Quelques contributions en reconnaissance automatique de la parole robuste

Christophe Cerisara

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

Quelques contributions en reconnaissance automatique de la parole robuste

MÉMOIRE

pour l'obtention de l'

Habilitation de l'Université Henri Poincaré – Nancy I

(Spécialité Informatique)

par

Christophe Cerisara

Composition du jury

Président :	Jean-Paul Haton
Rapporteurs :	Hervé Glotin Laurent Besacier Régine André-Obrecht
Examinateurs :	Anne Boyer Claire Gardent Hugo Van Hamme Yves Laprie

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thloria.

Remerciements

Je voudrais remercier en premier lieu Jean-Paul Haton, qui a toujours été pour moi (et pour beaucoup d'autres) le guide et l'exemple que j'ai tenté de suivre, tout au long de ces années passées à débroussailler tant bien que mal les innombrables arbres et autres jungles bayésiennes que j'ai dû décortiquer depuis la thèse et qui ne m'ont sans doute pas laissé indemne. Sans Jean-Paul, je ne serais pas ici, cela ne fait aucun doute. Merci mille fois !

Je m'étonne encore aujourd'hui d'ailleurs d'avoir eu la chance d'être entouré successivement par autant de collègues et amis de si grande qualité scientifique et humaine ; les plus importants pour moi et ceux qui m'ont le plus aidé étant certainement, chronologiquement, Yifan Gong, Jean-Claude Junqua et Yves Laprie. Je ne pense pas avoir mérité de tels soutiens à mes côtés, et je ne peux que les remercier chaleureusement de m'avoir accordé le leur, ainsi que leur amitié. Merci beaucoup !

Un grand merci aussi à Régine André-Obrecht, Hervé Glotin et Laurent Besacier, pour avoir accepté de consacrer un peu de leur temps à relire ce mémoire ; leurs avis et conseils me sont très précieux, et j'espère de tout cœur pouvoir continuer à collaborer avec eux bien après cette étape importante pour moi que constitue ce travail de synthèse.

Un grand merci également aux membres du jury : Anne Boyer, Claire Gardent et Hugo Van Hamme d'avoir accepté de participer à mon jury ainsi que pour leurs nombreux conseils toujours riches en enseignements pour moi.

Je ne peux que remercier chaleureusement toute la communauté scientifique qui m'a entouré pendant ces quelques années, et en particulier les membres de l'équipe Parole, mais aussi de nombreux chercheurs d'autres équipes avec qui j'ai échangé et collaboré.

Un merci tout particulier à Laurent Wendling, pour sa relecture bien sûr, mais surtout pour son amitié, les multiples défis intellectuels et les discussions scientifiques interminables qu'il m'accordait, qui ont sans aucun doute grandement influencé ma façon d'aborder un problème. Enfin, alors que j'écris ces lignes de la main droite, la gauche occupée à tenir sur mon genou Paul et son petit livre d'images, merci à Agnès, Théo, Nathan et Paul pour m'avoir accordé le temps nécessaire à rédiger ce mémoire, temps dont je leur suis redevable. Merci pour tous ces bonheurs partagés !

Table des matières

Dossier : Parcours personnel

Curiculum Vitæ

État Civil

Christophe CERISARA

14 A, Grande Rue
54610 SIVRY

03.83.31.51.82
Christophe.Cerisara@loria.fr

Né le 3 juin 1973 à Metz, nationalité française
Marié, trois enfants.

Situation actuelle

Chargé de Recherche 1ère classe au CNRS

Cursus professionnel

Janvier 2006	Promotion CR1 CNRS
Octobre 2002	Titularisation CR2 CNRS
Juin 2001	Candidatures : CR2 CNRS : admis CR2 INRIA : admis à Nancy Maître de Conférence : Classé 1er à l'UHP Nancy I Classé 1er à l'IUT de Grenoble Classé 1er à l'IUT du Mans
2000-2001	Attaché Temporaire d'Enseignement et de Recherche (ATER) au département informatique de l'Université Henri Poincaré de Nancy I
Juin 2000	Candidatures : Nancy I : classé 3ème sur un poste Nancy II : classé 7ème sur deux postes
Mars 2000	Obtention de la qualification pour les candidatures Maître de Conférence
1999-2000	Post-doctorat au Panasonic Speech Technology Laboratory aux USA (*voir descriptif ci-dessous*)
1996-1999	Moniteur à l'ESSTIN et allocataire de recherche MESR
Été 1996	Stage de recherche de DEA au CLIPS/GEOD à Grenoble *Thème de recherche :* reconnaissance automatique de la parole par des modèles hybrides combinant modèles stochastiques et réseaux de neurones *Directeur :* Jean Caelen
Été 1995	Stage de recherche au CRIN/INRIA Lorraine à Nancy *Thème de recherche :* reconnaissance automatique de la parole par les modèles sto-

chastiques de trajectoires
Directeur : Jean-Paul Haton

Cursus universitaire

1996-1999	Doctorat en informatique (Mention très honorable avec les félicitations du jury)
1995-1996	D.E.A. (Mention bien) en Informatique, Systèmes et Communications de l'Université Joseph Fourier à Grenoble
1993-1996	Diplôme d'ingénieur (Mention bien) de l'ENSIMAG (Informatique et Mathématiques Appliquées) à Grenoble
1991-1993	Mathématiques Supérieures et Spéciales (option M') à Fabert (Metz)

Descriptif du Post-doctorat

Laboratoire	Speech Technology Laboratory (Panasonic Inc.), 3888 State St. Suite 202, Santa Barbara, CA 93105, U.S.A
Responsable	Dr. Jean-Claude Junqua
Financement	Post-doctorat financé par une bourse INRIA de post-doctorat à l'étranger, ainsi que par un complément de salaire versé par le laboratoire d'accueil
Thèmatique	Reconnaissance robuste de la parole, application à la conception d'un système de navigation embarqué dans une voiture

Descriptif du Doctorat

Titre	Contribution de l'approche multi-bandes à la reconnaissance automatique de la parole.
Financement	Allocation de recherche du ministère MENESR et contrat d'encadrement doctoral.
Directeur	Pr. Jean-Paul Haton
Soutenance	Thèse soutenue le 20 septembre 1999, au Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) à NANCY.

Résumé des activités de recherche

1 Introduction

Dans ce chapitre, je résume très brièvement, en quelques pages seulement, mes principales contributions de recherche au cours de ces dernières années. Pour une présentation plus approfondie de mes contributions, j'invite le lecteur à se reporter à la partie description scientifique du mémoire, à partir de la page 29.

L'objectif de ce chapitre étant de présenter mes contributions principales, j'ai décidé d'isoler chacune d'entre elle en la présentant indépendamment des autres. Dans la partie suivante, mes contributions sont au contraire introduites, présentées en contexte et motivées. Le mémoire développera également une analyse globale du domaine selon un point de vue original, celui de l'information, ce qui permet de mieux comprendre les similitudes et différences entre les nombreuses pistes de recherche explorées au cours de ces dernières années et donc de mettre en perspectives les évolutions prévisibles et potentiellement intéressantes pour l'avenir.

2 Modèles acoustiques multi-bandes

Les modèles multi-bandes décomposent le signal acoustique en plusieurs bandes de fréquences, et modélisent chacune de ces bandes indépendamment les unes des autres. Les unités ou séquences d'unités reconnues dans chaque bande sont ensuite recombinées en un résultat unique.

Cette idée est issue de travaux de psycho-acoustiques qui suggèrent que l'audition humaine fonctionnerait selon le même principe, c'est-à-dire traiterait les informations acoustiques indépendamment dans chaque bande fréquentielle.

L'un des aspects les plus délicats de cette approche est la méthode de recombinaison choisie : de nombreuses approches de recombinaison ont été proposées, et j'en ai moi-même testé quelques-unes, comme la combinaison linéaire pondérée ou l'utilisation d'un réseau de neurones, mais aucune de ces approches n'a permis d'approcher, même imparfaitement, les résultats observés dans les expériences psycho-acoustiques, qui ont mis en évidence une recombinaison "idéale" du système auditif humain. En effet, il a été observé que les taux d'erreur de chaque bande fréquentielle se multiplient lors de la décision finale, ce qui consitute un résultat qui n'a jamais pu être reproduit artificiellement.

Afin de pallier aux limites de nos approches de combinaison, j'ai proposé d'augmenter l'information en entrée du système, en combinant les résultats de chaque bande et du spectre complet, ce qui permet de rendre au système multi-bandes l'information jointe perdue lors du découpage fréquentiel, tout en isolant les zones fréquentielles éventuellement bruitées. J'ai ainsi montré que la reconnaissance pouvait être améliorée significativement même en conditions non bruitées.

Toutefois, l'intérêt principal d'un système multi-bandes est sa capacité à désynchroniser les différentes bandes de fréquences. En effet, un système de reconnaissance traditionnel, qui modélise chaque vecteur acoustique par un mélange de gaussiennes, peut être également considéré comme un systèmes multi-bandes avec une recombinaison bayésienne, au moins lorsque les matrices de covariance sont supposées diagonales, hypothèse très commune. La différence principale entre les systèmes multi-bandes et les modèles classiques est donc bien la possibilité de désynchroniser les bandes. Pour explorer cette possibilité, j'ai testé la resynchronisation des bandes après chaque trame ou chaque modèle acoustique, comme cela avait déjà été proposé par ailleurs, mais j'ai également proposé de resynchroniser les bandes seulement à la fin de la

phrase, ce qui permet d'exploiter au mieux les avantages du modèle multi-bandes. Toutefois, l'avantage acquis lors de cette désynchronisation était en grande partie compensée par la perte de performances due à l'augmentation de l'espace de recherche.

Une autre contribution importante au domaine est la proposition et l'etude de nouvelles unités acoustiques mieux adaptées à l'information effectivement présente dans chaque bande considérée individuellement.

La modélisation multi-bandes a depuis évolué notamment vers les modèles "multi-flux" qui intègrent plusieurs sources d'information différente, comme la lecture labiale et l'acoustique. De nouvelles solutions ont également été proposées pour résoudre le problème de la recombinaison, la plus intéressante d'entre elles étant certainement l'utilisation de modèles bayésiens qui permettent à la fois d'intégrer la combinaison dans le formalisme du modèle et de modéliser explicitement dans le réseau le degré d'asynchronisme souhaité. Il est également raisonnable de considérer la reconnaissance avec données manquantes comme une évolution naturelle des modèles multi-bandes, car nous y retrouvons l'idée de masquer certaines parties du spectre et de marginaliser les contributions correspondantes lors du calcul de la vraisemblance des observations. C'est d'ailleurs dans cette voie que mes efforts se sont portés par la suite.

3 Adaptation des modèles acoustiques

Mes principales contributions dans le domaine de l'adaptation des modèles acoustiques sont d'une part les modèles auto-régressifs multi-échelles, et d'autre part les approches basées sur l'adaptation jacobienne.

3.1 Modèles auto-régressifs

Dans les approches classiques d'adaptation telles que MLLR ou MAP, un arbre de régression est généralement construit afin de regrouper hiérarchiquement les gaussiennes des modèles acoustiques en classes de plus en plus spécialisées. Cet arbre est très important, car il permet de modifier dynamiquement la précision de l'adaptation en fonction de la quantité d'adaptation disponible. Ainsi, lorsque seulement quelques dizaines de secondes d'adaptation sont disponibles, les paramètres de l'adaptation seront calculés pour la racine de l'arbre, qui contient toutes les gaussiennes des modèles. Par contre, si plusieurs minutes de signal étiquetté sont disponibles, alors le système descendra plus profondément dans l'arbre de régression, et utilisera une transformation différente par exemple pour les voyelles et les consonnes.

Il est donc fondamental de pouvoir descendre autant que possible dans l'arbre de régression, afin d'augmenter la précision de l'adaptation. Pour ce faire, j'ai proposé, en collaboration avec Khalid Daoudi, d'utiliser des modèles multi-échelles qui tiennent compte des relations de dépendance entre deux niveaux de l'arbre, ce qui permet de decsendre même des dans nœuds possédant peu de données d'adaptation. Je me suis donc appuyé sur la théorie des systèmes linéaires dynamiques qui modélisent cette dépendance par des distributions conditionnelles linéaires, et qui proposent des algorithmes d'inférence rapide, comme l'algorithme de Rauch-Tung-Striebel.

3.2 Adaptation jacobienne

J'ai beaucoup travaillé sur l'adaptation jacobienne, qui est une approximation linéaire dans le cepstre de la fonction d'adaptation utilisée dans la combinaison parallèle de modèles (PMC).

Le principale avantage de l'adaptation jacobienne est sont coût très réduit par rapport à une adaptation exacte, comme celle réalisée dans PMC. Ce faible coût permet de déployer cette approche dans des puces de faible capacité, comme celles des téléphones portables. J'ai exploré cet avantage de l'approche jacobienne en proposant une modification de l'algorithme d'adaptation permettant de régler avec précision le meilleur compromis possible entre complexité et qualité de l'approximation, en fonction de la puissance disponible et des exigences de performances de l'application. Cette capacité est obtenue grâce à une discrétisation plus ou moins fine de l'espace acoustique sur laquelle des transformations exactes sont calculées, et servent de points de référence sur lesquels s'appuie l'adaptation linéaire. Cette approche permet également de s'affranchir en partie de l'hypothèse de base de l'adaptation jacobienne, qui est la proximité acoustique des environnements d'apprentissage et de test. Dans ce même but, j'ai également

proposé de compenser le biais induit par l'adaptation jacobienne lorsque ces conditions diffèrent de plus en plus.

Une autre contribution importante au domaine a été d'étendre le formalisme jacobien aux bruits additifs et convolutifs.

Notons pour conclure que certaines conséquences induites par l'approximation linéaire réalisée par l'adaptation jacobienne peuvent se révéler particulièrement intéressante, en annulant certains défauts de l'adaptation PMC. Citons-en simplement deux ici. Tout d'abord le fait que PMC ne peut qu'ajouter du bruit aux modèles, ce qui impose d'avoir des conditions d'apprentissage aussi peu bruitées que possibles, conditions qui sont donc relativement éloignées des conditions réelles, alors que l'adaptation jacobienne fonctionne mieux lorsque les modèles sont enregistrés dans des environnements (relativement) bruités. Ainsi, l'adaptation jacobienne peut aussi bien ajouter qu'enlever du bruit aux modèles, grâce à la linéarité de la transformation. Enfin, toujours du fait de cette linéarité, l'adaptation jacobienne n'introduit pas de seuillage, comme il en existe dans PMC, et qui sont à l'origine du bruit musical de la soustraction spectrale.

4 Débruitage du signal

Ma principale contribution dans le domaine du débruitage du signal de parole concerne la proposition d'une méthode de débruitage bayésien nommée SPACE, pour *Stereo-based Picewise Affine Compensation for Environments*. J'ai proposé cette approche en collaboration avec Khalid Daoudi, à l'IRIT de Toulouse.

Le principe fondamental de cette approche consiste à modéliser respectivement les espaces de parole propre et bruitée par deux GMM (Gaussian Mixture Model) dont les gaussiennes sont en correspondance. Nous avons ensuite proposé deux variantes de SPACE :

- Dans la première, nous construisons une transformation affine de débruitage entre les espaces définis par ces deux GMM.
- Dans la seconde, nous remplaçons cette transformation affine par une technique de débruitage exploitant la corrélation croisée entre les observations bruitées et non bruitées.

La principale difficulté de SPACE est de préserver autant que possible la correspondance entre les gaussiennes propres et bruitées. Pour ce faire, nous avons testé plusieurs approches :

- La première préserve l'alignement des observations et des gaussiennes entre les apprentissages des deux GMM ;
- La deuxième s'appuie sur le critère de minimisation de l'erreur quadratique pour construire une gaussienne propre à partir d'une gaussienne bruitée ;
- La troisième réalise un apprentissage joint dans l'espace propre et bruité, puis marginalise le modèle résultant en ses deux composantes.

L'avantage de notre approche par rapport aux autres méthodes classiques de débruitage, comme SPLICE, est de s'appuyer sur deux ensembles de GMM qui sont mis en correspondance, ce qui permet d'adapter l'un ou l'autre des mélanges à de nouvelles conditions qui n'ont pas été vues à l'apprentissage. Dans ce cadre, nous avons proposé une méthode d'adaptation linéaire simple mais efficace, car elle garantit la conservation de la correspondance entre les GMM propres et bruités.

5 Reconnaissance avec données manquantes

5.1 Principes de la reconnaissance avec données manquantes

Plutôt que de bruiter les modèles ou débruiter le signal, la reconnaissance avec données manquantes suppose que les modèles de parole ne peuvent pas représenter tout le signal perçu, mais seulement une partie de celui-ci, le reste relevant de la contribution d'une ou plusieurs autres sources de bruits.

La séparation entre les observations qui correspondent respectivement à de la parole et à du bruit est réalisée dans le domaine spectral au moyen d'un masque qui, dans sa forme la plus simple, affecte à chaque coefficient spectral une valeur binaire selon qu'elle représente principalement la parole ou le bruit.

Il existe donc deux problèmes fondamentaux en reconnaissance avec données manquantes, problèmes qui peuvent être résolus conjointement, mais que nous séparons pour la clarté de l'exposé :

- Construire le masque, c'est-à-dire décider pour chaque coefficient spectral s'il est dominé par le bruit ou la parole ;
- Prendre en compte le masque pendant la reconnaissance.

Dans la littérature, deux approches principales ont été proposées pour résoudre le premier problème : la première s'appuie sur des techniques de traitement de signal pour estimer le rapport signal-sur-bruit local ou d'autres critères comme l'harmonicité, et en déduire une valeur de masque : c'est l'école de "Sheffield". La seconde s'appuie sur des modèles stochastiques de données respectivement masquées et non masquées : c'est l'école du "CMU". Mon travail prend place dans le cadre de cette deuxième approche.

De même, pour prendre en compte les masques pendant la reconnaissance, deux méthodes concurrentes peuvent être respectivement attribuées à ces deux universités : pour l'Angleterre, la méthode dite de marginalisation, qui intègre lors du calcul de la vraisemblance des observations les coefficients masqués sur toutes les valeurs que peut prendre la contribution de la parole seule ; et pour les Etats-Unis, la méthode d'imputation, qui tente de reconstruire la contribution de la parole seule. Mon travail s'inspire cette fois de l'approche anglaise.

Je résume ci-dessous mes principales contributions, respectivement pour les deux problèmes.

5.2 Étude sur la génération de masques

Ma première contribution concerne la rédaction d'un article de synthèse réalisant une étude et un état de l'art des différentes méthodes qui permettent d'estimer un masque de données manquantes. J'y passe en revue les approches basées sur l'analyse de scènes auditives computationnelle, la séparation de sources aveugle avec un seul capteur, les techniques de traitement de signal ainsi que les méthodes fondées sur différents types de modèles.

5.3 Modèles de masques contextuels et dépendant de l'environnement

Une autre contribution, qui a été réalisée dans le cadre du travail de thèse de Sébastien Demange, propose d'améliorer la qualité des modèles de masque en réduisant la variabilité du signal modélisé au moyen d'un nouvelle variable aléatoire discrète représentant différents types d'environnements bruités. Le système résultant estime donc d'abord quelle est la probabilité d'être dans un certain type d'environnement, puis pondère les contributions des modèles correspondant par cette probabilité.

De plus, toujours dans le but de réduire la variabilité du signal modélisé, nous avons proposé d'augmenter les observations acoustiques modélisées par leur contexte temporel et fréquentiel. Le premier est pris en compte grâce aux dérivées temporelles des observations et aux transitions des modèles de Markov. Le second l'est en modélisant un masque vectoriel pour tous les coefficients spectraux au sein d'un même vecteur acoustique, ce qui ne peut toutefois être réalisé directement à cause de l'explosion combinatoire du nombre de masques différents à modéliser. Nous avons donc réduit le nombre de modèles en discrétisant l'espace des masques, après avoir montré que cet espace est creux, ce qui autorise à ne modéliser que les quelques régions denses en masques.

5.4 Optimisation de la marginalisation

En ce qui concerne l'utilisation des masques de données manquantes pendant la reconnaissance, également dans le cadre de la thèse de Sébastien Demange, nous avons proposé d'optimiser l'intervalle de marginalisation en le réduisant autant que possible afin d'améliorer la qualité de l'estimation de la vraisemblance acoustique. Nous nous sommes appuyés d'une part sur la définition exacte du critère de seuillage du rapport signal-sur-bruit utilisé pour estimer les masques afin de fixer des premières bornes de l'intervalle de marginalisation, puis nous avons exploité l'hypothèse selon laquelle les observations sont gaussiennes pour déduire un intervalle beaucoup plus petit qui a 90 % de chance de contenir la contribution de la parole seule.

5.5 Optimisation du taux d'erreur

Depuis environ deux ans, parallèlement aux divers travaux résumés ci-dessus, je réfléchis à une nouvelle formulation des concepts fondamentaux de la reconnaissance avec données manquantes, qui vise à optimiser directement le taux d'erreur en mots plutôt que de passer par l'intermédiaire du rapport signal-sur-bruit. Je n'ai pas publié cette approche pendant plusieurs années, car j'estime que je n'avais pas le recul nécessaire pour cela. De plus, la seule tentative que j'ai faite en 2007 pour publier cette idée a été plutôt mal accueillie par le relecteur, à mon avis parce qu'elle se démarque trop du courant de réflexion principal du domaine, et sans doute aussi parce que je n'avais pas alors de résultats expérimentaux suffisants. Ces résultats commençant à apparaître, j'ai donc publié pour la première fois ces travaux en 2008. La proposition de masquer les observations en optimisant le taux d'erreur résoud de nombreux problèmes persistant du domaine de la reconnaissance avec données manquantes, dont la dépendance par rapport à la paramétrisation, ou encore le traitement des coefficients dynamiques. Toutefois, la modélisation de tels masques est beaucoup plus difficile que dans le cas classique. Je suis néanmoins finalement parvenu à inférer ces masques à partir des mesures de confiance estimées sur le résultat de la reconnaissance. Cette approche présente donc un autre avantage, qui est de faire le lien entre deux domaines de recherche : la reconnaissance avec données manquantes et l'estimation de mesures de confiance, et par là même propose une piste originale pour résoudre un problème récurrent du dernier domaine, qui est d'exploiter efficacement les mesures de confiance pour corriger les erreurs de la reconnaissance.

6 Reconnaissance de la parole en Intelligence Ambiante

6.1 Analyse du domaine

L'intelligence ambiante (AmI) étant un domaine de recherche beaucoup plus général que la reconnaissance automatique de la parole, elle n'a pas a priori de liens directs avec cette dernière. L'AmI s'appuie sur de nombreuses technologies différentes, comme les réseaux de capteurs, la gestion de l'énergie, les nouveaux terminaux, les réseaux ad hoc, la gestion du contexte, ou les interfaces homme-machine (IHM). Toutefois, l'utilisateur étant au cœur de l'AmI, ces dernières ont une place prépondérante dans le domaine.

L'application directe à ce nouveau domaine des technologies intéractives qui ont été développées depuis de nombreuses années pour les interfaces homme-machine (IHM) pose un cetain nombre de problèmes. En effet, une étude plus approfondie de l'informatique ubiquitaire montre que les besoins et les contraintes sont profondément différents de ceux des IHM classiques, au point q'un nouveau terme est apparu, celui d' "interactions implicites". Il y a donc une réflexion particulière à mener pour appliquer la reconnaissance de la parole à l'AmI, réflexion dans laquelle je me suis engagé depuis 2002, date de début du projet européen OZONE. D'autres chercheurs renommés en reconnaissance automatique de la parole, dont Sadaoki Furui, ont également menés très récemment une réflexion similaire.

J'ai travaillé sur cette question dans le cadre de deux projets européens intégrés, OZONE puis Amigo, ainsi que dans le cadre d'un groupe de travail de l'OFTA, qui a réunit une dizaine de chercheurs de différents domaines pendant 2 ans pour réfléchir au domaine et proposer des orientations futures. J'ai ainsi publié dans un chapitre d'un ouvrage de l'OFTA le fruit de cette analyse, du point de vue des interfaces homme-machine.

6.2 Interactions implicites

Les systèmes d'AmI doivent être déployés *partout*, fonctionner *en permanence*, et aider les utilisateurs autant que possible dans leurs tâches quotidiennes. Ils doivent donc interagir avec les utilisateurs, mais en les dérangeant le moins possible, c'est-à-dire en requérant un minimum d'effort cognitif et d'attention de leur part. Les interactions implicites sont destinées à réaliser ceci, et sont fondées sur une "observation" constante de l'utilisateur, qui permet au système de prévoir quelles sont les actions qui peuvent lui être utiles, sans forcément avoir recours à une interaction explicite avec lui. Ce domaine est également connu sous le nom d' "interfaces attentives".

Dans ce cadre, les interactions implicites par la parole consistent à écouter en permanence l'utilisateur et son environnement, notamment au cours de ses conversations avec d'autres personnes, médiatisées

(téléphone, vidéo-conférence, etc.) ou non. La technologie vocale requise pour cela diffère en plusieurs points importants des approches traditionnelles en reconnaissance automatique de la parole :

- L'objectif n'est (en général) pas de transcrire la parole, mais il est de la comprendre, ou de manière plus réaliste, d'extraire un certain nombre d'informations utiles au système d'AmI.
- Le domaine d'application n'est pas restreint comme dans la plupart des applications interactives, mais il est général et ouvert à tout ce qui peut être dit au cours d'une conversation entre personnes.
- L'utilisateur ne parle pas volontairement au système, ou du moins le destinataire principal de son discours n'est pas la machine : la parole est donc spontanée, et l'utilisateur s'appuyera sur les connaissances générales partagées entre les hommes, et sur le contexte interactif présent et passé qu'il a avec ses interlocuteurs. Toutes ces informations ne sont pas disponibles au système, qui ne pourra résoudre facilement les ellipses, allusions, et métaphores omni-présentes dans nos conversations, en supposant qu'il parvienne déjà à transcrire relativement exactement de la parole spontanée.
- La tâche paraît donc insoluble, mais une contre-partie importante à ces difficultés est que l'utilisateur n'attend rien a priori de particulier du système : il est donc en général préférable de ne pas déranger l'utilisateur en cas de doute sur ses souhaits, et d'attendre d'avoir une confirmation implicite, ou indirecte un peu plus tard.

Notons finalement que la parole n'est qu'une modalité particulière pour les interactions implicites, et j'ai donc travaillé en collaboration avec les équipes Langue&Dialoge, puis TALARIS, à intégrer la reconnaissance de la parole dans des architectures multi-modales, en fonction des contraintes imposées par les plate-formes d'intelligence ambiante, que j'ai prises en compte notamment en concevant des implémentations de faible complexité algorithmique et portables. De plus, une interaction implicite étant par définition discrète vis-à-vis de l'utilisateur, elle doit obligatoirement passer par un module de gestion du contexte, ce que nous avons réalisé en analysant d'une part le contexte environnemental pour décider de la modalité d'interaction la plus appropriée, et en complétant les informations contextuelles globales par les indices déduits de l'écoute des utilisateurs.

Au-delà des mots, j'ai également proposé d'extraire des informations de plus haut niveau du flux de parole, informations qui peuvent ainsi enrichir le module de contexte du système d'AmI, et qui sont décrites dans les deux paragraphes suivants.

Actes de dialogue

Dans le cadre de la thèse de Pavel Král, nous avons travaillé sur la reconnaissance automatique des actes de dialogue à partir d'un signal de parole. Un acte de dialogue distingue par exemple les questions des réponses, les affirmations des ordres, etc. Nous avons développé plusieurs modèles d'actes de dialogue combinant les informations lexicales et prosodiques. Nous avons également proposé des approches permettant de prendre en compte des informations syntaxiques globales, qui se sont révélées très utiles pour l'identification des actes de dialogue. Les applications directes de ces travaux dans le cadre de l'intelligence ambiante sont par exemple de reconnaître une question directe de l'utilisateur pour le système, et donc de discriminer entre interactions explicites et implicites, ou encore d'utiliser les actes de dialogue comme informations de base qui, combinées à d'autres, peuvent permettre de déduire des connaissances sur l'occupation de l'utilisateur, son implication dans le dialogue, son état émotionnel, etc.

Reconnaissance de thèmes

J'ai travaillé également sur la reconnaissance automatique du thème du discours ou des conversations engagées entre plusieurs utilisateurs. La reconnaissance de thème à partir de texte est un domaine de recherche bien connu, et mes contributions principales sont, d'une part, d'avoir développé un modèle bayésien de thème utilisable directement dans une plate-forme d'intelligence ambiante, et surtout d'avoir travaillé sur des méthodes de classification automatique de thèmes directement à partir du flux audio, c'est-à-dire sans passer par un système de reconnaissance de la parole à grand vocabulaire, mais en s'appuyant sur une reconnaissance phonétique suivie d'une méthode d'acquisition automatique du lexique et d'une classification hiérarchique des morphèmes acoustiques extraits en classes sémantiques au moyen d'une distance basée sur la co-occurence. L'intérêt principal de cette recherche est sa capacité d'adaptation

à de nouvelles langues et à de nouvelles tâches à moindre coût, ce qui correspond bien aux contraintes d'autonomie et d'adaptabilité de l'intelligence ambiante.

Encadrements

Post-Doctorat

J'ai accueilli un chercheur en post-doctorat (Dr. Sen Zhang) en 2002, dans le cadre du projet européen Ozone, pendant 18 mois, sur l'intégration de notre système de reconnaissance de la parole dans l'architecture multimodale développée dans ce projet.

J'encadre depuis octobre 2009 un post-doctorant (Dr. Frédéric Tantini) sur l'apprentissage semi-supervisé d'un analyseur syntaxique des transcriptions automatiques de l'oral.

Thèses

J'ai co-encadré avec un grand plaisir deux thésards de l'équipe Parole : Sébastien Demange et Pavel Král, et j'ai commencé à co-encadrer depuis octobre 2008 un nouveau thésard, Christian Gillot, sur l'amélioration de la reconnaissance automatique de la parole par des informations syntaxiques et sémantiques.

Pavel Král a commencé une thèse en co-tutelle en 2003, sous la direction de Yves Laprie et de moi-même à l'Université Henri Poincaré de Nancy, et sous la direction du Pr. Jana Kleckova à l'Université de Bohème de l'Ouest à Plzen, en République Tchèque. Son sujet de recherche concerne la reconnaissance automatique des actes de dialogue en Français et en Tchèque, sujet qui intègre harmonieusement nos compétences en modèles stochastiques à Nancy, et les études en analyse prosodique réalisées à Plzen. Pavel a soutenu sa thèse avec succès en 2007 et a obtenu dans la foulée un poste d'enseignant-chercheur permanent dans son Université tchèque.

Sébastien Demange a débuté sa thèse à l'Université Henri Poincaré de Nancy dans l'équipe Parole en 2004 sous la direction conjointe du Pr. Jean-Paul Haton et de moi-même. Le sujet de recherche que j'ai proposé en 2004 est directement issu de mes travaux sur la reconnaissance de la parole avec données manquantes. Sébastien a soutenu sa thèse avec succès en 2007, et est parti depuis en post-doctorat à l'Université Catholique de Leuven sur un financement européen.

J'ai également encadré, pour un court séjour d'un mois en février 2008, une enseignante de l'Université d'Oran (Algérie) qui était en cours de préparation de sa thèse, sur un sujet se raportant à la construction des masques de données manquantes.

DEA, Master 2

J'ai co-encadré Pavel Král également au cours de son stage de D.E.A dans l'équipe Parole en 2002/2003 sur un sujet proche de la reconnaissance des actes de dialogue.

Licences, maîtrises, Master 1, élèves ingénieurs

J'ai encadré trois stagiaires (Laëtitia Chipot, Fawzi Lahmar et Alex Pandourangam) en maîtrise d'informatique à l'Université Henri Poincaré sur un sujet de normalisation de la longueur du conduit vocal en reconnaissance automatique de la parole en 2002.

J'ai également encadré en 2002 pendant 3 mois un étudiant (Pierre-Alix Dancer) en deuxième année d'ingénieur ENSTA sur la séparation de sources sonores.

J'ai également encadré en 2005 un stagiaire de 1ère année de l'ENSEIHT sur la conception d'interface graphique pour un systeme de reconnaissance de la parole.

J'ai encadré un stagiaire de 2ème année d'ESIAL sur la conception d'un démonstrateur de reconnaissance de la parole en 2006.

Plus récemment, j'ai encadré un autre stagiaire de 2ème année d'ESIAL en 2008 sur le projet ALIGNE du CPER TALC concernant l'alignement semi-automatique texte/parole, à raison d'une après-midi par semaine pendant 4 mois.

J'ai également encadré un stagiaire de fin d'école d'ingénieur ENSI (Tunis) pendant 4 mois sur l'amélioration de la reconnaissance automatique de la parole par la prise en compte d'informations sémantiques en 2008.

J'ai encadré pendant deux mois à compter de juin 2009 un stagiaire de Master 1 sur un sujet de recherche sur la phonétisation automatique des mots inconnus exploitant un classifieur à base d'exemples, afin d'améliorer le logiciel d'alignement JTrans.

J'ai co-encadré également, avec Claire Gardent, pendant les deux mois d'été 2009, 4 stagiaires en licence linguistique pour annoter des corpus de transcriptions orales en dépendances syntaxiques. De même, j'ai également co-encadré pendant deux mois un stagiaire universitaire indien sur la classification automatique des verbes selon leur réalisation syntaxique.

Valorisation et transfert

Projet européen OZONE

OZONE est un projet intégré (IST-2000-30026) financé par la CEE portant sur l'intelligence ambiante. Il est porté par Philips Research Eindhoven, les autres partenaires étant : T-Systems Nova (Allemagne), IMEC (Belgique), Philips Research France, EPICTOID (Pays-Bas), TUE (Pays-Bas), INRIA et Thomson Multimedia. OZONE a débuté en novembre 2001 et s'est terminé avec succès en mai 2004. Le coût total du projet s'élevait à 12,21 millions d'euros.

Mon implication dans le projet concernait aussi bien les aspects scientifiques que la gestion du projet au sein du LORIA. J'étais en particulier responsable d'une tâche du projet concernant la multimodalité. J'assurais également la gestion du workpackage sur les interfaces utilisateurs avec Yves Laprie, qui en était le responsable. Je gérais également les aspects financiers concernant l'implication du LORIA dans le projet, en collaboration avec les SAF du LORIA et de l'INRIA Rocquencourt. J'animais le groupe de travail du LORIA composé des membres du laboratoire impliqués dans le projet, et j'assurais la liaison entre le LORIA et l'INRIA d'une part, et les partenaires extérieurs du projet d'autre part.

Projet européen Amigo

Amigo est un projet intégré (IST 004182) financé par la CEE portant sur l'intelligence ambiante. Il est porté par Philips Research Eindhoven, les autres partenaires étant : Philips Design et Philips Consumer Electronics, Fagor (Espagne), France Telecom, Fraunhofer IMS (Allemagne), Fraunhofer IPSI, Ikerlan (Espagne), INRIA, Italdesign Giugiaro (Italie), Knowledge (Grèce), Microsoft (Allemagne), Telin (Pays-Bas), ICCS (Grèce), Telefonica I+D (Espagne), Université de Paderborn (Allemagne) et VTT (Finlande). Amigo s'est poursuivi de septembre 2004 à mars 2008, pour un coût total de 24,03 millions d'euros.

Dans ce projet, nous collaborons avec l'équipe Langue&Dialogue du LORIA dans la poursuite de nos efforts débutés dans OZONE destinés à développer les interactions mutimodales et les interactions implicites dans les plate-formes d'intelligence ambiante. Tout comme dans le projet OZONE, je gère les aspects financiers concernant le LORIA en collaboration avec les SAF du LORIA et de l'INRIA Rocquencourt, ainsi que l'animation des personnels du LORIA impliqués dans le projet, qui sont des membres des équipes Parole et Langue&Dialogue (aujourd'hui TALARIS). J'assume également la responsabilité d'une sous-tâche du projet concernant les interfaces multimodales, et je représente généralement le LORIA vis-à-vis des autres partenaires du projet.

Projet européen MIAMM

Le projet MIAMM mets en collaboration les équipes PAROLE et LANGUE ET DIALOGUE du LORIA avec entre autre l'Université de Karlsruhe et la société SONY afin de développer des interfaces Homme-machine multi-modales exploitant notamment la reconnaissance de la parole et le retour d'effort. Mon implication dans ce projet a été relativement modeste et a consisté au mois d'octobre 2001 à des activités de conseil en tant que spécialiste en reconnaissance de la parole.

Projet européen HIWIRE

Le projet HIWIRE s'est déroulé de septembre 2004 à 2007. Ma participation dans ce projet est uniquement scientifique : je m'occupe des aspects de recherche et de transferts technologiques concernant la reconnaissance avec données manquantes et les réseaux bayésiens, en collaboration avec Irina Illina, Dominique Fohr et Sébastien Demange.

Campagne d'évaluation Technolangue

J'ai participé à la campagne d'évaluation Technolangue « ESTER », en collaboration avec les autres membres de notre équipe. Dans ce cadre, nous avons développé une plate-forme de reconnaissance automatique de la parole grand vocabulaire dédiée à la transcription automatique d'émissions radiophoniques. J'ai notamment mis en place et géré une grappe de PC, indispensable aux tâches lourdes d'apprentissage requises dans une telle application. Je me suis ainsi occupé de l'apprentissage parallèle des modèles acoustiques contextuels, et je suis également à l'origine de l'utilisation du décodeur grand vocabulaire basé sur le moteur de reconnaissance julius.

Je suis également impliqué dans la deuxième campagne d'évaluation ESTER2 qui a débuté en 2008. J'ai notamment conçu une nouvelle architecture modulaire pour la plate-forme ANTS, et ré-implémenté l'ensemble de nos scripts d'apprentissage sous la forme de composants indépendant, ce qui facilite grandement la maintenance de ces outils et leur réutilisabilité dans d'autres projets. Pour la même raison, et pour faire face au doublement de la taille du corpus d'apprentissage entre les deux campagnes ESTER, j'ai re-programmé un composant JAVA qui facilite le déploiement en parallèle de l'apprentissage et de l'optimisation des paramètres sur un corpus de développement sur un cluster de PC. Je me suis finalement engagé à participer à la campagne d'évaluation fin mars 2009 de la segmentation du flux audio en phrases avec Christian Gillot et Pavel Kràl. Nous avons conçu pour ce faire des méthodes originales exploitant le rythme prosodique combiné avec des informations syntaxiques.

Projet RAPSODIS

Je suis actuellement porteur du projet RAPSODIS qui est une ARC INRIA ayant débuté en 2008 et devant durer deux ans. Cette action a pour thème le calcul d'informations syntaxico-sémantiques afin d'améliorer la reconnaissance automatique de la parole. Il est réalisé en collaboration entre les équipes PAROLE et TALARIS du LORIA, les équipes METISS et TEXMEX de Rennes, et l'équipe LIC2M du CEA-LIST de Paris.

Projet ALIGNE

Je suis actuellement porteur du projet ALIGNE du Contrat Plan-Etat Région (CPER) MISN TALC qui a débuté en 2008 et qui concerne l'alignement texte / parole dans le but d'aider à la création de corpus linguistiques. Ce projet, d'une durée de deux ans a commencé en 2008 et est réalisé en collaboration avec les équipe PAROLE et TALARIS du LORIA, et l'équipe Corpus Oraux du laboratoire ATILF. Plus globalement, je partage la coordination du CPER MISN TALC avec Claire Gardent depuis 2008.

Divers

- J'ai été impliqué dans un certain nombre de propositions de projets qui n'ont finalement pas abouti, comme les réseaux d'excellence AIR@LARGE et AIR@SOFT, la proposition de projet européen Pierre et Marie Curie AmbiCom, une proposition de projet ITC-Asie, et plusieurs propositions nationales ARC et ANR.
- Je suis actuellement en train de préparer une collaboration nationale dans le cadre des ANR blanches, action qui devrait être déposée en 2010.

Enseignement

Mes principaux enseignements et responsabilités associées sont résumés ci-dessous.

- De 1996 à 1999, pendant mes trois années de thèse, j'ai réalisé mon monitorat à l'ESSTIN, où j'ai pris la responsabilité du cours de JAVA et d'UML. J'ai ainsi réalisé 64 heures equivalent TD par an de CM, TD et TP, soit au total 192 h.
- Après mon post-doctorat, pendant l'année scolaire 2000-2001, j'ai obtenu un poste d'ATER à l'UHP, au cours duquel j'ai enseigné les bases de données en Licence et en Maîtrise. J'ai également pris la responsabilité du cours de JAVA et d'UML à l'IUT de St-Dié.

Après l'obtention de mon poste de CR CNRS en septembre 2001, j'ai choisi de poursuivre les enseignements décrits ci-dessous.

- Pendant l'année scolaire 2002/2003, j'ai donné 32 h de TD/TP en DEUG en programmation objet et JAVA.
- Ensuite, pendant l'année scolaire 2003/2004, j'ai donné 37 h de CM, TD et TP en initiation à l'informatique en Maîtrise d'italien et d'espagnol à Nancy II.
- Ensuite, pendant l'année scolaire 2004/2005, j'ai donné 20 h de TD d'algorithme et de JAVA en DESS Chimie.
- Finalement, pendant l'année scolaire 2006/2007, j'ai pris la responsabilité du cours de reconnaissance automatique de la parole en Master recherche, pour un total de 15h de cours magistraux. J'ai donné également une version plus courte de ce cours (6h de CM) en novembre 2009.

Responsabilités et rayonnement

Participation à des jurys

J'ai participé à deux jurys de thèses :

1. Jury de thèse de Sébastien Demange en octobre 2007 ;
2. Jury de thèse de Pavel Král en novembre 2007 ;

J'ai participé à un jury de DEA, celui de Pavel Král en 2003.

Responsabilités

J'ai pris un certain nombre de responsabilités dans le cadre de la direction, de la proposition, de la gestion et du suivi de plusieurs projets européens et plusieurs projets nationaux : ces responsabilités sont résumées dans le paragraphe "Valorisation et transfert", et je ne les rappelle pas ici.

D'autre part, j'ai pris d'autres responsabilités d'administration de la recherche plus "traditionnelles", dont les suivantes :

- J'ai été élu au conseil de laboratoire du LORIA pendant 4 ans.
- J'ai été élu à la commission de spécialiste de l'UHP en 2006, mais malheureusement sans effet, car cette élection a été annulée suite à la loi sur l'autonomie des universités.

Rayonnement

Mon rayonnement scientifique repose essentiellement sur mes publications, dont la liste complète est donnée en Annexe B à la fin de ce dossier, ainsi que sur mon implication dans des projets de recherche, qui sont décrits préalablement. De plus, nous pouvons noter les quelques éléments suivants :

- J'ai été conférencier invité à la conférence internationale Acoustics en mai 2008.
- Je suis relecteur régulier des revues internationales IEEE Trans. on Speech and Audio Processing, Speech Communication, Computer Speech and Language, "Integration, the VLSI Journal", Pattern Recognition Letters, et de plusieurs conférences du domaine.
- J'ai été relecteur pour l'ANR en 2009.
- J'ai été relecteur pour le "Council of Research" de l'Université catholique de Leuven en 2009.
- En 2004, j'ai été invité comme membre du groupe de l'Observatoire Français des Techniques Avancées (OFTA `http://www.ofta.net`) concernant l'intelligence ambiante. Cette association, créée par la Société Amicale des Anciens Elèves de Polytechnique, organise tous les deux ans un groupe

de travail sur un domaine scientifique émergent. Ce groupe rassemble une vingtaine de membres, académiques et industriels, afin d'analyser l'importance et l'évolution future d'un domaine technologique émergent. Cette étude a duré 2 ans, à raison d'une réunion toutes les 6 semaines, à l'issu desquelles un livre de la série Arago est publié, qui résume l'état de l'art du domaine et émet des recommandations sur le domaine étudié à destination des décideurs industriels, politiques et financiers. Outre ma participation aux réunions du groupe, j'ai réalisé dans ce cadre un séminaire sur les interfaces homme-machines en intelligence ambiante et j'ai co-rédigé un chapitre du livre de synthèse.

