**	**57,9**	**17,1**	**28,4**
Base	test	29,6	39,3	31,2	11,4

TAB. 5.13 – Rappel, précision, taux d'erreurs introduites et taux de correction sur les participes passés homophones, en comparant la version de base avec la version enrichie de la méthode statistique.

La version de base de notre méthode statistique permet déjà d'obtenir des gains intéressants sur la correction des participes passés homophones. Néanmoins, la prise en compte de nouvelles informations, visible dans la version enrichie proposée, permet d'améliorer nettement ces résultats. Le gain au niveau du taux de correction est plus que doublé. Cette augmentation du taux de correction s'explique par les bons résultats obtenus au niveau de la précision de la méthode, où le gain relatif est de 25 %, et au niveau du taux d'erreurs introduites, qui baisse de 54 %. Pour bien comprendre ce qui, dans la méthode statistique, permet d'obtenir ces gains, nous présentons dans le tableau 5.14 les différentes étapes de la version enrichie de notre méthode sur le corpus de test. Quatre étapes successives seront présentées :

1. en utilisant seulement le classifieur général (quatre classes : masc/sing, masc/plu, fem/sing, et fem/plu) avec toutes les informations extraites (étiquetage grammatical, lemmatisation et découpage syntaxique) : *classifieur général*,
2. en ajoutant quatre classifieurs binaires spécialisés (masc/sing - autre, fem/sing - autre...). La décision sera prise uniquement si l'ensemble des classifieurs proposent la même hypothèse de classe : + *classif. spécial.*,
3. en y ajoutant la restriction au niveau des informations acoustiques fournies par le système de RAP : + *infos acous.*,

4. enfin, en restreignant le choix des participes passés à corriger (correction seulement si la mesure de confiance est inférieure à un certain seuil) : + *CM*.

Étape	**Rappel**	**Précision**	**Taux erreurs introduites**	**Taux correction**
classifieur général	49,4	47,2	36,4	11,3
+ classif. spécial.	48,4	47,5	30,4	17,4
+ infos acous.	42,6	50,6	23,4	22,9
+ CM	40,3	57,9	17,1	28,4

TAB. 5.14 – Rappel, précision, taux d'erreurs introduites et taux de correction sur les différentes étapes de la méthode statistique lors de la correction des erreurs d'accords des participes passés sur le corpus de test.

Lorsque nous comparons la première étape (*classifieur général*) avec l'ensemble des étapes (+ *CM*) de notre méthode statistique, nous pouvons voir très clairement que la précision ainsi que le taux d'erreurs introduites diminuent très fortement ; nous constatons un gain de 10,7 points en absolu pour la précision, et une baisse de 19,3 points en absolu pour le taux d'erreurs introduites. Nous constatons également que nous corrigeons moins d'erreurs puisque le rappel diminue. Cette baisse est normale, sachant que les différentes étapes cherchent à contrôler les corrections et non à corriger de nouvelles hypothèses de mot de la transcription. En détaillant un peu plus ce tableau de résultats, nous voyons que les classifieurs binaires spécialisés (*classif. spécial.*) permettent d'occulter un nombre important d'erreurs introduites, en baissant de manière négligeable le rappel. La baisse du nombre d'erreurs introduites se confirme avec l'utilisation d'informations acoustiques. Cette réduction confirme l'utilité de ces informations dans le processus de correction. Finalement, ces étapes successives permettent à chaque fois d'améliorer le taux de correction, passant de 11,3 % si aucune vérification sur les corrections n'est réalisée, à 28,4 %.

Au regard de ces résultats, nous avons choisi d'étendre la méthode aux adjectifs. Le tableau 5.15 présente les résultats obtenus sur la correction des adjectifs au moyen de la méthode statistique globale (version enrichie) sur les corpus de développement et de test. Nous constatons que bien que ces résultats soient positifs, notamment en terme de taux de correction, ils apparaissent moins bons que sur les participes passés. Cette baisse de performance peut s'expliquer par un nombre beaucoup plus important d'adjectifs à vérifier (voir tableau 5.11), mais surtout à un nombre d'homophones plus faible que celui des participes passés, comme le montre le tableau 5.16. Ces constats semblent avoir pour conséquence de complexifier le problème de correction des adjectifs. En effet, le fait d'avoir peu de mots homophones à corriger augmente le risque d'insertion d'erreurs.

Données	**Rappel**	**Précision**	**Taux erreurs introduites**	**Taux correction**
dev.	23,0	42,9	10,2	17,5
test	25,1	42,0	23,6	11,0

TAB. 5.15 – Rappel, précision, taux d'erreurs introduites et taux de correction sur les adjectifs au moyen de la méthode statistique.

Données	**Développement**	**Test**
participes passés	260 (36,6 %)	310 (42,2 %)
adjectifs	183 (18,4 %)	263 (29,5 %)

TAB. 5.16 – Nombre et taux d'erreurs dues à des erreurs d'homophones, par rapport au nombre total d'erreurs, sur les participes passés et les adjectifs dans le corpus de développement et de test.

Déploiement du processus de correction

La seconde expérience menée consiste à évaluer la portabilité de notre méthode statistique sur des transcriptions fournies par d'autres systèmes de RAP. Différents participants à la campagne d'évaluation ESTER 2 ont accepté de nous fournir leurs transcriptions générées par leurs systèmes sur le corpus de test. Une première difficulté concerne l'étape de vérification des informations acoustiques (prononciation des hypothèses de mot). En effet, seules les transcriptions textuelles nous ont été données.

Comme notre méthode nécessite l'utilisation de ces informations acoustiques pour obtenir les meilleurs résultats, nous proposons de réaliser un contrôle acoustique légèrement différent. Nous distinguons maintenant deux approches différentes, en fonction de l'information disponible :

- **Information acoustique précise :** la prononciation réellement choisie par le système de RAP pour reconnaître le mot est utilisée pour vérifier la concordance acoustique entre l'hypothèse de mot et la correction proposée.
- **Information acoustique approximative :** au moins une variante fournie par un dictionnaire de prononciations (où de multiples prononciations sont associées à un mot) ou un outil de phonétisation automatique (par exemple *lia_phon* [Béchet 2001]), doit être commune entre l'hypothèse et la correction proposée.

Afin de nous rendre compte des performances obtenues avec l'utilisation d'informations acoustiques approximatives au lieu d'informations acoustiques précises, le tableau 5.17 compare la méthode statistique sur ces deux approches. Ce tableau présente également les résultats

si aucune information acoustique n'est utilisée. Les résultats globaux sur la correction de participes passés et des adjectifs homophones sur les données de test sont ainsi présentés.

Nous constatons que les caractéristiques acoustiques précises sont très importantes pendant le processus de correction. Cependant, la simulation de ces caractéristiques, au moyen d'informations approximatives, permet d'obtenir de meilleurs résultats que si aucun contrôle acoustique n'était effectué. En effet, sans contrôle, le taux de correction devient négatif, ce qui veut dire que le système a ajouté plus d'erreurs qu'il n'en a corrigées. Nous pouvons donc en conclure que fournir au minimum une simulation de l'information acoustique est primordial pour obtenir un gain au niveau de taux de correction. Enfin, l'information acoustique issue du système de RAP est de bonne qualité puisqu'elle permet de mieux maîtriser la méthode statistique. Notons que dans ce tableau, les résultats ne prennent pas en compte les mesures de confiance, qui peuvent également être manquantes dans les transcriptions de sortie des systèmes de RAP.

Info. acous.	**Rappel**	**Précision**	**Taux erreurs introduites**	**Taux correction**
Aucune	46,6	38,7	40,9	-2,6
Approximative	43,3	44,9	30,7	13,6
Précise	35,1	50,5	25,1	17,6

TAB. 5.17 – Comparaison des résultats de correction des adjectifs et participes passés sur le corpus de test du système du LIUM, en utilisant les informations acoustiques précises, des informations acoustiques approximatives, ou aucune information acoustique.

Trois laboratoires nous ont généreusement fourni leurs transcriptions finales sur les données de test de la campagne d'évaluation ESTER 2 : le décodeur du LORIA, appelé ANTS et basé sur le décodeur *Julius* [Lee 2001], le décodeur du LIA, réalisé au sein du LIA et appelé *Speeral* [Nocéra 2002], et enfin le décodeur de l'IRISA, créé au sein du laboratoire IRISA et appelé *Sirocco* [Zweig 2002].

Pour chaque transcription de chaque laboratoire, nous voulons connaître l'impact de notre méthode statistique de correction sur les erreurs d'accord des participes passés et adjectifs, avec et sans l'utilisation des informations acoustiques précises de chaque mot. Le tableau 5.18 présente le taux de correction obtenu pour chaque système sur le corpus de test, en fonction de la nature de l'information acoustique disponible (approximative ou précise). De plus, lorsque les mesures de confiance sont disponibles, nous les utilisons. Les résultats obtenus avec les mesures de confiance fournies par les systèmes de RAP, qui améliorent notre méthode statistique, sont présentés séparément puisque seuls les décodeurs du LIUM et de l'IRISA ont fourni ces scores.