– J'ai participé au comité de lecture de la conférence MCCSIS'07.

Mémoire : Quelques contributions en reconnaissance automatique de la parole robuste

Chapitre 1

Introduction

1.1 Les hypothèses, sources d'information

En reconnaissance automatique de la parole, l'information fondamentale est celle qui est observée. Elle prend deux formes : d'une part, le signal de parole, et d'autre part, l'utilisation des mots dans la langue. Toute la reconnaissance est basée sur ces deux informations "primaires". Mais l'information acoustique est extrêmement variable, et il n'existe aucune grammaire qui puisse décrire une langue précisément. Il n'est donc pas possible d'établir une relation entre la parole et les mots à partir de ces deux seules observations, et d'autres sources d'information doivent être considérées.

Ces informations supplémentaires sont en fait des hypothèses qui reflètent nos connaissances liées aux phénomènes observés et qui contraignent ainsi l'espace de recherche des relations potentielles entre la parole et les mots.

Mais ces hypothèses peuvent aussi bien aider la reconnaissance que l'induire en erreur. En effet, dans le meilleur des cas, elles reflètent des connaissances que nous avons de bonnes raisons de croire véridiques, comme par exemple l'additivité des signaux acoustiques dans le domaine temporel. Mais nous savons que la plupart d'entre elles ne sont qu'approximatives, mais néanmoins utiles pour simplifier les calculs, comme par exemple l'hypothèse d'additivité des spectres de puissance des signaux acoustiques. Le pire cas, malheureusement très fréquent, est celui des hypothèses que nous savons fausses mais qui sont pourtant adoptées afin de faciliter les dérivations mathématiques, comme la forme conjuguée des densités *a priori* dans les dérivations bayésiennes ou l'indépendance des coefficients statiques et dynamiques (!) dans les vecteurs d'observation acoustiques.

Toute approche en reconnaissance de la parole repose donc sur un empilement d'hypothèses plus ou moins vraisemblables et discutables. Le nombre de méthodes existantes reflète la multiplicité des choix d'hypothèses réalisables. Portons au crédit des articles publiés dans la littérature du domaine le fait que les hypothèses réalisées sont en général relativement bien décrites et motivées. Toutefois, face à la multitude de méthodes proposées depuis plusieurs décennies, il est indispensable de pouvoir comparer ces méthodes entre elles. Cette comparaison est généralement réalisée selon un critère quantitatif, le plus souvent le taux de reconnaissance sur un corpus et une tâche donnée.

Je pense cependant qu'il est au moins aussi important de comparer les hypothèses sous-jacentes aux méthodes que leur taux de reconnaissance. En effet, au vu du nombre d'hypothèses réalisées aussi bien implicitement qu'explicitement, il n'est pas rare que des approches a priori motivées par des considérations très différentes se révèlent finalement semblables, voire identiques. Mais le plus important n'est pas là : chaque hypothèse établit en réalité un compromis entre les contraintes imposées au système de reconnaissance et ses capacités de généralisation à d'autres tâches et d'autres conditions. Les contraintes sont indispensables, car elles guident le choix de la solution dans l'espace quasiment infini des relations possibles. Mais elles limitent en contrepartie les domaines d'application (tâche, type de bruit, locuteurs, etc.) d'une méthode.

Comparer deux approches uniquement par leur taux de reconnaissance n'est donc pas fiable, car la méthode qui donne les meilleurs taux de reconnaissance peut aussi être celle qui est la plus contraignante

en terme d'hypothèses. L'approche alternative peut d'un autre côté compenser la faiblesse de ses performances par des capacités de généralisation qui la rende plus efficace dans d'autres conditions. Cet artifice est malheureusement trop rarement pris en compte dans les publications, ce qui affaiblit d'autant les conclusions souvent tirées hâtivement d'expériences trop spécifiques. La seule manière véritablement convaincante de valider une approche serait de la comparer aux autres dans de multiples conditions expérimentales, ou au moins sur un panel de conditions diverses. Ceci est bien entendu impossible à réaliser par une seule équipe de chercheurs dans des délais raisonnables.

Comment prouver alors qu'une méthode est meilleure qu'une autre, au-delà des strictes conditions expérimentales définies dans un article ? Il n'y a pas de réponse véritablement satisfaisante à court terme à cette question : il faudrait attendre que d'autres équipes de recherche analysent cette méthode dans d'autres conditions, ce qui peut prendre beaucoup de temps. C'est d'ailleurs pour cela que seules des approches relativement anciennes et qui ont fait leur preuve dans de multiples conditions, comme les MFCC ou la normalisation cepstrale, sont unanimement acceptées par la communauté du domaine, alors que d'autres approches plus récentes, même issues de campagnes d'évaluations internationales, comme l'ETSI AFE[1], ne le sont pas (encore).

Lorsqu'une nouvelle approche est proposée, il est impossible pour quelques chercheurs isolés de réaliser des expériences dans des conditions aussi variées. Il est néanmoins possible de mettre en évidence les hypothèses réalisées dans la méthode, et de comparer ces hypothèses avec celles des autres approches de référence afin d'identifier leurs capacités de généralisation respectives. Toute validation expérimentale devrait donc être complétée par une telle analyse, surtout lorsque la robustesse du système de reconnaissance est considérée. En effet, la robustesse est précisément la capacité des algorithmes de reconnaissance automatique de la parole à être efficaces lorsque les conditions d'utilisation diffèrent des conditions dans lesquelles ces méthodes ont été conçues. Retenons néanmoins que la validation expérimentale complétée par une analyse des hypothèses formulées n'est pas une solution entièrement satisfaisante, car elle ne remplacera jamais l'étude expérimentale à grande échelle dans de multiples conditions. Elle permettra toutefois de tempérer à juste titre les conclusions tirées des expérimentations, conclusions qui, par leur forme numérique, revêtent le caractère absolu des mathématiques, surtout lorsqu'elles sont confirmées statistiquement par des intervalles de confiance certains à 99 % ! N'oublions donc pas que toute cette certitude n'est valide que dans le cadre de conditions expérimentales très spécifiques...

1.2 Les informations utiles à la reconnaissance

Les types d'information les plus fréquemment exploités en reconnaissance automatique de la parole sont répertoriés ci-dessous :

- Le signal observé, même bruité, constitue bien entendu la première source d'information.
- Les modèles de parole et les modèles n-gram entraînés sur les bases d'apprentissage constituent la deuxième source d'information la plus importante.
- Des modèles de bruit, indépendant de la parole, peuvent également être entraînés sur des bases d'apprentissage.
- Les modèles et équations décrivant le mélange des signaux de parole et de bruit peuvent être interprétés comme des contraintes dont le but est de restreindre l'espace de recherche du décodage, et les connaissances qui sont à l'origine de ces équations sont donc également des sources de connaissance importantes pour le décodage.
- Les modèles paramétriques avec lesquels nous réalisons le décodage, comme les transformations MLLR ou les GMM modélisant conjointement la parole propre et bruitée dans SPACE, et dont la forme est souvent choisie pour des raisons pratiques de dérivabilité ou de complexité, jouent également un rôle de contraintes limitant l'espace de recherche. Ce sont alors des hypothèses simplificatrices et non plus des connaissances qui sont à l'œuvre, mais elles n'en constituent pas moins des informations qui affectent le décodage.

[1] cf. paragraphe 2.1.2

Voyons maintenant comment ces informations sont utilisées en pratique et dans quelles approches de l'état de l'art. La liste suivante instancie les types d'information généraux décrits ci-dessus et explicite les différents cas d'application de ces informations et les méthodes robustes les plus représentatives des catégories correspondantes. Cette typologie des approches robustes de reconnaissance est loin d'être exhaustive, et certains choix de classification réalisés sont certainement discutables. Toutefois, l'objectif principal est ici d'illustrer la réflexion menée au paragraphe précédent, et de montrer comment l'analyse des hypothèses réalisées peut parfois donner un nouvel éclairage sur des méthodes que nous connaissons tous depuis bien longtemps.

Les méthodes robustes sont simplement listées ici : les plus importantes d'entre elles, ou celles qui sont liées à mes propres travaux, sont décrites dans le reste du mémoire. Pour les autres, j'invite le lecteur à consulter des ouvrages de référence sur l'état de l'art du domaine, comme [Haton *et al.*, 2006].

Signal observé :

Toutes les méthodes exploitent cette information. Si nous supposons que le signal observé représente effectivement le signal de parole, alors la robustesse ne peut être améliorée qu'en focalisant au mieux la paramétrisation sur l'information phonétique présente dans le signal, ce que font toutes les approches de paramétrisations "simples", comme MFCC, Frequency Filtered ou PLP. Si par contre nous remplaçons cette hypothèse d'égalité du signal observé et du signal de parole par une autre hypothèse, moins forte, selon laquelle la parole n'est pas observée directement, alors de nouvelles contraintes doivent être appliquées, parmi lesquelles :

1. *Représentativité du bruit dans un corpus de parole bruitée :*
   Une première catégorie d'approches robustes exploite un corpus de parole bruitée, et fait l'hypothèse que les bruits présents dans ce corpus sont représentatifs des conditions de test, ou tout au moins qu'on peut approcher les seconds par les premiers. Les méthodes de ce type sont l'apprentissage bruité, multi-conditions, la sélection/interpolation/extrapolation de modèles, les voix (ou environnements) propres, le cluster adaptive training.

2. *Modèles bruités paramétriques :*
   Une deuxième catégorie d'approche n'utilise pas de corpus de parole bruitée, mais inclut dans la description des modèles acoustiques des paramètres libres qui permettent d'apprendre des modèles de parole bruitée sur le signal de test. Les approches de ce type sont MLLR, MAP, et leurs déclinaisons. Notons que la forme paramétrique adoptée est une contrainte forte qui constitue une information non négligeable ajoutée dans le système.

3. *Modèle d'environnement :*
   Une autre catégorie d'approches robustes utilise un modèle d'environnement, c'est-à-dire un modèle (ou des hypothèses) qui décrit comment plusieurs sources sonores se combinent. Ces modèles peuvent être de deux types :

   3.1. *Modèle d'environnement exact :*
   Dans cette catégorie, la plus fréquente, le modèle d'environnement décrit exactement la valeur observée en fonction des valeurs estimées du bruit et de la parole. Il peut prendre les formes suivantes :

   3.1.1. *Modèle d'environnement paramétrique :*
   Ce modèle inclut la description de la parole, provenant des modèles acoustiques, et du bruit, généralement sous la forme de paramètres dont la valeur est estimée sur la phrase de test en maximisant la vraisemblance des observations. Les approches de ce type sont par exemple VTS, CDCN, POF, ... La différence entre ces techniques et celles décrites ci-dessus sous l'intitulé "modèles bruités paramétriques" n'est pas évidente : en effet, dans les deux cas, des paramètres sont estimés sur le signal de test afin de rendre compte du bruit, souvent avec le même critère de maximisation de la vraisemblance.

Toutefois, MLLR par exemple, contraint effectivement les modèles bruités à être dérivés linéairement des modèles non bruités, mais elle n'impose pas de relation prédéfinie entre la parole et le bruit comme le fait VTS ; elle n'estime d'ailleurs pas le bruit, et MLLR est donc plus générale en ce sens que VTS. Ceci explique pourquoi MLLR peut traiter aussi bien le bruit que la variabilité due au locuteur. D'un autre côté, les contraintes plus fortes présentes dans VTS la rendent plus efficace (VTS a besoin d'un corpus d'adaptation plus petit que MLLR) pour le traitement du bruit.

3.1.2. Modèle d'environnement non paramétrique :
Le modèle d'environnement peut être une simple équation décrivant la relation entre différentes composantes du signal. Un modèle de bruit, qui représente une composante de l'équation, est alors généralement requis. La distinction porte alors sur la manière d'estimer le bruit :

3.1.2.1. Modèle de bruit paramétrique :
En plus de l'équation décrivant l'environnement, un modèle de bruit, indépendant de la parole, peut être construit sur le signal de test. Un module de détection des segments de parole/non parole est généralement utilisé pour sélectionner les segments de signal sur lesquels apprendre ce modèle. Les méthodes utilisant à la fois un modèle d'environnement et de bruit sont par exemple la soustraction spectrale, les filtres de Wiener, filtres de Kalman ainsi que la paramétrisation ETSI AFE qui s'appuie sur les filtres de Wiener.

3.1.2.2. Modèle de bruit non paramétrique :
Le modèle de bruit peut également être construit sur une base d'apprentissage. L'information "extérieure" importée dans le système est dans ce cas beaucoup plus importante que lorsque le modèle de bruit est entraîné sur le signal de test seul, ce qui justifie à mon avis de bien séparer les deux cas. L'approche la plus célèbre de ce type est PMC. L'adaptation jacobienne en fait également partie, même si les modèles de bruit sont généralement de simples centroïdes entraînés sur les corpus d'apprentissage et de test de la parole.

3.1.3. Modèle d'environnement statistique :
Le modèle d'environnement peut ne pas avoir de forme analytique, mais être plutôt un modèle statistique qui représente la combinaison de la parole et du bruit par une distribution de probabilité. C'est le cas des approches SPLICE ou SPACE.

3.2. Modèle d'environnement incomplet :
Les connaissances sur l'environnement peuvent être beaucoup moins riches, mais aussi moins contraignantes, et simplement donner des indications utiles au processus de reconnaissance sans imposer de valeur précise au signal de parole.

3.2.1. Connaissances sur le nombre de sources sonores :
Cette connaissance est utile surtout en conjonction avec une hypothèse d'indépendance des sources sonores, et l'utilisation de plusieurs microphones, ce qui permet alors de s'appuyer sur une analyse en composantes indépendantes (ICA) pour séparer la contribution de chaque source. Ces approches sont connues sous le nom de "séparation aveugle de sources" (BSS). Le fait d'utiliser plusieurs microphones permet de se passer totalement de contraintes sur la forme du signal. Ces approches sont ainsi également efficaces sur d'autres signaux que la parole.

3.2.2. Domination d'une source sonore :
Le modèle d'environnement peut également supposer qu'une seule des sources sonores domine le signal dans une zone limitée du spectre : dans ce cas, en l'absence d'hypothèses supplémentaires, aucune connaissance sur la valeur du signal de parole dans les zones dominées par le bruit n'est possible. C'est le cas notamment des approches de reconnaissance avec données manquantes.

3.2.3. Observations incertaines :
L'effet de l'environnement sur les observations peut être considéré comme inconnu. Dans ce cas, les observations ne sont plus des scalaires, ou des "observables" au sens de Bayes, mais des variables aléatoires dont la densité de probabilité peut être estimée : c'est ce que font les approches dites de "décodage incertain". La différence principale par rapport aux approches plus traditionnelles est que l'incertitude sur l'observation est transmise au module de décodage, alors qu'elle est résolue dès l'étape de paramétrisation pour le débruitage par exemple.

La première remarque concernant cette typologie concerne sa forme arborescente : c'est la forme "naturelle" qui s'est imposée d'elle-même lors de sa construction. Elle reflète les différents choix d'hypothèses qui peuvent être réalisés, les nœuds frères apparaissant dans un même niveau représentant les hypothèses alternatives modélisant un certain type d'information.

Cette structure arborescente peut parfois clarifier les hypothèses sous-jacentes à différentes méthodes. Toutefois, la comparaison des capacités de généralisation de deux méthodes dans l'arbre n'est pas toujours facile. Par exemple, l'apprentissage multi-conditions suppose que le corpus bruité d'apprentissage est représentatif du bruit de test, ce qui est une hypothèse a priori plus forte que la forme paramétrique utilisée dans l'adaptation MLLR, dont le rôle est également de prendre en compte la variabilité entre la parole d'apprentissage et de test, et qui est estimée sur le signal de test. Toutefois, l'hypothèse de linéarité de cette relation peut se révéler moins réaliste qu'un corpus d'apprentissage bruité bien conçu. Dans d'autres cas, la force des hypothèses[2] est plus facile à comparer. Par exemple, les approches de débruitage supposent que la forme du bruit est connue, ou peut être estimée, ce qui est une hypothèse plus forte que celle réalisée en reconnaissance avec données manquantes, qui suppose seulement que la localisation du bruit dans le plan temps-fréquence, et non sa contribution précise, est estimable. Nous pouvons donc en conclure que cette dernière approche est mieux à même de traiter des bruits ne respectant pas les contraintes de quasi-stationnarité nécessaires à l'estimation du bruit dans une phrase.

1.3 Positionnement du problème

Commençons par définir plus précisément le problème principal abordé dans ce mémoire. En effet, le thème général, qui est celui de l'étude des systèmes de reconnaissance automatique de la parole robustes au bruit, est très vaste, et il faut lui définir des limites raisonnables, en particulier concernant le type de bruit et les sources d'information qui vont jouer un rôle important dans les développements à venir.

Tout d'abord, nous considérons principalement le bruit additif, et en particulier, nous n'aborderons quasiment pas le bruit convolutif, ni la variabilité du signal provenant du locuteur lui-même (effet Lombard, voix extrêmes ou pathologiques, etc.), ni le phénomène de réverbération, ni le bruit introduit par un canal de communication artificiel (GSM, codage en mp3, etc.). De même, nous supposons qu'un unique microphone est disponible, et nous n'aborderons donc pas en particulier les techniques de débruitage basées sur les matrices de microphones, ou les méthodes utilisant des données visuelles.

[2] ou de manière équivalente, leurs contraintes

1.4 Les approches en reconnaissance robuste

La très grande majorité, si ce n'est la totalité, des systèmes de reconnaissance automatique de la parole comparent le signal acoustique avec des modèles acoustiques. Le bruit affecte le signal acoustique, et peut altérer la correspondance qui existe entre ce signal et le modèle qui le génère. Un système de reconnaissance automatique de la parole est robuste au bruit s'il parvient à retrouver la bonne séquence de modèles malgré le bruit.

Pour augmenter la robustesse au bruit, trois approches sont envisageables, et peuvent bien sûr être combinées :

- Modifier le signal afin de rétablir la correspondance avec les modèles ;
- Modifier les modèles ;
- Utiliser d'autres sources d'information, par exemple contextuelle, pour aider le processus de décodage.

Les deux premières approches ont été très largement étudiées au cours de ces deux dernières décennies, et la majeure partie des travaux présentés dans ce mémoire prennent également place dans ces deux catégories.

La troisième approche n'a encore été que très peu explorée, au moins d'un point de vue général. En effet, de nombreuses applications particulières, comme la réservation hôtelière, exploitent des grammaires et vocabulaires adaptés à la tâche, très contraints, qui peuvent être considérés comme des sources d'information contextuelle augmentant, parfois considérablement, la robustesse de la reconnaissance au bruit. Mais ces sources d'information sont généralement utilisées de manière ad hoc, et pour des cas particuliers comme celui de la réservation. Je pense qu'une meilleure prise en compte de ces informations contextuelles, évoluant progressivement des cas particuliers au cas général, constitue le meilleur moyen de repousser les limites, à mon avis fondamentales, sur lesquelles se heurtent aujourd'hui les systèmes de reconnaissance automatique de la parole. Selon cette vision, de multiples sources d'information doivent être utilisées conjointement pour retranscrire un message oral, et la contribution respective de chacune d'entre elles dépend du contexte. Ainsi, lorsque le bruit est très fort, le signal acoustique devient moins prépondérant, au profit d'autres connaissances, comme la sémantique.

Le chapitre 2 traite donc des méthodes qui modifient le signal de parole afin de rendre les paramètres acoustiques plus robustes. Successivement, les approches de type paramétrisation robuste, débruitage et finalement analyse de scènes auditives y sont présentées. Cette dernière approche est abordée car, bien que je n'ai pas vraiment proposé de nouvelles contributions dans ce domaine, j'ai néanmoins étudié l'analyse de scènes auditives en tant qu'étage de prétraitement aux méthodes de reconnaissance avec données manquantes. Or, il s'agit bien d'une technique de séparation de la contribution de la parole du bruit, et donc d'une méthode de débruitage qui a déjà été intégrée dans des systèmes de reconnaissance automatique de la parole. Le chapitre 2 termine par un résumé de mes contributions principales dans ce domaine.

Le chapitre 3 présente les approches de modification des modèles acoustiques. Il s'agit donc essentiellement des méthodes d'adaptation. J'y présente en particulier mes travaux sur les modèles multi-échelles pour l'adaptation, les modèles multi-bandes et l'adaptation jacobienne. Comme pour le chapitre 2, le chapitre 3 termine par un résumé de mes contributions principales, ce qui permet d'isoler plus clairement ce qui relève de mes travaux et ce qui provient de l'état de l'art.

Le chapitre 4 présente les approches de reconnaissance avec données manquantes. Après une brève description des principes fondamentaux et des problèmes du domaine, je décris plus précisément mes travaux passés et présents sur la généralisation de ces approches à des domaines de paramétrisation quelconques, ainsi que les travaux que j'ai menés avec Sébastien Demange pendant sa thèse. Le paragraphe 4.7 décrit alors mes travaux récents concernant l'intégration de la reconnaissance avec données manquantes et les mesures de confiance. Le paragraphe 4.8 termine en résumant également mes contributions principales du domaine.

Le chapitre 5 se distingue des précédents, car il ouvre le domaine de la reconnaissance automatique de la parole robuste à d'autres sources d'information que celles considérées jusqu'au chapitre 4. En particulier, la notion d'interaction implicite y est présentée : mon intérêt pour ce concept est issu de mes travaux sur l'application des techniques de reconnaissance de la parole aux plates-formes d'intelligence ambiante, qui constituent à mon avis un cadre idéal pour intégrer de nouvelles sources de connaissance aux systèmes

d'interaction vocale et multimodale. Je développe plus spécifiquement dans ce chapitre deux aspects de mes recherches liées à l'extraction d'information de haut niveau : les actes de dialogue, recherche menée conjointement avec Pavel Kral dans le cadre de sa thèse, et l'extraction d'information sémantique de la parole. Je ne réalise pas dans ce chapitre d'état de l'art, car, dans la mesure où j'y aborde des domaines aussi divers que l'intelligence ambiante, la structure du dialogue, la syntaxe et la sémantique des langues naturelles, un état de l'art, même réduit à son minimum, prendrait beaucoup trop de place par rapport au reste du mémoire, d'autant plus que ces notions y sont abordées essentiellement en tant que perspectives. Le chapitre 5 présentant exclusivement mes contributions, ne se termine pas par un rappel de celles-ci. Le chapitre 6 conclut le mémoire.

Une dernière remarque avant d'aborder le cœur du sujet : quelques résultats expérimentaux sont parfois donnés pour certaines de mes contributions afin d'illustrer les notions théoriques correspondantes. Ces résultats sont relativement peu nombreux, et ils ne constituent bien entendu qu'une petite fraction des expériences que j'ai pu menées au cours de ces dernières années et qui sont détaillées dans mes publications. Toutefois, afin d'éviter de répéter à travers le mémoire les caractéristiques des corpus et des systèmes de reconnaissance utilisés, les diverses conditions expérimentales que j'ai utilisées dans le mémoire sont toutes décrites et rassemblées dans l'annexe A.

Chapitre 2

La modification du signal

Le rapport signal-sur-bruit (SNR) est utilisé pour mesurer le niveau de bruit dans un signal. Théoriquement, il est défini par le rapport de la puissance du signal sur la puissance du bruit :

$$SNR[dB] = 10 \log_{10} \left(\frac{P_{signal}}{P_{bruit}} \right)$$

mais en pratique, il existe plusieurs manières de calculer ces puissances, selon qu'elles sont estimées globalement sur un corpus, une phrase, avec ou sans les segments de silence, etc. Le SNR le plus souvent utilisé est le SNR segmental, qui utilise une fenêtre glissante sur laquelle est calculée la puissance du signal et dont seule la valeur maximale (ou au niveau d'un quantile prédéfini) est retenue. Toutefois, ces valeurs ne sont que des approximations globales du SNR, qui en réalité varie constamment d'un instant à l'autre. On peut également essayer d'estimer le SNR local, à chaque instant et/ou pour chaque bande de fréquences, mais en l'absence de la donnée du bruit seul, une telle estimation est extrêmement difficile à réaliser et revient à résoudre le problème général de la robustesse. De nombreuses techniques intermédiaires estimant le SNR et ses modifications dynamiquement ont donc été élaborées, dont un bon exemple est donné dans [Plapous *et al.*, 2006].

L'amélioration du SNR global permet d'évaluer les performances d'un algorithme de débruitage, mais selon la finalité du système, il est plus intéressant d'évaluer soit l'amélioration du taux de reconnaissance automatique, soit l'amélioration de l'intelligibilité de la parole par des auditeurs humains.

Les approches robustes qui modifient le signal sont principalement les méthodes de :

- Paramétrisation
- Débruitage
- Analyse de scènes auditives et séparation de sources

2.1 Paramétrisation

La paramétrisation intervient dans tous les systèmes de reconnaissance de la parole, mais elle peut être plus ou moins robuste au bruit.

La paramétrisation MFCC (*Mel Feature Cepstral Coefficient*) est de loin la plus utilisée, aussi bien en reconnaissance automatique de la parole qu'en reconnaissance du locuteur ou de la langue. Cette préférence s'explique d'abord par ses performances appréciables dans le bruit : les coefficients MFCC sont réputés plus robustes que les coefficients spectraux, par exemple. Ceci s'explique en grande partie par le fait que les coefficients cepstraux sont décorrélés, et donc en particulier que les variations d'énergie n'affectent qu'un seul coefficient et que les matrices de covariance dans les modèles statistiques sont diagonales, ce qui permet de réduire de beaucoup le nombre de paramètres et prévient le risque que les modèles apprennent trop de détails du signal qui pourraient les rendre sensibles au bruit. Les dérivées des MFCC sont également plus robustes que les coefficients statiques, ce qui est lié au filtrage temporel qu'elles introduisent [Yang *et al.*, 2007]. Les paramètres cepstraux LPC (*Linear Predictive Coding*) sont assez semblables aux MFCC et obtiennent des résultats similaires.

Certains filtres temporels, comme l'analyse RASTA [Hermansky *et al.*, 1991], ou encore d'autres filtres temporels qui sont réalisés dans le domaine de l'autocorrélation et en association avec des coefficients MFCC Yuo99, sont parfois utilisés pour réduire encore la contribution du bruit dans le signal. L'analyse RASTA réduit l'influence du bruit à variation temporelle lente en filtrant les coefficients log-spectraux avec une durée d'intégration de l'ordre de 200 ms [Hermansky et Morgan, 1994]. Elle est souvent combinée aux paramètres PLP, donnant naissance à la paramétrisation RASTA-PLP, mais elle peut également être réalisée dans le domaine cepstral [Haeb-Umbach *et al.*, 1995], ou combinée avec un filtre de Kalman pour améliorer ses performances [Ma *et al.*, 2006]. Le filtre de Kalman est d'ailleurs souvent utilisé, seul ou combiné avec d'autres méthodes et algorithmes, pour débruiter le signal de parole [Kim *et al.*, 2005a].

Les techniques de normalisation (de la moyenne et de la variance du cepstre, de son histogramme [Hilger et Ney, 2006], de la longueur du conduit vocal [Afify et Siohan, 2007], ...) sont très fréquemment utilisées pour réduire la variabilité du signal, qui peut être attribuée pour l'essentiel au bruit. Mais de telles normalisations ont aussi leurs limites, car elles ignorent quelle part de variabilité est précisément due au bruit ou à la parole.

Un exemple simple, mais instructif, est celui de la normalisation de la moyenne du cepstre : celle-ci est généralement réalisée sur toute la phrase, segments sans parole inclus. Le coefficient $c0$, qui encode l'énergie du signal, est représenté sur la figure 2.1 pour un même segment de parole suivi par un silence plus ou moins long. Nous pouvons voir sur cette figure que la paramétrisation avec normalisation donne des

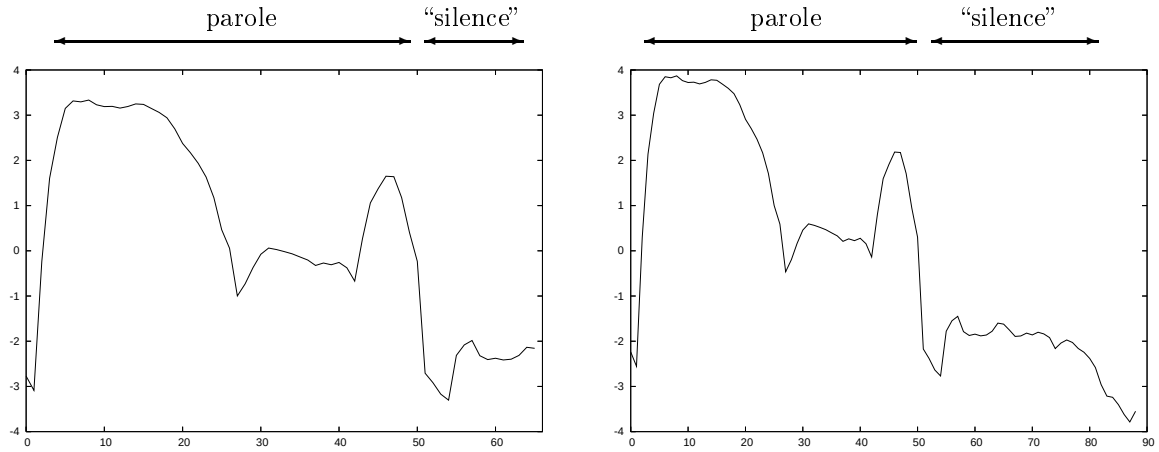


FIG. 2.1 – Effet de la normalisation cepstrale sur un même signal de parole, avec plus ou moins de silence.

valeurs significativement différentes, bien que le signal de parole soit strictement identique. Ces différences sont dues au silence qui perturbe le calcul de la moyenne du cepstre. Cette limitation se généralise dans une moindre mesure aux différents phonèmes qui constituent les phrases, et à leur distribution d'une phrase à l'autre.

2.1.1 Paramètres fréquentiels filtrés et ondelettes

Les paramètres fréquentiels filtrés (en anglais *Frequency Filtered Features*) ont été proposés afin d'obtenir un compromis unifiant les avantages respectifs du spectre et du cepstre, à savoir l'interprétation spectrale des coefficients et leur décorrélation [Nadeu *et al.*, 1995]. Ces coefficients sont calculés en filtrant les énergies fréquentielles contenues dans les filtres triangulaires traditionnellement utilisés dans les systèmes de reconnaissance, au moyen d'un filtre du second ordre, par exemple :

$$H(z) = z - z^{-1} \tag{2.1}$$

Cette paramétrisation permet d'obtenir des taux de reconnaissance comparables à ceux obtenus avec des coefficients MFCC, voire parfois meilleurs [Nadeu *et al.*, 2001]. Ces paramètres ont également été

utilisés pour reconnaître des signaux de parole échantillonnés à une fréquence différente de celle d'apprentissage [Bauerecker *et al.*, 2003] ou encore combinés avec l'adaptation jacobienne pour augmenter la robustesse des systèmes de reconnaissance au bruit [Abad *et al.*, 2003].

Murat Deviren a démontré dans sa thèse que les paramètres fréquentiels filtrés peuvent avantageusement être calculés au moyen d'une décomposition en ondelettes [Deviren, 2004]. En s'appuyant sur l'analogie entre les paramètres fréquentiels filtrés et l'ondelette de Haar, il a ainsi proposé d'améliorer la paramétrisation en déployant plusieurs échelles d'analyse, comme cela est réalisé dans les analyses par ondelettes. Cette nouvelle paramétrisation en ondelettes a été comparée favorablement avec les paramètres fréquentiels filtrés sur le corpus bruité Aurora3, mais sans pouvoir mettre en évidence un avantage consistant par rapport aux MFCC.

De nombreuses autres paramétrisations robustes basées sur les ondelettes ont été proposées : selon les cas, les coefficients en ondelettes sont utilisés directement en remplacement des coefficients MFCC [Kryze *et al.*, 1999], après avoir subies un débruitage [Gupta et Gilbert, 2001, Ghanbaria et Karimi-Mollaei, 200 pour calculer l'énergie dans les sous-bandes de l'échelle Mel [Sarikaya et Hansen, 2000, Kim *et al.*, 2000], pour approcher l'échelle Mel [Sarikaya et Hansen, 2000], ou encore en remplacement de la transformée en cosinus discrète [McCourt *et al.*, 2000, Gowdy et Tufekci, 2000]. Plus récemment, les auteurs de [Kotnik et Kačič, 2007] combinent une décomposition en ondelettes différenciée pour les segments voisés et non voisés avec des paramètres calculés par autocorrélation, qui présente des performances comparables avec celles de la paramétrisation ETSI AFE (décrite au paragraphe suivant).

2.1.2 ETSI Advanced Front-End

L'organisme de standardisation ETSI a conduit ces dernières années trois campagnes d'évaluation sur les bases de données Aurora2, 3 et 4, afin de définir une paramétrisation standard robuste au bruit pour la reconnaissance automatique de la parole. Le résultat de ces évaluations est la paramétrisation aujourd'hui connue sous le nom de *ETSI Advanced Front-End*, ou encore WI008, dont l'algorithme est le suivant :

Paramétrisation standard issue des évaluations Aurora

- Réduction du bruit : La première étape de l'algorithme débruite le signal en deux étages :
  1. Etage 1 : Une détection parole / non parole est réalisée, en se basant sur une estimée de l'énergie du bruit et de la parole. Ces estimées sont mises à jour à chaque trame au moyen d'un facteur d'oubli qui favorise les trames les plus récentes. La détection parole / non parole utilise alors un seuil sur le rapport signal-sur-bruit estimé.

     Les paramètres optimaux d'un filtre de Wiener sont ensuite calculés à partir du spectre de puissance du signal et d'une estimée du spectre de puissance du bruit, également calculée avec un facteur d'oubli. Ces paramètres sont alors transformés du spectre linéaire vers le spectre Mel, puis vers le domaine temporel. Le signal temporel est finalement filtré par le filtre de Wiener précédemment calculé.
  2. Etage 2 : Le deuxième étage du module de paramétrisation prend en entrée le signal débruité par le premier étage. Il réalise alors un second débruitage, qui suit le même schéma que le premier étage, à la différence que le filtre de Wiener est pondéré par un facteur dépendant du SNR estimé pour chaque trame : l'objectif est de débruiter plus agressivement les trames dominées par le bruit, et beaucoup moins les trames de parole. A la fin du deuxième étage, le biais du signal temporel débruité est enlevé afin de recentrer le signal sur zéro.
- Une fois débruité, le contour énergétique lissé du signal temporel est calculé au moyen de l'opérateur de Teager puis d'un filtre à réponse impulsionnelle finie (FIR). L'opérateur de Teager est un opérateur non linéaire pour calculer le contour énergétique d'un signal Teager. Dans le cas d'un signal discret, il vaut : $\Psi[x(n)] = x^2(n) - x(n+1)x(n-1)$. Ce contour sert à calculer les pics énergétiques correspondant à la fréquence fondamentale. La connaissance du nombre et de la position de ces

pics permet ensuite de pondérer chaque trame de parole en favorisant les trames proches des pics énergétiques.
- Des coefficients MFCC sont ensuite calculés à partir de ce signal.
- La dernière étape de paramétrisation est une égalisation aveugle, qui prend comme référence le cepstre du spectre plat.
- Ces paramètres sont ensuite codés, puis transmis au serveur, qui décode et restaure les paramètres du signal. Comme dans de nombreuses applications de téléphonie, un détecteur de parole / non parole est utilisé pour marquer les zones de non parole. Ce détecteur exploite l'accélération de l'énergie associée à l'attaque de la voix. Cette caractéristique est plus robuste que l'énergie ou que les harmoniques en présence de bruit.

Récemment, les auteurs de [Ishizuka *et al.*, 2007] ont montré qu'ils obtenaient de meilleures performances que l'ETSI AFE sur un corpus japonais semblable à Aurora2 en combinant la paramétrisation SPADE [Ishizuka *et al.*, 2006] avec un détecteur parole/non parole très précis basé sur l'analyse des composantes périodiques et apériodiques du signal. La paramétrisation SPADE utilise également une décomposition du signal en parties harmoniques/non harmonique. Bien que les résultats obtenus par cette nouvelle paramétrisation soient impressionnants, il convient de les considérer avec la même retenue que nous devons avoir en analysant l'ETSI AFE, ainsi qu'avec la majorité des autres méthodes de paramétrisation proposées dans la littérature, comme par exemple [Li *et al.*, 2007], qui exhibe également des résultats surpassant l'ETSI AFE sur un corpus dédié à la parole bruitée dans les voitures. En effet, chaque paramétrisation est conçue par rapport à un corpus particulier, comme Aurora2 3 et 4 pour ETSI AFE, et Aurora-2J pour SPADE. Il semble que nombre de ces paramétrisations soient très performantes sur "leur" corpus, et beaucoup moins sur d'autres. Notons à ce propos que la paramétrisation ETSI, qui était sortie vainqueur de la campagne d'évaluation sur Aurora2, a été ensuite à nouveau battue sur Aurora2 par son concurrent précédemment déchu, le consortium Qualcomm-ICSI-OGI [Adami *et al.*, 2002].

2.1.3 Brève conclusion sur la paramétrisation

Il est particulièrement difficile de comparer les paramétrisations entre elles car il faudrait les tester sur un échantillon aussi vaste que possible de tâches et de corpus divers et variés, ce qui peut par ailleurs expliquer pourquoi les MFCC, qui ont déjà fait leurs preuves dans toutes ces conditions imaginables, restent aussi difficiles à détrôner.

Néanmoins, l'étage de paramétrisation étant le premier de la chaîne de calcul, les performances de tout le système de reconnaissance reposent entièrement sur lui, ce qui explique que de nombreuses nouvelles paramétrisations continuent à être proposées en permanence, qui améliorent par exemple l'estimation spectrale [Dharanipragada *et al.*, 2007] ou encore qui proposent de nouveaux domaines de paramétrisation [Morgan *et al.*, 2005]. Les limites des méthodes de paramétrisation robustes dépendent en grande partie de l'absence de connaissance précise des caractéristiques des constituants du signal, et de nouveaux modèles ont donc été considérés, qui représentent cette différence entre la parole et le bruit : ce sont les méthodes de débruitage.

2.2 Débruitage

La différence entre paramétrisation robuste et débruitage n'est pas toujours évidente. Dans la classification proposée ici, les paramétrisations robustes ne considèrent pas le bruit en tant que tel, tandis que les méthodes de débruitage analysent le bruit et le prennent en compte comme une composante du signal à part entière. Ainsi, le filtrage temporel du signal ou la normalisation du cepstre sont considérés comme des techniques de paramétrisation, tandis que la soustraction spectrale relève du débruitage. Avant de présenter les principales approches de débruitage, je commence dans la suite par discuter des liens entre débruitage et adaptation, afin de mieux définir leurs caractéristiques respectives. Cette discussion introduit également la notion de modèles "pseudo-propres", notion qui est ensuite utilisée dans les approches

SPLICE et SPACE.

La première méthode de débruitage présentée est la soustraction spectrale, car ses principes et hypothèses de base sont réutilisés dans presque toutes les autres techniques de débruitage, puis je présente rapidement les autres approches principales de débruitage, en terminant par mes contributions.

2.2.1 De la dualité en débruitage et adaptation

Les approches de débruitage et d'adaptation sont souvent considérées comme duales, car d'un côté, il s'agit de supprimer le bruit du signal, et de l'autre, d'ajouter du bruit dans les modèles. Mais les relations entre ces deux classes d'approches sont plus subtiles, et il est possible de considérer que les approches de débruitage forment en fait un cas particulier des méthodes d'adaptation. Cette relation de subsomption peut se démontrer de la manière suivante : d'un côté, il est possible de reproduire à l'identique une approche de débruitage dans le cadre de l'adaptation, en translatant les moyennes de toutes les gaussiennes par le biais de débruitage, et en reproduisant cette adaptation dès que le biais est ré-estimé. Bien sûr, cet algorithme n'est pas efficace, en particulier lorsque le biais est ré-estimé à chaque trame, mais ceci est très rare en pratique, et cet exercice prouve au moins théoriquement que les approches de débruitage forment un sous-ensemble des méthodes d'adaptation. Il reste à démontrer que certains algorithmes d'adaptation ne peuvent pas être simulés par du débruitage, ce qui est vrai par exemple pour toutes les techniques d'adaptation qui utilisent une transformation différente pour plusieurs gaussiennes. En effet, dans ce cas, la topologie (ou répartition relative) des modèles dans l'espace acoustique a changé, ce qui ne peut se réduire à un simple biais du vecteur observé[3].

La combinaison des approches de débruitage et d'adaptation n'est donc pas forcément intéressante, car nous venons de voir que ces deux approches ne sont pas complémentaires. Toutefois, elle peut être intéressante pour des raisons purement algorithmiques, l'adaptation étant généralement plus coûteuse que le débruitage.

Cette combinaison peut être réalisée par exemple "en cascade", c'est-à-dire en débruitant tout d'abord les corpus d'adaptation et de test, puis en adaptant les modèles. Mais une telle combinaison débruitage/adaptation n'est que rarement réalisée en pratique, car le débruitage peut introduire de nouvelles distorsions non linéaires, dont l'exemple le plus connu est le célèbre bruit musical de la soustraction spectrale, qui s'accommode très mal des hypothèses de linéarité souvent requises par l'étage d'adaptation.

La combinaison débruitage/apprentissage est par contre beaucoup plus efficace, car elle s'affranchit des transformations linéaires de l'adaptation : ce sont les techniques d'apprentissage de modèles "pseudo-propres" . Ce procédé est basé sur l'addition de bruit, généralement artificiel, au corpus d'apprentissage, suivi d'une phase de débruitage de ce corpus, puis de l'apprentissage de modèles acoustiques "pseudo-propres" sur ce corpus débruité. Ensuite, le corpus de test est débruité à son tour, puis reconnu par les modèles pseudo-propres. Cette technique est notamment utilisée dans les approches SPLICE et SPACE décrites dans ce mémoire, et est représentée schématiquement sur la figure 2.2.

Ceci est généralement plus efficace que de débruiter simplement, car le débruitage est toujours imparfait : il s'appuie notamment sur une estimation erronée du bruit, la fonction de débruitage utilisée est paramétrique avec des valeurs estimées des paramètres, la couverture de cette fonction dans l'espace des fonctions possibles est limitée, et l'ajout de bruit étant une application non bijective qui augmente la confusion entre les modèles acoustiques, il est théoriquement très difficile de construire sa fonction inverse exacte. L'espace débruité est donc différent de l'espace propre, et réapprendre les modèles acoustiques dans l'espace débruité d'apprentissage permet de prendre en compte au moins une partie de ces imperfections dans les modèles.

2.2.2 Débruitage par filtrage linéaire

La soustraction spectrale est l'une des techniques de débruitage parmi les plus simples et les plus connues. La combinaison des signaux de parole et de bruit est linéaire dans le domaine temporel :

$$y(t) = x(t) + n(t) \tag{2.2}$$

[3] Sauf à débruiter différemment sur chaque chemin de décodage, autrement dit à intégrer le débruitage dans le décodage, auquel cas il ne s'agit plus vraiment de débruitage !

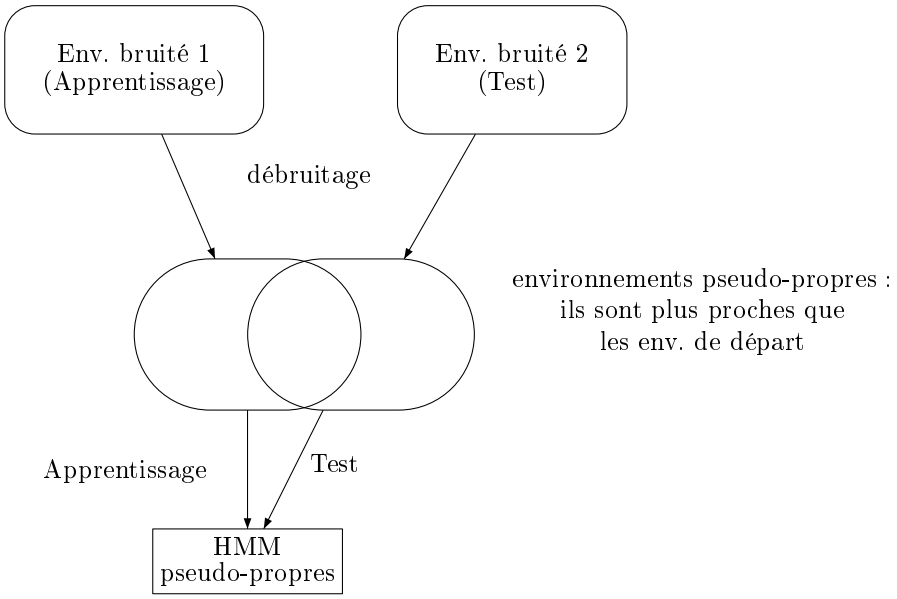


FIG. 2.2 – Schéma de principe de l'apprentissage de modèles acoustiques pseudo-propres.

où $y(t)$, $x(t)$ et $n(t)$ représentent respectivement l'amplitude de la parole bruitée, de la parole seule, et du bruit seul, à l'instant t. Le signal est ensuite fenêtré, puis transformé par une transformation de Fourier. La relation d'additivité est préservée dans le domaine spectral :

$$Y(\tau) = X(\tau) + N(\tau) \tag{2.3}$$

où $Y(\tau)$, $X(\tau)$ et $N(\tau)$ représentent respectivement les spectres complexes de la parole bruitée, de la parole seule, et du bruit seul, sur la fenêtre temporelle indicée par τ. Toutefois, cette relation d'additivité n'est plus préservée lors du passage au spectre de puissance. Néanmoins, l'hypothèse suivante est généralement admise pour simplifier les calculs :

$$|Y(\tau)| \simeq |X(\tau)| + |N(\tau)| \tag{2.4}$$

Cette dernière équation est utilisée pour la soustraction spectrale, mais également pour d'autres méthodes de débruitage spectral.

La soustraction spectrale revient donc à calculer une estimée du spectre de puissance du bruit $|\hat{N}(\tau)|$, puis à soustraire cette estimée du spectre de puissance du signal observé [Boll, 1979] :

$$|\hat{X}(\tau)| = |Y(\tau)| - |\hat{N}(\tau)| \tag{2.5}$$

Le spectre débruité $\hat{X}(\tau)$ peut alors être utilisé, soit pour construire des paramètres MFCC qui seront fournis à un système de reconnaissance automatique de la parole, soit pour reconstruire le signal temporel de la parole débruitée.

Une autre technique de débruitage très connue par filtrage linéaire est le filtre de Wiener [Wiener, 1949]. En supposant que les trames de parole et de bruit sont issues de processus gaussiens stationnaires de densités spectrales de puissances respectives $\sigma_X^2(f)$ et $\sigma_N^2(f)$, il est possible de dériver un estimateur $\hat{X}(t,f)$ du signal de parole seul, optimal au sens des moindres carrés, c'est-à-dire qui minimise :

$$E\left[(\hat{x}(t) - x(t))^2\right] \tag{2.6}$$

Cet estimateur prend la forme :

$$\hat{X}(t,f) = \frac{\sigma_X^2(f)}{\sigma_X^2(f) + \sigma_N^2(f)} Y(t,f) \tag{2.7}$$

où $Y(t,f)$ est le spectre de puissance observé à l'instant t et à la fréquence f.

Souvent, la densité spectrale de parole est estimée à partir du signal observé et d'une estimée du bruit :

$$\hat{X}(t,f) = \frac{\sigma_Y^2(f) - \sigma_N^2(f)}{\sigma_Y^2(f)} Y(t,f) \tag{2.8}$$

Le terme multiplicatif définit le filtre de Wiener, qui pondère les fréquences observées par le rapport signal-sur-bruit local à chaque instant t et fréquence f. Comme pour la soustraction spectrale, la principale difficulté est d'estimer $\sigma_N^2(f)$, ainsi que la variation du bruit au cours du temps lorsque celui-ci n'est pas stationnaire.

De nombreuses améliorations de ces deux approches de débruitage sont encore proposées aujourd'hui, comme [Chen *et al.*, 2006], qui quantifie le rapport entre les augmentations du SNR et de la distorsion de la parole à la sortie du filtre de Wiener, et qui propose plusieurs approches permettant d'améliorer ce rapport, dont l'utilisation de l'information d'autocorrélation du signal de parole, un réglage sous-optimal du filtre de Wiener et l'exploitation de plusieurs microphones. Parmi les autres articles récents sur ce sujet, nous retiendrons [Händel, 2007], qui analyse et compare d'un point de vue théorique et expérimental la soustraction spectrale et le filtre de Wiener.

2.2.3 Débruitage paramétrique non homogène

Le débruitage paramétrique non homogène utilise plusieurs transformations différentes pour différentes zones de l'espace acoustique. Ceci est accompli généralement en quantifiant l'espace acoustique de la parole, chaque classe acoustique étant mise en relation avec une classe correspondante dans l'espace acoustique bruité. Cette relation peut être apprise pendant la phase de construction du système en utilisant des bases de données dites « stéréo », c'est-à-dire possédant le même signal de parole bruitée et non bruitée. Les méthodes les plus connues sont :

- CDCN, ou *Codeword Dependent Cepstral Normalisation* [Acero, 1990], qui construit tout d'abord une quantification vectorielle de l'espace acoustique, puis apprend un biais correctif pour transformer chaque représentant bruité en son équivalent non bruité. Lors du test, une estimée de la parole non bruitée $\hat{x}_t$ est calculée à partir de chaque observation bruitée y_t par une combinaison linéaire des biais correctifs $b(i)$:

$$\hat{x}_t = y_t - \hat{h} - \sum_{i=1}^{I} f(i) b(i) \tag{2.9}$$

  $f(i)$ représente la probabilité *a posteriori* que l'observation soit générée par la classe i et $\hat{h}$ est une estimée du bruit convolutif constant. Les paramètres de la transformation sont appris itérativement par l'algorithme EM sur le corpus de test.
- SCDN (*SNR-dependent cepstral normalisation*) [Acero et Stern, 1991], une extension de CDCN, qui, pour éviter l'usage de EM pendant le test, apprend préalablement les vecteurs de correction sur une base de données bruitées puis estime seulement le rapport signal-sur-bruit lors du test.
- De la même manière que CDCN, l'algorithme POF [Neumeyer et Weintraub, 1994] (*Probabilistic Optimal Filtering*) construit une transformation linéaire paramétrée par $(A_{i,j}, b_i)$ entre des classes des espaces acoustiques bruité et non bruité. Toutefois, des modèles GMM g_i sont utilisés pour représenter ces transformations et la parole non bruitée s'estime de la manière suivante :

$$\hat{x}_t = \sum_{i=1}^{I} \left(W_i^T P(g_i|z_t) \right) Y_t \tag{2.10}$$

$$W_i^T = [A_{i,-p} \cdots A_{i,0} \cdots A_{i,p}\ b_i] \tag{2.11}$$

$$Y_t^T = [y_{t-p}^T \cdots y_t^T \cdots y_{t+p}^T\ 1] \tag{2.12}$$

  où z_n est un vecteur de conditionnement qui peut contenir plusieurs observations successives, voire des indices plus élaborés comme le rapport signal-sur-bruit.

2.2.4 SPLICE

SPLICE est une méthode de débruitage du signal basée sur un formalisme bayésien [Deng *et al.*, 2004]. C'est également une technique de débruitage paramétrique non homogène, mais je la décris plus en détail car elle s'apparente à la méthode SPACE que j'ai proposée avec Khalid Daoudi et qui est décrite ci-dessous. Le principe de SPLICE revient à estimer la probabilité *a posteriori* $p(x|y,n)$ du signal propre x connaissant le signal bruité y et une estimée du bruit n. Cette probabilité *a posteriori* est calculée à partir de la vraisemblance $p(y|x,n)$ et des probabilités *a priori* $p(x)$ et $p(n)$. $p(x)$ représente la parole non bruitée, et est estimée dans le domaine log-spectral par un mélange de distributions gaussiennes qui modélise les propriétés statiques et dynamiques (au moyen des dérivées premières des vecteurs statiques) du signal. $p(n)$ est une distribution de Dirac (non paramétrique) issue d'un algorithme d'estimation du bruit non stationnaire [Deng *et al.*, 2003]. La vraisemblance $p(y|x,n)$ est obtenue à partir d'une fonction d'environnement qui suppose l'additivité de la parole et du bruit dans le domaine temporel. Ceci se traduit par une équation non linaire dans le domaine log-spectral :

$$y = x + h + \log(1 + e^{n-h-x}) + \log\left[1 + \lambda \cdot/ \cosh\left(\frac{n-x-h}{2}\right)\right] \tag{2.13}$$

où h représente le terme de canal (bruit convolutif), l'opérateur $\cdot/$ est la division composante par composante de deux vecteurs, et λ est un vecteur dont les composantes sont comprises entre -1 et 1 et qui dépend du déphasage entre la parole et le bruit. Le dernier terme de cette équation représente l'erreur de la fonction d'environnement classique :

$$\hat{y} = x + h + \log(1 + e^{n-h-x}) \tag{2.14}$$

Ce terme d'erreur est modélisé par une densité de probabilité gaussienne centrée à l'origine. La vraisemblance $p(y|x,n)$ est donc également une gaussienne centrée en $\hat{y}$. Lors de la dérivation de la probabilité *a posteriori*, la fonction d'environnement est approchée par son développement de Taylor tronqué.

La méthode SPLICE a été évaluée sur le corpus standard Aurora2 dans [Droppo *et al.*, 2002]. Une comparaison récente de SPLICE avec d'autres approches robustes est réalisée dans [Deng *et al.*, 2005].

2.2.5 SPACE

J'ai travaillé en collaboration avec Khalid Daoudi, à l'IRIT de Toulouse, sur une méthode de débruitage paramétrique non homogène qui généralise l'approche SPLICE décrite ci-dessus, et que nous avons dénommée par référence SPACE, pour *Stereo-based Picewise Affine Compensation for Environments*. Le principe fondamental de cette approche consiste à modéliser les espaces de la parole propre et bruitée par deux GMM

$$P(x) = \sum_{i=1}^{I} \beta_i \mathcal{N}(x; \nu_i, R_i) \quad ; \quad P(y) = \sum_{i=1}^{I} \beta_i \mathcal{N}(y; \tilde{\nu}_i, \tilde{R}_i)$$

tout en garantissant que les gaussiennes propres et bruitées sont en correspondance.

Nous supposons ensuite que tout vecteur y appartient à la gaussienne $\mathcal{N}(\tilde{\nu}_i, \tilde{R}_i)$ et peut être débruité en $x \sim \mathcal{N}(\nu_i, R_i)$ par une transformation affine $x = A_i y + b_i$. En prenant l'espérance et la covariance de cette équation, nous montrons que :

$$\hat{x} = (R_i \tilde{R}_i^{-1})^{\frac{1}{2}}(y - \tilde{\nu}_i) + \nu_i$$

Dans une autre variante de cette approche, nous avons remplacé cette transformation affine par un débruitage exploitant la corrélation croisée $\hat{x} = E[x|y]$, mais ceci n'a pas permis d'améliorer significativement les résultats.

En supposant maintenant que l'appartenance d'une observation à une gaussienne est probabiliste, nous avons :

$$\hat{x} = \sum_i P(i|y) \left((R_i \tilde{R}_i^{-1})^{\frac{1}{2}}(y - \tilde{\nu}_i) + \nu_i \right) \tag{2.15}$$

avec

$$P(i|y) = \frac{\beta_i \mathcal{N}(y; \tilde{\nu}_i, \tilde{R}_i)}{\sum_j \beta_j \mathcal{N}(y; \tilde{\nu}_j, \tilde{R}_j)}$$

Nous avons testé plusieurs approches permettant de construire la correspondance entre les gaussiennes propres et bruitées. La première préserve l'alignement $P(i|x) = P(i|y)$ de chaque gaussienne i sur les corpus propres et bruités.

La deuxième s'appuie sur le critère de minimisation de l'erreur quadratique pour construire une gaussienne propre à partir d'une gaussienne bruitée. Formellement, la fonction objectif à minimiser est :

$$\begin{aligned} F &\stackrel{\triangle}{=} \sum_t E[(x_t - \tilde{x}(y_t))^2] \\ &= \sum_t \sum_i E[(x_t - \tilde{x}(y_t))^2 | y_t, i] P(i|y_t) \\ &= \sum_i \sum_t P(i|y_t)(x_t - A_i y_t - b_i)^2 \end{aligned}$$

où $A_i = (R_i \tilde{R}_i^{-1})^{\frac{1}{2}}$ et $b_i = (R_i \tilde{R}_i^{-1})^{\frac{1}{2}} \tilde{\nu}_i - \nu_i$. Sachant que les covariances sont diagonales, les A_i sont aussi diagonaux. Soit $A_i = diag(a_{i,1}, ..., a_{i,N})$. et $b_i = (b_{i,1}, ..., b_{i,N})^{'}$. Si $x_t = (x_{t,1}, ..., x_{t,N})^{'}$ et $y_t = (y_{t,1}, ..., y_{t,N})^{'}$, alors pour tout i, la fonction objectif à minimiser est :

$$F_i \stackrel{\triangle}{=} \sum_{n=1}^{N} F_{i,n}$$

où

$$F_{i,n} \stackrel{\triangle}{=} \sum_t P(i|y_t)(x_{t,n} - a_{i,n} y_{t,n} - b_{i,n})^2$$

Le problème est alors équivalent à minimiser $F_{i,n}$ par rapport à $a_{i,n}$ et $b_{i,n}$, pour tout i et n.

Soit

$$C_{i,n} \stackrel{\triangle}{=} (b_{i,n}, a_{i,n})^{'}$$

$$X_{i,n} \stackrel{\triangle}{=} (\sqrt{P(i|y_1)} x_{1,n}, ..., \sqrt{P(i|y_T)} x_{T,n})^{'}$$

et $Y_{i,n}$ la matrice de dimensions $T \times 2$ dont la première colonne est

$$(\sqrt{P(i|y_1)}, ..., \sqrt{P(i|y_T)})^{'}$$

et la seconde

$$(\sqrt{P(i|y_1)} y_{1,n}, ..., \sqrt{P(i|y_T)} y_{T,n})^{'}$$

Le problème devient alors :

$$\min (X_{i,n} - Y_{i,n} C_{i,n})^{'} (X_{i,n} - Y_{i,n} C_{i,n})$$

La solution de ce problème est donnée par :

$$C_{i,n} = (Y_{i,n}^{'} Y_{i,n})^{-1} Y_{i,n}^{'} X_{i,n}.$$

Nous obtenons ainsi A_i et b_i pour tout i. ν_i et R_i sont alors :

$$\nu_i = A_i \tilde{\nu}_i - b_i \ ; \ R_i = A_i^2 \tilde{R}_i.$$

La troisième méthode d'appariement des gaussiennes réalise un apprentissage joint dans l'espace propre et bruité. Nous commençons par apprendre sur une base stéréo les paramètres d'un GMM qui modélise la distribution jointe $P(x, y)$ grâce à l'algorithme EM :

$$P(x, y) = \sum_{i=1}^{I} \beta_i J_i(x, y)$$

où J_i est une gaussienne de moyenne m_i et de matrice de covariance diagonale Σ_i. Nous pouvons toujours écrire $m_i \triangleq (\nu_i, \tilde{\nu}_i)$ et $\Sigma_i \triangleq \begin{bmatrix} R_i & 0 \\ 0 & \tilde{R}_i \end{bmatrix}$ tels que

$$P(x) = \sum_{i=1}^{I} \beta_i \mathcal{N}(x; \nu_i, R_i) \quad ; \quad P(y) = \sum_{i=1}^{I} \beta_i \mathcal{N}(y; \tilde{\nu}_i, \tilde{R}_i)$$

Nous pouvons alors utiliser l'équation 2.15 pour débruiter la parole comme précédemment.

L'avantage de notre approche par rapport aux autres méthodes classiques de débruitage, comme SPLICE, est de s'appuyer sur deux ensembles de GMM qui sont mis en correspondance, ce qui permet d'adapter l'un ou l'autre des mélanges à de nouvelles conditions qui n'ont pas été vues à l'apprentissage. Ainsi, nous avons proposé d'adapter le GMM bruité à un nouvel environnement de test par la méthode d'adaptation MAP, et de remplacer dans l'équation 2.15 les paramètres du GMM bruité par les paramètres du GMM adapté. Nous avons de plus proposé une nouvelle méthode d'adaptation, très simple mais qui a l'avantage de garantir la conservation de la correspondance entre les GMM propres et bruités. Elle consiste à estimer un biais d'adaptation global δ entre l'environnement bruité d'apprentissage et celui de test, et à translater toutes les gaussiennes $\mathcal{N}(\tilde{\nu}_i + \delta, \tilde{R}_i)$ avec ce biais.

Nous avons confirmé expérimentalement le bon comportement de cette méthode de débruitage notamment dans [Daoudi et Cerisara, 2005, Daoudi et Cerisara, 2006, Cerisara et Daoudi, 2006]. Les figures 2.3, 2.4, 2.5 comparent les résultats obtenus avec respectivement les modèles multi-conditions (notre système de référence), les modèles SPACE avec modélisation jointe (SPACE-JM) et les modèles SPACE adaptés (B-SPACE-JM) sur Aurora2.

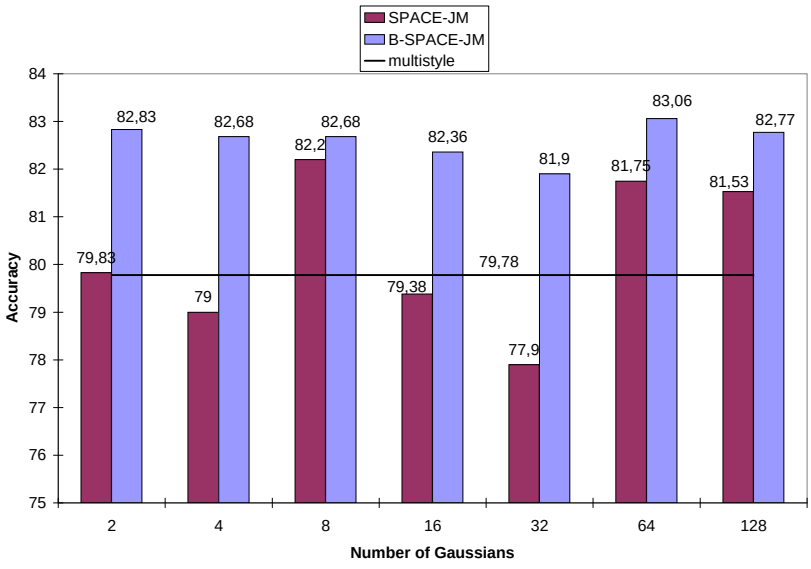


FIG. 2.3 – Taux de reconnaissance en mots moyen sur tous les environnements du test A d'Aurora2

Une approche basée sur le même principe a été proposée dans [Buera *et al.*, 2007].

2.3 Analyse de scènes auditives et séparation de sources

L'analyse de scènes auditives, ou ASA, désigne la faculté qu'ont les humains d'identifier et d'analyser, souvent inconsciemment, les différents sons qui composent notre environnement sonore. L'objectif de l'analyse de scènes auditives computationnelle, ou CASA, est donc de décomposer un environnement sonore en ses constituants individuels et de reconnaître ou d'identifier ces constituants. Dans ce cadre, la parole n'est qu'un élément parmi l'ensemble des sources sonores à analyser.

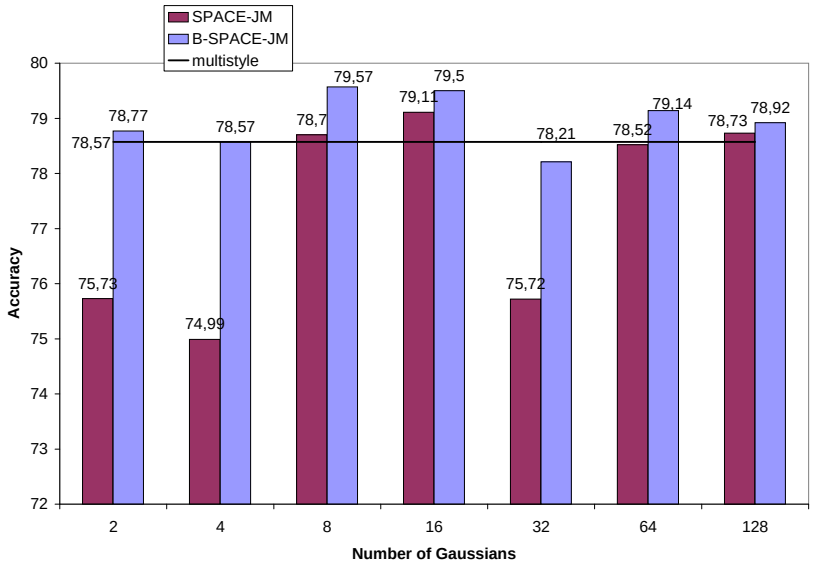


FIG. 2.4 – Taux de reconnaissance en mots moyen sur tous les environnements du test B d'Aurora2

La séparation de sources peut être vue comme un cas particulier de l'analyse de scènes auditives en ce sens qu'elle sépare les flux audio originaires de chacune des sources sonores et reconstitue un signal acoustique pour chaque source ne contenant que la contribution de cette source. Elle est souvent utilisée pour isoler la contribution du signal de parole du locuteur cible dans le but de la reconnaître, ce qui revient à débruiter le signal de parole. La majorité des approches en séparation de source utilisent des matrices de microphones et supposent qu'il y a plus de microphones que de sources à séparer. Toutefois, nous ne nous intéressons pas dans ce mémoire à cette classe d'approches, car nous supposons qu'un unique microphone est disponible.

L'analyse de scènes auditives computationnelle est à l'origine de la reconnaissance de la parole avec données manquantes, domaine auquel j'ai consacré une partie importante de mes recherches et qui est décrit au paragraphe 4. Je n'ai pas réalisé de véritables recherches en l'analyse de scènes auditives, mais ma contribution principale dans ce domaine a consisté à réaliser un état de l'art d'une partie de ce domaine qui est en rapport avec l'estimation des masques [Cerisara *et al.*, 2007]. D'autres travaux réalisent un état de l'art plus général de CASA, comme [Bronkhorst, 2000, M. Cooke et Ellis, 2001].

Le livre de Bregman [Bregman, 1990] définit les principes fondamentaux de CASA. Bregman a remarqué qu'en modifiant certaines caractéristiques des sons de la parole, ces sons pouvaient être perçus comme un son unique ou comme plusieurs sons simultanés. Cette expérience a montré qu'il existe un processus dans le cerveau qui collecte les sons et les groupe éventuellement. Sa théorie propose deux classes d'organisation. La première est la fusion simultanée : ainsi, des sinusoïdes qui sont corrélées harmoniquement sont groupées en un unique son. Altérer le rapport harmonique, ou retarder l'attaque ou la décrue du son peut aboutir à la perception de plusieurs sons séparés. L'attaque simultanée est particulièrement importante. Deux autres indices pour la fusion sont la localisation spatiale et la modulation. La seconde classe est l'organisation séquentielle : des événements successifs sont perçus dans des flux séparés selon les principes de la psychologie de Gestalt [Handel, 1986, Miller et Heise, 1950]. Bregman suggère que cette organisation primitive est pré-cablée dans l'appareil auditif. Il propose également un autre processus de fusion/ségrégation qui est dirigé par des schémas appris. Alors que la fusion primitive est indépendante de la source sonore, les schémas sont spécifiques à chaque type de source, comme la parole.

Brown et Wang réalisent un état de l'art des systèmes CASA les plus importants dans [Brown et Wang, 2005]. Les premiers systèmes CASA [Weintraub, 1985, Brown et Cooke, 1994] étaient essentiellement basés sur une approche ascendante, dans laquelle les caractéristiques des sons sont estimées, puis groupés ou séparés. Ainsi, le système de Brown construisait des cartes d'attaque et de décrue des sons, d'autocorrélation et de

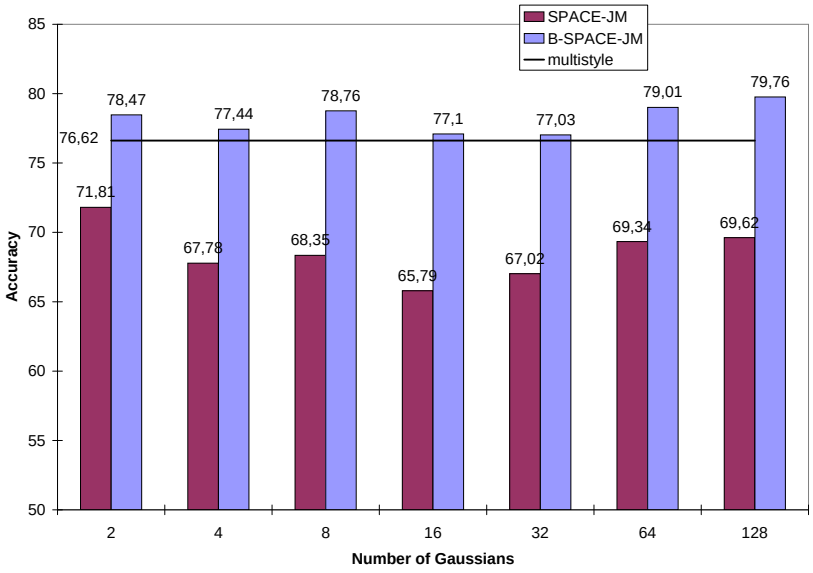


FIG. 2.5 – Taux de reconnaissance en mots moyen sur tous les environnements du test C d'Aurora2

transition en fréquence pour grouper les canaux qui correspondent au même composant acoustique. Un autre système CASA ascendant a été proposé dans [Masuda-Katsuse et Kawahara, 1999], où il est montré que la continuité des modifications spectrales, i.e. du timbre, est un critère important pour l'intégration séquentielle décrite par Bregman. Bien que l'existence d'un tel critère ait été écarté par Bregman, certaines expériences psycho-acoustiques décrites dans [Masuda-Katsuse et Kawahara, 1999] confirment son existence. Ce système exploite également l'harmonicité pour grouper les fréquences de plusieurs sources concurrentes. Ensuite, une fusion séquentielle est réalisée grâce à un suivi spectral, qui est implémenté par un filtre de Kalman autorégressif du second ordre. Une autre caractéristique acoustique est la modulation d'amplitude, exploitée dans [Hu et Wang, 2004] pour analyser les hautes fréquences, dont la largeur de bande élevée crée des battements et des combinaisons de tons qui modulent l'amplitude des flux audio concurrents.

À la suite de ces travaux, des systèmes CASA génériques ascendant et descendant ont été proposés. Ils peuvent être groupés en trois catégories, qui s'appuient sur une architecture de type tableau noir, multi-agents ou neuronale.

2.3.1 CASA sur un tableau noir

Les auteurs de [Godsmark et Brown, 1999] observent que l'organisation de l'oreille humaine prend en compte le contexte et la rétroactivité, ce qui suggère que le choix parmi plusieurs interprétations contradictoires du signal peut être reporté, en laissant aux étages supérieurs du cerveau le soin de désambiguïser les signaux. Ce système a été appliqué à l'analyse musicale et est basé sur une architecture en tableau noir à huit niveaux :

- Le signal est paramétrisé en "bandes synchronisés" [Cooke, 1993], chacune d'entre elles représentant un composant spectral dominant. D'une fenêtre d'analyse à l'autre, les bandes sont jointes par continuité temporelle, proximité fréquentielle et cohérence d'amplitude.
- Des bandes d'indices sont construites à partir des bandes synchronisées. Elles incluent l'attaque, la décrue, les fréquences initiales et finales et l'historique des transitions fréquentielles.
- Les bandes qui sont probablement issues de la même source sont groupées. Divers experts publient des hypothèses de fusion sur le tableau noir, en se basant sur la simultanéité des attaques et des décrues, la proximité temporelle et fréquentielle, et les mouvements harmoniques. Les hypothèses à cet étage correspondent aux notes.

- Les hypothèses sont évaluées par des experts et ordonnées. L'évaluation est basée sur certains principes, comme la proximité dans l'espace temps-fréquence ou la simultanéité des attaques.
- De nouvelles caractéristiques émergentes sont déduites des bandes groupées, comme la fréquence fondamentale et le timbre.
- Une nouvelle fusion est réalisée à partir de ces caractéristiques, en exploitant la proximité du pitch et la similarité du timbre. Le timbre est modélisé par les modifications du centre de gravité spectral, ce qui peut être interprété comme les variations de l'"éclat" du son, et l'enveloppe de l'amplitude. Les groupes obtenus correspondent aux lignes mélodiques.
- Ces nouvelles hypothèses sont également ordonnées.
- Au plus haut niveau, des connaissances dépendantes de chaque source sont utilisées pour calculer par exemple le rythme et les répétitions mélodiques. Ces experts prédisent de nouvelles et/ou renforcent de précédentes hypothèses aux étages inférieurs.

Un autre système CASA basé sur l'architecture en tableau noir est proposé dans [Ellis, 1996]. Ce système est basé sur des modèles du monde qui sont composés d'objets dont la combinaison explique et prédit le signal observé. Les modèles du monde contiennent trois types de sons : les nuages de bruit, les éléments harmoniques et les transitoires. Une hiérarchie repose sur ces éléments pour construire des motifs structurés de sons. Ainsi, à partir des caractéristiques du signal au temps t, comme l'enveloppe de l'énergie, les caractéristiques au temps $t+1$ sont prédites de manière probabilistes grâce aux modèles qui ont été sélectionnés lors des étapes précédentes. Ensuite, les observations sont comparées aux prédictions, et les modèles sont mis à jour, activés ou supprimés selon que les prédictions divergent ou non des observations.

Bien qu'il ne s'agisse pas réellement d'une architecture en tableau noir, mais plutôt d'une architecture "en cascades", le système décrit dans [Hu et Wang, 2004] réalise un traitement du signal qui est très proche de ceux décrits dans ce paragraphe. Ce système a été repris et amélioré dans [Li *et al.*, 2006] en lui adjoignant l'algorithme d'évaluation objective de la qualité du signal de parole P.563. Cet algorithme permet de simuler automatiquement les tests d'évaluation de la qualité d'un signal de parole réalisés par des utilisateurs. Il est donc utilisé comme critère objectif pour séparer et grouper les différentes contributions sonores du signal. Ce nouvel expert intervenant dans le processus de séparation est un autre exemple de processus descendant basé sur des schémas chers à Bregman.

2.3.2 Des agents CASA

L'architecture multi-agents proposée dans [Nakatani, 2002] est caractérisée par sa flexibilité. La paramétrisation exploite également l'harmonicité du signal. Des agents "génératifs" détectent le début des structures harmoniques et génèrent un nouvel agent de suivi. Ceci est réalisé par des agents détecteur de pitch, chacun d'entre eux évaluant l'intensité harmonique pour une fréquence fondamentale possible. L'agent avec la plus grande énergie génère un agent de suivi, qui s'appuie sur des critères de proximité fréquentielle. À chaque fois qu'un nouvel agent de suivi est activé, il soustrait sa contribution du signal global : le signal résiduel est alors analysé par d'autres agents. Des agents d'analyse sont également introduits pour détecter et éliminer les agents de suivis redondants ou défunts. Un tel sous-système qui extrait des flux audio à partir d'attributs sonores est appelé une agence. D'autres agences sont proposées, qui exploitent des indices visuels, spatiaux, stéréophoniques ou monophoniques. Chaque agence produit son propre ensemble de flux, et les agences peuvent interagir de différentes manières. Par exemple, les sorties de deux agences monophoniques peuvent constituer les entrées d'une agence stéréophonique, ou encore une agence qui détecte les segments de parole voisés peut remplacer ou modifier la sortie d'une agence stéréophonique.

2.3.3 CASA et réseaux de neurones

Un réseau de neurones oscillant à deux niveaux est implémenté pour séparer les sources sonores dans [Wang et Brown, 1999]. Chaque caractéristique auditive est encodée dans des oscillateurs, et les oscillateurs synchronisés sont fusionnés au sein d'un flux unique, tandis que les oscillateurs désynchronisés représentent plusieurs flux. Ce modèle, schématisé sur la figure 2.6 améliore le SNR d'un signal de parole

mélangé avec plusieurs types de bruits, mais obtient des performances mitigées lorsque plusieurs sources sont ajoutées [van der Kouwe *et al.*, 2001].

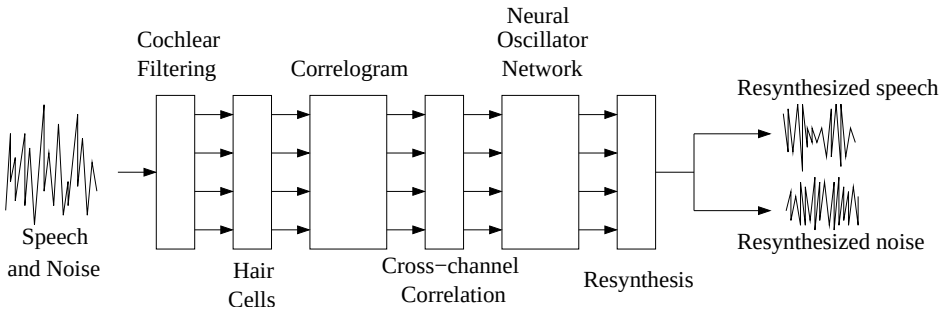


FIG. 2.6 – Architecture schématique, reproduit de [Wang et Brown, 1999], du système CASA basé sur des réseaux de neurones. Quatre étages sont utilisés pour traiter la parole bruitée : le premier simule l'appareil auditif périphérique humain (filtre cochléaire et cils cillés), le second calcule des représentations intermédiaires du signal dont un corrélogramme et une carte de corrélation inter-canaux, le troisième utilise un réseau oscillant pour grouper les composants acoustiques, et le dernier re-synthétise les signaux de parole et de bruit.

Une autre architecture neuronale, le réseau cortronique [Sagi *et al.*, 2001] a été proposée pour reconnaître des mots dans de la parole concurrente. Il est basé sur un réseau à mémoire associative composé de trois niveaux, respectivement pour paramétriser le son, le traiter et le reconnaître.

Plusieurs concepts importants concernant l'implémentation de CASA par des réseaux de neurones sont discutés dans [Haykin et Chen, 2005]. Ainsi, une architecture basée sur une nouvelle approche, appelée audition active, est proposée pour traiter le difficile problème de la parole concurrente. Cette architecture est composée de plusieurs niveaux :

- La localisation, qui infère la direction des signaux.
- La ségrégation, qui sépare plusieurs sources sonores et se focalise sur l'une d'entre elle.
- Le suivi, qui prédit les valeurs suivantes des caractéristiques du signal.
- L'apprentissage, qui mets à jour un modèle interne et l'adapte aux évolutions du monde perçu.

2.3.4 Séparation de source à un seul canal

L'objectif de la séparation de source est de séparer les signaux acoustiques de chacune des sources sonores simultanées. Ce domaine de recherche est très actif en reconnaissance automatique de la parole, mais aussi en analyse musicale [Virtanen, 2007].

À la différence de la séparation de source aveugle (BSS pour *Blind Source Separation*) qui utilise plusieurs microphones pour déconvoluer les sources sonores, des connaissances concernant la parole et/ou le bruit sont presque toujours utilisées lorsqu'un seul canal est disponible. Il est néanmoins possible d'appliquer, tout comme en BSS, une analyse en composantes indépendantes unifiée (ICA), comme le montre [Potamitis *et al.*, 2001]. Deux approches sont proposées dans ce cadre, la première calculant la matrice de démixage sur un grand corpus de parole propre et la seconde l'estimant sur la parole bruitée au moyen d'une approximation variationnelle. D'autres approches basées sur ICA sont résumées dans [Smaragdis, 2007] et [Molla et Hirose, 2007], qui propose également de calculer des paramètres issus du spectre de Hilbert avec une analyse en composante principale et ICA, paramètres qui sont ensuite classifiés par un algorithme des k-plus proches voisins exploitant la distance de Kullback-Leibler. Le spectre de Hilbert du signal combiné est projeté sur les deux bases ainsi créées, ce qui permet de reconstruire les deux signaux indépendants correspondant.

Une autre approche factorielle, qui est validée pour séparer deux locuteurs parlant en même temps, consiste à construire un modèle acoustique spécifique à chaque locuteur, puis à combiner ces deux modèles et chercher le meilleur chemin qui maximise la vraisemblance du modèle combiné. Roweis, dans [Roweis, 2003], utilise alors des modèles spectraux et fait l'hypothèse que chaque bande fréquentielle

est dominée à tout instant par l'un ou l'autre des locuteurs. Un schéma de principe de son approche est illustré sur la figure 2.7.

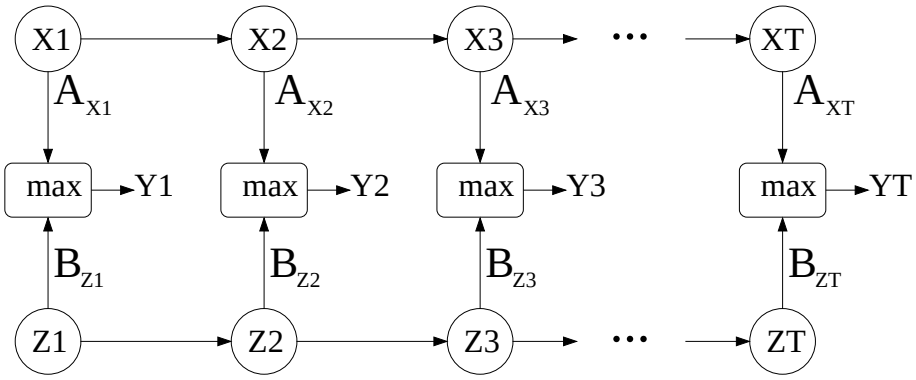


FIG. 2.7 – HMM factoriel combinant deux chaînes de Markov X_t et Z_t évoluant indépendamment. Les observations Y_t sont expliquées par l'une ou l'autre des chaînes $Y_t = \max[A_{xt}, B_{zt}]$ à laquelle est ajouté du bruit gaussien.

Un travail similaire, inspiré de la vision, est proposé dans [Reyes-Gomez *et al.*, 2004], où un réseau bayésien à plusieurs couches est utilisé, chaque couche modélisant une source de variabilité. Deux couches sont définies : la première pour les harmoniques de la fondamentale, et la seconde pour les résonances. Les auteurs calculent une transformation de chaque trame pour prédire la suivante. Les transformations sont encodées dans un modèle graphique génératif qui gère la dépendance entre les harmoniques, les filtres ou formants, leurs transformations et le spectre. Un tel réseau peut être appris pour chaque locuteur afin de séparer leur contribution (voir figure 2.8).

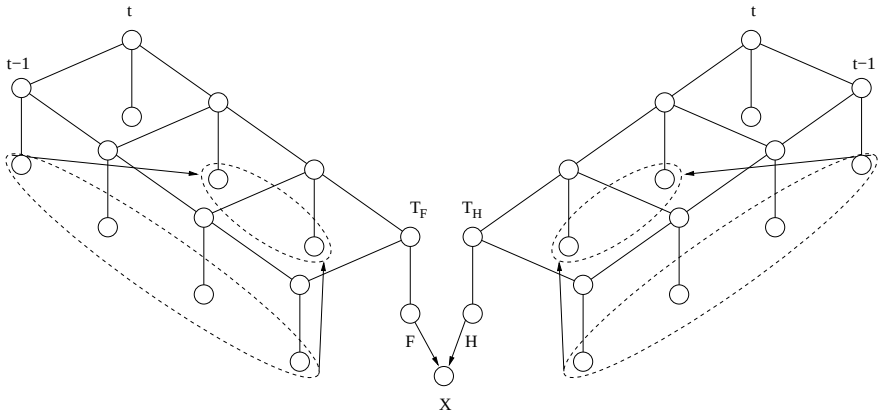


FIG. 2.8 – Modèle graphique génératif à deux couches pour séparer deux flux de parole concurrente : les trames $t-1$ et t sont représentées ; F et H représentent respectivement l'excitation et le filtre vocal, T_F et T_H représentent leur transformation de $t-1$ à t.

Une autre approche bayésienne est encore proposée dans [Kristjansson *et al.*, 2004], où les probabilités a posteriori des deux locuteurs $p(x_1|y)$ et $p(x_2|y)$ sont estimées à partir de la fonction de densité de probabilité $p(y|x_1, x_2)$. La moyenne de cette densité est linéarisée par rapport à x_1 et x_2, de sorte à modéliser $p(x_1, x_2|y)$ par une gaussienne. Le point de linéarisation est ajusté itérativement vers le mode de la véritable probabilité a posteriori. Deux GMM, respectivement pour $p(x_1)$ et $p(x_2)$ sont appris et utilisés finalement pour séparer les flux de chaque locuteur.

Les auteurs de [Bach et Jordan, 2005] apprennent un classifieur spectral qui groupe tous les points

du spectre en deux ensembles, un par locuteur. Ce segmenteur prend la forme de matrices paramétriques d'affinité qui codent les relations topologiques dans le spectre entre les observations, qui sont la continuité temporelle et fréquentielle, les variations temporelles communes, l'harmonicité et le timbre. Le segmenteur est entraîné sur de la parole mélangée d'un petit nombre de locuteurs et est testé sur d'autres locuteurs. L'avantage principal de cette approche est qu'elle ne requiert pas de modèle précis des différentes voix.

2.3.5 Conclusions

L'analyse de scènes auditives computationnelle n'est pas spécifiquement dédiée à la reconnaissance automatique de la parole robuste : c'est un domaine de recherche plus général qui s'intéresse aux processus cognitifs auditifs, et en particulier à l'identification et la séparation d'une source sonore parmi d'autres. Toutefois, la reconnaissance automatique de la parole robuste est certainement une application privilégiée de CASA. Mais fusionner ces deux domaines de recherche est une tâche ardue, car il faut maîtriser aussi bien les subtilités de CASA que la complexité des systèmes de reconnaissance, et pouvoir adapter les solutions proposées dans CASA aux spécificités de la reconnaissance robuste.

Ceci explique le faible nombre de tentatives d'intégration réalisées jusqu'à présent. Ces quelques tentatives méritent donc toute notre attention, car elles sont remarquables de par les difficultés que leurs auteurs ont su surmonter. Nous retiendrons en particulier le système CASA décrit dans [Brown *et al.*, 2001], qui est intégré avec un système de reconnaissance de la parole avec données manquantes et évalué sur la base de données TiDigits à laquelle du bruit enregistré dans une usine a été ajouté : les auteurs montrent que le système combiné donne de meilleurs résultats que la soustraction spectrale, avec un taux de reconnaissance de séquences de chiffres de 47 % à 0 dB. Le système cortronic [Sagi *et al.*, 2001] décrit précédemment a également été évalué sur une tâche de reconnaissance de mots, mais avec un système de reconnaissance spécifique et une tâche dédiée, ce qui rend l'interprétation des résultats difficile : les voix de 1 à 20 locuteurs sont mélangées à énergie égale, chacune d'entre elle prononçant un mot parmi 1024 mots possibles. Avec moins de 5 locuteurs simultanés, le taux de reconnaissance est supérieur à 98 %. Il décroit à 70 % avec 10 locuteurs et à 20 % avec 20 locuteurs.

Finalement, le modèle CASA le plus récent est celui des "fragments de parole" [Cooke, 2005], qui a été implémenté dans le cadre de la théorie des données manquantes. Cette intégration, certainement l'une des plus abouties aujourd'hui, est décrite au paragraphe 4.2.