Pour chaque transcription utilisée pendant ces expériences, nous obtenons un gain au niveau du taux de correction non négligeable sur les participes passés et les adjectifs. Nous nous

Info. acoustique	**LIUM**	**LIA**	**LORIA**	**IRISA**
approximative	13,6	17,9	17,0	14,3
précise	17,6	n/a	n/a	n/a
+ mesures de confiance	20,4	n/a	n/a	15,3

TAB. 5.18 – Comparaison des taux de correction (*n/a* si indisponible) sur les adjectifs et les participes passés, en utilisant la méthode statistique sur quatre transcriptions en sortie de systèmes de RAP.

apercevons que le taux de correction sur la sortie du système de l'IRISA est plus faible que sur les autres systèmes. Cette différence peut s'expliquer par l'intégration, à l'intérieur du système de l'IRISA, d'un traitement spécial visant à supprimer les erreurs grammaticales. Un modèle de langage 7-gramme de classes grammaticales sert à ré-estimer les n meilleures hypothèses. Le système de l'IRISA montre bien la difficulté de traiter ces types d'erreurs car notre système statistique permet encore de corriger des erreurs grammaticales. De plus, nous nous apercevons que si les informations acoustiques précises étaient disponibles, les gains pourraient être encore meilleurs, comme nous pouvons le constater en comparant les résultats sur le système du LIUM (testé au moyen de l'information acoustique précise et approximative).

Ensuite, si nous restreignons la correction des erreurs d'accord au moyen de mesures de confiance, nous pouvons encore améliorer le taux de correction. Cette amélioration est constatée sur les transcriptions corrigées en sortie des systèmes du LIUM et de l'IRISA, respectivement avec les informations acoustiques précises et les informations acoustiques approximatives. Notons que, comme les transcriptions fournies par les autres systèmes n'étaient que sur le corpus de test, le seuil appliqué sur les mesures de confiance a été optimisé sur les données de développement sur les transcriptions du système du LIUM. Ce seuil a ensuite été appliqué sur les sorties du système de l'IRISA. Il apparaît alors assez robuste pour être utilisé directement sur d'autres transcriptions de systèmes différents sans le ré-optimiser.

Enfin, nous avons cherché à connaître l'impact des erreurs d'accord sur le taux d'erreur-mot global, en utilisant la méthode statistique sur les participes passés et les adjectifs, ainsi que l'approche par règles grammaticales sur les homophones "*cent/cents*" et "*vingt/vingts*". Le tableau 5.19 présente les différents taux d'erreur-mot obtenus sur les transcriptions par les quatre systèmes avant (Baseline) et après correction sur les données de test de la campagne d'évaluation ESTER 2. Les résultats prennent en compte les meilleurs systèmes possibles, avec les mesures de confiance et les informations acoustiques précises si ces dernières sont disponibles.

Pour chaque transcription des différents décodeurs, le taux d'erreur-mot baisse, même lorsque l'on utilise une information acoustique approximative. Le gain peut sembler minimal mais nous

	LIUM	**LIA**	**LORIA**	**IRISA**
Baseline	19,06	26,86	26,27	25,75
Correction	18,91	26,72	26,17	25,62

TAB. 5.19 – Taux d'erreur-mot avant (*Baseline*) et après correction (*Correction*) sur quatre sorties de systèmes de RAP.

devons prendre en compte le fait que notre méthode n'est appliquée que sur un nombre réduit de classes de mots (le gain maximum potentiel est d'environ 0,5 points en absolu pour chaque décodeur testé).

5.3.4 Conclusion

Dans cette partie, nous avons proposé une stratégie de correction des mots homophones en post-traitement, en construisant des solutions spécifiques pour corriger des erreurs précises dans les transcriptions issues de systèmes de RAP. Cette stratégie est, dans un premier temps, composée de deux solutions possibles. La première est de modéliser directement les règles grammaticales lorsque ces règles sont simples à mettre en œuvre. La seconde propose une méthode statistique pour modéliser ces règles, lorsqu'elles sont trop difficiles. L'idée est de choisir les erreurs homophones les plus fréquentes des systèmes de RAP et de les corriger soit au niveau du mot, soit en les regroupant par classe. Plusieurs cas d'erreurs ont été étudiés : d'une part les erreurs d'accord des mots *cent* et *vingt* avec l'utilisation de leurs règles grammaticales, et d'autre part, les erreurs d'accord sur les participes passés et les adjectifs au moyen de la méthode statistique. La méthode a été appliquée sur des transcriptions fournies dans le cadre de systèmes de RAP devant traiter des journaux d'information en français.

Tout d'abord, les résultats obtenus en modélisant les règles linguistiques sur les mots *cent* et *vingt* montrent qu'un gain est possible. En effet, l'application de ces règles a permis de baisser de 86,4 % le taux d'erreurs pour ces mots sur le corpus de test de la campagne d'évaluation ESTER 1.

Ensuite, la méthode statistique proposée introduit l'idée d'apporter des informations supplémentaires à celles utilisées actuellement dans les systèmes de RAP. Le classifieur, permettant de choisir et de corriger les mots potentiellement erronés, prend alors en compte différentes informations telles qu'un étiquetage grammatical, une fenêtre lexicale étendue (deux 8-grammes, en historique gauche et droit du mot), un découpage syntaxique, et enfin une lemmatisation. Nous avons montré que la prise en compte d'informations acoustiques est très importante pendant le processus de correction. Nous avons également défini une nouvelle stratégie utilisant les informations acoustiques précises fournies par le système de RAP. Dans le cas contraire, une

approximation sur ces informations est réalisée en associant à un mot les différentes prononciations possibles contenues dans le dictionnaire d'un système de RAP. De plus, nous avons choisi d'effectuer la tâche de classification en utilisant un classifieur général (à quatre classes) et en construisant à côté des classifieurs binaires. Une décision ne sera prise que si un consensus existe entre ces classifieurs. Enfin, un seuil sur les mesures de confiance calculées par le système de RAP permet de limiter les corrections effectuées, et d'améliorer le taux de correction en réduisant fortement le nombre d'erreurs introduites. Toutes ces étapes permettent de réduire de manière drastique le taux d'erreur sur les erreurs d'accord des adjectifs et des participes passés, avec une baisse générale de 27 % sur ce type d'erreur.

La méthode a finalement été appliquée sur différents systèmes de RAP. Nous constatons que la méthode proposée est suffisamment générique pour fonctionner sur différentes transcriptions. Le gain obtenu sur le taux d'erreur-mot global d'une transcription automatique est pour l'instant minime mais peut s'expliquer par le nombre restreint d'erreurs corrigées. Cependant, nous avons remarqué qu'en ne prenant simplement que les 30 paires d'erreurs les plus fréquentes sur le système du LIUM dues à l'homophonie, un gain relatif (sur toute la transcription) d'environ 21,5 % sur les erreurs de substitution peut idéalement être possible (si toutes les confusions sont corrigées). Ce gain nous encourage dans l'idée d'étendre notre méthode de correction à un nombre plus vaste d'erreurs.

5.4 Résultats finaux des méthodes spécifiques

Les méthodes spécifiques, développées dans ce travail, sont, au final, des approches intervenant à deux niveaux différents du système de RAP : directement au niveau du système de RAP, pour la parole spontanée et en post-traitement des transcriptions fournies le système de RAP, dans le cas de l'homophonie. Il est possible d'appliquer chaque méthode de manière complémentaire. En effet, l'approche non-supervisée pour la parole spontanée fournit une transcription, qui sera ensuite corrigée au moyen du traitement spécifique de l'homophonie. Pour pouvoir connaître le taux d'erreur-mot final obtenu, nous avons pris comme point de départ la transcription obtenue au moyen de la technique de combinaison de systèmes (voir partie 5.2.4.2). Un rappel des données expérimentales manipulées peut se trouver à la partie 5.2.4. Le tableau 5.20 présente les résultats obtenus avec le système de base du LIUM (*Base*), en utilisant la méthode spécifique pour améliorer la parole spontanée (*Sponta*), et enfin en appliquant la correction d'homophones sur la transcription fournie avec *Sponta* sur les données de test

(*Sponta+Homoph*). Pour réaliser ces expériences, nous avons utilisé la version la plus aboutie du système du LIUM, correspondant à la fin du projet EPAC[85].

Base	18,85
Sponta	18,53
Sponta+Homoph	18,46 (-2,1 %)

TAB. 5.20 – Comparaison des taux d'erreur-mot du système de RAP en utilisant le système de base (*Base*), la méthode spécifique de la parole spontanée (*Sponta*), puis sa combinaison avec la méthode spécifique de correction d'homophones (*Sponta+Homoph*).

Au final, nous constatons que les méthodes spécifiques que nous proposons permettent, de manière complémentaire, une baisse globale du taux d'erreur-mot.

5.5 Perspectives

Dans cette partie présentant nos travaux, nous nous sommes intéressés dans un premier temps à la parole spontanée et aux différents traitements envisageables pour améliorer les systèmes de RAP sur ce type de parole. Dans un travail futur, il conviendra d'élargir l'ajout des variantes de prononciation spécialisées à la parole spontanée dans le dictionnaire. Cependant, ajouter de nouvelles prononciations ne sera pas suffisant. En effet, il sera indispensable de trouver des méthodes pour contrôler cet ajout. De plus, nos analyses ont montré un fort taux d'insertion lors de l'ajout de ces variantes spécifiques. Il serait intéressant de réaliser une optimisation sur la pénalité d'insertion des mots fournie au décodeur, puisque nous avons utilisé la valeur optimisée sur le dictionnaire de base. Au niveau de notre approche d'adaptation non-supervisée des modèles acoustiques et linguistiques, tout comme au niveau du dictionnaire de prononciations, nous devons concentrer nos efforts sur les choix des hypothèses de mot, en exploitant des méthodes plus sophistiquées, comme par exemple la Combinaison de Réseaux de Confusion (*Confusion Network Combination*, CNC) [Evermann 2000]. Le gain potentiel, estimé au moyen du score oracle, est de 12,6 points en relatif. Un taux d'erreur-mot plus bas peut donc être atteint. De plus, nous pouvons envisager d'utiliser les scores de spontanéité, obtenus au moyen de l'outil de détection de parole spontanée, dans le choix final de l'hypothèse.

Le caractère spécifique de la parole spontanée a, par la suite, orienté nos travaux vers le cas particulier de l'homophonie. Nous partions du principe que l'homophonie nécessitait également des travaux particuliers. Les perspectives de ce travail s'orientent vers le développement de

[85]Une réduction de plus de 25 % en relatif du taux d'erreur-mot a pu être constatée par rapport au système du LIUM au début du projet EPAC.

nouvelles solutions spécifiques de correction, afin de pouvoir couvrir beaucoup plus d'erreurs, tout en gardant l'idée de *ciblage* des corrections. Fournir de nouvelles informations à la méthode statistique pourrait également être une solution pour améliorer le processus de correction existant. En effet, de nombreuses étapes ont été proposées pour augmenter la précision du système. Or, il serait judicieux de concentrer nos efforts sur l'ajout d'informations permettant, notamment, d'augmenter le rappel, pour corriger une quantité plus importante de mots erronés. Bien que cette particularité du français se prête bien à notre approche, nous voulons appliquer cette stratégie sur d'autres langages, comme, par exemple, l'anglais.

Conclusion et perspectives

Le travail de thèse présenté dans ce manuscrit, s'inscrit dans le cadre du projet EPAC (Exploration de masse de documents audio pour l'extraction et le traitement de la PArole Conversationnelle). Un des principaux objectifs de ce projet est de proposer de nouvelles solutions pour améliorer les performances des systèmes de RAP sur la parole spontanée. En effet, comparativement à la parole préparée (proche de la parole lue), les systèmes ont de plus grandes difficultés à transcrire la parole spontanée, et ce, à cause de ses particularités (présence de nombreuses disfluences, une forte agrammaticalité au niveau des phrases, un registre de langage et un état émotionnel du locuteur différents...). Il convient donc de trouver des solutions pour aider les systèmes de RAP à traiter au mieux les spécificités de la parole spontanée et permettre, ainsi, une meilleure transcription de ce type de parole.

L'idée générale, que nous avons suivie, est de traiter la parole spontanée en tant que spécificité du langage, nécessitant des traitements particuliers. Afin de proposer de nouvelles solutions pour l'amélioration des systèmes de RAP sur ce type de parole, nous avons mis au point un détecteur automatique de parole spontanée. Ce détecteur nous a permis de mettre en place, dans la deuxième partie de cette thèse, une solution non-supervisée d'apprentissage des modèles acoustiques et des modèles de langage, spécifiques à la parole spontanée. Enfin, cette idée de spécificité nous a amenés à développer une solution de correction automatique des transcriptions des systèmes de RAP pour des erreurs particulières dues à l'homophonie.

1 Détecteur de la parole spontanée

Dans un premier temps, une étude préliminaire de différentes spécificités de la parole spontanée a été effectuée. Cette étude a été réalisée sur un corpus étiqueté manuellement en classes de spontanéité : parole *préparée*, *légèrement spontanée*, et *fortement spontanée*. Elle nous a permis d'extraire certaines caractéristiques, afin de proposer un détecteur de type de parole dont l'objectif principal est de trouver automatiquement les segments de parole *fortement spontanée*.

Le processus de détection automatique est composé de deux étapes :

- **Processus local** : la première étape consiste à extraire, pour chaque segment, différentes caractéristiques acoustiques et linguistiques, qui seront ensuite utilisées, avec un classifieur, pour fournir la classe de spontanéité la plus probable pour chaque segment de parole.
- **Processus global** : chaque segment étant étiqueté de manière indépendante dans l'étape précédente, nous proposons un deuxième niveau de classification, en prenant en compte les segments entourant chaque segment de parole. L'idée principale est de prendre en compte les scores obtenus pour chaque classe de spontanéité lors de la classification au niveau de chaque segment, puis de ré-estimer ces scores en prenant en compte les

segments précédent et suivant le segment courant. Une nouvelle hypothèse de classe de spontanéité peut alors être obtenue pour chaque segment, qui constitue le résultat final de détection du type de parole.

2 Modélisation spécifique des systèmes de RAP à la parole spontanée

2.1 Apprentissage non-supervisé des modèles acoustiques et linguistiques

Dans ce travail de thèse, nous avons cherché à améliorer les systèmes de RAP sur le traitement de la parole spontanée. De nombreuses études concluent qu'il est nécessaire de créer des modèles acoustiques et des modèles de langage spécialisés sur ce type de parole, puisque les modèles généraux ne sont pas adaptés à la parole spontanée. Cependant, obtenir des données d'apprentissage ne contenant que de la parole spontanée est très coûteux, en temps et en ressources. Nous avons alors proposé une méthode cherchant à fournir des modèles acoustiques et linguistiques spécialisés sur la parole spontanée, sans ajouter de données spécifiques supplémentaires. L'idée directrice est que les données d'apprentissage que nous possédons contiennent des passages de parole spontanée, qui peuvent être utilisés pour adapter les modèles existants à ce type de parole. La méthode d'adaptation proposée est complètement automatique ; le sous-corpus de parole spontanée est extrait au moyen de l'outil de détection de la parole spontanée que nous avons présenté dans ce mémoire. Ce sous-corpus permet ensuite d'adapter les modèles acoustiques, grâce à la méthode d'adaptation MAP, et d'interpoler le modèle de langage général avec les transcriptions de parole spontanée extraites.

2.2 Combinaison des systèmes

L'adaptation non-supervisée des modèles acoustiques et des modèles de langage a permis d'obtenir une nouvelle transcription. Nous avons alors proposé de combiner les transcriptions obtenues au moyen des modèles de base du système de RAP, avec les transcriptions obtenues avec les modèles adaptés à la parole spontanée. L'idée étant de choisir les meilleures hypothèses de mot de chaque système.

3 Correction d'erreurs spécifiques d'homophonie

Dans le cadre de ces travaux, nous avons également proposé une stratégie de correction des mots homophones hétérographes en post-traitement des systèmes de RAP. L'approche choisie

consiste à fournir des solutions spécifiques pour corriger des erreurs précises dans les transcriptions issues de systèmes de RAP. Deux stratégies différentes sont envisagées pour réaliser une correction :

– Modéliser directement les règles grammaticales, lorsque ces règles sont simples à mettre en œuvre.

– Utiliser une méthode statistique pour modéliser ces règles lorsqu'elles sont trop difficiles.

L'idée est alors de choisir les erreurs homophones les plus fréquentes des systèmes de RAP, et de les corriger soit au niveau du mot, soit en les regroupant par classes. Plusieurs cas d'erreurs ont été étudiés. Dans un premier temps, nous nous sommes focalisés sur les erreurs d'accord des mots "*cent*" et "*vingt*" avec l'utilisation de règles grammaticales. Puis, nous avons cherché à corriger les erreurs d'accord sur les participes passés et les adjectifs au moyen de notre méthode statistique. Notre méthode a été appliquée sur des transcriptions fournies dans le cadre de systèmes de RAP devant traiter des journaux d'information en français.

Enfin, la méthode a pu être déployée sur différents systèmes de RAP, ce qui nous a permis de conclure que cette méthode est suffisamment générique pour fonctionner sur différentes transcriptions (indépendamment du système de RAP). Nous avons montré, dans ce manuscrit, qu'il était possible de corriger des erreurs particulières en proposant des solutions spécifiques.