2.4 Contributions

Mes principales contributions dans le domaine du traitement du signal, sont au nombre de deux : tout d'abord, j'ai publié en collaboration avec Sébastien Demange et Jean-Paul Haton un état de l'art critique sur les approches CASA appliquées à la reconnaissance avec données manquantes, mais surtout, j'ai proposé avec Khalid Daoudi la nouvelle approche de débruitage bayésienne SPACE qui présente l'avantage de pouvoir s'adapter à de nouveaux environnements de test inconnus.

Chapitre 3

Adaptation des modèles

Après avoir passé en revue les principales approches qui altèrent le signal en vue de le débruiter ou d'en extraire des paramètres robustes, nous nous intéressons dans la suite à l'approche duale, qui modifie les modèles acoustiques afin de les rendre plus robustes ou de les adapter aux conditions de test.

3.1 De l'apprentissage à l'adaptation

La meilleure adaptation possible est de ré-entraîner les modèles acoustiques sur une nouvelle base de données correspondant aux conditions de test. Mais généralement, une telle base de données n'est pas disponible, et au mieux, seules quelques phrases représentant les conditions de test sont disponibles. Il est alors impossible d'apprendre tous les paramètres des modèles avec aussi peu de données. Le but des méthodes d'adaptation est donc de réduire le nombre de paramètres, tout en conservant une modélisation détaillée des variations phonétiques.

3.1.1 MAP, SMAP, MLLR, MAPLR, SMLLR

Les distributions issues de l'apprentissage des modèles acoustiques peuvent être considérées comme des distributions *a priori*, et les données d'adaptation servent alors à apprendre des distributions *a posteriori* : c'est l'adaptation selon le critère du maximum a posteriori (MAP) [Lee *et al.*, 1991, Gauvain et Lee, 1994]. Lorsque aucun exemple n'existe dans le corpus d'adaptation pour une distribution particulière des modèles acoustiques, cette distribution sera exactement la distribution *a priori*. Inversement, plus le nombre d'exemples augmente, plus l'estimation de cette distribution sur les nouvelles données contribue à sa nouvelle forme. L'équation 3.1 montre comment la moyenne d'une gaussienne μ est adaptée en $\hat{\mu}$ en fonction des données observées o_t :

$$\hat{\mu} = \frac{\tau\mu + \sum_{t=1}^{T} \gamma_t o_t}{\tau + \sum_{t=1}^{T} \gamma_t} \tag{3.1}$$

γ_t est la probabilité *a posteriori* que la gaussienne considérée soit alignée avec o_t et τ l'hyper-paramètre qui règle l'équilibre entre la distribution *a priori* et les données observées.

La technique dite MAP structurelle (ou SMAP) [Shinoda et Lee, 2001] augmente la vitesse de convergence de l'adaptation MAP en regroupant les gaussiennes hiérarchiquement dans un arbre de régression et en utilisant un unique biais d'adaptation pour chaque classe de régression dans cet arbre. A chaque niveau de l'arbre, l'estimée du biais au niveau supérieur est utilisée comme distribution *a priori*. Mes propres travaux décrits ci-dessous exploitent la structure de l'arbre de régression. Je décris donc ci-dessous plus en détails l'algorithme généralement utilisé pour construire un arbre de régression :

Algorithme de construction d'un arbre de régression

1. Initialement, toutes les gaussiennes sont regroupées dans une seule classe de régression : la racine de l'arbre, de niveau $i = 0$. Le vecteur représentant cette classe est calculé en moyennant toutes les moyennes des gaussiennes.
2. Pour chaque nœud n du niveau courant i de l'arbre, le représentant du nœud est divisé en deux vecteurs qui s'éloignent de part et d'autre d'une distance relative aléatoire. Ces deux vecteurs définissent les représentants des nœuds « fils » de n. Toutes les gaussiennes de n sont alors réaffectées à l'un ou l'autre de ses fils. Les représentants de chaque nœud-fils sont alors recalculés à partir de ses gaussiennes.
3. Ce processus est itéré au niveau $i+1$ à partir de l'étape 2. La division des nœuds s'arrête selon divers critères : nombre maximal de nœuds, gain minimum (calculé par exemple à partir des distances inter-gaussiennes) résultant de la division, nombre minimum d'exemples, etc. De même, les distances utilisées au cours de l'algorithme peuvent varier, par exemple en prenant en compte les variances des gaussiennes et/ou des classes de régression.

Une autre approche très connue pour diminuer le nombre de paramètres à apprendre est de ne pas apprendre directement les paramètres des modèles acoustiques (moyenne, variance), mais plutôt une transformation linéaire de ces paramètres : c'est l'approche dite d'adaptation par régression linéaire avec maximisation de la vraisemblance (en anglais *Maximum Likelihood Linear Regression*, MLLR). Généralement, ce sont les moyennes des gaussiennes qui sont adaptées [Leggetter et Woodland, 1995b] :

$$\hat{\mu}_i = A\mu_i + b \tag{3.2}$$

où les paramètres A et b de la transformation sont les mêmes pour toutes les gaussiennes semblables, c'est-à-dire appartenant à une même classe. La matrice A peut être pleine, ou diagonale par blocs, ce qui revient alors à apprendre une matrice différente pour les coefficients statiques, de vélocité et d'accélération, ou encore diagonale, en fonction de la quantité de données d'adaptation disponible. Les paramètres A et b sont appris sur les données d'adaptation par l'algorithme EM. Un autre algorithme d'apprentissage dérivé de EM optimise les variables d'adaptation afin qu'elles maximisent une vraisemblance réduite [Byrne et Gunawardana, 2001], dans le but de réduire l'effet du sur-apprentissage des paramètres de la transformation. Cette méthode est connue sous le nom de DLLR (pour *Discounted Likelihood Linear Regression*). Deux autres variantes de MLLR sont proposées dans [Goel *et al.*, 2002], qui utilise des transformations qui sont des combinaisons linéaires de matrices de rang unitaire, et dans [Kim et Chung, 2003], qui exploite des algorithmes de réduction de dimension, comme l'analyse en composantes principales ou en composantes indépendantes. Lorsqu'il y a suffisamment de données d'adaptation, les variances des gaussiennes peuvent être également adaptées, soit indépendamment des moyennes [Gales et Woodland, 1996], soit avec la même matrice d'adaptation que pour les moyennes (*CMLLR*) [Digalakis *et al.*, 1995, Gales, 1998].

Les paramètres des transformations linéaires peuvent être optimisés selon le critère de maximisation de la probabilité *a posteriori* (MAP) plutôt que selon le critère de maximisation de la vraisemblance. Dans ce cas, on parle de MAPLR [Chesta *et al.*, 1999, Chou, 1999], par similitude avec MLLR. Une distribution *a priori* des paramètres des transformations linéaires est alors choisie. Il est également possible d'adapter les variances des gaussiennes par cette technique [Chou et He, 2003].

Le regroupement des gaussiennes en classes peut être réalisé hiérarchiquement, comme cela a déjà été décrit pour l'adaptation MAP structurelle, plutôt qu'« à plat », comme cela est réalisé avec MLLR. Il s'agit alors d'adaptation MLLR structurelle (SMLLR) [Leggetter et Woodland, 1995a], ou encore MAPLR structurelle [Siohan *et al.*, 2000]. L'avantage principal de cette approche est de pouvoir adapter dynamiquement le nombre de classes de régression en fonction de la quantité de données d'adaptation disponible pendant le test. Ainsi, intuitivement, chaque trame observée dans le corpus d'adaptation est

comptabilisée dans toutes les classes de régression composant le chemin menant de la racine de l'arbre à la feuille contenant la gaussienne alignée avec cette trame. Ensuite, la profondeur maximale que l'on peut se permettre d'atteindre dans chaque branche est déterminée en choisissant un nombre d'observations minimal par classe.

SMLLR peut bien sûr être combinée avec d'autres techniques d'adaptation, comme l'approche dite des "voix propres" [Kuhn *et al.*, 2000], qui construit par une analyse en composantes principales (ou alternativement par le critère du maximum de vraisemblance [Nguyen *et al.*, 1999]) une base orthonormée de voix (ou de bruits) [Chen *et al.*, 2000], et la théorie des noyaux, qui généralise les voix propres à des transformations non linéaires, comme cela est proposé dans [Mak et Hsiao, 2007].

Notons que la plupart des techniques d'adaptation sont utilisées pour l'adaptation au locuteur, mais qu'elles sont également applicables en adaptation au bruit. Il existe de très nombreuses variantes des méthodes de base MAP et MLLR : je n'ai décrit ci-dessus que les principales. Je présente dans la suite ma contribution principale dans ce domaine.

3.1.2 Modèles multi-échelles

L'objectif fondamental de toute méthode d'adaptation est de maximiser le nombre de paramètres à adapter tout en garantissant que l'apprentissage de ces paramètres soit de bonne qualité, au vu de la petite taille habituelle du corpus d'adaptation. Pour ce faire, nous avons vu plusieurs hypothèses structurelles, dont la transformation linéaire et l'arbre de régression, qui permettent de diminuer et de partager les paramètres afin de les entraîner correctement. J'ai proposé, en collaboration avec Khalid Daoudi, une nouvelle hypothèse applicable aux arbres de régression et permettant de mieux exploiter leur topologie multi-échelles afin d'augmenter la précision des modèles sans perdre en qualité d'estimation : les modèles autorégressifs multi-échelles. L'intérêt de ces modèles est leur capacité à estimer rapidement et de manière optimale le vecteur d'état dans chaque nœud, au moyen de seules mesures peu nombreuses et bruitées réparties à différents endroits de l'arbre. Ce travail est basé sur l'approche développée par Kannan dans [Kannan, 1997].

Les processus autorégressifs multi-échelles (PAM) définissent une classe particulière de processus gaussiens indexés par un arbre [Luettgen *et al.*, 1993]. La relation entre deux échelles d'un PAM est définie par :

$$x(s) = A(s)x(s\bar{\gamma}) + \omega(s)$$

où $s\bar{\gamma}$ est le nœud parent de s, $A(s)$ est une matrice constante et $\omega(s)$ est un bruit blanc de covariance $Q(s)$. Les paramètres du PAM sont alors :

$$\{Q(s_0)\} \cup \{A(s), Q(s)\}_{s \neq s_0}$$

où s_0 est la racine de l'arbre qui indexe le processus $x(\cdot)$. Dans le cadre de l'adaptation, l'arbre concerné est l'arbre de régression et le processus $x(\cdot)$ est la transformation d'adaptation associée à la classe de régression s.

Un avantage important des PAM est qu'il existe des algorithmes d'inférence rapide de complexité linéaire en terme de nombre de nœuds qui estiment les valeurs de ce processus en tous les nœuds de l'arbre. Kannan suggère dans [Kannan, 1997] que l'adaptation SMAP n'est qu'un cas particulier des modèles PAM. La figure 3.1 représente une portion de l'arbre utilisée dans notre système.

Pendant le test, une adaptation MLLR classique est d'abord réalisée pour estimer les valeurs de $x(s)$ dans quelques classes de régression. Ensuite, une généralisation de l'algorithme de Rauch-Tung-Striebel (RTS) est appliquée pour estimer les valeurs de $x(s)$ en tous les nœuds de l'arbre, et en particulier aux feuilles. Ces valeurs aux feuilles sont ensuite utilisées pour adapter les modèles de parole.

L'apprentissage des paramètres du processus PAM est réalisé en calculant sur le corpus d'apprentissage les transformations d'adaptation dans tous les nœuds de l'arbre, ce qui permet de réaliser un apprentissage optimal complet, c'est-à-dire avec toutes les variables observées. Soient $o(s_0), \cdots, o(s_{N-1})$ les observations dans tout l'arbre pour un exemple du corpus d'apprentissage. Grâce à la propriété de Markov du PAM, nous pouvons calculer la vraisemblance ainsi :

$$P(O|\lambda) = P(x(s_0) = o(s_0)|\lambda) \times \prod_{i=1}^{N-1} P(x(s_i) = o(s_i)|s(s_i\bar{\gamma}) = o(s_i\bar{\gamma}), \lambda)$$

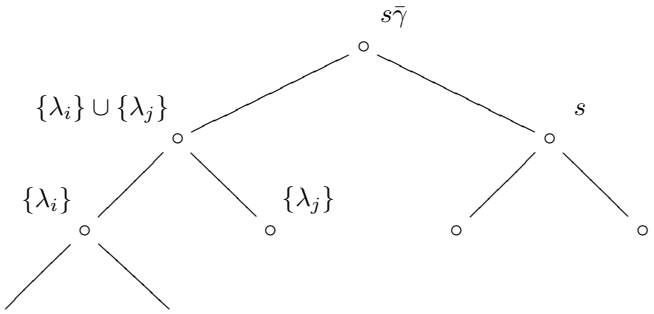


FIG. 3.1 – Exemple d'application des dépendances autorégressives multi-échelles sur un arbre de régression pour MLLR. λ_i représente les paramètres d'adaptation de la gaussienne i des modèles acoustiques.

L'apprentissage revient donc à trouver l'optimum des paramètres :

$$\frac{\partial P(O|\lambda)}{\partial \lambda} = 0$$

La solution est :

$$A(s) = P_{s,s\bar{\gamma}} \cdot P_s^{-1}$$

$$Q(s_0) = P_{s_0}$$

$$Q(s) = P_s - A(s) \cdot P_{s,s\bar{\gamma}}^T$$

où

$$P_s = \frac{1}{I} \sum_{i=1}^{I} o_{i,s} o_{i,s}^T$$

et

$$P_{s,s\bar{\gamma}} = \frac{1}{I} \sum_{i=1}^{I} o_{i,s} o_{i,s\bar{\gamma}}^T$$

avec $o_{i,s}$ la $i^{\text{ème}}$ observation de $x(s)$ et I le nombre total d'observations disponibles dans un nœud donné.

Nous avons comparé expérimentalement cette approche sur la base de données VODIS avec l'adaptation MLLR globale. Pour chacun des 18 locuteurs de VODIS, un nombre variable de phrases est utilisé en adaptation supervisée, et les moyennes des taux de reconnaissance sur tous les locuteurs sont résumées dans le tableau 3.1.

Nombre de phrases d'adaptation	MLLR global	MLLR + PAM
1	61.4	72.3
3	65.4	74.5
6	72.7	73.0
11	74.7	74.7
12	73.4	73.4
18	74.6	74.6

TAB. 3.1 – Taux de reconnaissance de MLLR global et de l'adaptation PAM sur VODIS en fonction du nombre de phrases utilisées par locuteur pour l'adaptation supervisée.

Nous voyons sur ces résultats que, comme prévu, l'utilisation du processus multi-échelles améliore les résultats de l'adaptation lorsque la quantité de données d'adaptation disponible est faible.

3.2 Modèles multi-bandes

Les modèles multi-bandes ont été proposés à la fin des années 90 pour rendre la modélisation acoustique plus robuste aux bruits en séparant les différentes bandes de fréquence. Sachant que le bruit n'affecte en général principalement qu'une zone limitée en fréquences, il existe donc des zones fréquentielles non corrompues par le bruit et qui peuvent être exploitées en priorité pour la reconnaissance. Une autre motivation pour l'utilisation du multi-bandes en reconnaissance automatique de la parole est d'ordre psycho-acoustique : en effet, Jont B. Allen a publié en 1994 un article retentissant [Allen, 1994] qui montre que l'oreille humaine semble traiter de manière indépendante les bandes critiques du spectre. Sans vouloir reproduire à l'identique les processus de haut niveau intervenant dans l'audition humaine, processus encore trop peu connus aujourd'hui, il ne peut qu'être intéressant de proposer des modèles artificiels tentant de reproduire quelques-uns des phénomènes observés.

La figure 3.2 montre un exemple de modèle multi-bandes.

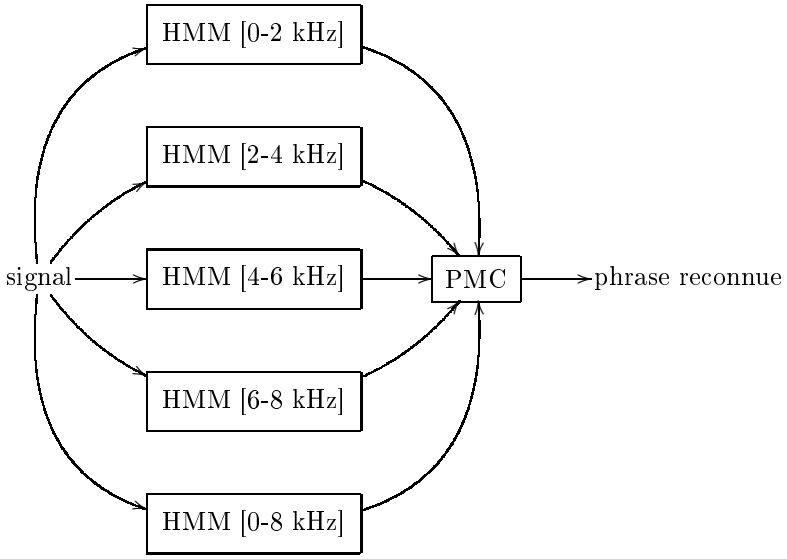


FIG. 3.2 – Exemple d'un modèle multi-bandes avec recombinaison par perceptron multicouches, PMC.

J'ai consacré ma thèse de doctorat aux modèles multi-bandes [Cerisara, 1999a]. Mes travaux ont été ensuite repris et complétés, notamment dans [Daoudi *et al.*, 2003]. Ils ont également été par la suite régulièrement cités dans d'autres travaux du domaine, comme par exemple [Hagen et Morris, 2000, Christensen *et al.*, 2000, Poh et Bengio, 2004].

L'une des caractéristiques principales d'un véritable système multi-bandes est sa capacité à supporter un certain degré d'asynchronisme entre les bandes fréquentielles. Cet asynchronisme résulte de la remise en cause de l'hypothèse commune à tous les autres systèmes de reconnaissance selon laquelle les événements acoustiques sont parfaitement synchrones sur l'ensemble du spectre. Le degré d'asynchronisme autorisé dépend de la longueur des segments temporels séparant deux recombinaisons successives, telle que celle illustrée figure 3.2. Les différents cas suivants peuvent être envisagés :

- Recombinaison après chaque trame.
- Recombinaison après chaque phonème.
- Recombinaison après chaque phrase.

Les recombinaisons après la trame et après le phonème ont été étudiées dans la littérature, par exemple respectivement dans [Tibrewala et Hermansky, 1997] et [Bourlard et Dupont, 1996]. J'ai proposé et étudié un algorithme permettant de recombiner les résultats de la reconnaissance à la fin de la phrase, en alignant les résultats des recombinaisons partielles. J'ai également proposé un nouvel algorithme basé sur l'algorithme du "Level Building" [Myers et Rabiner, 1981] et qui permet de recombiner les bandes après

chaque phonème [Cerisara, 1999b]. Cet algorithme présente plusieurs avantages par rapport aux autres solutions alors proposées pour résoudre ce problème, notamment :

- Sa complexité est linéaire en temps, tout comme Viterbi.
- Il est synchrone par rapport à la trame.
- Il est compatible avec les grammaires statistiques.

La recombinaison est un des éléments les plus importants dans l'architecture d'un système multi-bandes. C'est pourquoi plusieurs approches de recombinaison ont été proposées dans la littérature dans la période pendant laquelle je travaillais sur ce problème. Notons en particulier l'étude très poussée de la fusion de l'information provenant des différentes bandes de fréquence par l'approche Bayésienne avec diverses hypothèses simplificatrices réalisée dans [Besacier, 1998]. Une autre approche très intéressante, dénommée "Full Combination", propose plutôt d'entraîner un classifieur pour toutes les combinaisons de bandes possibles et intègre alors les décisions de ces classifieurs en les pondérant par la probabilité que chaque combinaison soit la meilleure à un instant donné du point de vue du taux de reconnaissance. Cette approche, ainsi qu'une approximation permettant de réduire drastiquement la taille de l'ensemble de classifieurs qu'il faut entraîner est décrite dans [Glotin et Berthommier, 2000].

J'ai également proposé, testé et comparé diverses recombinaisons, comme la recombinaison linéaire pondérée avec apprentissage des poids par le critère de minimisation de l'erreur de classification [Cerisara *et al.*, 1998] ou les réseaux de neurones. Les meilleurs résultats ont été obtenus avec des perceptrons à trois couches. J'ai ainsi montré que l'étude du perceptron après apprentissage permet de mesurer la quantité d'information présente dans chaque bande pour chaque phonème.

Ce résultat explique notamment les différences de taux de reconnaissance des différents phonèmes dans différentes bandes : en effet, il est logiquement plus difficile de reconnaître une fricative dans les basses que dans les hautes fréquences. J'ai donc également réalisé une étude détaillée des taux de reconnaissance de chaque phonème individuellement dans chaque bande pour mettre en évidence et mieux comprendre ce phénomène. J'ai ainsi proposé de remplacer les unités dans chaque bande par de nouvelles unités acoustiques qui représentent mieux l'information acoustique effectivement présente dans chaque bande de fréquences. La création de ces unités est réalisée automatiquement au moyen d'un apprentissage discriminant joint des modèles acoustiques et du module de recombinaison [Cerisara *et al.*, 1999, Cerisara et Fohr, 2001].

Aujourd'hui, les modèles multi-bandes ne sont plus souvent utilisés tels quels, mais ils ont évolué de différentes manières, et on retrouve leur principe dans plusieurs classes de modèles actuels. Ainsi, les modèles "multi-flux" constituent une généralisation des modèles multi-bandes qui intègrent l'information provenant de sources d'information variées, et non plus seulement du signal acoustique [Glotin, 2001a]. Il est ainsi envisageable de combiner information acoustique et visuelle par exemple. De plus, les réseaux bayésiens tels que proposés dans [Daoudi *et al.*, 2003] permettent de s'affranchir d'un certain nombre de problèmes difficiles à résoudre dans le cadre des modèles multi-bandes, comme la recombinaison des bandes fréquentielles.

Enfin, les travaux que j'ai réalisés sur la reconnaissance avec données manquantes, et qui sont décrits dans la suite de ce mémoire, peuvent être interprétés comme la continuation directe de mes travaux sur les modèles multi-bandes, car le principe fondamental reste le même, à savoir ne pas prendre en compte les zones fréquentielles bruitées du spectre. Un autre domaine de recherche en relation étroite avec les modèles multi-bandes et surtout la théorie des données manquantes consiste à sélectionner les meilleures trames du signal et à supprimer les autres trames [Besacier et Bonastre, 2002]. Cette approche est particulièrement importante en reconnaissance et identification du locuteur [Besacier, 1998], car pour de telles applications, la redondance de l'information acoustique permet plus facilement d'éliminer de l'information que pour la transcription automatique.

3.3 Modèles de bruit et modèles bruités

Tout comme avec la paramétrisation, il n'est pas possible de contrôler finement l'influence du bruit sans connaissances élaborées sur le bruit lui-même. Une solution consiste donc à modéliser le bruit, séparément ou conjointement à la parole. Dans le premier cas, nous parlons de modèle de bruit, et dans

le second, de modèle de parole bruitée.

3.3.1 Modèles de bruit

Les techniques les plus utilisées pour apprendre un modèle de bruit sont les suivantes :

- Utiliser un corpus de bruit, comme NOISEX [NOISE-ROM-0, 1990]. Il est alors possible d'apprendre des HMM de bruit, tout comme nous apprenons des HMM de parole sur un corpus de parole. Les HMM de bruit résultant peuvent être ensuite, par exemple, combinés avec des modèles de parole dans l'algorithme PMC décrit ci-dessous. Toutefois, cette approche nécessite de disposer de corpus qui contiennent les mêmes bruits que lors du test, et les modèles obtenus risquent d'encoder trop de variabilité dans les types de bruit et de ne pas être suffisamment précis pour le test.
- Pour estimer le bruit spécifique qui affecte une phrase de parole, la solution la plus commune est d'estimer un modèle simple du bruit, avec peu de paramètres comme un vecteur-moyenne et éventuellement une variance, pendant les segments de silence détectés dans la phrase de test. En pratique, dans les corpus enregistrés, les quelques dixièmes de secondes au début et à la fin de chaque fichier ne contiennent pas de parole et sont utilisés à cette fin. Mais cette approche ne permet de prendre en compte que les bruits quasi-stationnaires ou variant lentement, et n'est que de peu d'utilité pour les bruits très variables, comme la parole ou la musique.

La précision des modèles de bruit étant très importante pour les techniques d'adaptation, et au vu des limites des deux méthodes précédentes, de nombreuses recherches ont été menées pour proposer parfois des algorithmes très sophistiqués d'estimation du bruit. Toutefois, le plus souvent, de tels algorithmes émettent un certain nombre d'hypothèses, comme par exemple le fait que le SNR est supérieur à 0 dB, ce qui permet d'associer les trames à la parole ou au bruit en fonction de leur énergie. C'est l'approche suivie par exemple dans [Kim et Ruwisch, 2002]. Au contraire, les auteurs de [Srinivasan *et al.*, 2007] s'appuient plutôt sur une modélisation bayésienne de la parole et du bruit combinés pour estimer sur la phrase de test les variances de l'excitation du signal de parole et en déduire l'information du bruit.

Ces deux exemples sont représentatifs des divers algorithmes proposés pour estimer le bruit. Les modèles de bruit obtenus avec ces méthodes peuvent ensuite être utilisés pour reconnaître un type d'environnement bruité, ce qui permet de choisir ensuite des modèles spécifiques pour cet environnement, ou alors être combinés avec des modèles de parole pour construire des modèles bruités adaptés aux conditions de test.

3.3.2 Modèles bruités

Il existe plusieurs manières de bruiter les modèles afin de réduire les écarts entre les conditions d'apprentissage et de test. La plus simple est d'apprendre des modèles acoustiques sur une base d'apprentissage bruitée dans les mêmes conditions que celles de test, lorsque ces dernières sont connues, ou alors bruitée avec un ensemble représentatif de bruits qui risquent d'apparaître pendant le test. Cette technique, connue sous le nom d'apprentissage multi-conditions, ou multi-styles, est encore très souvent utilisée car relativement efficace malgré sa simplicité [Lippmann *et al.*, 1987].

Toutefois, l'apprentissage multi-conditions présente l'inconvénient d'augmenter la confusion entre les modèles acoustiques lorsque le nombre de bruits différents ajoutés dans la base d'apprentissage augmente. Pour éviter ce problème, il est possible d'apprendre différents ensembles de modèles, dans différents types de bruits, puis de sélectionner pendant le test les modèles les plus proches des conditions de test. Cette sélection peut être réalisée sur un critère de maximisation de la vraisemblance des observations, ou plus simplement en reconnaissant le type d'environnement de test [Akbacak et Hansen, 2007].

Une extension de ces approches est de combiner linéairement les modèles appris dans différentes conditions, plutôt que d'en sélectionner un seul, afin de former un nouvel ensemble de modèles adaptés aux conditions ou au locuteur de test. C'est le cas des méthodes de *Cluster Adaptive Training* [Gales, 2000] et de voix propres [Kuhn *et al.*, 2000] dont nous avons déjà parlé.

La combinaison parallèle de modèles (*Parallel Model Combination* [Gales et Young, 1993]) est encore une autre méthode permettant de construire des modèles bruités à partir d'un modèle de parole propre et d'un modèle de bruit seul, qui peut éventuellement être estimé sur le corpus de test. Cette méthode suppose l'additivité des différentes sources de bruit dans le spectre de puissance (voir l'équation 2.4).

L'algorithme revient alors à construire un modèle de Markov caché équivalent aux deux modèles initiaux (celui de parole et celui de bruit), comme celui représenté sur la figure 3.3.

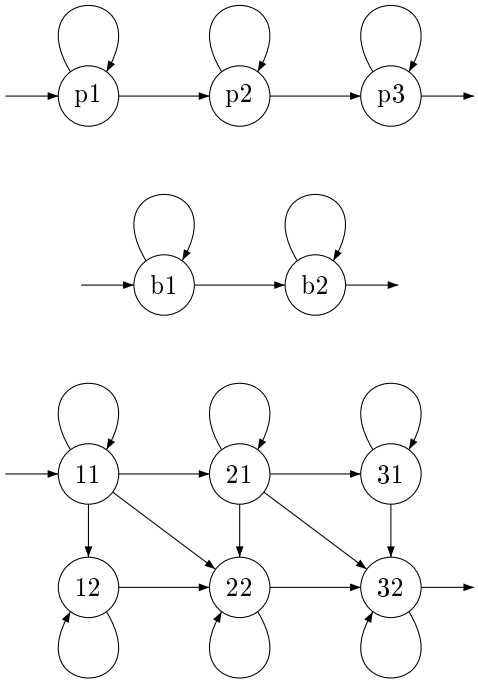


FIG. 3.3 – Exemple de combinaison parallèle de deux modèles ;
en haut : HMM de parole à 3 états ;
au milieu : HMM de bruit à 2 états ;
en bas : HMM combiné de parole et de bruit : les deux chiffres indiquent respectivement l'état de parole et l'état de bruit.

La topologie du modèle combiné est entièrement définie, ainsi que les probabilités de transition entre les états. Il reste à définir les fonctions de densité de probabilité d'émission des observations. Pour les modèles de parole et de bruit, celles-ci prennent généralement la forme d'un mélange de gaussiennes dans le cepstre. L'équation 2.4 définissant le résultat d'une telle combinaison dans le spectre de puissance, il faut tout d'abord projeter les densités de probabilité dans le domaine du spectre de puissance, puis les combiner, et enfin revenir au domaine cepstral. Cette suite de transformation est schématisée sur la figure 3.4. Cette transformation altère la forme normale des densités de probabilité et différentes approximations ont été proposées pour construire les lois de probabilité du modèle combiné [Gales et Young, 1993].

Les principaux problèmes posés par la combinaison parallèle de modèles sont les suivants :

- Un modèle du bruit ainsi que sa puissance relative par rapport à la parole doit être connu ou estimé.
- Il est nécessaire de (re)transformer les modèles acoustiques dans un espace dans lequel la contribution du modèle de bruit peut être ajoutée à celle du modèle de parole. Cette transformation est souvent coûteuse, et n'est parfois tout simplement pas réalisable, par exemple lorsque les paramètres acoustiques sont calculés par normalisation ou par projection (comme avec l'analyse discriminante linéaire ou LDA).
- Il n'est possible que d'ajouter du bruit aux modèles, et non d'en retirer. Ceci implique notamment que toutes les trames du signal de test soient plus bruitées que les modèles de parole. Si ce n'est pas le cas, la différence entre les conditions d'apprentissage et de test ne peut plus être réduite.

Pour pallier ces problèmes, les auteurs de [Zhao et Kleijn, 2007] proposent une extension de PMC dans laquelle les densités d'émissions dans les états sont combinés numériquement et le facteur de gain qui modélise la puissance relative de la parole et du bruit est estimé dynamique sur le test. D'autres techniques ont également été proposées pour bruiter (ou débruiter) les modèles directement dans le domaine cepstral, comme l'adaptation jacobienne (voir paragraphe 3.4) ou l'adaptation par séries de

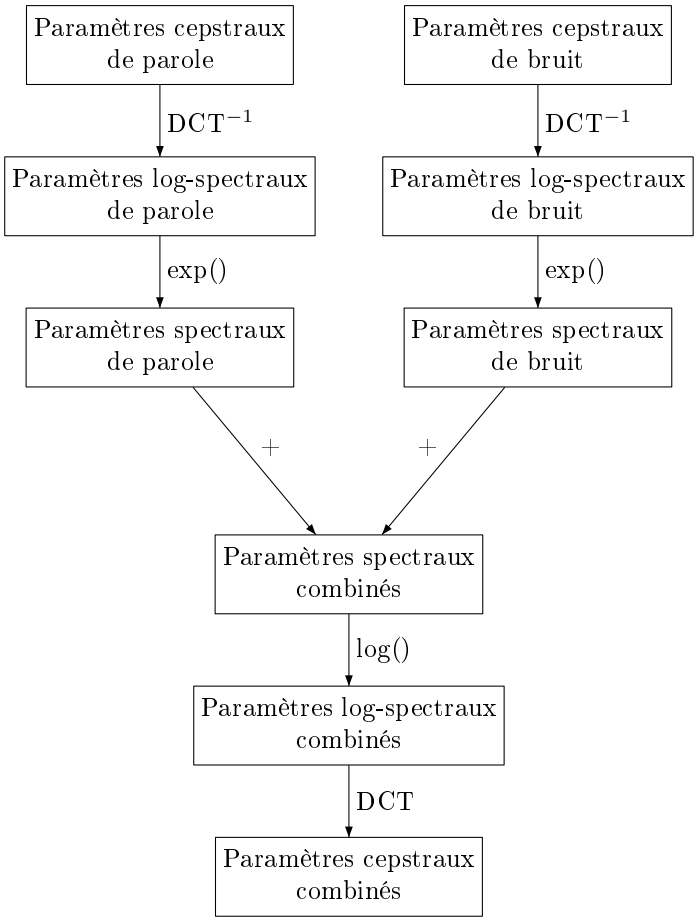


FIG. 3.4 – Combinaison des probabilités d'émission de deux états des modèles de parole et de bruit.

Taylor [Moreno *et al.*, 1996].

3.4 Adaptation jacobienne

L'adaptation jacobienne approche la transformation non linéaire du domaine cepstral vers le domaine spectral par une transformation linéaire, ce qui permet de réduire considérablement la complexité et donc le coût de l'adaptation [Sagayama *et al.*, 1997].

Soit F la matrice en cosinus discrète, dont les coefficients sont

$$F_{i,j} = \frac{1}{\sqrt{2N_{filt}}} \cos\left(\frac{i(j+0.5)\pi}{N_{filt}}\right)$$

avec N_{filt} le nombre de filtres triangulaires de l'analyse spectrale, c'est-à-dire le nombre de coefficients spectraux. Soit $C(\cdot)$ la transformation cepstrale d'un vecteur spectral X :

$$C(X) = F \cdot \log(X)$$

Soit S un vecteur spectral de parole. Supposons que la contribution spectrale du bruit à l'apprentissage et au test soit respectivement modélisée par un simple vecteur N^{ref} et N^{tar}. L'adaptation classique, au sens de PMC, est alors :

$$C(S + N^{tar}) = F \cdot \log\left(\exp\left(F^{-1} \cdot C(S + N^{ref})\right) + N^{tar} - N^{ref}\right)$$

Cette adaptation est coûteuse car elle requiert au minimum un exponentiel pour chaque vecteur moyenne. En appliquant l'approximation jacobienne à la fonction $f(C(N)) = C(S+N)$, nous avons :

$$C(S+N^{tar}) \simeq C(S+N^{ref}) + J_S \cdot \left(C(N^{tar} - C(N^{ref})\right) \tag{3.3}$$

où J_S est la matrice jacobienne :

$$\begin{aligned} J_S &= \frac{\partial C(S+N^{ref})}{\partial C(N^{ref})} \\ &= \frac{\partial C(S+N^{ref})}{\partial \log(S+N^{ref})} \frac{\partial \log(S+N^{ref})}{\partial (S+N^{ref})} \frac{\partial (S+N^{ref})}{\partial N^{ref}} \frac{\partial N^{ref}}{\partial \log(N^{ref})} \frac{\partial \log(N^{ref})}{\partial C(N^{ref})} \\ &= F \frac{N^{ref}}{S+N^{ref}} F^{-1} \end{aligned} \tag{3.4}$$

L'équation 3.3 permet d'adapter les modèles acoustiques à moindre coût à condition que $C(N^{tar}) - C(N^{ref})$ soit petit. J'ai proposé d'étendre le domaine de validité de cette adaptation linéaire en estimant différemment la pente J_S de la tangente d'adaptation, tout en minimisant l'impact sur le voisinage de $C(N^{ref})$. Deux solutions sont proposées dans [Cerisara *et al.*, 2004], respectivement l'adaptation alpha-jacobienne statique et dynamique, la seconde se différenciant de la première entre autre par le fait que la pente de la tangente d'adaptation est ré-estimée à chaque phrase. L'adaptation dynamique remplace J_S par $J'_S = \alpha_S J_S$, la valeur optimale de α étant

$$\tilde{\alpha}_S = \frac{C(S+N^{tar}) - C(S+N^{ref})}{J_S(C(N^{tar}) - C(N^{ref}))} \tag{3.5}$$

qui donne exactement la même adaptation que PMC.

J'ai proposé dans [Cerisara *et al.*, 2004] plusieurs algorithmes permettant de régler précisément le compromis entre l'approximation de $\tilde{\alpha}_S$ et le coût de l'adaptation. Le plus efficace est réalisé en regroupant plusieurs (α_S) au sein d'une même classes et en calculant un unique $\tilde{\alpha}_S$ optimal par classe. Les deux adaptations, statiques et dynamiques, sont également étendues aux bruits additifs et convolutifs, avec :

$$C(H^{tar} \times S + N^{tar}) \simeq C(H^{ref} \times S + N^{ref}) + C\left(\frac{H^{tar}}{H^{ref}}\right) + J_S\left(C(N^{tar}) - C(N^{ref}) - C\left(\frac{H^{tar}}{H^{ref}}\right)\right)$$

où H^e représente le bruit convolutif de l'environnement e. La précision et la complexité de l'adaptation approchée et exacte sont comparées sur deux corpus et montrent la supériorité de l'adaptation proposée, qui permet notamment de choisir le meilleur compromis précision/coût en fonction des besoins. En plus de ses performances très intéressantes, l'adaptation alpha-jacobienne répond à certaines faiblesses de PMC décrites au paragraphe 3.3.2, comme la possibilité d'enlever du bruit des modèles acoustiques.

Les principaux résultats concernant l'adaptation des modèles aux bruits additifs et convolutifs sur le corpus VODIS sont présentés figure 3.5.

Nous voyons qu'en jouant sur le paramètre du nombre de clusters de l'adaptation alpha-jacobienne, il est possible de régler le compromis entre complexité et précision de l'adaptation. Nous établissons dans [Cerisara *et al.*, 2004] le gain en complexité de l'adaptation jacobienne comparé à l'adaptation non linéaire exacte : dans le cas représenté sur la figure 3.5, la complexité de l'adaptation dynamique est inférieure à celle de l'adaptation exacte dès 5 clusters et moins.

3.5 Contributions

J'ai beaucoup travaillé au cours de ces dernières années dans le domaine de l'adaptation des modèles acoustiques, et mes contributions y sont donc nombreuses. Je résume simplement ici les plus importantes d'entre elles.

Mes deux principales contributions concernent les modèles multi-bandes, et l'adaptation jacobienne. Mes autres contributions sont les modèles autorégressifs multi-échelles [Cerisara et Daoudi, 2001], et plusieurs études originales réalisées dans le domaine de l'adaptation [Cerisara *et al.*, 2001, Cerisara et Illina, 2003]

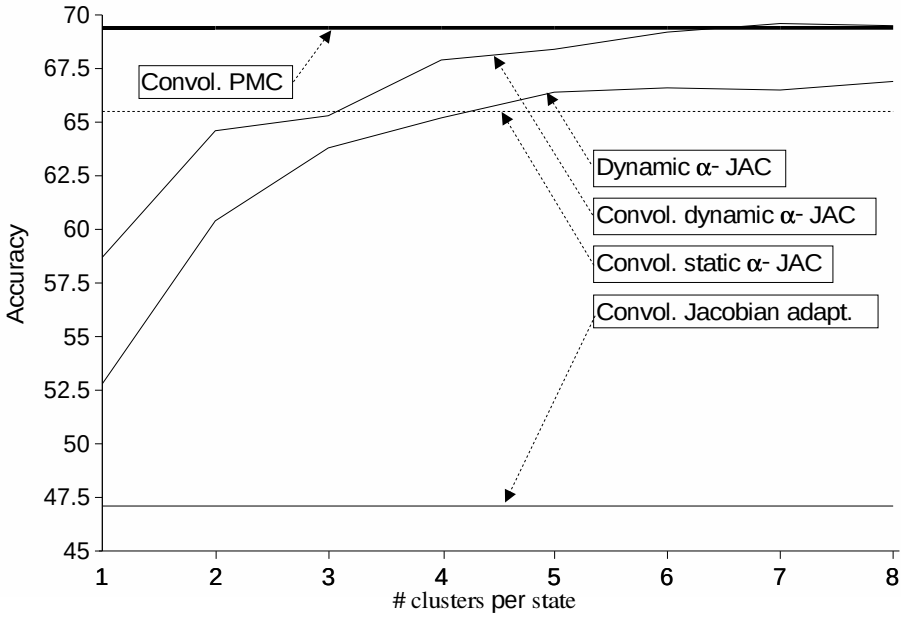


FIG. 3.5 – Taux de reconnaissance des algorithmes d'adaptation jacobienne aux bruits additif et convolutif sur le corpus VODIS, en fonction du nombre de classes de gaussiennes, nombre de classes qui sert à régler le compromis entre la précision de l'adaptation et son coût.

et qui m'ont permis de mieux comprendre les bases et problèmes fondamentaux de ce domaine, notamment en vue de préparer mon travail sur l'adaptation jacobienne et la reconnaissance avec données manquantes.

L'ensemble de mes contributions concernant les modèles multi-bandes sont détaillées dans ma thèse. Elles concernent par exemple la proposition d'une nouvelle architecture multi-bandes basée sur les modèles de Markov du second ordre, l'apprentissage global de l'ensemble du système par minimisation de l'erreur de classification, l'ajout du spectre complet aux côtés des sous-bandes, la conception de nouvelles unités phonétiques adaptées à l'information acoustique de chaque bande, l'application de ce principe aux tâches de reconnaissance de la parole et d'identification de la langue, etc.

En ce qui concerne l'adaptation jacobienne, mes contributions principales sont la proposition de deux modifications de l'algorithme d'adaptation de base permettant d'étendre la validité de l'approximation jacobienne au-delà du voisinage des conditions d'apprentissage, et de choisir précisément le meilleur compromis entre la précision de l'adaptation et son coût en fonction des besoins. J'ai également étendu ces deux méthodes aux bruits additifs et convolutifs, et j'ai montré l'intérêt de ces approches, à la fois en terme d'amélioration du taux de reconnaissance et de réduction des coûts, sur deux tâches de reconnaissance automatique de la parole.

Chapitre 4

Reconnaissance avec données manquantes

4.1 Introduction et principes fondamentaux

4.1.1 Introduction

Plutôt que de bruiter les modèles ou débruiter le signal, la reconnaissance avec données manquantes suppose que les modèles de parole ne peuvent pas représenter tout le signal perçu, mais seulement une partie de celui-ci, le reste relevant de la contribution d'une ou plusieurs autres sources de bruits.

La séparation entre les observations qui relèvent de la parole et celles du bruit est réalisée dans le domaine spectral au moyen d'un masque qui, dans sa forme la plus simple, affecte à chaque coefficient spectral une valeur binaire selon qu'elle représente principalement la parole ou le bruit.

4.1.2 Motivations et justifications

Cette hypothèse de masquage d'un son par un autre est fondée sur des observations psycho-acoustiques [Moore, 1982]. De plus, les expériences de bruitage artificiel montrent que, en pratique, la contribution d'une seule source sonore est dominante pour la plupart des coefficients spectraux, la contribution des autres sources étant souvent négligeable. Bien que cela soit peu fréquent, lorsque plusieurs sources sonores contribuent de manière non négligeable au même coefficient, le fait de masquer un tel coefficient détruit effectivement une partie de l'information de la reconnaissance. Toutefois, dans ce cas, l'information acoustique est corrompue, et il n'est pas certain, même avec des techniques de débruitage et d'adaptation, de pouvoir récupérer seulement l'information utile. En fait, nous pouvons penser qu'il est préférable de masquer ce coefficient, car il est trop risqué de l'utiliser en reconnaissance, d'autant plus que l'information acoustique est largement redondante dans le spectre : certaines expériences montrent en effet que l'on peut masquer aléatoirement, en l'absence de bruit, jusqu'à 80 % du spectre sans dégrader le taux de reconnaissance.