4 Perspectives

Les solutions spécifiques, apportées pour traiter plus efficacement la parole spontanée dans les systèmes de RAP, permettent d'améliorer les transcriptions finales de ces systèmes. Cependant, des améliorations sont encore nécessaires pour obtenir de meilleurs résultats. En effet, les performances obtenues au moyen de la méthode d'adaptation non-supervisée sur la parole spontanée dépendent des performances du détecteur automatique de parole. Des travaux pour améliorer notre outil de détection de la parole spontanée ont été proposés dans [Rouvier 2010], en combinant notre approche avec des méthodes issues de l'identification des langues. La faible quantité de données pour adapter certains modèles acoustiques peut également être une raison des performances limitées de notre approche sur les segments de parole spontanée. La mise en place de solutions permettant d'augmenter automatiquement ce corpus d'adaptation peuvent alors être envisagées. L'outil de détection de la parole spontanée ne nécessitant pas obligatoirement des transcriptions de référence pour fonctionner, nous espérons utiliser des données audio brutes (par exemple les données non-annotées de la campagne ESTER 1) pour enrichir notre corpus de parole spontanée. Ces nouvelles données pourraient être utiles à deux niveaux : pour le détecteur de type de parole (augmentation du corpus d'apprentissage) et pour l'adaptation des modèles acoustiques. Bien entendu, une transcription manuelle de ces données devra être

effectuée pour l'adaptation des modèles acoustiques, mais cette phase permettrait de cibler les conditions ayant la plus faible quantité de données (les locuteurs femmes et bande de fréquence étroite par exemple). Les efforts nécessaires pour augmenter ce corpus spécialisé seraient alors limités.

Nous pensons étendre l'utilisation de l'outil de détection de la parole spontanée au-delà du cadre de la transcription automatique de la parole. Nous voulons l'utiliser dans d'autres applications, comme, par exemple, la catégorisation de documents. Cette nouvelle utilisation peut être particulièrement intéressante pour la catégorisation des vidéos, domaine très étudié actuellement suite à l'essor des sites Internet de partage de vidéos. En poursuivant cette idée de catégorisation des documents, nous pensons également transposer notre méthode d'adaptation des modèles acoustiques et de langage à des phénomènes spécifiques autres que la parole spontanée. En effet, en reprenant l'exemple des sites de partage de vidéos sur Internet, nous constatons que les documents sont très hétérogènes : registre de langage parfois courant, soutenu ou, au contraire, très familier. . . Avec un travail similaire à celui réalisé pour la parole spontanée, des modèles adaptés pourraient être créés pour ces particularités.

Au niveau de la combinaison de systèmes, nous devons concentrer nos efforts sur les choix des hypothèses de mot, grâce à des méthodes plus sophistiquées, comme par exemple la Combinaison de Réseaux de Confusion (*Confusion Network Combination*, CNC) [Evermann 2000]. En effet, le gain potentiel, estimé au moyen du score oracle, est de 12,6 points en relatif, donc un taux d'erreur-mot plus bas peut être atteint. De plus, une autre idée serait d'utiliser les scores de spontanéité, obtenus au moyen de notre outil de détection automatique de parole spontanée, dans le choix final de l'hypothèse.

Pour la correction des mots homophones, les perspectives de travail s'orientent vers le développement de nouvelles solutions spécifiques de correction, afin de pouvoir couvrir beaucoup plus d'erreurs. Nous envisageons également d'enrichir les informations utilisées par la méthode statistique afin d'améliorer le processus de correction existant. En effet, les étapes successives proposées ont principalement cherché à augmenter la précision du système, au détriment du nombre d'erreurs corrigées. Nous voulons étendre notre méthode à la liste de paires de confusion les plus fréquentes d'un système. Ainsi, au moyen d'un processus automatique, la méthode pourrait choisir les erreurs potentiellement corrigibles et appliquerait les méthodes adaptées. Enfin, nous voulons appliquer cette stratégie sur d'autres langues contenant des homophones, comme l'anglais par exemple.

Bibliographie personnelle

Conférences d'audience internationale avec comité de sélection

Conférences et workshops majeurs du domaine

- Dufour R., Bougares F., Estève Y. et Deléglise P., Unsupervised model adaptation on targeted speech segments for LVCSR system combination, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japon, Septembre 2010.
- Dufour R. et Favre B., Semi-supervised Part-of-speech Tagging in Speech Applications, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japon, Septembre 2010.
- Rouvier M., Dufour R., Linarès G. et Estève Y., A Language-identification inspired method for spontaneous speech detection, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japon, Septembre 2010.
- Dufour R., Estève Y., Deléglise P. et Béchet F., Local and global models for spontaneous speech segment detection and characterization, dans *Automatic Speech Recognition and Understanding (ASRU)*, Merano, Italie, Décembre 2009.
- Dufour R. et Estève Y., Correcting ASR outputs : specific solutions to specific errors in French, dans *Workshop Spoken Language Technology (SLT)*, Goa, Inde, Décembre 2008.

Autres conférences

- Dufour R., Estève Y. et Deléglise P., Automatic indexing of speech segments with spontaneity levels on large audio database, dans *Workshop Searching Spontaneous Conversational Speech (SSCS)*, Florence, Italie, Octobre 2010.
- Estève Y., Deléglise P., Meignier S., Petit-Renaud S., Schwenk H., Barrault L., Bougares F., Dufour R., Jousse V., Laurent A. et Rousseau A., Some recent research work at LIUM based on the use of CMU Sphinx, dans *CMU SPUD Workshop*, Dallas (Texas).
- Dufour R., Jousse V., Estève Y., Béchet F. et Linarès G., Spontaneous speech characterization and detection in large audio database, dans *International Conference on Speech and Computer (SPECOM)*, Saint-Petersbourg, Russie, Juin 2009.
- Dufour R., From prepared speech to spontaneous speech recognition system : a comparative study applied to french language, dans *International Conference on Soft Computing as Transdisciplinary Science and Technology (CSTST)*, pages 595–599, Cergy-Pontoise, France, Octobre 2008.

Conférences d'audience nationale avec comité de sélection

- Dufour R., Estève Y., Deléglise P. et Béchet F., Utilisation conjointe de modèles locaux et globaux pour la caractérisation et la détection de segments de parole spontanée, dans *Journées d'Étude sur le Parole (JEP)*, Mons, Belgique, Mai 2010.

– Dufour R., Estève Y. et Deléglise P., Corrections spécifiques du français sur les systèmes de reconnaissance automatique de la parole, dans *Rencontre des Jeunes Chercheurs en Parole (RJCP)*, Avignon, France, Novembre 2009.

Bibliographie

[Adda-Decker 1999] Adda-Decker M., Boula de Mareüil P. et Lamel L., Pronunciation variants in French : Schwa & Liaison, dans *International Congress of Phonetic Sciences (ICPhS)*, pages 2239–2242, San Francisco (Californie), États-Unis, Août 1999.

[Adda-Decker 2008] Adda-Decker M., Gendrot C. et Nguyen N., Contributions du traitement automatique de la parole à l'étude des voyelles orales du français, dans *Revue Traitement Automatique des Langues (TAL)*, volume 49, pages 13–46, 2008.

[Adda-Decker 2003] Adda-Decker M., Habert B., Barras C., Adda G., Boula de Mareüil P. et Paroubek P., A Disfluency Study for Cleaning Spontaneous Speech Automatic Transcripts and Improving Speech Language Models, dans *Workshop Disfluency In Spontaneous Speech (DISS)*, pages 67–70, Göteborg, Suède, Septembre 2003.

[Adda-Decker 2004] Adda-Decker M., Habert B., Barras C., Adda G., Boula de Mareüil P. et Paroubek P., Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage, dans *Journées d'Étude sur le Parole (JEP)*, Fès, Maroc, Avril 2004.

[Allauzen 2007] Allauzen A., Error Detection in Confusion Network, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1749–1752, Anvers, Belgique, Août 2007.

[Allauzen 2004] Allauzen A. et Gauvain J.-L., Construction automatique du vocabulaire d un système de transcription, dans *Journées d'Étude sur le Parole (JEP)*, Fès, Maroc, Avril 2004.

[Anastasakos 1997] Anastasakos T., McDonough J. et Makhoul J., Speaker Adaptive Training : A Maximum Likelihood Approach to Speaker Normalization, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 2, pages 1043–1046, Munich, Allemagne, Avril 1997.

[Anastasakos 1996] Anastasakos T., McDonough J., Schwartz R. et Makhoul J., A Compact Model for Speaker-Adaptive Training, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1137–1140, Philadelphie (Pennsylvanie), États-Unis, Octobre 1996.

[Ariki 2003] Ariki Y., Shigemori T., Kaneko T., Ogata J. et Fujimoto M., Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1453–1456, Genève, Suisse, Septembre 2003.

[Atal 1971] Atal B. S. et Hanauer S. L., Speech analysis and synthesis by Linear Prediction of the Speech Wave, dans *Journal of Acoustical Society of America*, volume 50, pages 637–655, Août 1971.

[Auran 2004] Auran C., Bouzon C., Hirst D., Lévy C. et Nocéra P., Algorithme de prédiction d'élisions de phonèmes et influence sur l'alignement automatique dans le cadre du projet Aix-MARSEC, dans *Journées d'Étude sur le Parole (JEP)*, Fès, Maroc, Avril 2004.

[Baum 1972] Baum L. E., An inequality and associated maximization technique in statistical estimation for probabilistic functions on markov processes, dans *Inequalities III*, volume 3, pages 1–8, 1972.

[Bazillon 2007] Bazillon T., Le codage de la parole spontanée pour la reconnaissance automatique de la parole, dans *Rencontre des Jeunes Chercheurs en Parole (RJCP)*, pages 16–19, Paris, France, Juillet 2007.

[Bazillon 2008a] Bazillon T., Estève Y. et Luzzati D., Manual vs assisted transcription of prepared and spontaneous speech, dans *Language Resources and Evaluation (LREC)*, Marrakech, Maroc, Mai 2008a.

[Bazillon 2008b] Bazillon T., Jousse V., Béchet F., Estève Y., Linarès G. et Luzzati D., La parole spontanée : transcription et traitement, dans *Revue Traitement Automatique des Langues (TAL)*, volume 49, pages 47–67, 2008b.

[Bear 1992] Bear J., Dowding J. et Shriberg E., Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog, dans *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 56–63, Newark (Delaware), États-Unis, Juillet 1992.

[Béchet 2001] Béchet F., LIA–PHON : Un système complet de phonétisation de textes, dans *Revue Traitement Automatique des Langues (TAL)*, volume 42, pages 47–67, 2001.

[Béchet 1999a] Béchet F., Nasr A., Spriet T. et De Mori R., Large Span Statistical Language Models : Application to Homophone Disambiguation for Large Vocabulary Speech Recognition in French, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1763–1766, Budapest, Hongrie, Septembre 1999a.

[Béchet 1999b] Béchet F., Nasr A., Spriet T. et De Mori R., Modèles de langage à portée variable : Application au traitement des homophones, dans *Traitement Automatique des Langues Naturelles (TALN)*, Cargèse, France, Juillet 1999b.

[Bertoldi 2001] Bertoldi N., Brugnara E., Cettolo M., Federico M. et Giuliani D., From broadcast news to spontaneous dialogue transcription : portability issues, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 37–40, Salt Lake City, États-Unis, Mai 2001.

[Biais 2005] Biais M., GRAC : GRAmmar Checker, pages 1–6, `http://grac.sourceforge.net/grac_architecture.pdf`, Février 2005.

[Bigi 2000] Bigi B., De Mori R., El-Bèze M. et Spriet T., A fuzzy decision strategy for topic identification and dynamic selection of language models, dans *Signal Processing Journal*, volume 80, pages 1085–1097, 2000.

[Boula de Mareüil 2005] Boula de Mareüil P., Habert B., Bénard F., Adda-Decker M., Barras C., Adda G. et Paroubek P., A quantitative study of disfluencies in French broadcast interviews, dans *Workshop Disfluency In Spontaneous Speech (DISS)*, Aix-en-Provence, France, Septembre 2005.

[Bustamante 1996] Bustamante F. R. et León F. S., GramCheck : A Grammar and Style Checker, dans *Conference on Computational Linguistics (COLING)*, pages 175–181, Copenhague, Danemark, Août 1996.

[Byrne 1997] Byrne B., Finke M., Khudanpur S., McDonough J., Nock H., Riley M., Saraclar M., Wooters C. et Zavaliagkos G., Pronunciation modelling for conversational speech recognition - A status report from WS97, dans *Automatic Speech Recognition and Understanding (ASRU)*, pages 26–33, Santa Barbara (Californie), États-Unis, Décembre 1997.

[Byrne 1998] Byrne W., Finke M., Khudanpur S., McDonough J., Nock H., Riley M., Saraclar M., Wooters C. et Zavaliagkos G., Pronunciation modelling using a hand-labelled corpus for conversational speech recognition, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 313–316, 1998.

[Byrne 2001] Byrne W., Venkataramani V., Kamm T., Zheng F., Song Z., Fung P., Liu Y. et Ruhi U., Automatic generation of pronunciation lexicons for Mandarin spontaneous speech, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 569–572, Salt Lake City (Utah), États-Unis, Mai 2001.

[Bürki 2008] Bürki A., Gendrot C., Gravier G., Linarès G. et Fougeron C., Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa, dans *Revue Traitement Automatique des Langues (TAL)*, volume 49, pages 165–197, 2008.

[Caelen-Haumont 2002a] Caelen-Haumont G., Perlocutory values and functions of melisms in spontaneous dialogue, dans *First International Conference on Speech Prosody*, pages 195–198, Aix-En-Provence, France, Avril 2002a.

[Caelen-Haumont 2002b] Caelen-Haumont G., Prosodie et dialogue spontané : valeurs et fonctions perlocutoires du mélisme, dans *Revue Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, volume 21, pages 13–24, Aix-en-Provence, France, 2002b.

[Campione 2004] Campione E. et Véronis J., Pauses et hésitations en français spontané, dans *Journées d'Étude sur le Parole (JEP)*, Fès, Maroc, Avril 2004.

[Candéa 2000] Candéa M., Les *euh* et les allongements dits "d'hésitation" : deux phénomènes soumis à certaines contraintes en français oral non lu, dans *Journées d'Étude sur le Parole (JEP)*, pages 73–76, Aussois, France, Juin 2000.

[Chen 1996] Chen F. et Goodman J., An empirical study of smoothing techniques for language modeling, dans *34th annual meeting on Association for Computational Linguistics*, pages 310–318, Santa Cruz (Californie), États-Unis, Juin 1996.

[Chen 2003] Chen L., Gauvain J.-L., Lamel L. et Adda G., Unsupervised Language Model Adaptation for Broadcast News, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 220–223, Hong Kong, Chine, Avril 2003.

[Chen 1998] Chen S. S. et Gopalakrishnan P. S., Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion, dans *DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, Lansdowne (Pennsylvanie), États-Unis, Février 1998.

[Clément 2009] Clément L., Gerdes K. et Marlet R., Grammaires d'erreur – correction grammaticale avec analyse profonde et proposition de corrections minimales, dans *Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France, Juin 2009.

[Cohen 1960] Cohen J., A coefficient of agreement for nominal scales, dans *Educational and Psychological Measurement*, volume 20, pages 37–46, 1960.

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

[Deléglise 2005] Deléglise P., Estève S. Yannick an Meignier et Merlin T., The LIUM speech transcription system : a CMU Sphinx III-based System for French Broadcast News, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1653–1656, Lisbonne, Portugal, Septembre 2005.

[Deléglise 2009] Deléglise P., Estève Y., Meignier S. et Merlin T., Improvements to the LIUM French ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ?, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2123–2126, Brighton, Angleterre, Royaume-Uni, Septembre 2009.

[Dempster 1977] Dempster A. P., Laird N. M. et Rubin D. B., Maximum likelihood from incomplete data via the EM algorithm, dans *Journal of the Royal Statistical Society*, volume 39, pages 1–38, 1977.

[Devillers 2004] Devillers L. et Vasilescu I., Détection des émotions à partir d'indices lexicaux, dialogiques et prosodiques dans le dialogue oral, dans *Journées d'Étude sur le Parole (JEP)*, Fès, Maroc, Avril 2004.

[Di Eugenio 2004] Di Eugenio B. et Glass M., The Kappa statistic : A second look, dans *Computational Linguistics*, volume 30, pages 95–101, 2004.

[Digalakis 1995] Digalakis V., Rtischev D. et Neumeyer L., Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures, dans *Transactions Speech and Audio Processing*, pages 357–366, Septembre 1995.

[Duez 1982] Duez D., Salient pauses and non salient pauses in three speech style, dans *Language and Speech*, volume 25, pages 11–28, 1982.

[Dufour 2008a] Dufour R., From prepared speech to spontaneous speech recognition system : a comparative study applied to french language, dans *International Conference on Soft Computing as Transdisciplinary Science and Technology (CSTST)*, pages 595–599, Cergy-Pontoise, France, Octobre 2008a.

[Dufour 2010] Dufour R., Bougares F., Esteve Y. et Deleglise P., Unsupervised model adaptation on targeted speech segments for LVCSR system combination, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japon, Septembre 2010.

[Dufour 2008b] Dufour R. et Estève Y., Correcting ASR outputs : specific solutions to specific errors in French, dans *Workshop Spoken Language Technology (SLT)*, pages 213–216, Goa, Inde, Décembre 2008b.