Toutes ces considérations motivent l'utilisation de masques pour la reconnaissance, mais il existe une dernière motivation qui justifie encore plus clairement le paradigme de la reconnaissance avec données manquantes : il s'agit d'une expérience de reconnaissance, réalisée dès les premières recherches dans ce domaine, et qui s'appuie sur des masques oracles, c'est-à-dire des masques qui ont été calculés en connaissant précisément la contribution de la parole et du bruit pour chaque coefficient spectral. Les taux de reconnaissance alors obtenus sont excellents même en conditions très bruitées [Cooke *et al.*, 2001, Morris *et al.*, 2001a]. Bien entendu, le défi reste entier, car il faut pouvoir maintenant estimer les masques avec pour seule information disponible le signal bruité, mais ces expériences prouvent néanmoins que la redondance de l'information acoustique est très souvent suffisante pour compenser le fait de masquer certaines parties du spectre, même en conditions bruitées.

4.1.3 Définition d'un masque

Un masque peut être défini formellement par la fonction M suivante :

$$M : (\mathbb{N}, \mathbb{N}, \Omega) \rightarrow D \tag{4.1}$$

où les deux premiers paramètres sont respectivement l'instant t et la bande fréquentielle f, tous deux étant supposés discrétisés, et Ω est l'ensemble des informations permettant d'estimer le masque. Lorsque Ω contient l'information exacte du SNR local, alors nous avons des masques oracles, mais dans le cas le plus courant, Ω est simplement le signal observé. Il est cependant possible d'ajouter dans Ω d'autres informations, par exemple des contraintes concernant l'étendue minimale et maximale d'une zone masquée. Un cas particulier existe, lorsque nous contraignons tous les coefficients d'un même vecteur spectral à partager la même valeur de masque : ce cas est l'objet d'un domaine de recherche à part entière, celui dit de la parole utile, ou *usable speech* [Yantorno *et al.*, 2003]. Il est également utilisé dans certains algorithmes basés sur la détection de la parole [Wang, 2004].

D est le domaine du masque. Il peut être discret ou continu :

- $D = \{0, 1\}$: ces masques binaires sont appelés dans la littérature *hard masks*.
- $D = [0, 1]$: ces masques réels sont appelés *soft masks*. Ils représentent la probabilité qu'un coefficient soit masqué.

4.1.4 Utilisation des masques

Il existe deux grandes familles d'approches pour effectuer la reconnaissance automatique de la parole avec des données masquées : l'imputation et la marginalisation des données.

B. Raj a beaucoup travaillé au CMU sur l'imputation des données [Raj, 2000]. Il a proposé plusieurs approches d'imputation, dont l'imputation conditionnée par la classe, qui reconstruit le signal de parole des coefficients masqués pour chaque hypothèse de reconnaissance en maximisant la probabilité a posteriori des observations reconstruites sachant les données non masquées et l'état du HMM aligné avec la trame courante. Il a aussi proposé la reconstruction géométrique, qui interpole les valeurs manquantes à partir des données observées précédent et suivant (en temps et en fréquence) les coefficients masqués par une fonction polynomiale ou un rapport de polynômes. La reconstruction peut également être probabiliste, auquel cas des mixtures de gaussiennes modélisant la parole sont utilisés pour inférer les données masquées, ou statistique, lorsque le valeurs manquantes sont estimées au moyen de la corrélation croisée entre les coefficients masqués et observés. Toutes ces techniques d'imputation peuvent également être améliorées pour prendre en compte des informations a priori concernant le signal de parole propre, comme l'énergie maximale de ce signal qui ne peut être supérieur à l'énergie observée.

Les approches de marginalisation ont été essentiellement développées à l'Université de Sheffield. Le principe de base revient à considérer que les observations masquées ne sont plus scalaires, mais deviennent des variables aléatoires et sont représentées par des densités de probabilité. En général, dans le cadre de la reconnaissance avec données manquantes, des densités uniformes sont utilisées, ce qui signifie que la contribution de la parole pour ces coefficients est supposée totalement inconnue, sous les contraintes toutefois que l'énergie de la parole soit comprise entre 0 et l'énergie observée. Ainsi, la log-vraisemblance $P([Y_m(t), y_{\bar{m}}(t)]|\lambda)$ des observations complètes (masquées $Y_m(t)$ ou non $y_{\bar{m}}(t)$) est calculée sur l'espérance des valeurs manquantes, ce qui revient à l'équation suivante en supposant les matrices de covariance diagonales :

$$P([Y_m(t), y_{\bar{m}}(t)]|\lambda) = P(y_{\bar{m}}(t)|\lambda)\frac{1}{y_m(t)}\int_{x=0}^{y_m(t)} P(x|\lambda)dx \tag{4.2}$$

D'autres formes de marginalisation peuvent être obtenues, par exemple en bornant ou non l'intégrale, en utilisant des relations approchées dans le cas des matrices non diagonales, ou alors en utilisant des distributions uniformes également pour la parole non masquée. Ces différentes marginalisations et leur dérivation sont détaillées notamment dans [Cooke *et al.*, 2001, Demange, 2007]. Il est également montré dans [Morris *et al.*, 1998] que la marginalisation est une solution optimale au sens de Bayes en présence de données manquantes.

4.2 Estimation des masques

Les masques sont en général estimés avant la reconnaissance, mais ils peuvent également l'être au cours de la reconnaissance. Dans ce dernier cas, il est possible d'estimer d'abord plusieurs ensembles de masques possibles, puis de laisser la reconnaissance décider du meilleur chemin parmi les masques possibles en maximisant la vraisemblance des observations masquées. J'ai proposé un système basé sur ce principe qui utilise un HMM ergodique pour modéliser l'ensemble des masques possibles et qui combine ce modèle avec celui de la parole [Cerisara et Illina, 2003]. Le décodeur multi-sources proposé dans [Barker *et al.*, 2005, Coy et Barker, 2007] est également basé sur ce principe, et constitue probablement aujourd'hui un des meilleurs systèmes de reconnaissance de la parole avec données manquantes. Dans ce système, des fragments de parole sont d'abord détectés par une approche ascendante, puis ils sont utilisés comme masques potentiels pendant la reconnaissance, qui choisit la meilleure combinaison de ces fragments au sens de la vraisemblance des observations masquées. Certains systèmes de séparation de sources présentés au paragraphe 2.3.4, tel que celui développé par Roweis [Roweis, 2003], combinent également deux modèles pour sélectionner pendant la reconnaissance les données masquées ou non. Toutefois, mis à part certains cas particuliers comme celui du mélange de deux locuteurs connus dans [Roweis, 2003], il n'est pas possible de simplement tester toutes les combinaisons possibles de masques, car l'espace des combinaisons possibles est de taille exponentielle. De plus, les vraisemblances obtenues avec des masques différents sont difficilement comparables, même si les normalisations introduites dans [Coy et Barker, 2007] permettent de le faire, dans une certaine mesure.

Il est donc important de disposer dans le cas général d'une analyse indépendante de la reconnaissance afin de construire soit des masques fixés a priori, soit des masques potentiels qui seront résolus pendant la reconnaissance. Pour ce faire, les méthodes CASA décrites au paragraphe 2.3 ont d'abord été utilisées, avec un succès relatif [Brown *et al.*, 2001]. Notons de même que, dans le cadre du rapprochement entre les approches de reconnaissance avec données manquantes et multi-bandes discuté au paragraphe 3.2, les méthodes CASA ont également été exploitées dans des systèmes multi-flux pour estimer la fiabilité respective des différentes combinaisons possibles [Glotin *et al.*, 1998]. Les techniques de séparation de sources sont également parfois exploitées pour construire les masques [Christensen *et al.*, 2007]. Mais les approches les plus utilisées sont basées sur des méthodes d'analyse du signal, et en particulier d'estimation du SNR local et d'harmonicité. En fait, tout algorithme permettant d'estimer la contribution du bruit à un instant donné et dans chaque bande fréquentielle peut servir à générer des masques, en comparant simplement cette contribution avec l'énergie observée [Renevey, 2001, Dupont et Ris, 2001]. En supposant que la parole est harmonique, il est possible de détecter les échelles harmoniques pour lesquelles l'énergie est la plus grande et de masquer les autres fréquences [Barker *et al.*, 2001, van Hamme, 2004b]. Enfin, d'autres indices, comme la modulation d'amplitude et de fréquence [Tchorz et Kollmeier, 2002], ou encore les réseaux de neurones [Potamitis *et al.*, 2000] peuvent être calculés à des fins d'identification des données corrompues.

Tous ces indices extraits par des techniques de traitement du signal peuvent être considérés ensuite comme des paramètres dont la distribution peut être apprise par un modèle stochastique. Cette approche basée sur la modélisation a été développée en particulier au CMU, d'abord dans [Seltzer *et al.*, 2004], puis dans [Kim *et al.*, 2005b] et [Kim et Stern, 2006]. Les premiers paramètres utilisés étaient simplement les énergies spectrales [Raj, 2000], puis des paramètres spécifiques au problème considéré ont été proposés [Seltzer, 2000]. Ces paramètres sont :

- Le *comb-filter ratio*, qui représente l'énergie des harmoniques ;
- Le rapport de "pic d'autocorrélation", qui mesure le degré de périodicité du signal ;
- Le rapport de l'énergie en sous-bande sur l'énergie totale, qui encode la forme spectrale globale ;
- Le kurtosis, qui mesure la gaussianité du signal ;
- La profondeur des vallées spectrales, qui est liée au SNR ;
- Le rapport entre l'énergie et le seuil de bruit en sous-bande, qui est basée sur une estimation du seuil de l'énergie du bruit ;
- L'estimation du SNR basée sur la soustraction spectrale.

Un état de l'art de l'ensemble des techniques permettant d'estimer les masques des données manquantes est réalisé dans [Cerisara *et al.*, 2007].

4.3 Reconnaissance dans le domaine cepstral

Toutes les techniques de reconnaissance avec données manquantes requièrent un espace acoustique dans lequel les contributions respectives de la parole et du bruit sont séparables. C'est pour cette raison que la reconnaissance avec données manquantes a été réalisée principalement dans le domaine spectral ou avec des modèles simulant l'oreille interne [Cooke, 1993]. Cependant, nous avons vu que d'autres paramétrisations, comme le cepstre ou l'ETSI AFE, sont beaucoup plus robustes au bruit que le spectre. Les premiers travaux explorant la possibilité d'appliquer la reconnaissance avec données manquantes dans le cepstre sont [Häkkinen et Haverinen, 2001, Renevey, 2001].

Les approches basées sur l'imputation des données permettent d'apprendre des modèles acoustiques cepstraux, car l'utilisation des masques pour estimer les coefficients débruités est découplée du calcul de la vraisemblance acoustique [Raj, 2000]. Mais nous avons également vu que l'estimation des masques est meilleure lorsqu'elle est réalisée conjointement avec le processus de décodage. J'ai donc étudié dans [Cerisara, 2003] une approche dans laquelle les modèles acoustiques cepstraux sont transformés en modèles log-spectraux. Les deux ensembles de modèles, cepstraux et log-spectraux sont conservés en mémoire, et les masques sont appliqués dans le log-spectre. Les observations masquées sont alors remplacées par la moyenne des modèles log-spectraux avec lesquels elles sont alignées, ce qui correspond à la technique dite d'imputation conditionnée par la classe. Une fois débruités, ces vecteurs sont transformés à nouveau dans le domaine cepstral, où ils sont comparés aux modèles acoustiques. C'est aussi la première fois qu'une approche avec données manquantes est appliquée sur une tâche grand vocabulaire.

Simultanément, dans [van Hamme, 2003], Hugo van Hamme propose une solution basée sur un principe similaire, mais qui diffère de la mienne sur les points suivants :

- Tandis que je propose de conserver deux ensembles de modèles en mémoire, un log-spectral et un cepstral, [van Hamme, 2003] impute directement les données manquantes depuis les modèles cepstraux en maximisant la vraisemblance des observations cepstrales sous la contrainte que le spectre de la parole est inférieur au spectre observé. En d'autres termes, mon approche décompose l'imputation en deux processus indépendants : la transformation du domaine cepstral au domaine log-spectral, puis l'imputation des données manquantes, tandis que [van Hamme, 2003] se passe de cette hypothèse d'indépendance et travaille depuis le domaine cepstral ;
- Je propose de lisser l'imputation des données en introduisant deux paramètres correctifs qui limitent le nombre de coefficients masqués et l'impact de l'imputation.

Le coût de l'approche proposée dans [van Hamme, 2003] est par ailleurs considérablement réduit lorsque les paramètres PROSPECT sont utilisés à la place du cepstre [van Hamme, 2004a]. Un vecteur de paramètres PROSPECT $p = \begin{bmatrix} c \\ d \end{bmatrix}$ est composé des vecteurs cepstraux classiques c et du résiduel log-spectral $d = s - C'_K c$, où s est le vecteur log-spectral et C_K est la matrice en cosinus effectuant un lissage cepstral sur K paramètres. La vraisemblance d'une mixture gaussienne est alors :

$$f(p|\mu, \Sigma) = \mathcal{N}(c|\mu_c, \Sigma_c)\mathcal{N}(d|\mu_d, \Sigma_d)^\beta$$

qui se réduit dans le log-spectre à une loi normale de covariance inverse :

$$C'_K \Sigma_c^{-1} C_K + \beta P_{\frac{1}{K}} \Sigma_d^{-1} P_{\frac{1}{K}}$$

où $P_{\frac{1}{K}} = I_D - C'_K C_K$. Ainsi, il devient possible d'appliquer des masques spectraux sur ces paramètres "cepstraux". Les paramètres PROSPECT sont également appliqués avec la normalisation de la longueur du conduit vocal dans [Jansen et van Hamme, 2005], afin de masquer les informations spectrales perdues lors de la réduction de la longueur du conduit vocal.

Une autre approche pour appliquer la reconnaissance avec données manquantes dans le cepstre est décrite dans [Ge et Song, 2004], où une mesure de confiance pour chaque coefficient est estimée à partir du SNR. Cette confiance est utilisée pour pondérer la contribution d'un coefficient au calcul de la log-vraisemblance des observations. Cette pondération, réalisée dans le domaine log-spectral, est ensuite convertie dans le domaine cepstral par la transformation en cosinus classique. De même, un SNR pour les coefficients dynamiques est défini, mais ce SNR est un artifice pour permettre de masquer les dérivées des coefficients cepstraux.

Une autre approche intéressante est celle décrite dans [Srinivasan et Wang, 2007] : les auteurs proposent d'utiliser les similarités entre la reconnaissance avec données manquantes et le décodage incertain pour appliquer conjointement des masques spectraux et des modèles cepstraux. Leur idée est de transformer les masques binaires calculés dans le spectre en fonctions de densité de probabilité (pdf) qui représentent l'incertitude des coefficients dans le cepstre. Ces pdf sont alors utilisées au cours de la reconnaissance pour marginaliser la vraisemblance des observations dans le cadre du décodage incertain.

J'ai récemment proposé une nouvelle approche pour appliquer la reconnaissance avec données manquantes dans le domaine cepstral, qui optimise le taux de reconnaissance plutôt que le rapport signal-sur-bruit pour calculer des masques cepstraux. Cette approche est détaillée au paragraphe 4.7.

4.4 Modèles de masques dépendant de l'environnement

J'ai travaillé avec Sébastien Demange dans le cadre de sa thèse de doctorat sur la conception de modèles de masques dépendant de l'environnement. L'objectif est d'isoler la variabilité des modèles due au bruit en explicitant leur dépendance par rapport aux conditions environnementales. La variable explicative, qui classifie les différentes conditions bruitées en plusieurs types, est inférée pendant le test par un GMM modélisant l'environnement :

$$p(y_c|e_k) = \sum_i w_i \mathcal{N}(y_c; \mu_{e_k,i}, \Sigma_{e_k,i})$$

où e_k représente un environnement possible et y_c les trames cepstrales de parole bruitée. Nous pouvons alors estimer la probabilité que chaque environnement e_k vu à l'apprentissage se réalise pendant le test :

$$p(e_k|y_c) = \frac{p(y_c|e_k)p(e_k)}{\sum_j p(y_c|e_j)p(e_j)}$$

Les environnements (e_k) $\forall k$ sont supposés équiprobables a priori :

$$p(e_k) = p(e_j) \qquad \forall (k,j)$$

Cette étape de détection de l'environnement est semblable à ce qui est réalisé dans [Akbacak et Hansen, 2007].

Une fois la probabilité de chaque environnement connue, nous avons proposé deux approches pour estimer la probabilité $p(m_i|y_c)$ que la bande fréquentielle i soit masquée. La première approche utilise une dérivation bayésienne classique :

$$p(m_i|y_c) = \sum_k p(m_i|y_c, e_k)p(e_k|y_c)$$

tandis que la seconde approche la somme par l'environnement d'apprentissage le plus proche :

$$\begin{aligned} \hat{e}_k &= \arg\max_{e_k} p(e_k|y_c) \\ p(m_i|y_c) &= p(m_i|y_c, \hat{e}_k) \end{aligned}$$

Dans les deux cas, nous avons besoin d'un modèle de masque dépendant de l'environnement qui calcule $p(m_i|y_c, e_k)$. Ce modèle est constitué d'un GMM par environnement qui retourne :

$$p(y_c|m_i, e_k) = \sum_j \omega_{i,k,j} \mathcal{N}(y_c; \nu_{i,k,j}, Q_{i,k,j})$$

La probabilité a posteriori du masque sachant l'environnement est alors simplement calculée comme suit :

$$p(m_i|y_c, e_k) = \frac{p(y_c|m_i, e_k)p(m_i|e_k)}{p(y_c|m_i, e_k)p(m_i|e_k) + p(y_c|\bar{m}_i, e_k)p(\bar{m}_i|e_k)}$$

où m_i et $\bar{m}_i$ sont respectivement les notations utilisées pour indiquer que la bande fréquentielle i est masquée et non masquée. $p(m_i|e_k)$ et $p(\bar{m}_i|e_k)$ sont les probabilités a priori qu'un coefficient soit masqué dans un environnement donné, ce qui peut s'interpréter comme le masque "moyen" pour un type de bruit.

4.5 Estimation contextuelle des masques

Les modèles de masque tels que ceux présentés au paragraphe précédent ou dans [Kim et Stern, 2006] supposent l'indépendance des masques d'une trame à l'autre et d'une bande fréquentielle à l'autre. Cette hypothèse simplificatrice est courante car elle réduit considérablement la dimension de l'espace modélisé et permet donc d'apprendre des modèles de meilleure qualité. Toutefois, la précision des modèles en souffre, car cette hypothèse est bien entendu erronée. Nous avons donc proposé une nouvelle approche permettant d'accroître la précision des modèles en prenant en compte le contexte, tout en limitant la taille de l'espace des masques.

La prise en compte du contexte est réalisée selon deux dimensions : le temps et les fréquences.

En ce qui concerne l'axe temporel, le contexte est pris en compte en construisant un HMM ergodique dont chaque état représente un masque possible à un instant donné. Les transitions entre les états encodent les dépendances temporelles $p(m_t|m_{t-1})$.

Les dépendances fréquentielles sont quant à elles considérées sur la totalité du spectre : un masque devient donc un vecteur spectral composé de N_{bande} coefficients booléens :

$$m_t = [m_1, \cdots, m_{N_{bande}}]^T$$

Toutefois, cette approche est inapplicable telle quelle, car la dimension de l'espace des masques croît exponentiellement avec le nombre de bandes (il y a $2^{N_{bande}}$ masques possibles). Nous avons cependant observé que le sous-ensemble des masques effectivement utilisés est clairsemé et localisé dans un petit nombre de régions dans cet espace. Nous avons donc créé des classes de masques centrées autour des K masques oracles les plus fréquents dans la base bruitée d'apprentissage. K est choisi selon un critère de couverture minimal. Les autres masques sont rapprochés de leur classe la plus proche selon la distance euclidienne.

Le tableau 4.1 montre le nombre de masques K et le taux de reconnaissance obtenu sur une partie du corpus d'apprentissage bruité de Aurora2 avec des modèles acoustiques spectraux 12 bandes en fonction du seuil de couverture α.

α	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
Nombre de classes	7	12	18	31	54	109	244	3 044
Tx de reco oracle	92.2	94.4	94.7	95.9	96.1	96.4	96.6	96.7

TAB. 4.1 – Nombre de masques après regroupement et taux de reconnaissance oracle en fonction du seuil de couverture α.

Ce tableau montre que les masques sont effectivement localisés dans l'espace total et que seulement 31 masques permettent de reproduire 80 % des masques réels, avec une dégradation plutôt faible du taux de reconnaissance (-0.8 %).

Il serait également intéressant de poursuivre les recherches dans cette voie en proposant une nouvelle définition de masque oracle, qui intègre un compromis entre l'amélioration du SNR et le rapport couverture sur nombre de représentants. Ainsi, il serait sans doute possible de réduire encore la perte en taux de reconnaissance pour un même nombre de masques.

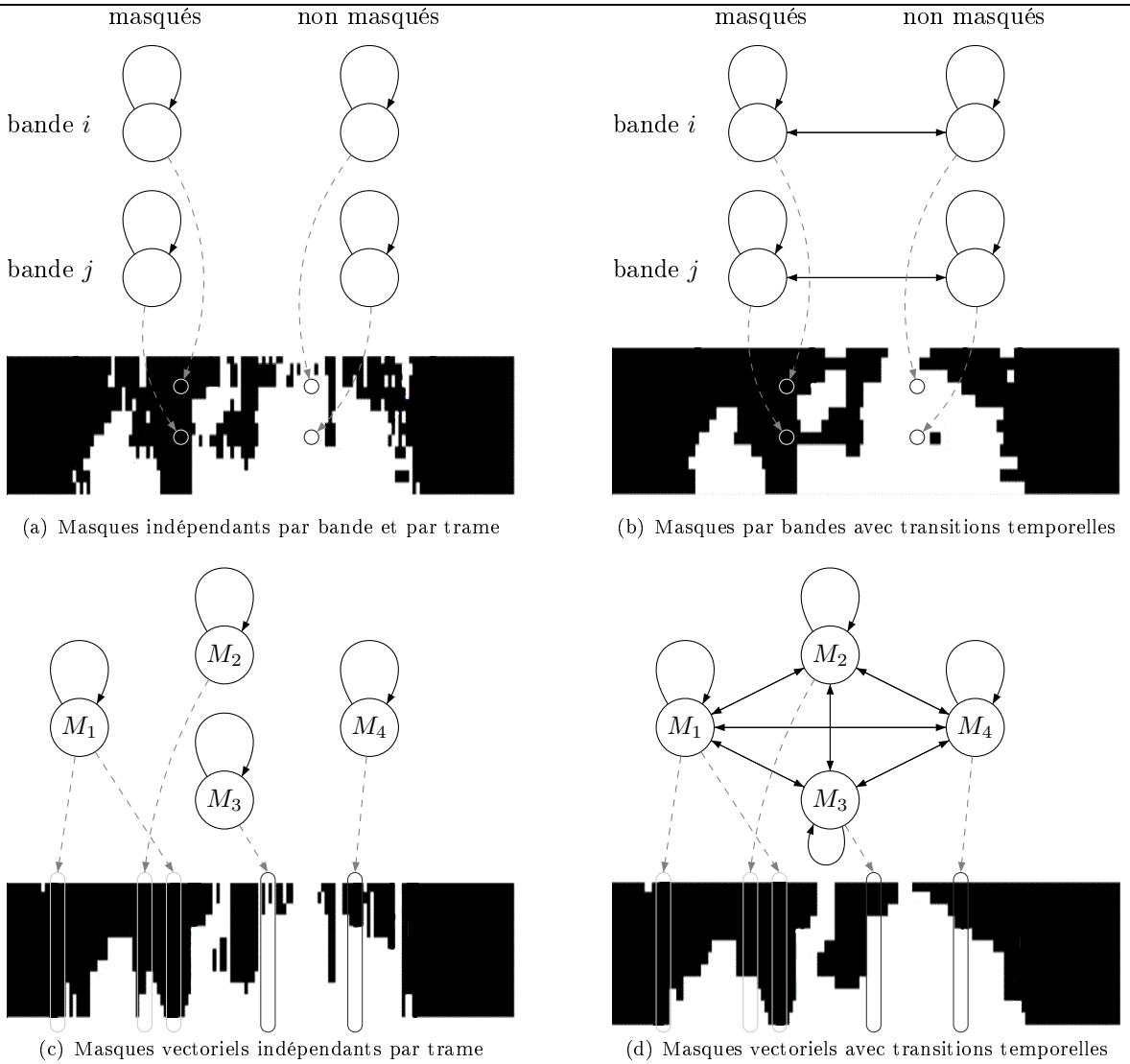
Les quatre modèles proposés, respectivement sans dépendance, avec dépendances temporelles, avec dépendances fréquentielles, et avec les deux types de dépendances sont schématisés sur la figure 4.1.

Le premier modèle, sans dépendance, affecte un masque au coefficient spectral $y^S_{t,i}$ en fonction de la vraisemblance des observations calculée avec les GMM par bande masqués $\lambda_{m,i}$ et non masqués $\lambda_{\bar{m},i}$:

$$m_{t,i} = 1 \text{ ssi } p(y^C_t|\lambda_{m,i})p(\lambda_{m,i}) > p(y^C_t|\lambda_{\bar{m},i})p(\lambda_{\bar{m},i}) \quad (4.3)$$

$$= 0 \text{ sinon} \quad (4.4)$$

Le deuxième modèle, avec dépendance temporelle, affecte un masque au coefficient spectral $y^S_{t,i}$ en fonction de la log-vraisemblance calculée par un décodage de Viterbi réalisé avec le HMM ergodique à 2



(a) Masques indépendants par bande et par trame

(b) Masques par bandes avec transitions temporelles

(c) Masques vectoriels indépendants par trame

(d) Masques vectoriels avec transitions temporelles

FIG. 4.1 – Quatre modèles de masque : indépendant du contexte (a), incluant le contexte temporel (b), fréquentiel (c), et les deux types de contexte (d).

états correspondant à chaque bande λ_i :

$$s_i = (s_{1,i}, \cdots, s_{T,i}) \tag{4.5}$$

$$s_i = \arg\max_s p(s|y^C, \lambda_i) \tag{4.6}$$

$$= \arg\max_s \prod_{t=1}^{t=T} p(s_t|y_t^C, \lambda_i) p(s_t|s_{t-1}, \lambda_i) \tag{4.7}$$

$$m_{t,i} = s_{t,i} \in \{0, 1\} \tag{4.8}$$

Le troisième modèle, avec dépendance fréquentielle, affecte un vecteur-masque m_t à la trame spectrale y_t^S en fonction du meilleur modèle GMM pour chaque masque possible λ_m :

$$m_t = \arg\max_m p(y_t^C|\lambda_m) p(\lambda_m) \tag{4.9}$$

Le quatrième modèle, avec les deux types de dépendance, affecte un vecteur-masque m_t à la trame spectrale y_t^S en fonction de la vraisemblance calculée par Viterbi avec le HMM ergodique global λ :

$$\hat{s} = \arg\max_s p(s|y^C, \lambda) \tag{4.10}$$

$$= \arg\max_s \prod_{t=1}^{t=T} p(s_t|y_t^C, \lambda) p(s_t|s_{t-1}, \lambda) \tag{4.11}$$

$$m_t = \hat{s}_t \in \{1, \cdots, K\} \tag{4.12}$$

Nous pouvons remarquer que les masques sont affectés à des coefficients spectraux $y_{t,i}^S$, car les contributions de plusieurs sources sonores doivent être séparables dans le domaine des masques, mais les modèles de masque sont entraînés sur des trames cepstrales y_t^C. Autrement dit, les modèles acoustiques sont spectraux tandis que les modèles de masque sont cepstraux.

En ce qui concerne les modèles acoustiques spectraux, nous avons déjà vu qu'un domaine de paramétrisation ayant des propriétés similaires au domaine spectral est un pré-requis pour pouvoir appliquer la reconnaissance avec données manquantes. Mais en ce qui concerne les modèles de masque, aucune contrainte n'est imposée quant au domaine à utiliser. Choisir le domaine cepstral présente plusieurs avantages : tout d'abord, nous supposons que l'ensemble du spectre apporte de l'information utile pour déterminer si une bande fréquentielle quelconque est masquée ou non. Il est donc préférable de considérer les vecteurs d'observation au complet pour tous les modèles de masque. En ce sens, même pour le premier modèle de la figure 4.1, le contexte fréquentiel de la variable observée (les observations) est pris en compte, mais le contexte fréquentiel de la variable cachée (les masques) ne l'est pas. De plus, le vecteur d'observation cepstral décorrèle les coefficients, ce qui permet d'utiliser sans aucun préjudice une matrice de covariance diagonale dans les GMM. Enfin, le domaine cepstral est réputé plus robuste au bruit que le domaine spectral.

Le tableau 4.2 résume les principaux résultats obtenus avec cette approche sur Aurora2 et Aurora4.

	Aurora2			Aurora4						
	test A	test B	test C	test 1	test 2	test 3	test 4	test 5	test 6	test 7
Système de référence	Cepstre : 12 MFCC + E + Δ + $\Delta\Delta$									
Sans bruit	61.1	55.6	66.7	85.1	34.8	30.8	36.9	27.7	30.6	26.8
Multi-conditions	87.3	85.5	83.1	79.4	76.8	65.6	59.9	61.8	65.3	58.7
Données manquantes	Log-spectre : 12 Mel spectral coefficients + Δ									
Sans dépendance	82.7	78.1	73.6	82.5	69.4	56.3	44.7	51.5	53.9	51.0
Dép. temporelles	85.1	81.1	77.9	83.1	72.7	59.2	49.9	53.8	55.7	53.0
Dép. fréquentielles	82.6	78.0	73.2	83.7	71.6	56.0	46.1	49.6	55.0	50.4
Dép. T + F	85.0	78.5	80.3	81.5	73.2	59.3	49.4	52.8	57.7	54.0
Oracle	96.2	96.6	93.9	83.1	79.1	70.7	66.6	64.9	73.1	64.8

TAB. 4.2 – Taux de reconnaissance sur Aurora2 (moyenne de 0 dB à 20 dB) et Aurora4. "Sans bruit" et "Multi-conditions" font référence respectivement aux modèles acoustiques entraînés sur le corpus d'apprentissage non bruité et multi-conditions. Ces deux systèmes opèrent dans le domaine cepstral tandis que les systèmes avec données manquantes fonctionnent dans le domaine log-spectral.

Nous voyons que les dépendances temporelles sont les plus importantes en présence de bruit, tandis que les résultats des dépendances fréquentielles seules sont beaucoup plus mitigées. Ceci est certainement dû à l'approximation réalisée lors de la discrétisation de l'espace des masques possibles qui compense le gain en performances obtenu en ajoutant l'information contextuelle fréquentielle. Toutefois, la combinaison de ces deux types de dépendance permet dans certains cas d'obtenir les meilleurs résultats, comme sur le test C de Aurora2 ou les tests 2, 3, 6 et 7 de Aurora4.

4.6 Optimisation de l'approche basée sur la marginalisation

Nous avons vu que l'approche classique de marginalisation suppose que les observations non masquées ne sont pas corrompues tandis que les observations masquées ont une énergie indéterminée comprise entre 0 et l'énergie observée (cf. équation 4.2). Lorsque le masque est basé sur le SNR, une bande fréquentielle est masquée lorsque son SNR est inférieur à 0 dB, et non masquée lorsque son SNR est supérieur à 0 dB (ou un autre seuil proche de 0 dB). Or, affirmer que le SNR est supérieur à 0 dB est strictement équivalent à dire que la contribution énergétique de la parole est comprise entre la moitié de l'énergie observée et cette dernière. Il est donc erroné de supposer que les observations non masquées ne sont pas corrompues, et il faut donc également les marginaliser. C'est ce qui a été proposé (mais pas testé) dans [Morris, 2001].

Formellement, soit $\gamma(\cdot)$ la fonction utilisée pour compresser le spectre de puissance : il s'agit généralement d'un logarithme ou d'une racine cubique. Lorsque le SNR vaut exactement 0 dB, l'énergie du bruit et de la parole sont toutes deux égales à la moitié de l'énergie observée :

$$X_i = \gamma(\gamma^{-1}(Y_i)/2)$$

Lorsque le SNR est inférieur à 0 dB (coefficient masqué), on a :

$$\gamma(0) \leq X_i \leq \gamma(\gamma^{-1}(Y_i)/2)$$

et lorsque le SNR est supérieur à 0 dB (coefficient non masqué), on a :

$$\gamma(\gamma^{-1}(Y_i)/2) \leq X_i \leq Y_i$$

Notons $Y_{i,snr0} = \gamma(\gamma^{-1}(Y_i)/2)$. L'équation de marginalisation devient alors :

$$P([Y_m(t), Y_{\bar{m}}(t)]|\lambda) = \frac{1}{Y_{m,snr0}(t)} \int_{x=0}^{Y_{m,snr0}(t)} P(x|\lambda)dx + \frac{1}{Y_{\bar{m}}(t) - Y_{\bar{m},snr0}(t)} \int_{x=Y_{\bar{m},snr0}(t)}^{Y_{\bar{m}}(t)} P(x|\lambda)dx \tag{4.13}$$

Dans cette approche, la même procédure est utilisée pour calculer la log-vraisemblance de tous les coefficients, qu'ils soient masqués ou non : seules les bornes de l'intervalle de marginalisation diffèrent.

Or, la largeur de l'intervalle de marginalisation représente notre incertitude concernant la véritable contribution de la parole dans le signal observé. Nous sommes donc face à un compromis : plus l'intervalle de marginalisation est grand, moins nous faisons d'erreur pour estimer les valeurs possibles de la véritable contribution de la parole, mais plus nous perdons d'information dans le calcul de la vraisemblance. Les intervalles de marginalisation proposés ci-dessus représentent uniquement les informations fournies par le module d'estimation des masques, et ne sont probablement pas optimaux du point de vue de la reconnaissance de la parole.

Nous avons donc proposé une nouvelle méthode pour calculer des bornes plus précises qui limitent la perte d'information due à la marginalisation. Pour cela, un modèle GMM M^k du rapport :

$$\frac{X_i}{Y_i} = \frac{1}{1 + 10^{-\frac{\text{SNR}_{\text{local}}}{20}}}$$

est entraîné. Ce modèle sert à discrétiser l'espace acoustique, et chaque gaussienne est un représentant d'une partie de cet espace. Il est ainsi possible d'estimer un masque plus précis pour chaque représentant, en considérant que la vraisemblance d'observer un rapport X_i/Y_i est :

$$P\left(\frac{X_i}{Y_i} = \alpha | M^k\right) = \mathcal{N}(\alpha; M^k)$$

Puisque nous supposons que cette densité de probabilité est modélisée par une gaussienne, alors nous savons que 95 % des occurrences de $\frac{X_i}{Y_i}$ sont à une distance inférieure à deux fois l'écart type de la moyenne, c'est-à-dire :

$$\mu_i^k - 2\sigma_i^k \leq \frac{X_i}{Y_i} \leq \mu_i^k + 2\sigma_i^k$$

Donc

$$Y_i(\mu_i^k - 2\sigma_i^k) \leq X_i \leq Y_i(\mu_i^k + 2\sigma_i^k) \tag{4.14}$$

Nous supposons que les 5 % d'occurrences de $\frac{X_i}{Y_i}$ qui sont en-dehors de cet intervalle ont une influence négligeable sur le taux de reconnaissance.

Nous avons ainsi défini de nouvelles bornes de marginalisation, qui sont beaucoup moins larges que les autres intervalles utilisés dans la littérature, dont la *bounded marginalisation*, qui utilise l'intervalle $[0, Y_i]$, ou l'intervalle défini dans l'équation 4.13.

Les figures 4.2 et 4.3, extraites de la thèse de Sébastien Demange, illustrent bien la différence entre les intervalles de marginalisation obtenus avec la méthode proposée et l'approche classique sur un exemple de HIWIRE. Nous voyons ainsi sur la figure 4.2 que les intervalles de marginalisation sont beaucoup plus précis que ceux de la figure 4.3, et malgré quelques erreurs, comme par exemple pour la trame 160 de la figure 4.2 à -5 dB, les intervalles proposés sont presque aussi souvent centrés autour de la véritable contribution de la parole que les masques classiques.

Les principaux résultats expérimentaux obtenus avec cette approche sont résumés dans le tableau 4.3.

		Clean	LN	MN	HN
ETSI AFE		3.6	16.2	34.8	93.0
Masques de référence (eq. 4.13)	Oracle	5.6	25.6	32.6	67.4
	Estimés	18.0	41.9	41.8	73.8
Masques proposés (eq. 4.14)	Oracle	10.8	9.8	12.1	19.7
	Estimés	10.8	14.8	24.4	72.1

TAB. 4.3 – Taux d'erreur en mots sur le corpus HIWIRE pour les masques proposés et les masques de référence. Les résultats obtenus avec un système de reconnaissance standard basé sur les paramètres ETSI AFE sont donnés pour comparaison.

Nous voyons dans ce tableau que les nouveaux intervalles de marginalisation donnent des résultats très intéressants sur cette expérience : la diminution du taux d'erreur par rapport aux masques estimés classiques est particulièrement élevée lorsque le niveau de bruit est faible. Les performances obtenues sont même meilleures qu'avec la paramétrisation ETSI AFE, mais ceci peut également s'expliquer par le fait que cette paramétrisation est moins efficace sur le corpus HIWIRE que sur le corpus Aurora2, comme cela a été discuté dans l'introduction.

4.7 Optimisation du taux de reconnaissance

La principale limitation de la reconnaissance avec données manquantes est sa dépendance au domaine de paramétrisation : en effet, cette approche n'est applicable que sur certains domaines bien définis, comme le spectre, les ondelettes, des domaines fondés sur un modèle d'audition, ou encore des paramètres dérivés du cepstre comme les paramètres PROSPECT.

Depuis environ deux ans, parallèlement aux divers travaux résumés ci-dessus, je réfléchis à une nouvelle formulation des concepts fondamentaux de la reconnaissance avec données manquantes, qui vise à optimiser directement le taux d'erreur en mots plutôt que de passer par l'intermédiaire du rapport signal-sur-bruit.

La proposition de masquer les observations en optimisant le taux d'erreur résout de nombreux problèmes persistant du domaine de la reconnaissance avec données manquantes, dont la dépendance de l'approche par rapport à la paramétrisation, ou encore le traitement des coefficients dynamiques. Toutefois, la modélisation de tels masques est beaucoup plus difficile que dans le cas classique. J'ai proposé une approche d'inférence de ces masques à partir des mesures de confiance estimées sur le résultat de la reconnaissance. Cette approche présente un autre avantage, qui est de faire le lien entre deux domaines de recherche : la reconnaissance avec données manquantes et l'estimation de mesures de confiance, et par là même propose une piste originale pour résoudre un problème récurrent du dernier domaine, qui est d'exploiter efficacement les mesures de confiance pour corriger les erreurs de la reconnaissance.

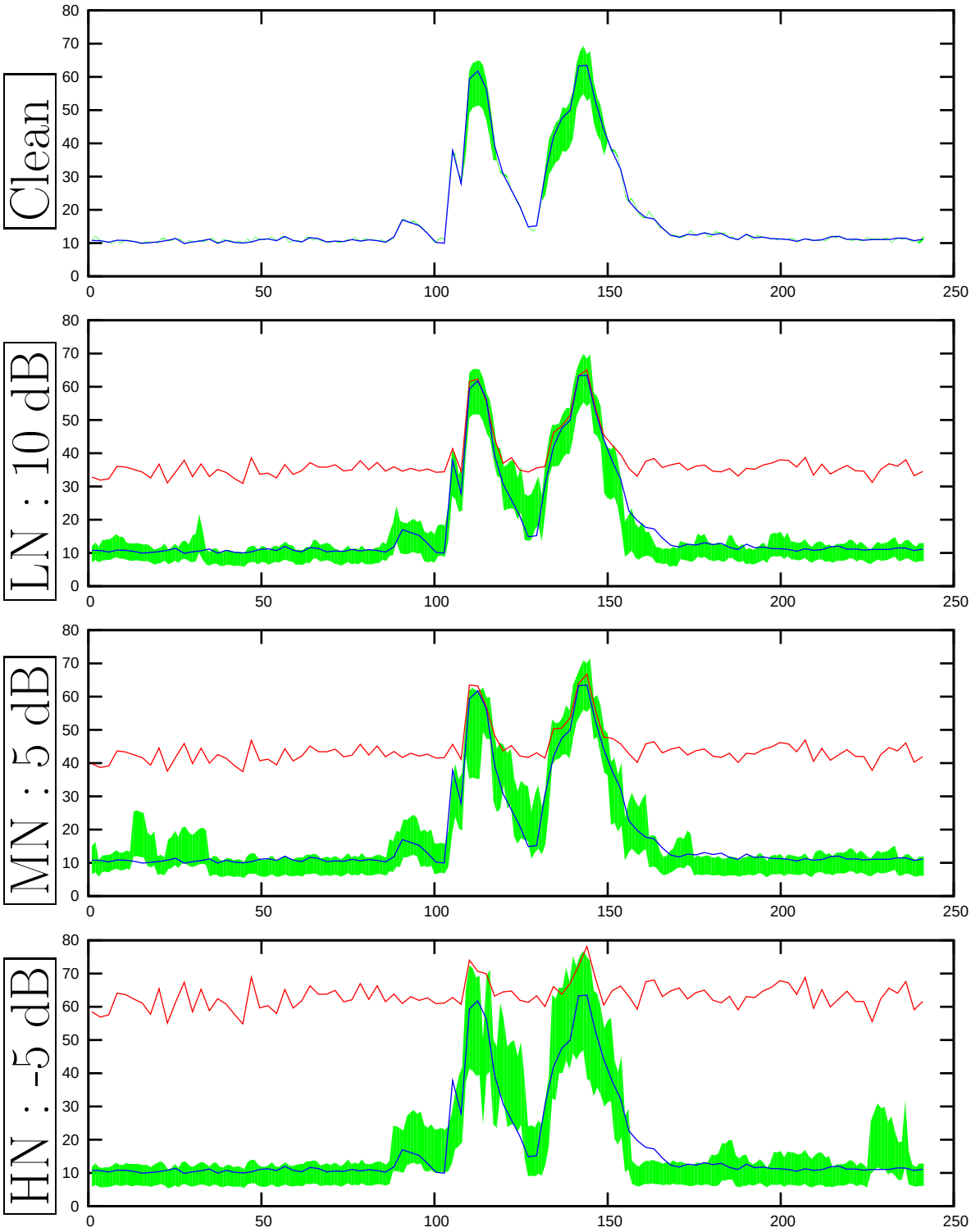


FIG. 4.2 – Intervalles de marginalisation dérivés de la méthode proposée dans l'équation 4.14. Les courbes bleue et rouge représentent respectivement l'évolution énergétique des signaux de parole seule et de parole bruitée pour la $6^{\text{ième}}$ bande de fréquence. Les intervalles de marginalisation sont représentés par les aires vertes.