[Dufour 2009a] Dufour R., Estève Y., Deléglise P. et Béchet F., Local and global models for spontaneous speech segment detection and characterization, dans *Automatic Speech Recognition and Understanding (ASRU)*, Merano, Italie, Décembre 2009a.

[Dufour 2009b] Dufour R., Jousse V., Estève Y., Béchet F. et Linarès G., Spontaneous speech characterization and detection in large audio database, dans *International Conference on Speech and Computer (SPECOM)*, Saint-Pétersbourg, Russie, Juin 2009b.

[Ellis 2000] Ellis D., Improved recognition by combining different features and different systems, dans *AVIOS Speech Developers Conference and Expo*, pages 236–242, San José (Californie), États-Unis, Mai 2000.

[Estève 2009] Estève Y., Traitement automatique de la parole : contributions, dans *Habilitation à Diriger des Recherches (HDR)*, LIUM, Université du Maine, France, 2009.

[Estève 2010] Estève Y., Bazillon T., Antoine J.-Y., Béchet F. et Farinas J., The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news, dans *Language Resources and Evaluation (LREC)*, La Valette, Malte, Mai 2010.

[Estève 2004] Estève Y., Deléglise P. et Jacob B., Système de transcription automatique de la parole et logiciels libres, dans *Revue Traitement Automatique des Langues (TAL)*, volume 45, pages 15–39, 2004.

[Evermann 2000] Evermann G. et Woodland P., Posterior Probability Decoding, Confidence Estimation And System Combination, dans *The NIST 2000 Speech Transcription Workshop*, College Park (Maryland), États-Unis, Mai 2000.

[Favre 2007] Favre B., Hakkani-Tür D. et Cuendet S., Icsiboost, `http://code.google.com/p/icsiboost`, 2007.

[Federico 1998] Federico M. et De Mori R., Language modelling, dans *Spoken Dialogues with Computers*, pages 204–210, 1998.

[Fiscus 1997] Fiscus J. G., A Post-Processing System To Yield Reduced Word Error Rates : Recognizer Output Voting Error Reduction (ROVER), dans *Automatic Speech Recognition and Understanding (ASRU)*, pages 347–352, Santa Barbara (Californie), États-Unis, Décembre 1997.

[Forsberg 2003] Forsberg M., Why is Speech Recognition Difficult ?, dans *Chalmers University of Technology*, 2003.

[Fosler-Lussier 1999] Fosler-Lussier E. et Morgan N., Effects of speaking rate and word frequency on pronunciations in conversational speech, dans *Speech Communication*, volume 29, pages 137–158, Mai 1999.

[Fougeron 1999] Fougeron C. et Steriade D., Au delà de la syllabe : le rôle des informations articulatoires stockées dans le lexique pour l'analyse de la chute de schwa, dans *Journée d'Etudes Linguistiques*, pages 122–127, Nantes, France, 1999.

[Freund 1995] Freund Y. et Schapire R. E., A decision-theoretic generalization of on-line learning and an application to boosting, dans *European Conference on Computational Learning Theory (EUROCOLT)*, Barcelone, Espagne, Mars 1995.

[Furui 2003] Furui S., Recent advances in spontaneous speech recognition and understanding, dans *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pages 1–6, Tokyo, Japon, Avril 2003.

[Furui 2000] Furui S., Hori C. et Shinozaki T., Toward the realization of spontaneous speech recognition – Introduction of a Japanese Priority Program and preliminary results, dans *International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 518–521, Beijing, Chine, Octobre 2000.

[Furui 2005] Furui S., Nakamura M., Ichiba T. et Iwano K., Why Is the Recognition of Spontaneous Speech so Hard ?, dans *Text, Speech and Dialogue*, volume 3658, pages 9–22, 2005.

[Gales 1998] Gales M. J. F., Maximum likelihood linear transformations for HMM-based speech recognition, dans *Computer Speech and Language*, volume 12, pages 75–98, Avril 1998.

[Galliano 2005] Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J. et Gravier G., The ESTER phase II evaluation campaign for the rich transcription of French broadcast news, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1149–1152, Lisbonne, Portugal, Septembre 2005.

[Galliano 2009] Galliano S., Gravier G. et Chaubard L., The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2583–2586, Brighton, Angleterre, Royaume-Uni, Septembre 2009.

[Gauvain 2005] Gauvain J.-L., Adda G., Adda-Decker M., Allauzen A., Gendner V., Lamel L. et Schwenk H., Where Are We In Transcribing French Broadcast News ?, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1665–1668, Lisbonne, Portugal, Septembre 2005.

[Gauvain 1994a] Gauvain J.-L., Lamel L., Adda G. et Adda-Decker M., Speaker-Independent Continuous Speech Dictation, dans *Speech Communication*, volume 15, pages 21–27, Octobre 1994a.

[Gauvain 1994b] Gauvain J.-L., Lamel L., Adda G. et Adda-Decker M., The LIMSI Continuous Speech Dictation System : Evaluation on the ARPA Wall Street Journal Task, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 557–560, Adélaïde, Australie, Avril 1994b.

[Gauvain 1994c] Gauvain J.-L. et Lee C.-H., Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, dans *Transactions on Speech and Audio Processing*, volume 2, pages 291–298, Avril 1994c.

[Gendner 2002] Gendner V. et Adda-Decker M., Analyse comparative de corpus oraux et écrits français : mots, lemmes et classes morpho-syntaxiques, dans *Journées d'Étude sur le Parole (JEP)*, Nancy, France, Juin 2002.

[Goldwater 2010] Goldwater S., Jurafsky D. et Manning C. D., Which words are hard to recognize ?, dans *Speech Communication*, volume 52, pages 181–200, 2010.

[Goto 1999] Goto M., Itou K. et Hayamizu S. A., A Real-time Filled Pause Detection System for Spontaneous Speech Recognition, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 227–230, Budapest, Hongrie, Septembre 1999.

[Hazen 2000] Hazen J. T., Burianek T., Polifroni J. et Seneff S., Recognition Confidence Scoring for Use in Speech Understanding Systems, dans *Computer Speech and Language*, volume 16, pages 49–67, 2000.

[Heeman 1996] Heeman P. A., Loken-Kim K.-h. et Allen J. F., Combining the Detection and Correction of Speech Repairs, dans *International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 362–365, Philadelphie (Pennsylvanie), États-Unis, Octobre 1996.

[Henry 2004] Henry S., Campione E. et Véronis J., Répétitions et pauses (silencieuses et remplies) en français spontané, dans *Journées d'Étude sur le Parole (JEP)*, pages 261–264, Fès, Maroc, Avril 2004.

[Henry 2002] Henry S. et Pallaud B., étude des répétitions en français parlé spontané pour les technologies de la parole, dans *Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, pages 467–476, Nancy, France, Juin 2002.

[Hermansky 1990] Hermansky H., Perceptual linear predictive (PLP) analysis of speech, dans *Journal of Acoustical Society of America*, volume 87, pages 1738–1752, Avril 1990.

[Honal 2003] Honal M. et Schultz T., Correction of Disfluencies in Spontaneous Speech using a Noisy-Channel Approach, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2781–2784, Genève, Suisse, Septembre 2003.

[Honal 2005] Honal M. et Schultz T., Automatic disfluency removal on recognized spontaneous speech – rapid adaptation to speaker dependent dislfuencies, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 969–972, Philadelphie (Pennsylvanie), États-Unis, Mars 2005.

[Huet 2007] Huet S., Gravier G. et Sébillot P., Morphosyntactic processing of n-best lists for improved recognition and confidence measure computation, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, volume 12, pages 663–684, Anvers, Belgique, Août 2007.

[Huet 2010] Huet S., Gravier G. et Sébillot P., Morpho-syntactic post-processing of N-best lists for improved French Automatic Speech Recognition, dans *Computer Speech and Language*, volume 12, pages 663–684, 2010.

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

[Jelinek 1977] Jelinek F., Mercer R., Bahl L. et Baker J., Perplexity – A Measure of Difficulty of Speech Recognition Tasks, dans *94th meeting of the Acoustical Society of America*, volume 62, page S63, Décembre 1977.

[Jelinek 1987] Jelinek F. et Mercer R. L., Interpolated estimation of Markov source parameters from sparse data, dans *International Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, Pays-Bas, 1987.

[Jiang 2005] Jiang H., Confidence measures for speech recognition : A survey, dans *Speech Communication*, volume 45, pages 455–470, 2005.

[Johnson 2004] Johnson M., Charniak E. et Lease M., An improved model for recognizing disfluencies in conversational speech, dans *Rich Transcription 2004 Fall Workshop (RT-04F)*, Palisades (New York), États-Unis, Novembre 2004.

[Jousse 2008] Jousse V., Estève Y., Béchet F., Bazillon T. et Linarès G., Caractérisation et détection de parole spontanée dans de larges collections de documents audio, dans *Journées d'Étude sur le Parole (JEP)*, Avignon, France, Juin 2008.

[Katz 1987] Katz M. S., Estimation of probabilities from sparse data for the language model component of a speech recognizer, dans *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 35, pages 400–401, 1987.

[Kawahara 2003] Kawahara T., Nanjo H., Shinozaki T. et Furui S., Benchmark Test For Speech Recognition, dans *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pages 135–138, Tokyo, Japon, Avril 2003.

[Kessens 1999] Kessens J., Wester M. et Strik H., Improving the performance of a dutch CSR by modeling within-word and cross-word pronunciation, dans *Speech Communication*, volume 29, pages 193–207, 1999.

[Kipp 1997] Kipp A., Wesenick B. et Schiel F., Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1023–1026, Rhodes, Grèce, Septembre 1997.

[Kneser 1993] Kneser R. et Ney H., Improved clustering techniques for class-based statistical language modeling, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 2, pages 973–976, Berlin, Allemagne, Septembre 1993.

[Kneser 1995] Kneser R. et Ney H., Improved backing-off for M-gram language modeling, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 181–184, Détroit (Michigan), États-Unis, Mai 1995.

[Kuhn 1990] Kuhn R. et De Mori R., A cache-based natural language model for speech recognition, dans *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 12, pages 570–582, 1990.

[Lafferty 2001] Lafferty J., McCallum A. et Pereira F., Conditional random field : Probabilistic models for segmenting and labeling sequence data, dans *International Conference on Machine Learning (ICML)*, pages 282–289, Williamstown (Maryland), États-Unis, Juin 2001.

[Lavecchia 2006] Lavecchia C., Smaïli K. et Haton J.-P., How to handle gender and number agreement in statistical language models ?, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1854–1857, Pittsburgh (Pennsylvanie), États-Unis, Septembre 2006.

[Lease 2006] Lease M., M. J. et Charniak E., Recognizing Disfluencies in Conversational Speech, dans *IEEE Transactions on Audio, Speech and Language Processing*, volume 14, pages 1566–1573, Septembre 2006.

[Lecouteux 2008] Lecouteux B., Linarès G., Estève Y. et Gravier G., Combinaison de systèmes par décodage guidé, dans *Journées d'Étude sur le Parole (JEP)*, Avignon, France, Juin 2008.

[Lee 2001] Lee A., Kawahara T. et Shikano K., Julius – an open source real-time large vocabulary recognition engine, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1691–1694, Septembre 2001.

[Lee 1990] Lee K.-F., Hon H.-W. et Reddy R., An Overview of the SPHINX Speech Recognition System, dans *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 38, pages 35–45, Janvier 1990.

[Leggetter 1995] Leggetter C. J. et Woodland P. C., Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, dans *Computer Speech and Language*, volume 9, pages 171–185, Avril 1995.

[Lickley 1994] Lickley R. J., Detecting disfluency in spontaneous speech, dans *Thèse de doctorat*, Université d'Édimbourg, Écosse, Royaume-Uni, 1994.

[Liu 2003] Liu Y., Shriberg E. et Stolcke A., Automatic disfluency identification in conversational speech using multiple knowledge sources, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 957–960, Genève, Suisse, Septembre 2003.

[Liu 2005] Liu Y., Shriberg E., Stolcke A. et Harper M., Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3033–3036, Lisbonne, Portugal, Septembre 2005.

[Luzzati 2004] Luzzati D., Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané, dans *Workshop Modélisation pour l'Identification des Langues (MIDL)*, pages 13–17, Paris, France, Novembre 2004.

[Mangu 2000] Mangu H., Brill E. et Stolcke A., Finding Consensus in Speech Recognition : Word Error Minimization and other Applications of Confusion Networks, dans *Computer Speech and Language*, volume 14, pages 373–400, 2000.

[Martin 2006] Martin P., Intonation du Français : Parole spontanée et parole lue, dans *Estudios de Fonética Experimental*, volume 15, pages 133–162, Université d'Édimbourg, Écosse, Royaume-Uni, 2006.

[Mauclair 2006] Mauclair J., Mesures de confiance en traitement automatique de la parole et applications, dans *Thèse de doctorat*, LIUM, Université du Maine, Le Mans, France, Décembre 2006.

[Meignier 2010] Meignier S. et Merlin T., LIUM_SpkDiarization : An open source toolkit for diarization, dans *CMU Sphinx Users and Developers Workshop*, Dallas (Texas), États-Unis, Mars 2010.

[Metze 2000] Metze E., Kemp T., Schaaf T., Schultz T. et Soltau H., Confidence measure based language identification, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1827–1830, Istanbul, Turquie, Juin 2000.

[Mohri 2002] Mohri M., Pereira F. C. N. et Riley M., Weighted Finite-State Transducers in Speech Recognition, dans *Computer Speech and Language*, volume 16, pages 69–88, 2002.

[Mudge 2009] Mudge R. S., After the Deadline – Language Checking Technology, `http://open.afterthedeadline.com`, 2009.

[Naber 2003] Naber D., A Rule-Based Style and Grammar Checker, dans *Thèse de Masters*, Université de Bielefeld, Allemagne, Décembre 2003.

[Nakamura 2008] Nakamura M., Iwano K. et Furui S., Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance, dans *Computer Speech and Language*, volume 22, pages 171–184, 2008.

[Nemoto 2008a] Nemoto R., Vasilescu I. et Adda-Decker M., Mots fréquents homophones en français : analyse acoustique et classification automatique par fouille de données, dans *Journées d'Étude sur le Parole (JEP)*, Avignon, France, Juin 2008a.

[Nemoto 2008b] Nemoto R., Vasilescu I. et Adda-Decker M., Speech Errors on Frequently Observed Homophones in French : Perceptual Evaluation vs Automatic Classification, dans *Language Resources and Evaluation (LREC)*, Marrakech, Maroc, Mai 2008b.

[Nocéra 2002] Nocéra P., Linarès G., Massonié D. et Lefort L., Phoneme lattice based A* search algorithm for speech recognition, dans *Text, Speech and Dialogue*, volume 2448, pages 83–111, 2002.

[Normandin 1994] Normandin Y., Lacouture R. et Cardin R., MMIE training for large vocabulary continuous speech recognition, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1367–1370, Yokohama, Japon, Septembre 1994.

[O'Shaughnessy 1992] O'Shaughnessy D., Recognition of hesitations in spontaneous speech, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 521–524, San Francisco (Californie), États-Unis, Mars 1992.

[O'Shaughnessy 1993] O'Shaughnessy D., Analysis and automatic recognition of false starts in spontaneous speech, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 2, pages 724–727, Minneapolis (Minnesota), États-Unis, Avril 1993.

[Pallaud 2004] Pallaud B. et Henry S., Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé, dans *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Louvain-la-Neuve, Belgique, Mars 2004.

[Pallaud 2007] Pallaud B. et Xuereb R., Les troncations et les répétitions de mots chez un locuteur bègue, dans *Revue Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, volume 26, pages 93–113, Aix-en-Provence, France, 2007.

[Pallet 1990] Pallet D. S., Fisher W. M. et Fiscus J. G., Tools for the Analysis of Benchmark Speech Recognition Tests, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 97–100, Albuquerque (Nouveau-Mexique), États-Unis, Avril 1990.

[Pérennou 1987] Pérennou G. et de Calmès M., BDLex lexical data and knowledge base of spoken and written French, dans *European Conference on Speech Technology (ECST)*, pages 1393–1396, Édimbourg, Écosse, Royaume-Uni, Septembre 1987.

[Polzin 1998] Polzin S. T. et Waibel A., Pronunciation variations in emotional speech, dans *ESCA Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 103–108, Kerkrade, Pays-Bas, Mai 1998.

[Povey 2002] Povey D. et Woodland P., Minimum phone error and i-smoothing for improved discriminative training, dans *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 105–108, Orlando (Floride), États-Unis, Mai 2002.

[Rabiner 1989] Rabiner L. R., A tutorial on hidden Markov models and selected applications in speech recognition, dans *Proceedings of the IEEE*, volume 77, pages 257–286, Février 1989.

[Ravishankar 2000] Ravishankar M., Singh R., Raj B. et Stern R. M., The 1999 CMU 10x real time broadcast news transcription system, dans *DARPA Workshop on Automatic Transcription of Broadcast News*, Washington DC, États-Unis, Mai 2000.

[Riley 1999] Riley M., Byrneb W., Finkec M., Khudanpurb S., Ljoljea A., McDonough J., Nockd H., Saraclarc M., Wooterse C. et Zavaliagkosf G., Stochastic pronunciation modelling from hand-labelled phonetic corpora, dans *Speech Communication*, volume 29, pages 209–224, Novembre 1999.