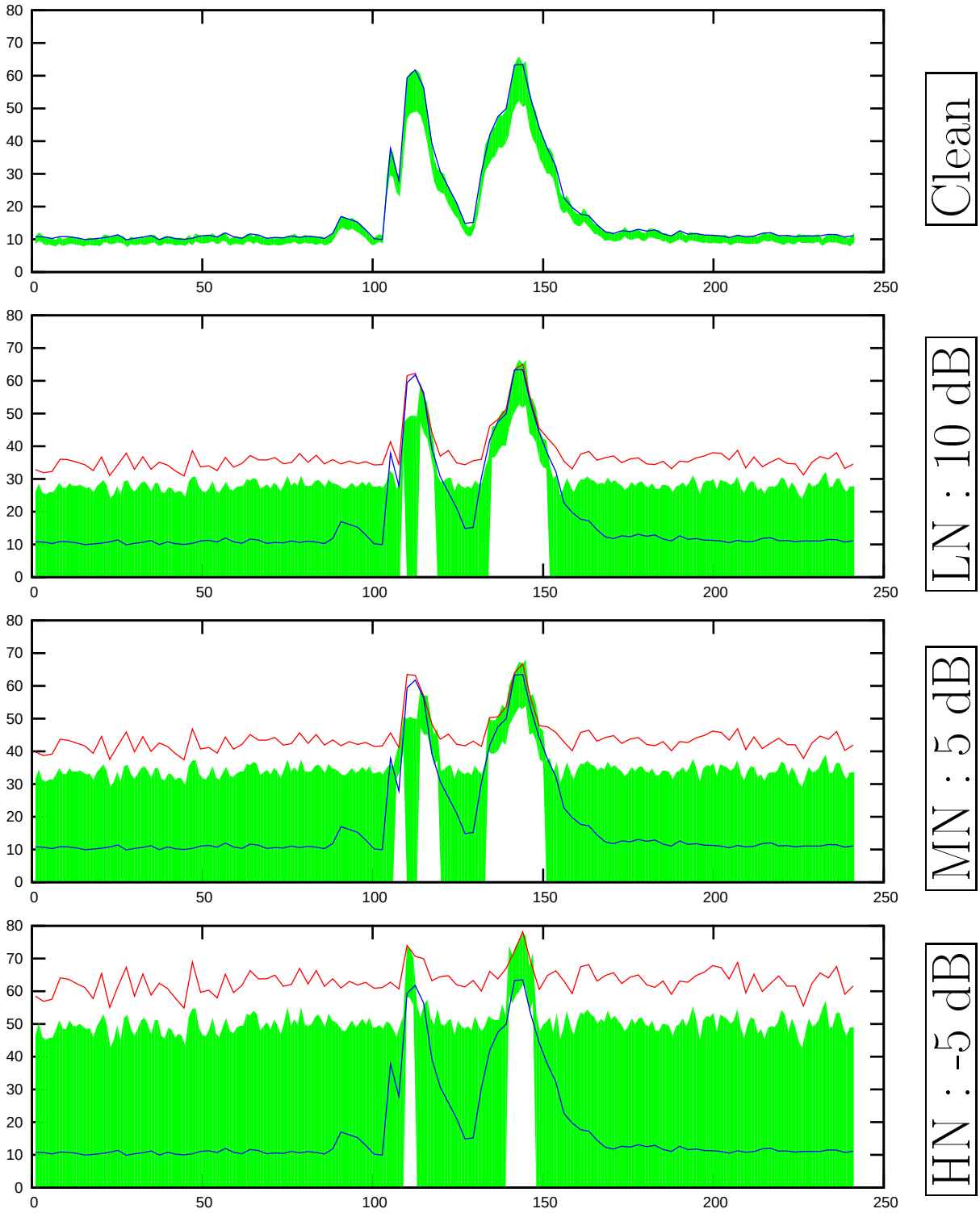


FIG. 4.3 – Intervalles de marginalisation dérivés des masques "classiques" de l'équation 4.13. Les courbes bleue et rouge représentent respectivement l'évolution énergétique des signaux de parole seule et de parole bruitée pour la $6^{\text{ième}}$ bande de fréquence. Les intervalles de marginalisation sont représentés par les aires vertes.

L'approche par marginalisation en reconnaissance avec données manquantes s'appuie sur trois hypothèses fondamentales :

1. *La différence entre la parole d'apprentissage et la parole de test est due à du bruit additif.* Cette hypothèse implique que les spectres de puissance respectifs de la parole et du bruit sont additifs et positifs, ce qui permet de réduire l'intervalle de marginalisation, mais d'un autre côté contraint la paramétrisation au domaine spectral.
2. *Les observations dans chaque bande de fréquence sont considérées comme indépendantes lors du calcul de la vraisemblance globale*, ce qui permet, avec l'hypothèse de Markov traditionnelle, d'isoler la contribution de chaque "point" de l'espace temps-fréquence (TF).
3. *Il est possible de localiser les points de cet espace qui sont dominés par le bruit de ceux dominés par la parole.* Cette hypothèse, la plus importante, permet de grouper les points du plan TF en deux classes disjointes : les zones masquées et non masquées.

Dans l'approche proposée, je remplace cette dernière hypothèse par la suivante : *Il est possible de localiser les points (t, i), où t représente un instant donné et i une dimension dans l'espace des paramètres du signal, qui ont un impact négatif sur le taux de reconnaissance global du système.*

Cette nouvelle hypothèse permet toujours de classifier les points de l'espace des paramètres en deux classes : masqués et non masqués, mais elle n'impose aucune contrainte quant aux propriétés de cet espace. Ainsi, la première hypothèse n'est plus nécessaire, ce qui implique que toute sorte de variabilité (locuteur, bruit convolutif, stress, etc.) peut être prise en compte par l'approche proposée. Par contre, une marginalisation totale doit être réalisée sur les points masqués.

Un masque "oracle" selon cette nouvelle hypothèse doit donc être calculé idéalement en masquant toutes les combinaisons de point possible et en retenant celle qui maximise le taux de reconnaissance. Ce calcul étant impossible à réaliser du fait de l'explosion combinatoire, j'ai proposé une heuristique d'approximation qui consiste à identifier les points qui favorisent plus l'alignement fourni par la reconnaissance que l'alignement forcé sur la bonne suite de mots. Formellement, la contribution de la $i^{\text{ème}}$ dimension du vecteur cepstral $y(t)$ à la vraisemblance des observations sur la transcription automatique est donnée par :

$$p(y_i(t)|e(t)) = \int \cdots \int p(y(t)|e(t)) dy_1(t) \cdots dy_{k \neq i}(t)$$

où $e(t)$ représente l'état aligné avec $y(t)$ dans la meilleure solution fournie par la reconnaissance. De même, la contribution de la $i^{\text{ème}}$ dimension du vecteur cepstral $y(t)$ à la vraisemblance des observations sur la transcription de référence est donnée par :

$$p(y_i(t)|e_0(t)) = \int \cdots \int p(y(t)|e_0(t)) dy_1(t) \cdots dy_{k \neq i}(t)$$

Un masque oracle au sens de la nouvelle hypothèse masque le coefficient $y_i(t)$ si et seulement si

$$p(y_i|e) - p(y_i|e_0) > 0$$

L'approche proposée s'appuie sur une nouvelle hypothèse : *Masquer les points définis par l'heuristique précédente améliore le taux de reconnaissance.* Avant de décrire la méthode d'estimation des masques en conditions de test, j'ai souhaité vérifier cette hypothèse en comparant les performances obtenues, d'une part avec ces masques oracles optimisant le taux de reconnaissance, et d'autre part, avec les masques oracles classiques qui optimisent le rapport signal-sur-bruit. Les deux articles de référence [Cooke *et al.*, 2001] et [Morris *et al.*, 2001a] dans le domaine de la reconnaissance avec données manquantes publient des résultats obtenus avec un masque oracle optimisant le SNR sur Aurora2. J'ai donc reproduit au mieux les conditions expérimentales décrites dans ces articles, mais en remplaçant les "masques oracles SNR" par mes "masques oracles WER"[4]. Les deux comparaisons sont données sur les figures 4.4 et 4.5.

Les résultats de ces expériences montrent sans doute possible que l'heuristique définie ci-dessus est suffisamment efficace pour poursuivre les développements dans cette voie.

[4] WER pour *Word Error Rate*, c'est-à-dire les masques optimisant le taux de reconnaissance

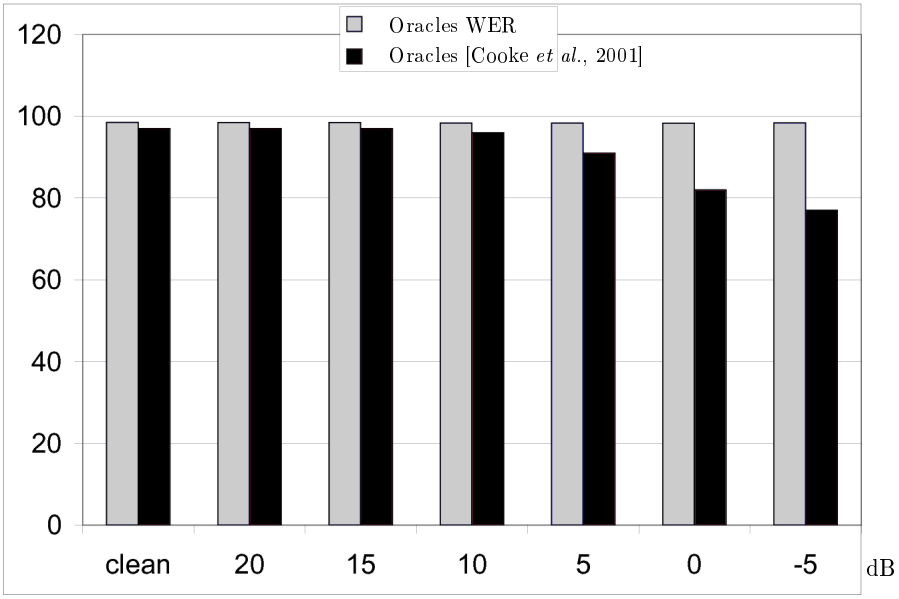


FIG. 4.4 – Taux de reconnaissance en mots sur Aurora2 avec le masque oracle WER tel que décrit au paragraphe 4.7, et le masque oracle SNR publié dans [Cooke *et al.*, 2001]

Le véritable problème est alors de localiser ces points : en reconnaissance avec données manquantes classique, une estimation du rapport signal-sur-bruit est généralement utilisée. Mais une telle donnée n'a plus de sens avec la nouvelle approche. L'objectif étant maintenant le taux de reconnaissance, il faut utiliser des indices qui peuvent être considérés comme des indicateurs de cet objectif. J'ai donc proposé de construire les masques à partir d'une mesure de confiance calculée sur le résultat de la reconnaissance.

En effet, une mesure de confiance peut être assimilée à la probabilité qu'un mot reconnu soit faux. Si nous supposons qu'un tel mot est effectivement faux, alors, lors de la reconnaissance, les points qui ont contribué le plus à choisir ce mot sont très probablement corrompus par une source de variabilité quelconque : nous devons donc masquer ces points.

J'ai ainsi réalisé une première implémentation de ce principe de base en fixant un seuil sur la mesure de confiance de la reconnaissance : un mot est supposé faux si et seulement si la mesure de confiance de ce mot est inférieure à ce seuil. Tous les coefficients des vecteurs d'observation du segment alignés avec ce mot sont alors triés par log-vraisemblance décroissante, indépendamment de leur indice dans le vecteur de paramètres, et les $N\%$ premiers coefficients sont alors masqués [Cerisara, 2008].

Une première expérience réalisée sur une heure du corpus ESTER a donné des résultats encourageants, comme le montre le tableau 4.4. Pour cette expérience, j'ai utilisé une implémentation réalisée par Joseph Razik [Razik, 2007] d'une des meilleures mesures de confiance de l'état de l'art décrite dans [Wessel *et al.*, 2001]. Cette mesure estime la probabilité a posteriori des mots par le biais d'une adaptation de l'algorithme forward-backward sur le graphe de reconnaissance.

Système de reconnaissance	WER [%]
Référence	18.6
Masques basés sur la mesure de confiance	17.5

TAB. 4.4 – Taux d'erreur en grand vocabulaire, sur une heure du corpus ESTER, avec des masques basés sur une mesure de confiance dans le domaine cepstral.

Le principe fondamental de cette méthode, qui est basée sur la sélection automatique d'un certain

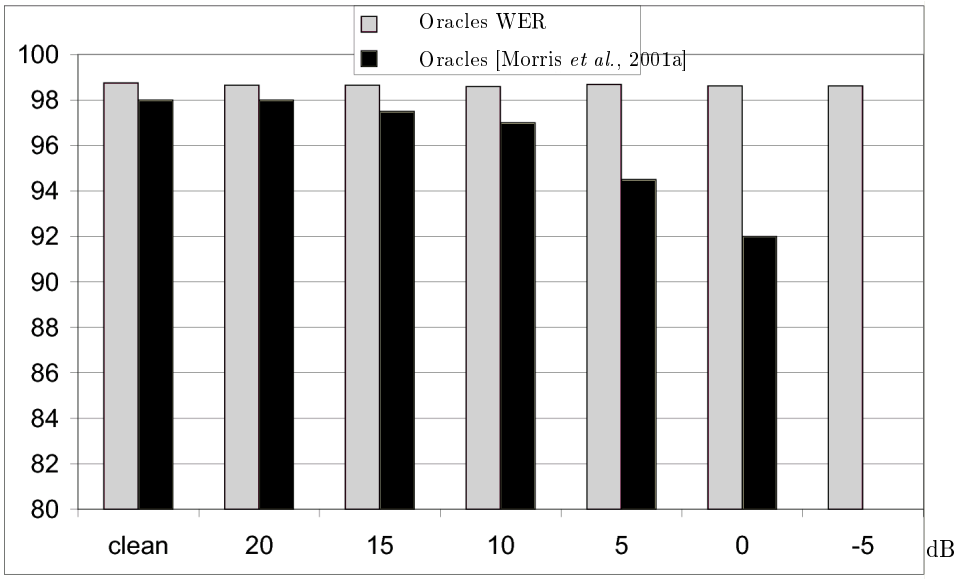


FIG. 4.5 – Taux de reconnaissance en mots sur Aurora2 avec le masque oracle WER tel que décrit au paragraphe 4.7, et le masque oracle SNR publié dans [Morris *et al.*, 2001a]

nombre d'observations à écarter lors du calcul de la vraisemblance, trouve un écho dans d'autres travaux du domaine de la reconnaissance de la parole robuste au bruit :

– Le "Union Model" [Ming et Smith, 2002]
– Le "Frame-skipping Viterbi Algorithm" [Siu et Chan, 2006]

Le "Frame-skipping Viterbi" est basé sur l'observation bien connue que la simple moyenne arithmétique d'une variable aléatoire est sensible aux valeurs extrêmes observées[5]. Face à des données corrompues, pour résoudre ce problème, d'autres moyennes sont utilisées, et en particulier la moyenne tronquée, qui exclut du calcul les observations dont les valeurs dépassent des quantiles prédéfinis, comme 25 % à 75 % des trames. De même, le calcul de la vraisemblance des observations par simple cumul des log-vraisemblances des trames est également très sensible aux trames extrêmes, et Y. Chan propose de supprimer les trames dont la vraisemblance a une valeur extrême afin de rendre la vraisemblance globale plus robuste au bruit. Cette méthode a été combinée dans [Lai, 2003] avec une approche de normalisation de la vraisemblance des observations afin de compenser à la fois le bruit quasi-stationnaire et le bruit impulsif.

L' "Union Model" est basée sur les mêmes principes fondamentaux que la reconnaissance avec données manquantes, à savoir la sélection des paramètres les moins corrompus pour la reconnaissance, mais sans nécessiter la localisation préalable des zones masquées. Le modèle proposé utilise pour ce faire la combinaison disjonctive des vraisemblances des observations entre toutes les bandes fréquentielles, ce qui peut s'interpréter comme le choix de la meilleure combinaison des différentes bandes parmi toutes les combinaisons possibles. Ce modèle est donc beaucoup moins contraint que les modèles de reconnaissance avec données manquantes présentés jusqu'ici, ce qui en fait un candidat mieux à même de traiter des conditions diverses et des bruits imprévisibles. Toutefois, cette absence de contrainte rend la méthode moins performante sur des tâches dont les conditions sont mieux maîtrisées.

4.8 Contributions

Mes contributions principales sont les suivantes :

– Etat de l'art fourni de l'estimation des masques.
– Développement d'une approche de reconnaissance avec données manquantes dans le domaine cepstral.

[5] *Outliers* en anglais.

- Etude de la combinaison de modèles de masques et de parole
- Développement de modèles de masques dépendant de l'environnement.
- Proposition d'une nouvelle approche bayésienne pour apprendre les modèles de masque, qui discrétise l'espace de recherche des masques possibles afin de réduire sa taille.
- Dérivation d'une nouvelle méthode minimisant la largeur de l'intervalle de marginalisation
- Définition d'un nouveau formalisme de calcul des masques minimisant directement le taux d'erreur de reconnaissance. J'ai aussi montré que ces masques peuvent être estimés à partir des mesures de confiance.

Chapitre 5

Perspectives : vers d'autres sources d'information

5.1 Introduction et motivations

J'ai tenté de montrer tout au long de cette première partie que la robustesse de la reconnaissance automatique de la parole est tributaire de l'information utilisée dans le système. Jusqu'à présent, seules des informations de bas niveau ont été considérées, comme différents modèles de bruit, de l'environnement, etc. Malgré les nombreuses recherches menées au cours de ces dernières années, recherches qui ont manipulé et transformé ces informations de base de toute les manières possibles, le problème de la robustesse n'est toujours pas résolu. Il est fort probable qu'une des causes principales de ces difficultés vienne du fait que, si les modèles changent, les informations qu'ils modélisent restent systématiquement les mêmes. Une des perspectives les plus prometteuses serait donc d'élargir le champ de connaissances à intégrer dans nos systèmes de reconnaissance robustes. Des informations de plus haut niveau peuvent ainsi être utilisées pour contraindre l'espace de recherche, évaluer la pertinence des réponses du système, voire diriger le décodage acoustico-phonétique.

Nous utilisons déjà fréquemment des contraintes de ce type lorsque les conditions environnementales sont très dégradées, par l'intermédiaire des grammaires de reconnaissance dépendantes de la tâche, qui n'autorisent qu'un petit nombre de phrases possibles et limitent ainsi les risques d'erreur. Mais cette approche réduit considérablement l'expressivité de l'utilisateur et ses possibilités de communication, et il serait beaucoup plus judicieux de guider la reconnaissance plutôt que de la limiter.

Le cœur du problème est donc de calculer des informations de haut niveau et de les intégrer au système de reconnaissance. C'est un véritable défi scientifique que je n'ai encore fait qu'effleurer, mais qui me paraît essentiel pour l'avenir du domaine.

Au-delà des informations dérivant directement du signal, comme celles que nous avons présentées en détail précédemment, trois grandes sources d'information peuvent être utilisées en reconnaissance de la parole : la syntaxe, la sémantique, et le contexte. J'ai très modestement commencé à m'intéresser aux aspects sémantiques, en débutant par le niveau zéro de la compréhension, c'est-à-dire la reconnaissance du thème abordé, et aux aspects contextuels par le biais de l'intelligence ambiante, qui est un domaine de recherche dans lequel le contexte occupe une place centrale et prépondérante. Plus spécifiquement, je me suis intéressé à la manière dont un système de reconnaissance pouvait s'intégrer dans une plate-forme d'intelligence ambiante multimodale.

Je décris dans la suite ces deux travaux. Le sujet abordé s'éloigne un peu du cœur du mémoire, qui est la robustesse de la reconnaissance automatique de la parole, mais mon objectif reste à long terme d'intégrer ces nouvelles connaissances dans la reconnaissance, afin d'apporter d'autres informations utiles aux aspects de robustesse de la reconnaissance, voire de la compréhension.

5.2 La parole en intelligence ambiante

5.2.1 Brêve présentation de l'intelligence ambiante

L'intelligence ambiante est un domaine applicatif visant à intégrer de multiples technologies, et en particulier des réseaux de terminaux, capteurs et divers appareils électroniques tels que ceux existant actuellement dans nos habitats, afin de fournir aux utilisateurs de nouveaux services au quotidien. Toutefois, l'utilisateur étant au cœur de l'intelligence ambiante, les interactions Homme-machine et les interactions Homme-Homme médiatisées ont une place prépondérante dans le domaine.

L'application directe des technologies interactives qui ont été développées depuis de nombreuses années à ce nouveau domaine n'est malheureusement pas possible. En effet, une étude plus approfondie de l'informatique ubiquitaire montre que les besoins et les contraintes sont profondément différents de ceux des interactions Homme-machine classiques, au point qu'un nouveau terme est apparu, celui d' "interactions implicites". Il y a donc une réflexion particulière à mener pour appliquer la reconnaissance de la parole à l'intelligence ambiante, réflexion dans laquelle je me suis engagé depuis 2002, date de début du projet européen OZONE. De grands noms du domaine de la reconnaissance automatique de la parole, dont Sadaoki Furui, se sont également penchés sur cette question.

J'ai mené cette réflexion dans le cadre de deux projets européens intégrés, successivement OZONE et Amigo, dont les conclusions sont résumées dans les paragraphes suivants, ainsi que dans le cadre d'un groupe de travail de l'OFTA[6], qui a réuni une dizaine de chercheurs de différents domaines pendant 2 ans pour réfléchir au domaine et proposer des orientations futures. J'ai ainsi publié dans un chapitre d'un ouvrage de l'OFTA le fruit de cette analyse, du point de vue des interfaces Homme-machine.

5.2.2 Interactions implicites

Les systèmes d'intelligence ambiante doivent être déployés *partout*, fonctionner *en permanence*, et aider les utilisateurs autant que possible dans leurs tâches quotidiennes. Ils doivent donc interagir avec les utilisateurs, mais en les dérangeant le moins possible, c'est-à-dire en requérant un minimum d'effort cognitif et d'attention de leur part. Les interactions implicites sont destinées à réaliser ceci, et sont fondées sur une "observation" constante de l'utilisateur, qui permet au système de prévoir quelles sont les actions qui peuvent lui être utiles, sans forcément avoir recours à une interaction explicite avec lui. Ce domaine est également connu sous le nom d' "interfaces attentives".

Dans ce cadre, les interactions implicites par la parole consistent à écouter en permanence l'utilisateur et son environnement, notamment au cours de ses conversations avec d'autres personnes, médiatisées (téléphone, vidéoconférence, etc.) ou non. La technologie vocale requise pour cela diffère en plusieurs points importants des approches traditionnelles en reconnaissance automatique de la parole :

- L'objectif n'est (en général) pas de transcrire la parole, mais il est de la comprendre, ou de manière plus réaliste, d'en extraire un certain nombre d'informations utiles au système d'intelligence ambiante.
- Le domaine d'application n'est pas restreint comme dans la plupart des applications interactives, mais il est général et ouvert à tout ce qui peut être dit au cours d'une conversation entre personnes.
- L'utilisateur ne parle pas volontairement au système, ou du moins le destinataire principal de son discours n'est pas la machine : la parole est donc spontanée, et l'utilisateur s'appuiera sur les connaissances générales partagées entre les hommes, et sur le contexte interactif présent et passé qu'il a avec ses interlocuteurs. Toutes ces informations ne sont pas disponibles au système, qui ne pourra résoudre facilement les ellipses, allusions, et métaphores omniprésentes dans nos conversations, en supposant qu'il parvienne déjà à transcrire relativement correctement de la parole spontanée.
- La tâche paraît donc insoluble, mais une contrepartie importante à ces difficultés est que l'utilisateur n'attend rien a priori de particulier du système : il est donc en général préférable de ne pas déranger l'utilisateur en cas de doute sur ses souhaits, et d'attendre d'avoir une confirmation indirecte un peu plus tard.

Notons finalement que la parole n'est qu'une modalité particulière pour les interactions implicites, et j'ai donc travaillé dans ce cadre, en collaboration avec les équipes Langue&Dialogue, puis TALARIS,

[6] Observatoire Français des Techniques Avancées

à intégrer la reconnaissance de la parole dans des architectures multimodales, en tenant compte des contraintes imposées par les plates-formes d'intelligence ambiante. Par exemple, le fait de devoir déployer un service sur une multitude de terminaux de capacités variées, souvent mobiles et donc pauvres en ressources énergétiques, est une contrainte forte qui implique de développer des algorithmes portables et de la plus faible complexité possible. De plus, une interaction implicite étant par définition *discrète* vis-à-vis de l'utilisateur, afin de ne pas accaparer son attention, elle doit obligatoirement passer par un module de gestion du contexte, ce que nous avons réalisé en analysant dans OZONE le contexte environnemental pour décider de la modalité d'interaction la plus appropriée, et en interagissant avec un service centralisé de gestion du contexte[7] afin de compléter les informations contextuelles globales par les indices déduits de l'écoute des utilisateurs : dans ce dernier cas, les informations de haut niveau extraites de la parole, dont les actes de dialogue et l'identification du thème décrits ci-dessous, sont des sources d'informations contextuelles au même titre que, par exemple, la localisation spatiale.

5.3 Actes de dialogue

Dans le cadre de la thèse de Pavel Král [Kral, 2007], nous avons travaillé sur la reconnaissance automatique des actes de dialogue à partir d'un signal de parole. Les applications directes de ces travaux dans le cadre de l'intelligence ambiante sont par exemple la reconnaissance d'une question directe de l'utilisateur pour le système, et donc la discrimination entre interactions explicites et implicites, et plus généralement l'enrichissement du contexte par l'information des actes de dialogue.

Au-delà du sens porté par les phrases, un dialogue est toujours composé d'une succession d'éléments fondamentaux, comme la mise en relation ("bonjour !"), les affirmations, les questions ouvertes ("Que faites-vous ici ?") ou fermées ("Il fait beau n'est-ce pas ?"), les réponses aux questions, les hésitations, les interruptions, les ordres, les suggestions, etc. Chacun de ces éléments définit un acte de dialogue [Austin, 1962]. Il n'existe pas de consensus concernant la liste des actes de dialogue existant dans une langue, car ceux-ci dépendent du type de dialogue et de la tâche. Ainsi, l'interview d'un journaliste et une discussion à la terrasse d'un café ne seront pas constitués des mêmes éléments. Toutefois, il existe quelques typologies d'actes de dialogue définies pour le traitement automatique des langues qui sont devenues des "standards" sur lesquels la plupart des travaux existant se basent. Parmi ceux-ci, nous trouvons en particulier :

- DAMSL (Dialogue Act Markup in Several Layers) [Allen et Core, 1997], qui a été conçu en 4 niveaux pour être le plus générique possible ;
- SWBD-DAMSL [Jurafsky *et al.*, 1997], qui est une adaptation de DAMSL au corpus de reconnaissance de la parole Switchboard ;
- Meeting-Recorder [Dhillon *et al.*, 2004] ;
- VERBMOBIL [Alexandersson *et al.*, 1995] ;
- Map-Task [Carletta *et al.*, 1997].

Dans notre travail, nous avons évalué expérimentalement nos approches sur deux corpus : un corpus de dialogue en tchèque pour la réservation de billets de train, et le corpus radiophonique français ESTER. Dans les deux cas, nous avons extrait un sous-ensemble réduit des actes définis dans DAMSL en fonction du type de dialogue présent dans chacun des deux corpus. Ainsi, nous avons utilisé 4 et 7 actes de dialogue respectivement pour les corpus tchèque et français. Ces actes de dialogues sont annotés au niveau des phrases, la segmentation des corpus en phrases étant dans la suite réalisée manuellement et supposée connue par le système.

Nos contributions principales au domaine sont :

- Prise en compte de l'information syntaxique globale de position des mots dans la phrase dans les modèles d'actes de dialogue à travers trois nouvelles approches : la *position multi-échelles*, la *combinaison non linéaire* et la *meilleure position au sens de Bayes* ;
- Proposition d'un nouveau modèle basé sur le regroupement des mots ;
- Analyse et comparaison de plusieurs méthodes de combinaison de classifieurs ;

[7] Le *Context Management System* déployé dans Amigo

– Proposition de deux mesures de confiance pour l'apprentissage semi-supervisé des modèles d'actes de dialogue ;

Je détaille dans la suite les méthodes basées sur la position des mots dans la phrase, les résultats expérimentaux obtenus en combinant les informations prosodiques et syntaxiques, et les mesures de confiance utilisées pour l'apprentissage semi-supervisé sur le corpus ESTER.

5.3.1 Modélisation de la position des mots dans la phrase

Dans la littérature, l'information syntaxique est souvent modélisée dans les modèles d'actes de dialogue par des n-grams. Toutefois, ces n-grams ne prennent en compte qu'une information locale autour de chaque mot de la phrase. Un analyseur syntaxique formel fournissant un arbre syntaxique de la phrase pourrait être utilisé pour extraire une information globale sur la phrase, mais de telles analyses ne sont malheureusement que rarement possibles sur des corpus oraux. Nous avons donc proposé d'inclure l'information relative à la position du mot dans la phrase, qui est une information syntaxique globale, facile à calculer et qui nous paraît l'une des informations les plus discriminantes pour la reconnaissance des actes de dialogue[8].

Le nombre de positions possibles des mots dans la phrase est fixé à $N_p = 8$. La correspondance entre les positions réelles des mots dans une phrase et ces N_p positions est linéaire : en fonction de la longueur de la phrase, un mot peut donc être affecté à plusieurs positions, et inversement. L'approche la plus simple pour modéliser ces positions consiste à ajouter une variable aléatoire représentant la position dans le modèle bayésien, et à construire les densités de probabilité conditionnelles correspondantes. Toutefois, cette approche n'est pas envisageable, car elle réduit considérablement la taille du corpus servant à estimer les paramètres des modèles. Nous avons donc proposé les trois approches suivantes :

Position multi-échelles [Král et al.*, 2006] :* Le nombre de positions N_p peut varier entre 1 et 8, selon la quantité de données d'apprentissage disponible. Ces différents nombres de position sont codés dans un arbre, comme celui représenté sur la figure 5.1.

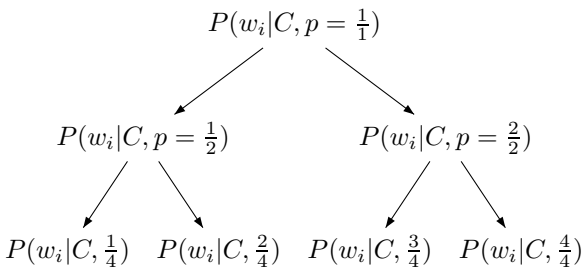


FIG. 5.1 – Arbre de calcul de la vraisemblance du mot w_i pour l'acte de dialogue C et à la position p.

La profondeur dans l'arbre pour calculer la vraisemblance $P(w_i|C)$ du mot w_i pour l'acte de dialogue C est choisie en descendant la branche le plus profondément possible jusqu'à ce que le nombre d'occurrences utilisées pour estimer cette probabilité soit inférieur à un seuil prédéfini. Finalement, l'acte de dialogue est choisi en maximisant la probabilité a posteriori :

$$\begin{aligned} \hat{C} &= \arg\max_C P(C|w_1, \cdots, w_T, p_1, \cdots, p_T) \\ &= \arg\max_C P(C) \prod_{i=1}^{T} P(w_i|C, p_i) \end{aligned} \tag{5.1}$$

où $p_i \in \{1, 2, \cdots, N_p\}$ et T est le nombre de mots de la phrase.

Combinaison non linéaire : Cette approche utilise les modèles unigrammes indépendant de la position. L'information de position est fournie à un réseau de neurones qui va ainsi apprendre la relation non linéaire

[8] Par exemple, les phrases interrogatives commencent souvent par "Que/Qui/Où/..."

entre les probabilités unigramme et les positions des mots d'une part, et les probabilités a posteriori $P(C|w_1, \cdots, w_T, p_1, \cdots, p_T)$ d'autre part.

Meilleure position : Dans cette approche [Král *et al.*, 2007b], la variable p a une signification légèrement différente de celle utilisée précédemment : soit p la meilleure position parmi toutes les positions possibles, c'est-à-dire la position telle que le taux de reconnaissance est maximal lorsque tous les mots qui ne sont pas à cette position sont supprimés. Notre objectif est de maximiser la probabilité a posteriori :

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)} \tag{5.2}$$

$$= \frac{P(C)\sum_p P(W,p|C)}{P(W)} \tag{5.3}$$

$$= \frac{P(C)\sum_p P(W|C,p)P(p|C)}{P(W)} \tag{5.4}$$

Une fois la meilleure position p connue, le classifieur choisit l'acte de dialogue en utilisant seulement le mot à cette position[9] :

$$P(W|C,p) = P(w_p|C)$$

Donc,

$$P(C|W) = \frac{P(C)\sum_p P(w_p|C)P(p|C)}{P(W)} \tag{5.5}$$

Finalement,

$$\hat{C} = \arg\max_C P(C)\sum_p P(w_p|C)P(p|C) \tag{5.6}$$

De la sorte, les probabilités "lexicales" $\prod_i P(w_i|C)$ utilisées jusqu'à présent sont remplacées par la somme pondérée des vraisemblances des mots, les poids représentant intuitivement l'importance d'une position dans la phrase pour un acte de dialogue donné. Ces poids peuvent être appris sur un corpus de développement, mais ils peuvent également être prédéfinis, par exemple sous la forme de probabilités a priori dans le cadre bayésien, grâce à des connaissances expertes. Par exemple, nous savons que le premier mot est important pour reconnaître une question.

Dans cette approche, le modèle de position est découplé du modèle lexical, ce qui permet d'apprendre les modèles lexicaux sur tout le corpus disponible. Ainsi, le nombre de positions n'est plus limité à 8, et nous avons donc utilisé 20 positions différentes pour cette approche.

Dans nos expériences, nous avons d'abord calculé sur le corpus de développement le taux de reconnaissance en fonction de la position p choisie. Les résultats sont représentés sur la figure 5.2.

Les poids sont donc initialisés à $P(p = 1|C) = 1$ et $P(p > 1|C) = 0$ pour tout C. Ensuite, un apprentissage des poids par descente de gradient est réalisé. Après l'apprentissage, les poids résultants sont représentés sur la figure 5.3.

Les résultats de cet apprentissage confirment notre intuition selon laquelle le premier mot est le plus important pour les questions ouvertes, et, dans une moindre mesure, le dernier mot plus important pour les ordres que pour les autres actes de dialogue.

Les résultats obtenus avec ce modèle sont donnés dans le tableau 5.1.

5.3.2 Combinaison de classifieurs

Nous avons proposé de combiner les informations lexicales et de position décrites ci-dessus avec des informations prosodiques qui, intuitivement, doivent avoir de l'importance pour la reconnaissance des actes de dialogue [Král *et al.*, 2005]. Deux caractéristiques prosodiques ont été calculées :

[9] À cause de la correspondance linéaire entre les positions, plusieurs mots peuvent être alignés avec cette position, mais le raisonnement exposé reste vrai sans perte de généralité.

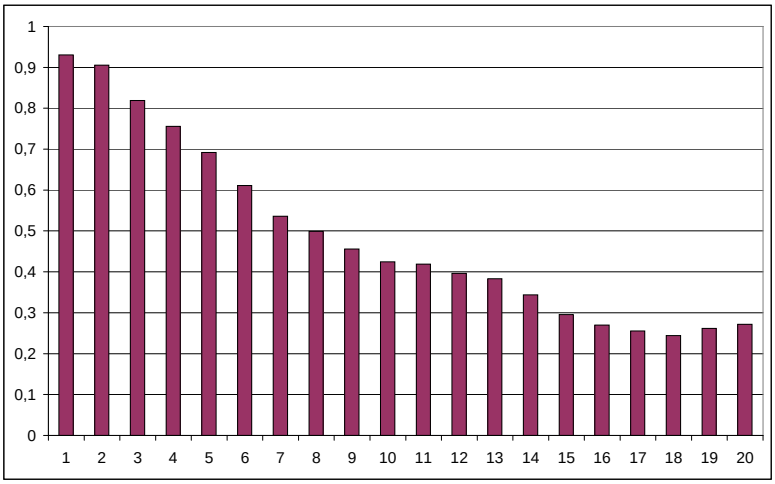


FIG. 5.2 – Taux de reconnaissance sur le corpus de développement lorsqu'une seule position p est considérée.

1. La courbe de la fréquence fondamentale, discrétisée sur 20 points équi-répartis le long du segment temporel de la phrase ;
2. La courbe énergétique, également discrétisée sur 20 points.

Un modèle prosodique par acte de dialogue est alors entraîné sur ces données. Nous avons choisi pour cela un GMM qui modélise $P(F|C)$ où F est le vecteur des 40 valeurs prosodiques.

Pendant le test, les vraisemblances retournées par le modèle prosodique sont normalisées en probabilités a posteriori, puis combinées avec les probabilités a posteriori des modèles syntaxiques décrits précédemment. Nous avons testé plusieurs combinaisons. Les premières sont les plus génériques, car elles ne nécessitent pas de corpus de développement. Il s'agit de :

- Naïve Bayes : Les classifieurs sont considérés indépendants et leurs probabilités multipliées ;
- Statistiques d'ordre supérieur : La probabilité associée à une classe est respectivement la plus grande (maximum), la plus petite (minimum) et la médiane des probabilités des classifieurs individuels.

Les autres combinaisons envisagées requièrent un corpus de développement, et donnent donc logiquement de meilleurs résultats, mais sont moins susceptibles d'être performantes dans des conditions inattendues :

- Combinaison linéaire pondérée :

$$P(C|W,F) = \alpha P(C|W) + (1-\alpha)P(C|F)$$

- Réseau de neurones :

$$P(C|W,F) = f(P(C|W), P(C|F))$$

Les meilleurs résultats ont toujours été obtenus avec le réseau de neurones (MLP). Le tableau 5.1 compare les performances des différents modèles présentés jusqu'à présent sur le corpus tchèque, qui est composé de 2173 phrases. Toutes ces expériences sont réalisées en double validation croisée, 10 % du corpus étant réservé au développement et 10 autres % au test.

Nous pouvons remarquer que les taux de reconnaissance sont très élevés : ceci est dû au corpus qui est composé essentiellement de phrases bien structurées, et au nombre réduit d'actes de dialogue à reconnaître. L'information lexicale est de loin la plus importante : l'information prosodique peut effectivement apporter de nouveaux indices, qui se traduisent pas une augmentation du taux de reconnaissance lorsqu'elle est

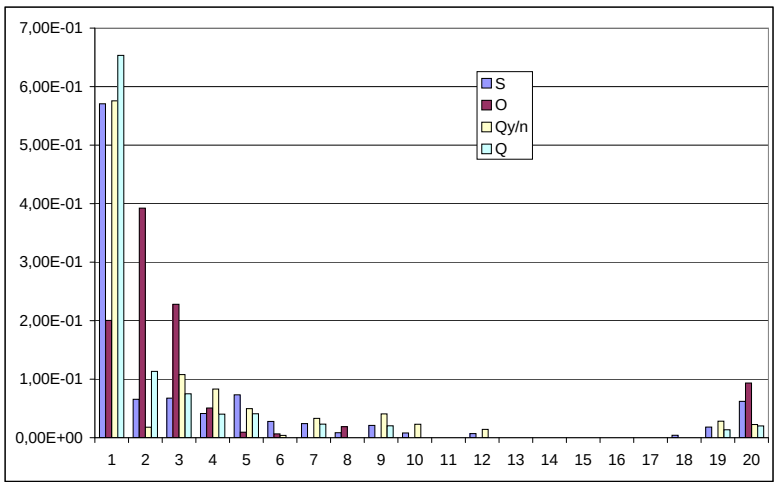


FIG. 5.3 – Poids obtenus après la descente de gradient pour chacun des 4 actes de dialogue : affirmations (S), ordres (O), questions oui/non (Qy/n), autres questions(Q).

	Taux de reco. en [%]				
Approche	S	O	Qyn	Q	Global
0. Méthode purement lexicale (référence)					
Unigrammes	93.5	77.6	96.5	89.9	**91.0**
1. Méthodes utilisant la position					
Multi-échelles	94.7	70.4	96.1	95.3	**93.8**
Non linéaire	90.3	83.2	91.1	98.8	**94.7**
Meilleure position	93.6	95.2	97.2	94.3	**95.8**
2. Classifieur prosodique					
GMM	47.7	43.2	40.8	44.3	**44.7**
3. Combinaison "Meilleure pos/prosodie"					
MLP	94.0	95.6	97.0	95.2	**96.9**

TAB. 5.1 – Taux de reconnaissance pour (1.) les approches basées sur la position, (2.) le classifieur prosodique, et (3.) la combinaison du GMM prosodique avec la méthode "meilleure position" par un MLP.

combinée avec l'information lexicale, mais la prosodie seule ne peut en aucun cas rivaliser avec les modèles lexicaux. L'information de position est également très importante, comme nous nous y attendions, et permet de réduire le nombre d'erreurs du système de plus de 50 %.

5.3.3 Mesures de confiance pour l'apprentissage semi-supervisé

L'un des principaux problèmes du domaine de la reconnaissance automatique des actes de dialogue est la création du corpus d'apprentissage. En effet, à la différence des phonèmes d'une langue qui sont définis une fois pour toutes pour annoter les corpus de reconnaissance de la parole, l'ensemble des actes de dialogue à prendre en compte dépend pour une grande part de l'application : il faut donc annoter manuellement un nouveau corpus d'apprentissage à chaque nouvelle application.

Nous avons étudié la possibilité d'appliquer les techniques d'apprentissage semi-automatique issues de la reconnaissance automatique de la parole au domaine de la reconnaissance des actes de dialogue.

Ces techniques sont très souvent dérivées de l'algorithme EM (Expectation/Maximization), qui itère successivement les phases d'estimation des étiquettes des exemples du corpus, et d'apprentissage des paramètres des modèles. Toutefois, cet algorithme est sensible aux conditions initiales (les paramètres initiaux des modèles), et peut diverger lorsque les modèles font beaucoup d'erreurs de reconnaissance. Pour réduire ce problème, des mesures de confiance sont parfois utilisées pour filtrer les exemples probablement erronés avant la phase de maximisation.

Nous avons appliqué cette approche à l'apprentissage semi-supervisé des modèles d'actes de dialogues [Král *et al.*, 2007a] sur le corpus radiophonique français ESTER. La première étape a consisté à analyser une partie de ce corpus afin d'identifier les actes de dialogue pertinents. Ce travail a mené à la définition de 21 actes de dialogue. Ensuite, le nombre d'occurrences de ces actes de dialogue a été estimé, et certains actes ont été regroupés car ils apparaissaient trop peu souvent. Au final, les 7 actes de dialogue listés dans le tableau 5.2 ont été retenus.

N°	Acte de dialogue	Code	Exemples
1.	Affirmation	gs	*Je pense que...*
2.	Question O/N	qy	*Est-ce que vous ... ?*
3.	Autre question	gq	*Quand partez-vous ?*
4.	Limites de dialogue	goc	*Bonjour !* *À vous les studios !*
5.	Accords	ga	*oui oui...* *hmm hmm...*
6.	Prise de parole	h	*euh... je...*
7.	Radio spécifique	gg	*France-Inter !* *La météo.* *Joël Colado.*

TAB. 5.2 – Actes de dialogue définis sur le corpus ESTER.

Une fois les actes de dialogue définis, un étiquetage manuel d'une partie du corpus ESTER a été réalisé pour construire le corpus de test, qui est composé d'approximativement 1000 actes de dialogue. De même, environ 600 actes de dialogue ont été étiquetés manuellement pour créer un corpus d'apprentissage initial. Ce corpus d'apprentissage a ensuite été complété par un corpus construit à partir de règles définies manuellement. Ces règles correspondent à des propriétés générales de la langue française. Par exemple, les phrases commençant par "est-ce que" sont annotées en tant que "Question O/N", et les phrases terminant par un point d'interrogation et commençant par "combien de" ou "comment" sont des "Autres questions". Finalement, le corpus d'apprentissage initial est composé de 1650 phrases. Un modèle lexical unigramme par acte de dialogue est alors entraîné sur ce corpus.

L'algorithme d'apprentissage semi-supervisé se résume alors à :

1. Les modèles à l'itération i servent à annoter les phrases du reste du corpus ESTER ;
2. Une mesure de confiance est calculée pour chacun de ces exemples ;
3. Les exemples les plus fiables selon cette mesure sont intégrés au corpus d'apprentissage ;
4. Les modèles sont ré-entraînés sur le corpus d'apprentissage complété ;
5. Ce processus est itéré depuis l'étape 1 jusqu'à un nombre donné d'itérations.

Deux mesure de confiance ont été testées :
– La première est une estimation de la probabilité a posteriori des classes :

$$P(C|W) = \frac{P(W|C).P(C)}{\sum_{D \in \mathcal{DA}} P(W|D).P(D)} \tag{5.7}$$

où $\mathcal{DA}$ est l'ensemble des 7 actes de dialogue possibles.