[Rouas 2004] Rouas J.-L., Farinas J. et Pellegrino F., Évaluation automatique du débit de la parole sur des données multilingues spontanées, dans *Journées d'Étude sur le Parole (JEP)*, pages 437–440, Fès, Maroc, Avril 2004.

[Rouvier 2010] Rouvier M., Dufour R., Linarères G. et Estève Y., A Language-identification inspired method for spontaneous speech detection, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japon, Septembre 2010.

[Sagawa 2004] Sagawa H., Mitamura T. et Nyberg E., Correction Grammars for Error Handling in a Speech Dialog System, dans *Human Language Technology conference (HLT-NAACL)*, pages 61–64, Boston (Massachusetts), États-Unis, Mai 2004.

[Schapire 2003] Schapire R. E., The Boosting Approach to Machine Learning : An Overview, dans *Nonlinear Estimation and Classification*, 2003.

[Schapire 2000] Schapire R. E. et Singer Y., BoosTexter : A boosting-based system for text categorization, dans *Machine Learning*, volume 39, pages 135–168, 2000.

[Servan 2006] Servan C., Raymond C., Béchet F. et Nocéra P., Conceptual decoding from word lattices : application to the spoken dialogue corpus MEDIA, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1614–1617, Pittsburgh (Pennsylvanie), États-Unis, Septembre 2006.

[Shinozaki 2001] Shinozaki T., Hori C. et Furui S., Towards Automatic Transcription of Spontaneous Presentations, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 1, pages 491–494, Aalborg, Danemark, Septembre 2001.

[Shriberg 1999] Shriberg E., Phonetic consequences of speech disfluency, dans *International Congress of Phonetic Sciences (ICPhS)*, pages 619–622, San Francisco (Californie), États-unis, Août 1999.

[Shriberg 2005] Shriberg E., Spontaneous speech : How people really talk and why engineers should care, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1781–1784, Lisbonne, Portugal, Septembre 2005.

[Siu 1999] Siu M. et Gish H., Evaluation of word confidence for speech recognition systems, dans *Computer speech and language*, volume 13, pages 299–318, 1999.

[Siu 1996] Siu M.-H. et Ostendorf M., Modeling disfluencies in conversational speech, dans *International Conference on Spoken Language Processing (ICSLP)*, volume 1, Philadelphie (Pennsylvanie), États-Unis, Octobre 1996.

[Sloboda 1996] Sloboda T. et Waibel A., Dictionary learning for spontaneous speech recognition, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 2328–2331, Philadelphie (Pennsylvanie), États-Unis, Octobre 1996.

[Souque 2007] Souque A., Conception et développement d'un formalisme de correction grammaticale automatique - Application au français, dans *Mémoire de Master 2 Recherche Sciences du Langage*, Université Stendhal, Grenoble, France, Juin 2007.

[Souque 2008] Souque A., Vers une nouvelle approche de la correction grammaticale automatique, dans *Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, Avignon, France, Juin 2008.

[Spilker 2000] Spilker J., Klarner M. et Görz G., Processing Self-Corrections in a Speech-to-Speech System, dans *Conference on Computational Linguistics (COLING)*, volume 2, pages 131–140, Sarrebruck, Allemagne, Août 2000.

[Stolcke 2002] Stolcke A., SRILM - An extensible language modeling toolkit, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver (Colorado), États-Unis, Septembre 2002.

[Stolcke 1997] Stolcke A., König Y. et Weintraub M., Explicit Word Error Minimization In N-Best List Rescoring, dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 163–166, Rhodes, Grèce, Septembre 1997.

[Stolcke 1998] Stolcke A., Shriberg E., Bates R., Ostendorf M., Hakkani D., Plauche M., Tur G. et Lu Y., Automatic Detection Of Sentence Boundaries And Disfluencies Based On Recognized Words, dans *International Conference on Spoken Language Processing (ICSLP)*, volume 5, pages 2247–2250, Sydney, Australie, Novembre 1998.

[Stolcke 2004] Stolcke A., Wang W., Vergyri D., Ramana V., Gadde R. et Zheng J., An efficient repair procedure for quick transcriptions, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1961–1964, Île de Jeju, Corée du Sud, Octobre 2004.

[Strik 1999] Strik H. et Cucchiarini C., Modeling pronunciation variation for ASR : A survey of the literature, dans *Speech Communication*, volume 29, pages 225–246, Novembre 1999.

[Suhm 1994] Suhm B. et Waibel A., Towards Better Language Models For Spontaneous Speech, dans *International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 831–834, Yokohama, Japon, Septembre 1994.

[Vasilescu 2009] Vasilescu I., Adda-Decker M., Lamel L. et Halle P., A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English, dans *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 144–147, Brighton, Angleterre, Royaume-Uni, Septembre 2009.

[Vasilescu 2004] Vasilescu I., Candea M. et Adda-Decker M., Hésitations autonomes dans 8 langues : une étude acoustique et perceptive, dans *Workshop Modélisation pour l'Identification des Langues (MIDL)*, pages 25–30, Paris, France, Novembre 2004.

[Viterbi 1967] Viterbi A. J., Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, dans *IEEE Transactions on Information Theory*, volume 13, pages 260–269, 1967.

[Walker 2000] Walker M. et Langkilde I., Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system, dans *International Conference on Machine Learning (ICML)*, pages 1111–1118, San Francisco (Californie), États-Unis, Juin 2000.

[Walker 2004] Walker W., Lamere P., Kwok P., Raj B., Singh R., Gouvea E., Wolf P. et Woelfel J., Sphinx-4 : A flexible open source framework for speech recognition, Sun Microsystems Laboratories, Novembre 2004.

[Ward 1989] Ward W., Understanding spontaneous speech, dans *Workshop on Speech and Natural Language*, pages 137–141, Philadelphie (Pennsylvanie), États-Unis, Octobre 1989.

[Wester 2000] Wester M. et Fosler-Lussier E., A Comparison Of Data-Derived And Knowledge-Based Modeling Of Pronunciation Variation, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 270–273, Beijing, Chine, Octobre 2000.

[Witten 1991] Witten I. H. et Bell T. C., The zero-frequency problem : estimating the probabilities of novel events in adaptive text compression, dans *IEEE transactions on information theory*, volume 37, pages 1085–1094, 1991.

[Zweig 2002] Zweig G., Yvon F. et Saon G., Sirocco, un système ouvert de reconnaissance de la parole, dans *Journées d'Étude sur le Parole (JEP)*, pages 273–276, Nancy, France, Juin 2002.

Résumé

Les systèmes de Reconnaissance Automatique de la Parole (RAP) atteignent actuellement des performances suffisantes pour être intégrés dans diverses applications (dialogue homme-machine, recherche d'information, indexation automatique...). Cependant, dans le cadre de la reconnaissance automatique de la parole continue à grand vocabulaire, que l'on utilise par exemple pour transcrire des émissions radiophoniques d'information, la qualité des transcriptions varie selon le type de parole contenu dans les documents. En effet, les systèmes de RAP ont beaucoup plus de facilité à transcrire de la parole préparée, proche d'un texte lu, que de la parole spontanée, caractérisée par de nombreuses spécificités (disfluences, agrammaticalité, baisse de la fluidité de la parole...).

Le travail de cette thèse vise le traitement de la parole spontanée et s'inscrit dans le cadre du projet EPAC (Exploration de masse de documents audio pour l'extraction et le traitement de la PArole Conversationnelle). L'objectif principal est de proposer des solutions pour améliorer les performances des systèmes de RAP sur ce type de parole. Nous avons choisi d'aborder, dans notre travail, la parole spontanée en tant qu'objet d'étude particulier nécessitant des traitements spécifiques.

Ainsi, dans un premier temps, nous proposons un outil de détection automatique de la parole spontanée, basé sur les spécificités de ce type de parole. Cet outil est très important puisqu'il nous permet, dans un deuxième temps, de proposer une approche d'adaptation des modèles acoustiques et des modèles de langage du système de RAP à la parole spontanée sans ajout de données, en sélectionnant automatiquement les segments contenant ce type de parole. La transcription résultant de cette adaptation propose des hypothèses de reconnaissance différentes de celles fournies par le système de base. La combinaison de ces deux propositions de transcription permet d'observer une réduction significative du taux d'erreur-mot.

Ce besoin de solutions spécifiques a finalement orienté une partie de notre travail vers la correction d'un problème particulièrement présent en français : l'homophonie. Nous cherchons alors à corriger les transcriptions, fournies par un système de RAP, au moyen d'une méthode proposant des solutions spécifiques à certains problèmes particuliers de l'homophonie. L'approche se focalise sur la correction de certaines erreurs, auxquelles une solution particulière est proposée. Cette méthode, en post-traitement des systèmes de RAP, corrige certains mots et classes de mots homophones, indépendamment du système de RAP utilisé.

Mots-clés: Reconnaissance automatique de la parole, parole spontanée, homophonie, classification automatique.