– La seconde est la différence des probabilités a posteriori de la classe gagnante et de la meilleure classe en compétition :

$$\hat{C} = \arg\max_{C}(P(C|W)) \tag{5.8}$$

$$\Delta P = P(\hat{C}|W) - \max_{C \neq \hat{C}}(P(C|W)) \tag{5.9}$$

Cette mesure accorde un peu plus de crédit aux classes qui dominent les autres, même lorsque leurs probabilités a posteriori sont largement inférieures à 1.

Les meilleures performances sont résumées dans le tableau 5.3. Elles ont été obtenues avec la première mesure de confiance et un seuil de filtrage très élevé (0.999), ce qui suggère que les erreurs de reconnaissance sont trop nombreuses. La taille du corpus d'apprentissage obtenu est alors seulement de l'ordre de 2150 exemples après 9 itérations.

	Taux de reco. en [%]							
Itér.	s	qy	q	oc	a	h	g	glob.
0	72.4	70.3	62.9	66.1	51.4	100	41.6	**70.6**
1	76.4	58.6	62.9	66.1	51.4	100	42.7	**73.7**
2	81.8	58.0	62.5	66.1	65.3	100	45.7	**78.6**
3	83.8	52.3	65.5	66.1	65.3	100	41.0	**80.0**
4	82.6	51.1	66.5	66.1	62.5	100	43.1	**79.1**
5	81.9	47.1	68.2	66.1	62.5	100	43.1	**78.4**
6	81.8	51.1	68.2	66.1	62.5	100	43.1	**78.5**
7	81.8	46.8	68.2	66.1	62.5	100	43.1	**78.4**
8	82.2	46.8	68.8	66.1	62.5	100	43.1	**78.7**
9	81.9	46.8	68.8	66.1	62.5	100	43.1	**78.5**

TAB. 5.3 – Taux de reconnaissance obtenu par apprentissage semi-supervisé sur le corpus ESTER en fonction du nombre d'itérations de EM.

5.3.4 Conclusions sur la reconnaissance des actes de dialogue

Pour reconnaître les actes de dialogue, nous avons montré que l'information syntaxique globale que constitue la position des mots dans la phrase est particulièrement importante. Nous avons également montré que l'information prosodique améliore significativement les taux de reconnaissance, mais que l'information lexicale est de loin la plus importante, au moins globalement, lorsque toutes les catégories d'actes de dialogue sont considérées.

L'approche que nous avons choisie est l'approche bayésienne, qui s'appuie sur un corpus d'apprentissage préalablement étiqueté. Les expériences les plus convaincantes ont été réalisées sur un corpus tchèque de réservation de billets de train, corpus qui est composé de phrases plutôt bien formées, ce qui explique les bons taux de reconnaissance obtenus, de l'ordre de 90 % pour l'approche unigramme de base. Lorsque nous avons appliqué cette même approche sur un corpus beaucoup plus difficile, le corpus radiophonique français ESTER, nous avons obtenu des taux de reconnaissance de l'ordre de 78 %. La différence de performances entre les deux corpus est grande, mais elle s'explique avant tout par le nombre presque double d'actes de dialogue considérés sur ESTER, et bien sûr par la difficulté respective des tâches.

Le bénéfice apporté par l'apprentissage semi-supervisé est très limité, ce qui est probablement dû à un corpus initial manuel trop restreint. Le problème de la conception des corpus annotés en actes de dialogue pour de nouvelles tâches est donc encore loin d'être résolu, et il reste beaucoup de travail pour améliorer les techniques d'apprentissage semi-supervisé dans ce domaine et les rendre suffisamment efficaces.

Toutefois, ces études ont montré que la reconnaissance automatique de certains types d'actes de dialogue est possible, avec relativement peu d'informations, à savoir seulement les informations lexicales et leur position dans la phrase. D'autres expériences que nous avons réalisées sur le corpus tchèque

montrent également que cette conclusion reste vrai lorsque la reconnaissance des actes de dialogue est réalisée non plus sur une transcription manuelle, mais à la sortie d'un système de reconnaissance grand vocabulaire. Dans ce cas, avec un taux de reconnaissance en mots de 83 %, le taux de reconnaissance des modèles unigrammes d'actes de dialogue diminue seulement de 90 % à 88 %, ce qui reste très raisonnable et confirme le potentiel applicatif des approches proposées.

La reconnaissance des actes de dialogue peut enrichir le contexte interactif des systèmes d'intelligence ambiante et se révéler particulièrement utile par exemple pour identifier les questions explicites que l'utilisateur pose au systèmes et les simples commentaires implicites qui ne requièrent pas nécessairement de réactions immédiates de la part du système. Ce problème est d'ailleurs général et se pose dans tous les systèmes interactifs actuels, qui doivent faire appel à des artifices comme les boutons "push-to-talk" ou à des mots-clefs prédéfinis pour le contourner. Les actes de dialogue peuvent ainsi constituer un embryon de réponse plus "intelligente" à ce problème particulièrement difficile.

5.4 Reconnaissance de thèmes

5.4.1 Principe et contributions

J'ai commencé à travailler dans le projet Amigo sur la reconnaissance automatique du thème, qui constitue une information sémantique "à contour large", beaucoup moins précise que les relations sémantiques "fines" que nous pouvons définir à la main, mais qui a l'avantage de pouvoir être calculée automatiquement à partir des relations de co-occurrences. La reconnaissance de thème à partir de texte est un domaine de recherche bien connu, mais je me suis plus particulièrement intéressé à la reconnaissance du thème dans la parole. Les approches traditionnelles procèdent en deux étapes :

1. Transcription automatique de la parole en texte ;
2. Reconnaissance du thème à partir du texte.

Ma contribution principale dans ce domaine a été de proposer une nouvelle approche permettant de découvrir et de classifier automatiquement un corpus de parole en thèmes sans connaissance préalable d'un lexique, et donc sans système de transcription automatique [Cerisara, 2009].

Cette approche "pauvre en connaissance" possède plusieurs avantages importants par rapport à l'approche traditionnelle :

1. Les mots hors-vocabulaires, comprenant notamment les noms propres, sont très nombreux et malheureusement très importants pour identifier le thème. Cette information est systématiquement perdue dans les systèmes utilisant une transcription automatique.
2. À la différence de l'anglais, la très grande majorité des langues du monde ne possède pas de systèmes de transcription automatique de bonne qualité, car ces systèmes nécessitent des corpus textuels gigantesques et des corpus de parole annotés de très grande taille, ce qui est extrêmement coûteux.

La méthode que j'ai proposée procède en plusieurs étapes :

1. Le signal de parole est tout d'abord transformé en une suite de phonèmes grâce à un reconnaisseur phonétique indépendant de l'application ;
2. Un pseudo-lexique est extrait automatiquement à partir de ces séquences de phonèmes ;
3. Les entrées de ce lexique sont regroupées hiérarchiquement en classes représentant les différents thèmes abordés.

Le lexique produit est en fait un lexique "sémantique" car chaque entrée correspond à une branche du dendogramme résultant de la classification thématique. Toutefois, n'ayant pas abordé le problème de la conversion d'une suite phonétique en graphie, aucune forme orthographique n'est associée aux entrées du lexique. Une entrée du lexique est appelée dans la suite un "morphème acoustique", par référence aux travaux de Levit et al. [Levit *et al.*, 2002], qui utilisent des unités semblables pour la tâche de routage téléphonique.

Le lexique est construit par détection des séquences récurrentes dans la suite de phonèmes du corpus. Cette détection est réalisée par un algorithme parcourant tout le corpus au moyen d'une fenêtre glissante de taille variable entre 7 et 11 phonèmes, et recherchant les répétitions de la fenêtre courante dans le

reste du corpus grâce à un algorithme de comparaison approximative de chaînes basé sur une distance de Levenshtein [Wu et Manber, 1992], de complexité linéaire. La complexité totale de l'algorithme de construction du lexique est quadratique par rapport au temps.

Soit f_i le $i^{\text{ème}}$ morphème acoustique, et soit $f_{i,j}$ la $j^{\text{ème}}$ occurrence de f_i dans le corpus. Soit $p_{i,j}$ la position de $f_{i,j}$ dans le corpus, et N_i le nombre d'occurrences de f_i. J'ai défini la mesure de distance suivante entre morphèmes acoustiques basée sur la co-occurrence :

$$d(f_i, f_j) = \frac{1}{2N_i} \sum_{k=1}^{N_i} \min_l (p_{i,k} - p_{j,l})^2 + \frac{1}{2N_j} \sum_{k=1}^{N_j} \min_l (p_{j,k} - p_{i,l})^2 \tag{5.10}$$

Cette distance est alors utilisée dans un algorithme de classification hiérarchique ascendante standard. Le résultat de cette classification est un dendogramme dont la racine et les feuilles contiennent respectivement toutes et chacune des entrées du lexique. Les différents niveaux du dendogramme représentent une classification thématique possible des entrées du lexique.

5.4.2 Validation expérimentale

La méthode proposée a été évaluée sur deux corpus français : le corpus OGI de parole téléphonique, et le corpus radiophonique ESTER.

Évaluation sur OGI

Seule la partie française de OGI a été considérée; dans cette partie, seuls les enregistrements de parole spontanée pouvant être classés par thème ont été choisis. Il s'agit des réponses aux quatre requêtes suivantes :

- **thème R :** Décrivez la pièce dans laquelle vous êtes.
- **thème M :** Décrivez votre dernier repas.
- **thème HTL :** Dites-nous ce que vous aimez dans la ville où vous habitez.
- **thème HTC :** Parlez-nous du climat dans la ville où vous habitez.

Le tableau 5.4 donne quelques exemples de morphèmes acoustiques découverts dans chacune de ces quatre classes :

R	c'est un laboratoire, j'me trouve, des étagères ...
M	bien francais, j'ai mangé, dernier repas, déjeuné ...
HTL	ma ville natale, de la France, j'habite à, près de Paris ...
HTC	assez froid, climat tempéré, les étés sont chauds, continental ...

TAB. 5.4 – Exemples de morphèmes acoustiques.

Sachant que le corpus est ordonné de sorte que toutes les phrases appartenant à une même classe sont consécutives, les figures 5.4 et 5.5 montrent une visualisation en "dotplot" [Reynar, 1998] du résultat de la classification ascendante. Ce type de visualisation représente chaque paire de répétition $(f_{i,j}, f_{i,l})$ d'un même morphème acoustique f_i par un point de coordonnées $(p_{i,j}, p_{i,l})$. Pour construire les quatre figures, les morphèmes acoustiques ont été regroupés hiérarchiquement jusqu'à 8 classes. Parmi ces 8 classes, 4 classes peuvent être associées aux 4 classes de référence, tandis que les 4 autres classes contiennent les morphèmes acoustiques qui ne sont pas spécifiques à un thème : ce sont en quelque sorte les équivalents des "mots-outils" de nos lexiques, et ils apparaissent en général loin de la diagonale dans le dotplot. Les quatre premiers groupes de morphèmes acoustiques sont respectivement représentés sur les figures 5.4 et 5.5, et les carrés dessinés sur ces figures représentent les limites de référence entre deux thèmes, dans l'ordre : R, M, HTL et HTC.

Nous pouvons voir sur ces figures que la classification produit des groupes de mots qui sont visuellement corrélés avec les quatre thèmes de référence.

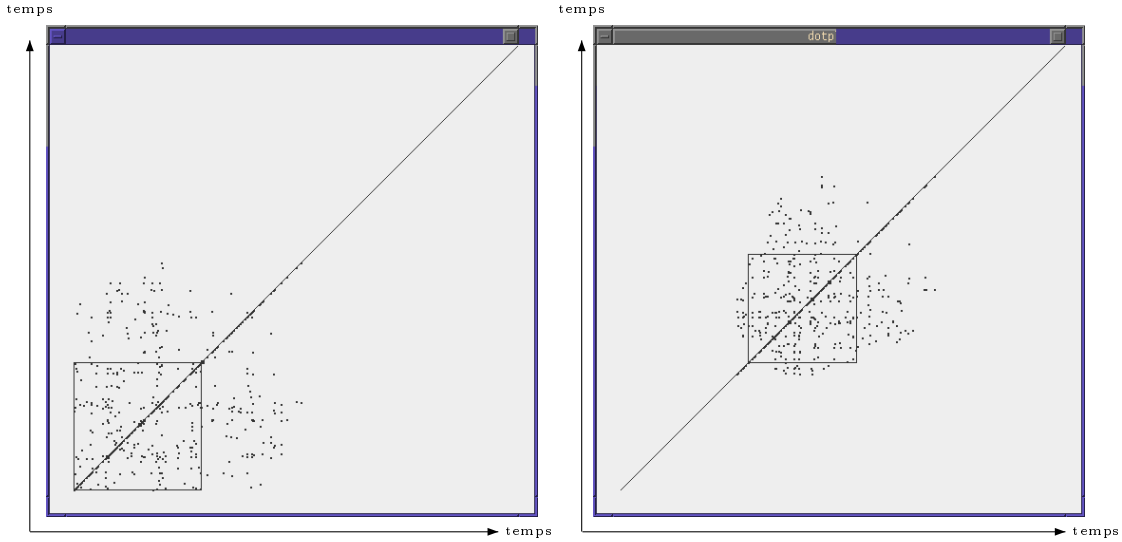


FIG. 5.4 – Visualisation par dotplot des thèmes R (à gauche) et M (à droite). Les deux axes représentent le temps. Les points représentent les répétitions des morphèmes acoustiques découverts par la méthode proposée et regroupés dans la même classe par la classification hiérarchique. Les carrés représentent la zone de référence d'un thème donné.

Pour évaluer quantitativement la classification des morphèmes acoustiques en groupes thématiques, le taux de classification correcte est calculé comme suit. Tout d'abord, un thème de référence doit être associé avec chaque morphème acoustique découvert. La segmentation initiale du corpus nous donne un thème de référence pour chaque segment de parole S_t :

$$\Phi(S_t) = \mathcal{T} \in \{\mathrm{R}, \mathrm{M}, \mathrm{HTL}, \mathrm{HTC}\}$$

Toute occurrence $f_{i,j}$ $j \in \{1, \cdots, N_i\}$ du morphème acoustique f_i appartient au segment $S_{s(i,j)}$: le thème $\Phi(S_{s(i,j)})$ est affecté à $f_{i,j}$. La distribution de référence de f_i sur les quatre thèmes est :

$$P(\mathcal{T}|f_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{\Phi(S_{s(i,j)})=\mathcal{T}} N_i$$

avec δ le symbole de Kronecker. J'affecte donc à f_i le thème de référence qui maximise cette probabilité :

$$\Phi(f_i) = \arg\max_{\mathcal{T}}(P(\mathcal{T}|f_i))$$

Un thème de référence étant affecté à chaque morphème acoustique, il devient possible de comparer les groupes obtenus par la classification hiérarchique avec cette référence. Pour un nombre de classes donné, nous calculons, pour chaque correspondance possible entre une classe découverte et une classe de référence, leur intersection. Nous associons alors chaque classe de référence avec la classe découverte dont l'intersection est maximale. Cette affection est stricte, en ce sens qu'une classe de référence est affectée à une et une seule classe découverte : les morphèmes acoustiques appartenant aux classes découvertes restantes (lorsqu'il y a plus de 4 classes découvertes) sont systématiquement comptés comme erronés. La figure 5.6 affiche le taux de bonne classification en fonction du niveau du dendogramme duquel les classes sont extraites.

Les taux de classification obtenus sont largement supérieurs à ceux correspondant à une classification aléatoire, ce qui prouve qu'il est possible de découvrir des relations thématiques sans aucune information lexicale a priori. Une évaluation plus poussée de la qualité de cette classification est réalisée sur ESTER.

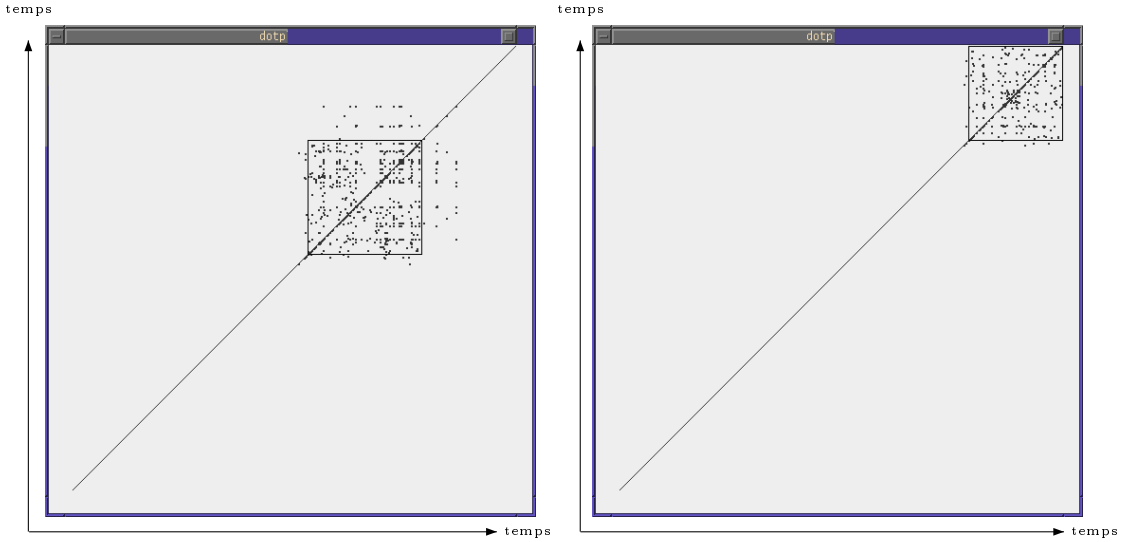


FIG. 5.5 – Visualisation par dotplot des thèmes *HTL* (à gauche) et *HTC* (à droite). Les deux axes représentent le temps. Les points représentent les répétitions des morphèmes acoustiques découverts par la méthode proposée et regroupés dans la même classe par la classification hiérarchique. Les carrés représentent la zone de référence d'un thème donné.

Évaluation sur ESTER

Le critère d'évaluation utilisé sur ESTER est une mesure du degré de correspondance entre la classification hiérarchique en classes thématiques et une classification de référence plate réalisée manuellement. J'ai choisi pour cela la statistique *Kappa* [Sim et Wright, 2005]. Cette statistique est plus intéressante que le taux de reconnaissance[10], car elle mesure l'amélioration du taux de reconnaissance par rapport à un taux de reconnaissance aléatoire, obtenu en choisissant les réponses au hasard en fonction de la probabilité de chaque classe.

L'évaluation a cette fois été réalisée sur une tâche standard de classification de documents [Lee *et al.*, 2006] : chaque phrase du corpus ESTER a été manuellement annotée avec un thème parmi 5 thèmes possibles : international, politique, économie, culture et sport. Les phrases ne pouvant être affectées sans ambiguïté à l'un de ces thèmes sont supprimées. Un document est défini comme une séquence de parole sans interruption de durée totale inférieure à 5 minutes appartenant à un seul thème.

Il est difficile de comparer une classification non supervisée avec une classification de référence, car la correspondance entre ces deux classifications est inconnue. J'ai donc adopté la solution, à mon avis plus fiable, qui consiste à comparer les deux ensembles composés de toutes les relations "thématiques" entre deux documents que l'on peut extraire d'une part de la classification non supervisée, et d'autre part de la classification de référence. En d'autres termes, supposons que la classification automatique fournisse les deux classes suivantes :

$$A = \{e_1, e_2\} \quad \text{et} \quad B = \{e_3, e_4\}$$

et la classification de référence les deux classes :

$$C = \{e_1\} \quad \text{et} \quad D = \{e_2, e_3, e_4\}$$

Alors, nous pouvons dire que la classification automatique a identifié les relations sémantiques entre les paires suivantes :

$$((e_1, e_2), (e_3, e_4))$$

et la classification de référence les relations entre paires suivantes :

$$((e_2, e_3), (e_3, e_4), (e_2, e_4))$$

[10] ou plus précisément, l'accuracy

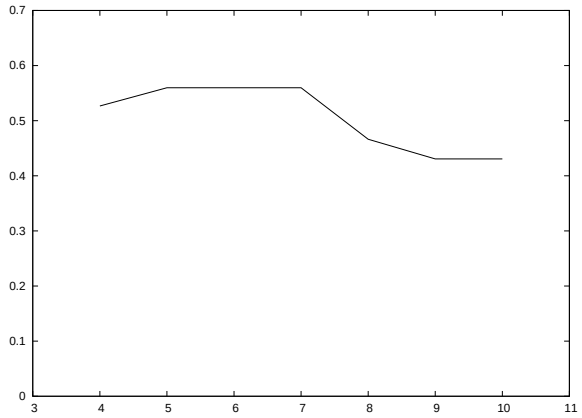


FIG. 5.6 – Taux de classification correct des morphèmes acoustiques dans les 4 classes de référence en fonction du nombre de classes découvertes.

Nous pouvons alors calculer le nombre de fausses acceptations (FA) et de faux rejets (FR) : (e_1, e_2) est une fausse acceptation tandis que (e_2, e_3) et (e_2, e_4) sont des faux rejets. Les mesures d'évaluation classiques peuvent alors être dérivées de ces quantités comme décrit par exemple dans [LingPipe, 2007]. L'avantage principal acquis en comparant les relations par paire plutôt que directement les classes est qu'aucune correspondance entre classes n'est requise : le résultat de la classification n'est plus interprété comme un groupe d'éléments, mais comme un ensemble de relations sémantiques découvertes entre ces éléments.

Trois systèmes sont évalués ainsi :

- **Référence :** Notre système de référence utilise la transcription manuelle de ESTER : il s'agit d'un système état de l'art en classification automatique de documents, qui réalise une classification hiérarchique ascendante à lien simple[11] des documents, en se basant sur la distance cosine dans l'espace des vecteurs termes-documents avec un filtrage TF-IDF des fréquences des termes : cet algorithme est décrit en détails et implémenté en JAVA dans [LingPipe, 2007].
- **Référence sans mots :** Il s'agit du même système que la référence, mais sans utiliser la transcription manuelle. À la place, les morphèmes acoustiques découverts par notre algorithme constituent les éléments lexicaux, ou les "termes" du document, tous les phonèmes ne faisant pas partie d'un morphème acoustique étant supprimés.
- **Lexique sémantique :** Ce système remplace de plus la distance TF-IDF cosine par la distance sémantique proposée. Il applique donc la méthode proposée pour construire des classes de morphèmes acoustiques, puis il affecte chaque document à une de ces classes en procédant à un simple vote majoritaire pondéré des morphèmes acoustiques composant le document. Le poids utilisé pour ce vote est inversement proportionnel au nombre de documents différents dans lequel le morphème acoustique apparaît.

Les figures 5.7, 5.8 et 5.9 montrent les performances respectives de ces trois systèmes en fonction du nombre de classes retenues.

Tout d'abord, notons que le nombre de classes est beaucoup plus grand pour le troisième système que pour les deux premiers, ce qui est normal, car il s'agit de classes de morphèmes acoustiques et non plus de classes de documents. Le tableau 5.5 résume les performances des trois systèmes pour un nombre de classes donné.

Une statistique *Kappa* égale à 1 représente un accord parfait entre deux classifications. Une valeur de 0 ou moins est considéré comme mauvais, tandis qu'un *Kappa* de 0.3 s'interprète comme un "accord raisonnable" [Sim et Wright, 2005]. Nous pouvons remarquer que la mesure *Kappa* ne décroît "que" de 0.31 à 0.22 lorsque l'information sur les mots est supprimée, ce qui reste relativement encourageant, étant donné que cette information est évidemment primordiale pour la classification thématique.

L'un des atouts de la méthode proposée est sa généricité, car elle ne s'appuie que sur très peu de

[11] single-link clustering

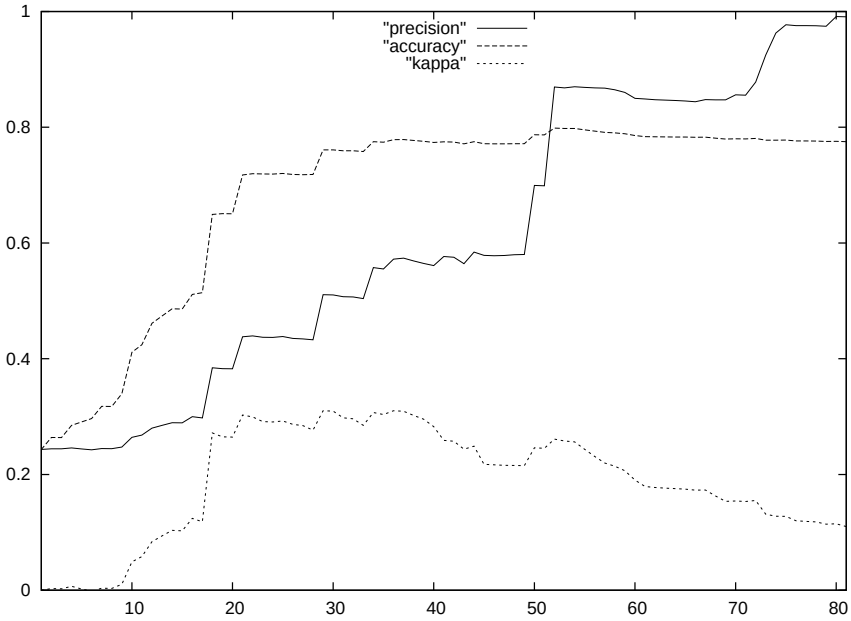


FIG. 5.7 – Performances du système de référence utilisant la transcription manuelle et la distance classique cosine TF-IDF sur une tâche de classification non supervisée de documents, en fonction du nombre de classes retenues.

	nb de classes	κ	Accuracy	Precision
Syst. de référence	36	0.31	77.8 % ± 0.7 %	57.2 %
Référence sans mots	36	0.22	69.3 % ± 1.1 %	41.2 %
Lexique sémantique	224	0.22	63.8 % ± 1.1 %	37.7 %

TAB. 5.5 – Comparaison des trois systèmes sur la tâche de classification non supervisée de documents.

contraintes. Mais cet atout constitue également sa faiblesse sur une tâche donnée, car, malgré les très nombreuses manières d'améliorer ces résultats (utiliser les n-meilleures chaînes de phonèmes, utiliser des coûts différents pour les insertions/suppressions/substitutions lors de la découverte des morphèmes acoustiques, choisir les morphèmes à prendre en compte selon des critères d'information mutuelle ou d'entropie, appliquer plusieurs passes de détection pour supprimer les morphèmes acoustiques dupliqués ou erronés, etc.), je ne pense pas qu'il soit possible de lui faire atteindre les mêmes performances sur une tâche donnée qu'un système utilisant un lexique connu, ne serait-ce qu'à cause de certains phénomènes linguistiques qu'il est très difficile de prendre en compte sans lexique prédéfini, comme les homophones qui portent un sens différent mais ont une acoustique semblable.

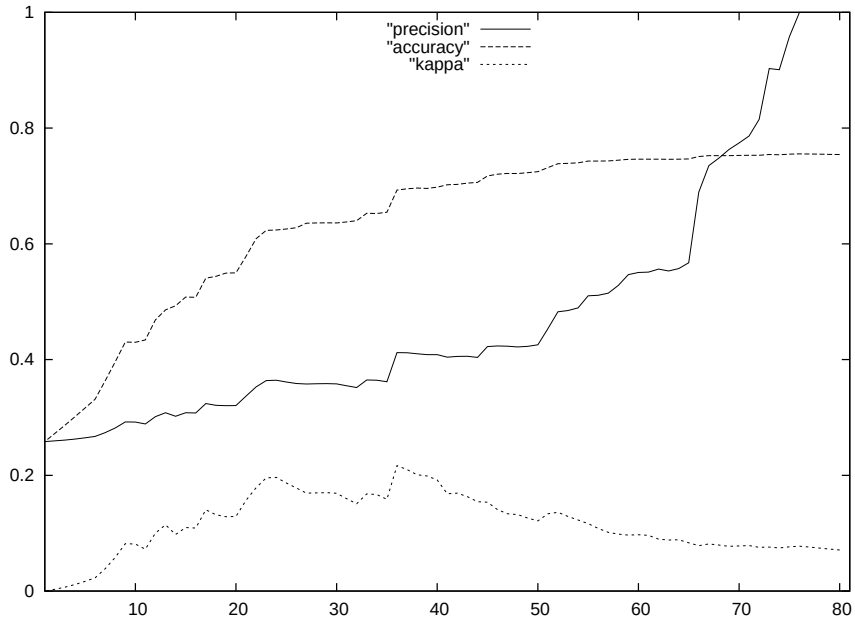


FIG. 5.8 – Performances du système utilisant les morphèmes acoustiques découverts et la distance classique cosine TF-IDF sur une tâche de classification non supervisée de documents, en fonction du nombre de classes retenues.

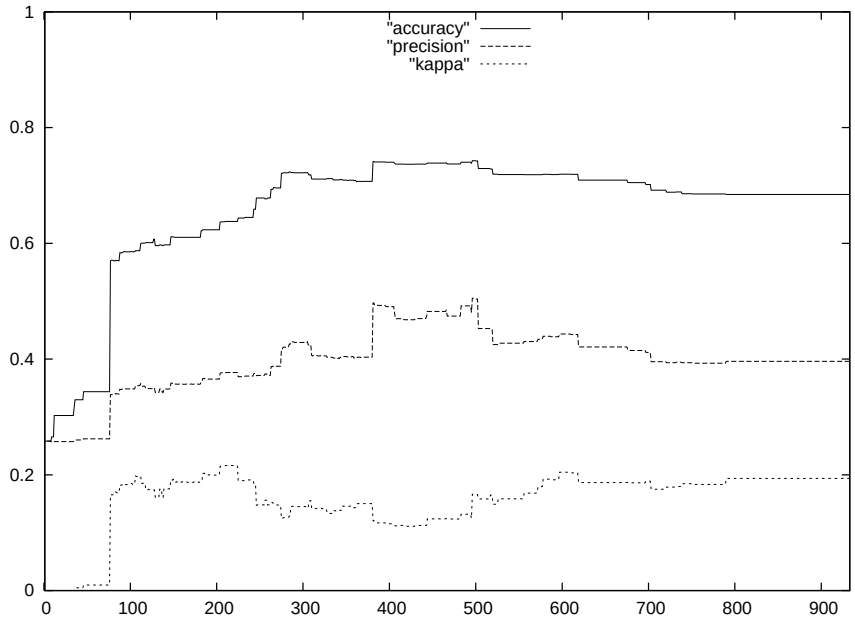


FIG. 5.9 – Performances du système utilisant les morphèmes acoustiques découverts et la distance proposée, en fonction du nombre de classes de morphèmes acoustiques retenues.

Projet de recherche

Intégrer reconnaissance automatique de la parole, syntaxe et sémantique

5.5 Présentation du projet de recherche

5.5.1 De Lao Zi à Turing

Depuis trente ans, la recherche en reconnaissance automatique de la parole développe des modèles stochastiques entraînés sur des corpus de parole de plus en plus grands. De nombreux types de modèles ont été appliqués à ce domaine de recherche : HMM, modèles de trajectoire, modèles bayésiens dynamiques, modèles multi-bandes, modèles hybrides, SVM, réseaux de neurones, boosting, etc. Tous ces modèles diffèrent par leur capacité de généralisation, leur algorithme d'apprentissage, ou la forme de découpage de l'espace des observations, mais ils possèdent tous un point commun fondamental : ils modélisent uniquement les observations acoustiques et lexicales. En d'autres termes, les approches actuelles ne possèdent pas une once d'intelligence, car elles ne font qu'apprendre à associer directement des observations à des classes prédéfinies, sans utiliser en aucune façon la signification ou l'interprétation que l'on peut donner à ces classes. Pourtant, les objets manipulés appartiennent au langage naturel, dont la caractéristique principale est l'extrême richesse et diversité des notions et concepts qu'il permet de manipuler. Ne pas prendre en compte les informations qui constituent le sens et le fondement du langage limite fortement à mon avis les perspectives des approches développées jusqu'à présent.

Dans mon projet de recherche à long terme, je veux rompre avec cette tradition ancrée dans le domaine de recherche de l'apprentissage automatique, qui considère la parole comme une donnée semblable à toute autre, en omettant sciemment ses caractéristiques qui la rendent pourtant si particulière et qui en font l'instrument de communication privilégié entre les Hommes et le moteur principal de l'évolution de l'humanité grâce au partage des connaissances qu'elle permet. Il me paraît donc essentiel de redonner du sens à la parole, car, d'une part, il est impossible de reconnaître parfaitement sans comprendre, et d'autre part, le langage oral est à l'origine de toute autre forme de langage évolué et en particulier de l'écrit.

Mais une multitude de chercheurs ont déjà tenté de modéliser la compréhension du langage naturel, sans grand succès il faut bien l'avouer, ce qui n'est guère étonnant, car la compréhension est une fonction intimement liée à l'intelligence humaine. À mon avis, l'étude de cette capacité que constitue la compréhension du langage naturel peut être abordée d'au moins deux façons duales, que je me permets d'attribuer dans la suite respectivement à Turing et à Lao Zi, en m'autorisant quelque liberté dans l'interprétation de leurs écrits.

L'un des fondateurs de notre discipline, Turing, place cette fonction au cœur du dialogue : selon ses préceptes, l'intelligence se définit par la capacité à soutenir une discussion avec autrui. C'est un point de vue qui est largement répandu, et qui a suscité de nombreux projets de recherche. Récemment, avec la montée en puissance de l'informatique, un regain d'intérêt s'est développé pour le test de Turing, et des dizaines de milliers d' "agents conversationnels" ont vu le jour sur Internet[12]. Plusieurs compétitions reproduisant le célèbre test de Turing ont également vu le jour, la plus célèbre d'entre elles étant le prix *Loebner*, qui organise et désigne annuellement le meilleur logiciel évalué par le test de Turing depuis le début des années 1990[13]. En 2009, cette compétition a d'ailleurs eu lieu conjointement avec INTERSPEECH, l'une des plus grandes conférences en reconnaissance automatique de la parole. Mais aucun de ces agents conversationnels ne peut prétendre modéliser réellement la compréhension du langage naturel, aux dires mêmes de leurs concepteurs. Leur principe de base consiste à reconnaître des *patterns*, éventuellement très complexes, dans les phrases de l'interlocuteur, et à réagir à ces *patterns* de manière prédéfinie. Le cœur de ces systèmes est donc une base de données de *patterns* la plus grande possible. Cette base constitue-t-elle un modèle du monde[14] primitif ? Peut-être, mais sa structure semble néanmoins bien pauvre pour remplir cette tâche convenablement.

S'il est envisageable de considérer que le test de Turing apporte la preuve de la compréhension, il ne dit rien quant au modèle sous-jacent et à son apprentissage qui autorise la compréhension. L'approche décrite ci-dessus consistant à développer des systèmes conversationnels dédiés au test de Turing est donc fondamentalement biaisée, car elle suppose implicitement que le dialogue, outil proposé par Turing pour mesurer l'intelligence, est également le moteur de son apprentissage. Or cette hypothèse est loin d'être

[12] Voir par exemple `http://www.chatterbots.com`

[13] `http://www.loebnerprize.com`

[14] Le modèle du monde est une des briques de base d'un modèle de compréhension, avec le graphe sémantique. Le "socle commun de connaissances" parfois également identifié dans les situations de dialogue fait partie du modèle du monde.

aussi convaincante. En effet, si nous envisageons la compréhension sous l'angle de la connaissance, car il faut connaître le monde et les concepts manipulés par le langage pour le comprendre, alors ce n'est pas tant le dialogue que l'écoute qui favorise l'apprentissage, ainsi que le stipule l'un des préceptes de la philosophie de Lao Zi, selon laquelle "celui qui sait ne parle pas, celui qui parle ne sait pas." Ce principe issu de la sagesse populaire, qui privilégie l'apprentissage par l'écoute, n'est plus tellement à la mode aujourd'hui, alors que de nouvelles théories cognitives développementales mettent en exergue l'importance de l'exploration active dans l'apprentissage humain [Kaplan *et al.*, 2008, Smith et Gasser, 2005]. Mais sans remettre en question ces découvertes, l'interaction dans ce cadre peut être interprétée avant tout comme un moyen pour générer de nouveaux stimuli accompagnés d'une certaine forme de supervision, l'essentiel de l'apprentissage restant la construction d'un modèle de compréhension issu du traitement des perceptions.

Je propose dans ce projet de développer les grandes lignes d'un tel modèle de compréhension de la parole, en accordant une importance particulière à la supervision "indirecte", non pas dérivée de l'observation des réactions de l'environnement comme pour les approches développementales, mais plutôt expoitant les commandes et actions des utilisateurs.

5.5.2 Ambition

Ce projet est (trop) ambitieux, car l'idée d'utiliser la compréhension pour aider la reconnaissance de la parole n'est évidemment pas nouvelle, et de nombreux chercheurs ont déjà tenté de modéliser la compréhension humaine, avec plus ou moins de succès. De même, les connaissances qui ont pu être intégrées efficacement dans les systèmes de reconnaissance automatique de la parole sont limitées aux caractéristiques physiques, directement mesurables, de notre oreille interne, comme par exemple la décomposition du signal acoustique en spectre fréquentiel et l'intégration des énergies acoustiques dans des bancs de filtres logarithmiques. Mais au-delà de ces pré-traitements de bas niveau, la manière dont le signal est interprété dans le cerveau nous est totalement inconnue, et aucune connaissance de cet ordre n'a jamais pu être modélisée automatiquement.

Pourtant, nous connaissons depuis longtemps l'existence de canaux afférents, qui contrôlent la manière dont l'oreille interne perçoit les sons en fonction des informations globales modélisées dans le cerveau. Ce contrôle descendant est mis en évidence par ailleurs dans de nombreuses expériences psycho-acoustiques sur lesquelles est bâtie l'analyse de scène auditive computationnelle (CASA), et qui ont abouti à la définition du modèle de Bregman décrit au paragraphe 2.3 de ce mémoire. Les illusions auditives, comme l'effet Mac-Gurk, illustrent également de manière spectaculaire l'importance des représentations abstraites prédites par le cerveau lors de l'interprétation des signaux auditifs perçus. Il existe donc des indices convaincants de l'importance de comprendre pour bien percevoir, mais ce phénomène n'a encore jamais pu être exploité avec succès dans les systèmes automatiques.

Il faudrait pour cela commencer par comprendre ce qu'est la compréhension. À défaut d'unanimité, de nombreuses définitions de la compréhension existent, qui dépendent du domaine de recherche considéré. En informatique, et plus spécifiquement en intelligence artificielle, le critère de Turing mentionné précédemment est resté une référence, et a abouti à un domaine de recherche dédié qui développe aujourd'hui ce que l'on appelle les « agents conversationnels », qui sont des algorithmes capables d'entretenir un dialogue avec quelqu'un. Mais ces agents ne comprennent pas vraiment ce qui est dit, et les techniques utilisées dans ces approches sont basées sur des heuristiques et sur des méthodes d'apprentissage automatique qui, au final, ne font que répéter des morceaux de phrases préalablement observées.

Les systèmes dits de « compréhension » de la parole s'appuient quant à eux sur un ensemble de concepts prédéfinis manuellement : ces concepts sont modélisés par une variable aléatoire, de la même manière que les mots ou les observations acoustiques, et les équations de base de la reconnaissance sont alors simplement modifiées pour intégrer ces nouvelles variables : au final, le système comprend effectivement ce qui a été dit, mais seulement pour un domaine d'application très restreint. Cette approche est incapable d'apprendre de nouveaux concepts, de déduire leur relation aux autres concepts, d'imaginer et de concevoir de nouvelles significations, et est inapplicable en l'état pour modéliser le monde dans sa globalité.

De même, en interrogation de base de données, la compréhension consiste simplement à traduire correctement une requête en langage naturel en la requête correspondante en SQL, puis à retrouver la

réponse à cette requête dans une base de données. Cette définition de la compréhension est éminemment pratique mais trop éloignée de la compréhension humaine.

La définition de compréhension en traitement automatique du langage naturel basé sur la logique s'appuie également sur des variables ou ontologies définies manuellement et représentant les concepts manipulables de notre monde, mais mets l'accent sur les capacités de généralisation, de subsumption et d'inférence, qui caractérisent de fait la cognition humaine. Toutefois, malgré de récentes avancées en création automatique d'ontologie, les facultés d'apprentissage permettant de construire un modèle du monde personnel mais global ne sont pour l'instant pas prises en compte dans ces approches.

Au sein de cette vaste littérature, mon projet de recherche développe des spécificités notamment liées à la combinaison entre reconnaissance et compréhension, un apprentissage peu supervisé et un modèle de compréhension combinant approches statistiques et formelles.

L'intérêt de combiner reconnaissance et compréhension est multiple : il s'agit tout d'abord d'exploiter un flux d'information beaucoup plus riche que le texte classiquement utilisé en compréhension. De plus, l'interaction orale avec des utilisateurs permet d'apporter une forme de supervision qui n'existe pas dans l'écrit traditionnel[15]. Enfin, le taux de reconnaissance est un objectif mesurable qui permet d'utiliser une méthodologie d'évaluation précise et élaborée depuis plus de trente ans. Cette fusion entre reconnaissance et compréhension devrait d'ailleurs permettre de libérer cette dernière des contraintes qui la limitent à des domaines d'application restreints et bien définis, grâce en particulier aux approches d'apprentissage non supervisé et de découverte de la sémantique. Mais pour ce faire, il faudra également développer des modèles de compréhension évolutifs qui s'appuient d'une part sur un graphe sémantique définissant les concepts manipulables et leurs relations, et d'autre part sur un modèle du monde qui intègre le résultat de la compréhension et lui fournit en retour un contexte riche pour interpréter les phrases.

Finalement, le contexte scientifique actuel n'est plus du tout le même que celui qui accompagnait les tentatives précédentes en intelligence artificielle, et il est à mon avis bien plus favorable aujourd'hui pour travailler sur l'intégration entre les différents domaines de recherche concernés. Un élément fort du projet de recherche est donc la multidisciplinarité, qui concerne d'abord le domaine de la reconnaissance de la parole, lui-même issu du domaine plus général de reconnaissance des formes et de l'intelligence artificielle. Le deuxième domaine fortement concerné dans ce projet de recherche est celui du traitement automatique du langage naturel (TAL), qui se différencie du premier essentiellement par son formalisme : réécriture et logique pour TAL, modèles bayésiens et stochastiques pour la reconnaissance. Un troisième domaine concerné est celui de l'apprentissage, et plus particulièrement de l'apprentissage non-supervisé et semi-supervisé. Un quatrième domaine connexe est celui de l'extraction automatique d'information à partir de corpus de parole et en particulier de l'acquisition automatique du lexique, de la syntaxe et de la sémantique des langues naturelles.

5.5.3 Positionnement au cœur des priorités nationales

Mon projet de recherche est fédérateur, car il s'inscrit dans les objectifs de recherche prioritaires du CNRS, de l'INRIA et du LORIA comme suit :

- CNRS (quatrième rupture scientifique du département ST2I) : "Traiter des données peu structurées et en extraire des connaissances : son, images, propriétés, valeurs numériques, langues, web sémantique..."
- INRIA (deuxième défi prioritaire) : "Développer le traitement des informations et données multimédia"
- LORIA (plan quadriennal 2009-2012) : premier thème scientifique "Traitement automatique des langues et des connaissances"

Le besoin scientifique en reconnaissance et compréhension automatique de la parole est également plus fort que jamais, du fait de l'explosion de la quantité de contenus audio et video sur Internet, comme l'a bien perçu Google en lançant son service de transcription automatique *Gaudi*[16].

[15] Une forme d'interaction écrite commence à apparaître dans les "chats" sur Internet, mais elle reste beaucoup moins développée qu'à l'oral, qui pourrait d'ailleurs progressivement la remplacer sur Internet.

[16] http ://labs.google.com/gaudi

Au-delà des liens avec la recherche d'information, le travail que je préconise sur l'apprentissage automatique est également un atout majeur pour le développement de nouveaux systèmes multilingues, pour lesquels il n'est pas question de construire manuellement des corpus très coûteux, comme cela a été fait pour l'anglais. Or, une conséquence de la mondialisation des communications est précisément de créer un besoin en technologies de la communication pour toutes les langues dites "pauvres", besoin qui tend à se renforcer de plus en plus vite.

5.5.4 Contexte international

Le contexte international devient de plus en plus favorable à la convergence des recherches en reconnaissance automatique de la parole et en linguistique. Mais il revient de loin : tous les chercheurs du domaine de la reconnaissance se souviennent de la sentence radicale prononcée par le célèbre professeur Jelinek en 1988, pendant le *Workshop on Evaluation of NLP Systems* à Wayne, en Pensylvanie : "every time I fire a linguist, the performance of our speech recognition systems goes up". A l'époque, cette phrase, volontairement provocatrice, reflétait surtout la déception et la frustration ressentie par les chercheurs du domaine, face aux échecs répétés de leurs tentatives d'apporter des connaissances expertes dans les systèmes de reconnaissance : il semblait alors que seules les solutions les moins "intelligentes", tels les apprentissages automatiques, soient efficaces.

Mais le contexte a profondément changé, et la vision actuelle de l'intégration de connaissances linguistiques en reconnaissance a également évolué. Il ne s'agit plus aujourd'hui de formaliser nos connaissances expertes et de les utiliser telles quelles pour simuler la compréhension humaine, mais bien plus de guider l'apprentissage automatique de modèles statistiques appris sur de grands corpus.

Deux exemples illustrent particulièrement bien cette évolution :

- l'analyse par sémantique latente (ou LSA) est une méthode apparue relativement récemment pour apprendre automatiquement des liens sémantiques entre certains mots (synonymes, antonymes, etc.) ;
- les analyseurs grammaticaux stochastiques apprennent automatiquement sur des corpus à identifier les relations syntaxiques entre les mots. Certaines approches récentes découvrent également ces relations de manière non-supervisée à partir de textes bruts, non annotés.

Ces exemples sont détaillés aux paragraphes 5.7.4 et 5.7.5.

Mon projet de recherche a naturellement une place dans le domaine du traitement automatique des langues, qui a toujours été, et reste un domaine de recherche prioritaire en Europe, comme le montre par exemple l'objectif 2.2 de l'appel à projet N°4 du 7ème programme-cadre sur les interactions par le langage et la traduction automatique[17], ou encore l'initiative *i2010 pour une Société de l'Information Européenne*, dans laquelle les technologies des langues ont une place primordiale[18].

5.6 Développement d'un projet collaboratif

Je ne peux évidemment pas mener à bien seul le projet de recherche décrit dans ce mémoire. C'est donc un projet que je compte mener au sein de l'équipe, mais également en tissant et en renforçant des liens avec d'autres chercheurs et à travers des projets collaboratifs nationaux et internationaux.

5.6.1 Un nouveau thème au sein de l'axe TAL du LORIA

Je travaille actuellement dans l'équipe PAROLE du LORIA. L'équipe PAROLE est relativement jeune (2001), mais elle a un passé important, car elle est directement issue de l'équipe RFIA, et certains de ses membres travaillent sur la reconnaissance automatique de la parole depuis bientôt trente ans. L'expérience de l'équipe concernant les technologies classiques de traitement de la parole constitue donc une source de connaissance importante qu'il faut exploiter autant que possible.

Depuis 2001, un nouvel élan lui est donné en ré-orientant la thématique de recherche vers la pluridisciplinarité reconnaissance de la parole / phonétique / production de la parole. Cette combinaison

[17] http://cordis.europa.eu/fp7/ict/language-technologies/fp7-call4_en.html
[18] http://europa.eu/languages/en/chapter/46

unique de chercheurs de différents horizons est également une richesse, car elle facilite les interactions entre les membres de l'équipe, et facilite ainsi la collaboration entre les disciplines.

Je voudrais participer pleinement à développer et à accentuer encore le caractère pluri-disciplinaire de l'équipe, non pas en recrutant de nouveaux membres de nouvelles disciplines, mais en m'associant plus fortement avec certains chercheurs de l'équipe et en collaborant plus fortement avec des chercheurs d'autres équipes et labratoires. L'impact de mon projet de recherche dans l'équipe sera donc la création d'un nouvel axe pluri-disciplinaire, dont l'objectif est l'intégration de connaissances syntaxiques et sémantiques en reconnaissance automatique de la parole et regroupant, en plus de moi-même, trois autres membres permanents intéressés par le thème développé dans mon projet de recherche : Dominique Fohr (CR1 CNRS), Odile Mella (MdC Nancy-1) et Irina Illina (MdC Nancy-2). Je serai ravi de porter et de développer ce nouvel axe de recherche qui, à mon avis, ouvre de nouvelles perspectives d'évolution de la priorité thématique TAL du LORIA.

5.6.2 Collaborations

La thématique développée dans mon projet de recherche s'appuie sur les compétences de chercheurs en linguistique et traitement automatique des langues. J'ai déjà mis en place et initié les collaborations suivantes, tout d'abord dans les projets européens OZONE et Amigo, et plus récemment dans l'ARC RAPSODIS et le CPER MISN TALC :

- collaboration au sein du LORIA avec l'équipe TALARIS, et plus spécifiquement Claire Gardent et Matthieu Quignard ;
- collaboration régionale avec le laboratoire ATILF, et plus spécifiquement avec Bertrand Gaiffe, Virginie André, Christophe Benzitou, Emmanuelle Canue, Jeanne-Marie Debaisieux et Evelyne Jaquet ; ces deux collaborations ont lieu dans le cadre du CPER TALC[19].
- collaboration nationale avec les équipes METISS et TEXMEX à Rennes, et plus spécifiquement avec Guillaume Gravier et Pascale Sébillot, qui s'intéressent également fortement à l'intégration d'information de haut niveau en reconnaissance, notamment depuis la thèse de Stéphane Huet ;
- collaboration nationale avec l'équipe LIST du CEA à Paris, et plus spécifiquement avec Guillaume Pitel et Olivier Ferret, qui travaillent sur les espaces sémantiques. ces deux collaborations ont lieu dans le cadre de l'ARC INRIA RAPSODIS[20].

J'ai également des contacts internationaux avec l'Université de Bohème de l'Ouest, qui est un candidat potentiel intéressé pour proposer de futurs projets internationaux.

5.6.3 Projets nationaux et internationaux

J'ai commencé à préparer le mise en œuvre concrète de ce projet de recherche à travers diverses actions, et en particulier :

- J'ai dirigé la thèse de Pavel Kral qui s'est terminée fin 2007 sur la reconnaissance automatique des actes de dialogue, combinée en cascade avec un système de reconnaissance automatique de la parole. Les actes de dialogue constituent concrètement les premières informations de haut niveau auxquelles je me suis intéressé. Aujourd'hui, le travail de Pavel Kral va nous servir de base pour regrouper les groupes de souffle issus de la reconnaissance de la parole en phrases analysables syntaxiquement.
- Un étudiant en thèse, Christian Gillot, financé par une bourse du ministère, a commencé à travailler en octobre 2008 sur le calcul et l'intégration des informations syntaxiques et sémantiques en reconnaissance automatique de la parole.
- J'ai également commencé des recherches personnelles sur l'extraction non supervisée d'un lexique sémantique, recherches qui ont été publiées dans une revue début 2009 [9].
- Nous avons demandé à l'INRIA un financement de post-doctorat pour 2010, sur les questions plus spécifiques d'apprentissage semi-supervisé pour l'analyse syntaxique de l'oral transcrit.
- J'ai proposé fin 2007 l'action de recherche concertée (ARC) INRIA RAPSODIS[20], qui a été acceptée sur la même thématique. Je coordonne donc depuis février 2008 cette ARC qui nous a permis, aux

[19] http://wikitalc.loria.fr/dokuwiki/doku.php?id=operations:aligne
[20] http://rapsodis.loria.fr

équipes de Rennes, de Nancy et du CEA à Paris, de défricher le terrain en étudiant différentes approches d'extraction d'information syntaxique et sémantique.
- Nous réfléchissons actuellement à la suite de cette ARC qui se terminera fin 2009, avec éventuellement la proposition d'une ANR ou d'un projet européen.

5.7 Développement scientifique

5.7.1 Perspectives à court et moyen terme

La spécialisation des domaines de recherche a permis la conception de théories très poussées, mais elle est également la cause du découpage de l'ensemble des informations qui participent à la perception humaine en catégories indépendantes : acoustique, phonétique, psycho-acoustique, linguistique computationnelle, sémantique, pragmatique, etc. Toutefois, nous savons que les processus cognitifs mis en œuvre sont étroitement imbriqués, et nos technologies "spécialisées" sont aujourd'hui fortement limitées à cause de cette approche focalisée sur un seul type d'information.

Ainsi, de multiples sources d'information participent à la perception de la parole. Les sources d'information que je souhaite aborder en priorité dans mon projet de recherche sont les suivantes :
- Evidemment, en premier lieu, les informations acoustique et lexicale qui sont utilisées dans tous les systèmes de reconnaissance automatique de la parole existants.
- L'information syntaxique, qui structure les phrases bien formées et permet la généralisation à partir d'exemples préalablement observés ;
- L'information thématique, qui indique le thème général abordé dans un discours, ou un ensemble de phrases, par exemple "météo", "politique", etc. J'associe cette information à une information sémantique "à large spectre", car, bien qu'elle soit dérivée de tous les mots prononcés, elle donne une indication générale, globale sur le sens porté par ces mots.
- Le sens précis porté par une phrase, que j'associe par opposition au point précédent à de la sémantique "fine". Par exemple, le thème de la phrase "il fait beau" est la météo, mais son sens est que, à un instant donné et pour un lieu donné, qui peuvent être déduits de l'historique du dialogue ou, lorsqu'ils sont implicites, du contexte plus général, le temps est beau.
  Pour travailler avec ce type d'information, je m'appuie sur les deux hypothèses suivantes, très fréquemment admises :

  1. L'hypothèse distributionnelle, selon laquelle les mots qui apparaissent dans des contextes similaires ont également des propriétés similaires (sens, fonction, ...) [Harris, 1985].
  2. L'hypothèse de compositionnalité, qui stipule que le sens d'une phrase est défini par la combinaison des sens des mots de la phrase.

- L'influence du contexte sur le sens d'une phrase est également très importante, et n'est pas capturée par les deux hypothèses ci-dessus. L'influence du contexte sur la sémantique est même parfois considérée comme prépondérante, comme par exemple dans le cadre de la sémantique interprétative de Rastier, selon laquelle le sens d'une phrase est avant tout interprété par la personne qui la perçoit ; ainsi, deux interlocuteurs différents interprèteront le sens d'une même phrase de deux manières différentes. Sans aller jusqu'à mettre en œuvre cette définition telle quelle, ce qui serait beaucoup trop ambitieux pour notre objectif, il est bon de conserver à l'esprit que le contexte influe sur le sens d'une phrase. Toutefois, le contexte est un concept extrêmement large, et recouvre de multiples notions qu'il serait déraisonnable de vouloir prendre en compte dans ce projet de recherche. Aussi, pour le moment, je réduis la notion de contexte à quelques éléments seulement, et notamment aux actes de dialogue décrits dans la suite, ou encore à l'information prosodique, qui contribue également largement à construire le sens d'une phrase, par exemple au travers de l'ironie, du sarcasme, de l'ennui, de la déférance, de l'insolence, etc.

Je détaille dans la suite les différents aspects sur lesquels j'ai déjà commencé à travailler, en commençant par les actes de dialogue, puis l'identificaion du thème, les espaces sémantiques, et l'analyse syntaxique, toutes ces recherches étant à chaque fois réalisées dans le cadre de la reconnaissance automatique de la parole. Les autres perspectives à plus long terme terminent ce chapitre.

5.7.2 Actes de dialogue

Les actes de dialogue influencent le sens de la phrase, et donc les conclusions que doit tirer le système après avoir perçu une phrase. Par exemple, selon le contexte et l'intonation, la phrase "il fait beau à Paris" peut être associée à une affirmation ou une interrogation. De même, lorsqu'un système automatique "entend" une phrase prononcée par un locuteur identifié, il est important que le système parvienne également à identifier l'interlocuteur "cible" de la phrase, par exemple pour décider s'il doit répondre ou non à une sollicitation. Enfin, différencier un ordre d'une simple affirmation aide considérablement le système automatique dans sa prise de décision. J'ai travaillé sur ces aspects depuis 2003 avec Pavel Kràl dans le cadre de sa thèse. Ce travail est décrit plus en détail au paragraphe 5.3. Les perspectives de ce travail sont importantes pour mon projet de recherche, en particulier en ce qui concerne le problème de la segmentation du flux audio en phrases, prélude nécessaire à l'analyse syntaxique. De même, à plus long terme, la reconnaissance automatique de l'ironie et des émotions dans un dialogue est importante pour mieux comprendre un dialogue.

5.7.3 Reconnaissance automatique du thème

Le thème du discours est l'information sémantique la plus simple qu'il est possible d'analyser. Nous avons travaillé sur ce problème depuis plusieurs années déjà dans l'équipe, notamment avec la thèse d'Armelle Brun sur la reconnaissance automatique du thème dans le texte [Brun, 2003], et plus récemment, avec la proposition d'une approche non supervisée qui classifie automatiquement un flux de parole perçue en segments thématiques cohérents, comme cela est décrit au paragraphe 5.4.

Mais ce travail ne fait qu'effleurer le problème de l'identification du thème sur l'oral, et laisse ouvertes de nombreuses questions qu'il conviendra d'aborder à plus ou moins long terme, comme l'appariement de thèmes découverts avec des thèmes existants, le niveau de précision requis et suffisant dans la hiérarchie des thèmes pour l'application visée, la remise en cause même de cette structure hiérarchique, qui peut parfois sembler trop rigide, et l'utilisation efficace de cette information pour la reconnaissance.

5.7.4 Espaces sémantiques

L'hypothèse distributionnelle suppose que des informations relatives à la fonction des mots, c'est-à-dire à la syntaxe et à la sémantique, peuvent être déduites automatiquement en analysant de grands corpus de texte brut. Elle est à l'origine d'une classe de modèles, appelés "modèles vecteur-espace", qui associent à chaque mot du lexique un vecteur caractéristique construit à partir de l'analyse des documents dans lesquels le mot apparaît. Intuitivement, cette hypothèse revient donc à déduire des informations par l'étude des collocations des mots, le type d'information extraite précisément dépendant de la définition de la notion de document.

Le représentant le plus connu de cette famille de modèle est l'analyse par sémantique latente (LSA), mais il en existe de nombreuses autres, comme l'analyse par sémantique latente probabiliste, l'indexation aléatoire, l'allocation de Dirichlet latente, certaines approches basées sur l'information mutuelle, etc.

Je présente brièvement ci-dessous l'analyse par sémantique latente, qui décrit intuitivement les principes fondamentaux de toutes ces approches, puis deux approches sur lesquelles j'ai travaillé récemment : l'indexation aléatoire et la carte sémantique.

Analyse par sémantique latente

Le paradoxe de la "pauvreté du stimulus" a été énoncé depuis longtemps [Quine, 1960], et constitue l'une des motivations de l'invention des grammaires de profondeur par Chomsky [Comsky, 1991]. Ce paradoxe s'appuie sur l'observation de l'apprentissage de nouveaux mots par les enfants d'environ deux ans. Certaines études ont montré que les enfants de cet âge pouvaient apprendre jusqu'à huit nouveaux mots par jour, ce qui paraît beaucoup. Le paradoxe vient du fait que les stimuli supervisés auxquels sont soumis ces enfants sont insuffisants pour expliquer une telle capacité d'apprentissage. Ce paradoxe peut être résolu par l'hypothèse de l'existence d'une structure syntaxique profonde innée, comme l'a supposé Chomsky ; mais des études plus récentes suggèrent une autre explication, liée à l'apprentissage des nouveaux mots par leur contexte : intuitivement, l'utilisation d'un mot inconnu est déduite de l'observation

de ce mot dans différents contextes, contextes qui rappellent d'autres mots connus qui peuvent alors servir de modèle pour le nouveau mot.

C'est le principe même de l'analyse par sémantique latente. L'approche de base est la suivante :

- Dans un premier temps, le corpus de texte est découpé en documents, chaque document représentant par exemple un paragraphe.
- Une matrice de grande dimension est alors créée, dans laquelle chaque ligne correspond à un mot du vocabulaire et chaque colonne à un document du corpus.
- Cette matrice est remplie en comptant simplement le nombre d'occurences d'un mot donné dans un document donné.
- Chaque mot est ainsi associé à un vecteur caractéristique de très grande dimension (le nombre de documents). Cette dimension est finalement réduite par une décomposition en valeurs singulières qui ne conserve que les dimensions de plus grandes valeurs propres.

La réduction de la dimensionnalité des vecteurs des mots est essentielle, car elle permet d'éliminer une grande part de la variabilité qui n'est pas celle contenant le plus d'information. Les vecteurs de mots résultants les plus proches, selon la distance en cosinus classiquement utilisée, sont également ceux qui apparaissent dans des contextes similaires, c'est-à-dire en compagnie des mêmes autres mots. Il est donc possible, grâce à cette distance, de créer automatiquement des classes de mots qui sont sémantiquement "proches". Ainsi, lorsque les documents considérés sont grands, la distance entre les mots représente la proximité thématique, tandis que lorsque seul un contexte linéaire limité est considéré (quelques mots à gauche et à droite), les classes représentent plutôt des synonymes ou des antonymes. La LSA a notamment été évaluée sur un test de synonymie issu du TOEFL, et le système automatique a obtenu un score équivalent à celui d'un apprenant non-natif moyen.

Indexation aléatoire

Le principal problème de la LSA est la difficulté voire l'impossibilité de réaliser une décomposition en valeurs singulières lorsque le corpus devient trop grand. L'indexation aléatoire [Kanerva *et al.*, 2000, Sahlgren, 2005] est une approximation incrémentale de la LSA qui s'appuie sur le théorème de Johnson-Lindenstrauss. Ce théorème démontre que lorsque l'on projette des vecteurs de dimension n dans un sous-espace de k dimensions choisies aléatoirement tel que k est logarithmique par rapport à n, alors la distance entre les vecteurs est préservée avec une précision inférieure à un facteur arbitraire [Achlioptas, 2001].

Cette propriété est exploitée ici en définissant *a priori* l'espace de projection, à la différence de la LSA qui le calcule *a posteriori*. Il est ainsi possible de mettre à jour les vecteurs caractéristiques des mots au fur et à mesure que les documents sont observés, sans avoir recours à des inversions de matrice.

Une description plus détaillée de l'algorithme de base est le suivant :

- Choisir la dimension de l'espace final, généralement de une à quelques centaines de dimensions.
- Associer à chaque mot du lexique un vecteur d'index de la dimension choisie, constitué d'une faible proportion de -1 et de +1, le reste du vecteur étant composé de 0 ; les dimensions nulles, positives ou négatives sont choisies aléatoirement.
- Associer à chaque mot du lexique un vecteur de contexte de la même dimension et initialement nul.
- Pour chaque mot d'un nouveau document, cumuler dans le vecteur de contexte de ce mot les vecteurs d'index des mots co-occurents dans le document. C'est au cours de cette étape que l'information distributionnelle sur les mots est extraite.
- Les vecteurs de contexte sont les vecteurs caractéristiques des mots désirés.

Il existe bien sûr de nombreuses variantes de cet algorithme en fonction du type de normalisation choisie : nous avons choisi pour notre part de normaliser par la fréquence des termes.

Nous avons réalisé des études préliminaires dans ce domaine dans le cadre du projet RAPSODIS, avec un stagiaire en 2008, dont le rapport est consultable en ligne[21]. Le travail réalisé s'inspire d'un article de Stephen Cox [Cox et Dasmahapatra, 2002], qui a entraîné la LSA sur des paragraphes du Wall Street Journal afin de mesurer la cohérence thématique des phrases reconnues par un système automatique. Cox a ainsi construit une mesure de confiance thématique et l'a comparée à une mesure de confiance fondée

[21] `http://rapsodis.loria.fr/reunions.html`

sur les vraissemblances acoustico-linguistiques fournies par le système de reconnaissance pour une liste des N-meilleures solutions.

Nous avons reproduit ces expériences, mais sur le français (quotidien "Le Monde") et en utilisant l'indexation aléatoire à la place de la LSA. Nous avons également comparé et combiné notre mesure de confiance thématique avec la mesure de confiance de Wessel-Ney [Wessel *et al.*, 2001], qui est une référence dans ce domaine. Les premiers résultats obtenus avec cette approche sont comparables à ceux présentés dans les travaux de Cox : nous avons ainsi pu valider nos algorithmes en extrayant automatiquement des listes de mots-outils, c'est-à-dire des mots peu chargés sémantiquement, et nous avons montré que l'information thématique modélisée ainsi est potentiellement intéressante, car complémentaire de celle issue du système de reconnaissance. Toutefois, en pratique, les approches testées pour combiner les deux mesures de confiance n'ont pas abouti à une amélioration des taux de détection des mots faux. Ceci est certainement dû en partie à la simplicité des méthodes de combinaison testées (linéaire et en cascade), mais il existe d'autres problèmes plus importants liés à l'approche thématique qui concernent par exemple le potentiel relativement faible d'identification des erreurs de reconnaissance sur base purement thématique.

Pour espérer un gain conséquent en reconnaissance, il faut donc à mon avis modéliser une information sémantique plus fine que la seule thématique, ce qui est envisageable en combinant indexation aléatoire et analyse syntaxique, comme présenté ci-dessous dans la "carte sémantique".

La carte sémantique

Toujours dans la cadre du projet RAPSODIS, nous avons collaboré avec l'équipe LIST du CEA à Paris, qui a une forte compétence en analyse distributionnelle. Cette équipe a ainsi développé très récemment un modèle de "carte sémantique", qui correspond à un modèle vecteur-espace dans lequel les documents sont le produit d'une analyse syntaxique. Ainsi, deux termes sont considérés comme co-occurrents lorsqu'ils appartiennent à un même "triplet" syntaxique (gouverneur, dépendant, relation). Par exemple, dans la phrase "Je suis vivant", il y a 2 documents, les triplets (suis, je, sujet) et (suis, vivant, attribut du sujet).

Avec cette définition de contexte, l'indexation aléatoire produit un espace dans lequel deux termes sont proches s'ils se comportent similairement par rapport aux triplets syntaxiques. Ainsi, intuitivement, tous les objets comestibles qui sont complément d'objet direct de verbes comme "manger", "avaler" ou "dévorer" devraient être proches dans cet espace. Nous voyons que l'information sémantique ainsi capturée est plus précise que l'information thématique préalablement considérée.

Le CEA a donc construit un tel espace en aspirant des millions d'URL du web francophone, et a dénommé cet espace la "carte sémantique". Elle a déjà été utilisée pour désambiguiser les mots [22] et pour reconnaître les émotions dans du texte [Pitel et Grefenstette, 2008].

Cette approche est très intéressante car elle définit un formalisme de base pour manipuler des informations sémantiques précises. Toutefois, certains problèmes comme sa dépendance par rapport à un analyseur syntaxique donné et le biais sémantique introduit par le choix du corpus initial, limitent pour le moment son efficacité. Nous souhaitons néanmoins surmonter ces difficultés et continuer à travailler dans cette voie.

Les perspectives de ces travaux pour mieux comprendre la résultat de la reconnaissance automatique de la parole sont nombreuses. Par exemple, la reconnaissance des émotions pourrait être certainement améliorée pour l'oral en incluant également nos travaux de l'équipe sur la prosodie, ce qui constituerait une extension de nos travaux précédents sur la reconnaissance des actes de dialogue ; je reviendrai sur d'autres évolutions de ces travaux dans la suite. Une autre perspective sur laquelle nous envisageons de travailler à moyen terme est l'application de cette information sémantique à la détection des mots mal reconnus et à leur correction. Mais pour ce faire, il faut préalablement étudier plus en détails les difficultés et les spécificités liées à l'analyse syntaxique sur de l'oral reconnu automatiquement, recherches que je décris brièvement dans la suite.

5.7.5 Prise en compte de la syntaxe

Une application directe de la syntaxe à la reconnaissance de la parole s'appuie sur l'hypothèse selon laquelle la syntaxe des phrases correctes est plus probable que la syntaxe des phrases erronées. C'est une

[22] rapport disponible sur le site `http://rapsodis.loria.fr`

hypothèse raisonnable car l'espace de toutes les combinaisons de mots possibles est exponentiel, et les phrases effectivement utilisées n'en occupent qu'une portion très réduite.

Le défi à relever revient donc à calculer un score de "grammaticalité" d'une phrase quelconque, puis à ajouter par exemple ce score au score acoustique du système de reconnaissance. Une autre possibilité particulièrement intéressante pour exploiter un tel score est de l'interpréter comme une mesure de confiance sur les mots ; cette mesure identifierait ainsi les mots qui, par leur inadéquation syntaxique, seraient vraisemblablement erronés. Il serait alors possible d'appliquer une reconnaissance avec données manquantes comme celle présentée au paragraphe 4.7. Mais nous n'en sommes pas là, car il convient d'étudier auparavant la validité et les conditions d'application de l'hypothèse mentionnée précédemment.

J'ai commencé à étudier les connections possibles entre le domaine de la syntaxe et celui de la reconnaissance, et je résume très brièvement quelques-unes de ces études dans la suite. Merci de bien vouloir noter que les quelques résultats expérimentaux présentés dans la suite ne sont donnés qu'à titre indicatif et ne constituent que des études exploratoires relativement peu poussées qu'il faut donc interpréter avec la plus grande précaution.

Score dérivé d'une analyse stochastique

L'analyse syntaxique traditionnelle est réalisée grâce à un ensemble de règles, généralement construites manuellement. Un tel analyseur syntaxique du français est par exemple le logiciel Syntex, que j'ai utilisé pour réaliser mes premières expériences. Notons que Syntex ne réalise pas d'analyse en structures de phrases au sens traditionnel, mais en dépendances. Je ne veux pas entrer ici dans une explication sur les différences et les points communs entre ces deux types d'analyse, mais pour simplifier, l'analyse en structures de phrases crée un arbre dont les feuilles représentent les mots et les nœuds internes des non-terminaux comme "groupe nominal", tandis que l'analyse en dépendance peuple tous les nœuds de l'arbre avec des mots.

La figure 5.10 montre une analyse en dépendance issue de Syntex sur deux phrases reconnues par le système de reconnaissance avec lequel nous travaillons dans l'équipe.

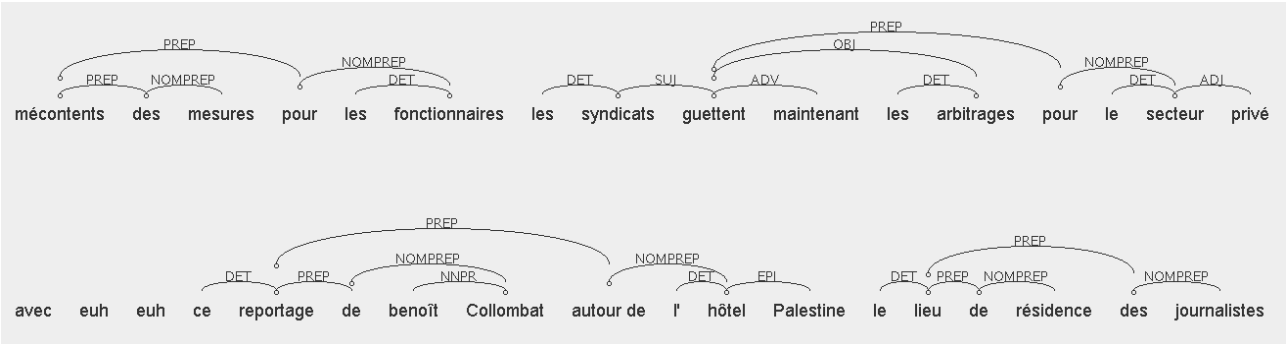


FIG. 5.10 – Exemple d'analyse syntaxique en dépendances réalisée par Syntex. Les ronds aux extrémités des arcs indiquent le gouverneur de la relation.

Le principal problème avec les analyseurs syntaxiques à base de règle est qu'ils ne calculent pas de score. Pour cela, je me suis intéressé à d'autres types d'analyseurs syntaxiques : les analyseurs stochastiques. Ces derniers n'utilisent généralement pas de règles prédéfinies, mais entraînent des modèles stochastiques sur de grands corpus de textes arborés, comme le "Penn-Tree-Bank" anglais. Ils apprennent ainsi automatiquement pour chaque phrase d'apprentissage une succession d'opérations à réaliser pour reproduire l'arbre d'analyse. Les modèles impliqués dans cet apprentissage peuvent être des modèles à base de mémoire, des modèles bayésiens, des SVM, etc, l'important étant qu'ils fournissent une (ou plusieurs) opération à réaliser à un instant donné en fonction d'un contexte observé ainsi qu'un score associé à cette opération. Lors du test, ils peuvent ainsi rechercher la suite d'opérations la plus vraisemblable pour toute phrase.

Un analyseur stochastique très connu est le "Standford parser", mais il produit des structures de phrase, et par cohérence avec Syntex, j'ai plutôt travaillé avec le "Malt parser". Cet outil empile les mots d'une phrase dans une pile de mots "à traiter" (PIN), et utilise une deuxième pile de travail (PW). Il définit alors quatre opérations de base, qui sont :

- SHIFT : transfert du mot en haut de PIN vers PW ;
- REDUCE : élimination du mot en haut de PW ;
- LEFTARC : création d'une dépendance du mot en haut de PW vers le mot en haut de PIN, et élimination du premier ;
- RIGHTARC : création d'une dépendance du mot en haut de PIN vers le mot en haut de PW, et transfert du premier ;

Malheureusement, les ressources arborées en français ne sont pas comparables en taille et en qualité aux ressources anglophones, et je me suis donc appuyé sur Syntex pour analyser une partie du corpus "Le Monde", et apprendre ensuite sur ce corpus avec l'analyseur Malt quelles dépendances sont le plus souvent créées par Syntex en fonction du contexte observé (mots en haut des piles, leurs classes morphosyntaxiques, etc.).

Les études que j'ai menées jusqu'à présent suggèrent que le score global issu de Malt ne semble pas totalement corrélé à la grammaticalité des phrases, car Syntex produit de nombreux petits groupes syntaxiques isolés, qui favorisent exagérément la probabilité d'apparition du SHIFT dans Malt. En d'autres termes, l'analyse de Syntex n'est pas de qualité suffisante pour évaluer correctement la grammaticalité d'une phrase.

L'un des biais que nous envisageons de traiter en priorité est l'absence de frontières de phrases dans le résultat de la reconnaissance, qui sont indispensables pour réaliser une bonne analyse syntaxique. Nous [23] travaillons donc en particulier sur la détection automatique des fins de phrase, qui est réalisée grâce à des indices prosodiques et lexicaux. Ce travail constitue un prolongement des recherches menées pendant la thèse de Pavel Kràl sur la reconnaissance automatique des actes de dialogue.

Au lieu de calculer le score global de la phrase, il est également possible d'exploiter le score local calculé dans Malt pour construire une mesure de confiance en mots basée sur une information syntaxique. Succintement, le modèle utilisé dans Malt retourne, pour un vecteur d'observation particulier associé aux deux mots en haut des deux piles, un score corrélé au nombre de fois où un tel vecteur a été vu dans le corpus d'apprentissage. Chaque mot peut être impliqué dans plusieurs vecteurs de ce type, et l'hypothèse de base est que les vecteurs associés aux mots faux apparaissent moins souvent que ceux associés aux mots corrects. Pour chaque mot, il est ainsi possible de définir une mesure de confiance simple qui est le rapport du nombre de vecteurs déjà vus sur le nombre vecteurs jamais observés.

La figure 5.11 montre la courbe DET de détection des mots mal reconnus obtenue avec cette mesure syntaxique.

L'erreur égale (EER) est de 40 %, ce qui est nettement moins bon que les 22 % obtenus avec la mesure de confiance de Wessel-Ney, mais les mots erronés sont détectés avec notre mesure uniquement sur des considération syntaxiques lexicalisées, et sont donc probablement complémentaires des erreurs détectées par la mesure de Wessel-Ney. De plus, ces résultats ne sont que des résultats préliminaires obtenus en n'entraînant Malt que sur un seul mois du corpus "Le Monde", destinés à valider les hypothèses sous-jacentes et à étudier les pistes de recherche les plus prometteuses.

Autres indices statistiques

Les pistes de recherche présentées jusqu'à présent sont essentiellement génératives, car elles analysent de grandes quantités de données bien formées syntaxiquement ou sémantiquement, et détectent les erreurs de reconnaissance par leurs différences avec ces données d'apprentissage. Une autre perspective intéressante est par contraste discriminante, car il s'agit d'apprendre des descripteurs positifs et négatifs, de phrases bien reconnues et erronées, dans un modèle qui choisira ensuite l'une de ces deux classes en fonction des observations.

Le principal problème est alors de construire des descripteurs syntaxiques qui différencient au mieux les mots ou phrases corrects des mots ou phrases erronés.

[23] Christian Gillot et moi-même, mais aussi Pavel Kràl dans le cadre d'une collaboration informelle avec son université en Tchéquie et Guillaume Gravier à Rennes.

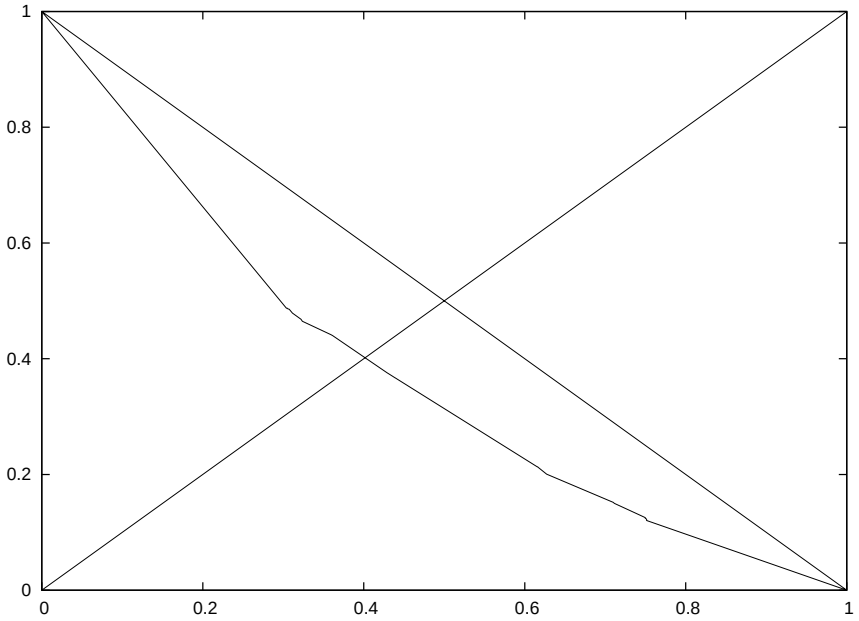


FIG. 5.11 – Courbe DET pour la mesure de confiance syntaxique.

Un tel descripteur envisagé est la taille des groupes syntaxiques générés. L'intuition nous suggère en effet qu'un analyseur syntaxique va avoir plus de mal à trouver des dépendances entre des mots faux qu'entre des mots justes. Cette intuition est en grande partie vérifiée, comme le montrent les figures 5.12 et 5.13, qui affichent la distribution de la taille des groupes syntaxiques incluant un mot donné, respectivement lorsque ce mot est faux et lorsqu'il est juste.

J'ai également proposé un autre descripteur syntaxique : le "score de recouvrement syntaxique" qui est simplement, pour chaque phrase, le rapport du nombre de mot ayant un gouverneur et du nombre de mots n'en ayant pas. Pour étudier si ce score permet de discriminer les phrases correctes des phrases incorrectes, j'ai mis en compétition chaque phrase reconnue avec la phrase de référence, et j'ai remplacé la phrase reconnue par la phrase correcte lorsque cette dernière a un score de recouvrement syntaxique strictement supérieur. Alors, le taux de reconnaissance passe de 88.87 % à 97.45 %, ce qui prouve qu'un tel indice est caractéristique, dans une certaine mesure, de la correction grammaticale des phrases.

5.7.6 Perspectives à long terme

Mon projet de recherche décrit précédemment propose d'étudier les informations syntaxiques et sémantiques pour améliorer la reconnaissance automatique de la parole. J'ai présenté ci-dessus plusieurs pistes de recherche différentes, dans la mesure du possible accompagnées de quelques études préliminaires, afin de montrer que ce projet de recherche n'est pas construit à partir de seules idées abstraites, mais qu'il est le fruit d'une réflexion menée depuis plusieurs mois voire plusieurs années (notamment depuis la thèse de Pavel Kràl), réflexion étayée par le résultat de mes recherches présentées dans la première partie de ce mémoire et par plusieurs expériences dédiées spécifiquement aux nouveaux aspects que je compte aborder. De plus, il me semble important de préciser que, malgré les apparences, la rupture avec le domaine de la reconnaissance automatique de la parole n'en est pas une, car, au contraire, l'intégration finale avec mes travaux sur la reconnaissance avec données manquantes par exemple est planifiée, de même que mes travaux précédents sur les actes de dialogue constituent naturellement une base pour développer ces nouvelles idées.

L'objectif de mon projet de recherche, de par son ambition, offre des perspectives à long terme. Les quelques éléments présentés ci-dessus sont des pistes de recherche concrètes pour progresser pendant les quelques années à venir. En ce qui concerne les développements à long terme de mon projet de recherche,

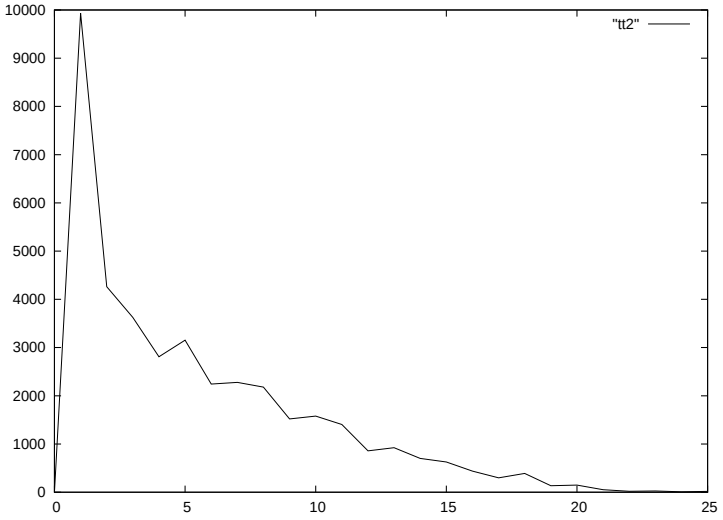


FIG. 5.12 – Distribution de la taille des groupes syntaxiques pour les mots faux

il ne serait pas raisonnable de les décrire avec le même niveau de précision, et je me contente donc ici de dépeindre les grandes lignes qui me paraissent potentiellement les plus prometteuses.

Tout d'abord, il me paraît important de ne pas privilégier les approches complètement automatiques, mais de toujours inclure les utilisateurs et leurs interactions dans nos modèles. Cette vision anthropocentrique m'est notamment inspirée par mes travaux dans le domaine de l'intelligence ambiante, mais également par mes contacts récents avec les linguistes qui m'ont conforté dans mon opinion que le "tout automatique" n'est pas une voie d'avenir. Concrètement, cela signifie par exemple que les algorithmes d'apprentissage tels qu'ils sont utilisés aujourd'hui présentent à mon avis des défauts importants qu'il conviendra de résoudre : ainsi, l'apprentissage "supervisé" suppose l'existence d'une base étiquettée d'une manière bien précise, ce qui est extrêmement coûteux en temps humain. De même, l'apprentissage "non-supervisé" livré à lui-même, comme le clustering de données brutes, ne découvre que les variabilités les plus importantes des données et ne peut pas positionner, interpréter et adapter ses découvertes par rapport à son environnement. Les solutions prétendues intermédiaires, comme l'apprentissage "semi-supervisé", consistent le plus souvent à itérer l'apprentissage en filtrant les exemples de moindre confiance, éventuellement en demandant à l'utilisateur de corriger les étiquettes de certaines données dans le cas de l' "active learning". Mais dans tous les cas, l'apprentissage impose à l'utilisateur son format et une tâche très précise, ce qui est contraignant d'un point de vue cognitif. Il me semble important dans ce domaine d'adapter au contraire l'apprentissage à l'utilisateur, en concevant par exemple un apprentissage "guidé" ou "opportuniste", qui exploite au mieux les informations fournies par l'utilisateur, quel que soit le moment où elles arrivent et même si elles ne respectent pas le format des données initiales. Un tel apprentissage requièrerait beaucoup moins d'effort de la part des utilisateurs, et s'insèrerait naturellement dans les aspects collaboratifs mis en exergue dans le WEB 2.0. Il permettrait également de faire grandement progresser quasiment tous les aspects de mon projet de recherche, par exemple les analyseurs syntaxiques stochastiques ou les graphes sémantiques.

Ainsi, je souhaite en particulier explorer dans mes perspectives à long terme ces aspects d'apprentissage automatique de la syntaxe, comme proposé par exemple dans les thèses de Clark [Clark, 2001], puis de Klein [Klein, 2005] et de Solan [Solan, 2006], ou encore dans l'approche [Bod, 2006]. Ces approches restent incontestablement inférieures aux grammaires apprises sur des corpus arborés manuellement, mais elles sont plus facilement adaptables à de nouvelles tâches, ou aux spécificités du langage oral qui ne sont pas couvertes actuellements par les corpus arborés écrits. De plus, l'extension de ces méthodes dans le paradigme de l'apprentissage "opportuniste" décrit précédemment devrait permettre d'améliorer l'utilité des structures découvertes en fonction des besoins de l'application, c'est-à-dire dans mon cas la transcription de la parole.

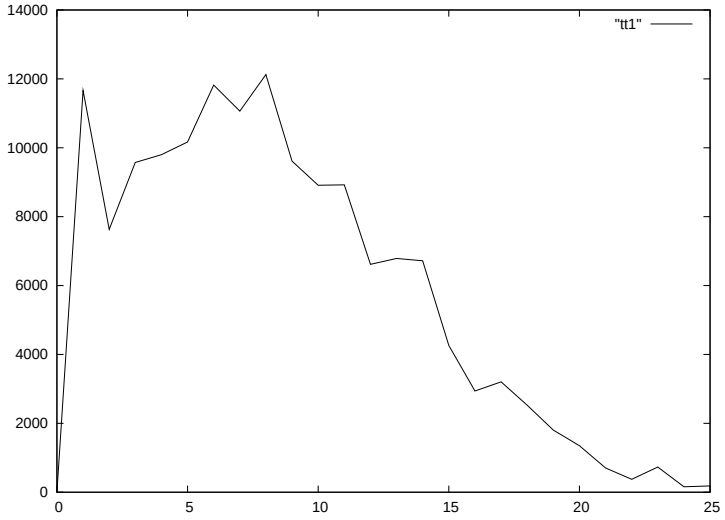


FIG. 5.13 – Distribution de la taille des groupes syntaxiques pour les mots justes

Enfin, il me semble que tous ces aspects d'apprentissage bénéficieraient largement d'une meilleure prise en compte de la mémoire à long terme des observations. Le courant de recherche concernant les classifieurs à base de mémoire a pris récemment beaucoup d'importance, notamment depuis les travaux sur la mémoire "épisodique" dans le cerveau, mais aussi grâce aux nouvelles capacités de stockage disponibles aujourd'hui et surtout à la possibilité de distribuer à grande échelle les données et les calculs, possibilité qui s'inscrit parfaitement dans le paradigme de l'apprentissage basée sur la mémoire. Ces nouvelles dispositions sont également des éléments supplémentaires qui viennent renforcer ma proposition d'apprentissage opportuniste décrit ci-dessus, car dans un tel apprentissage, les modèles évoluent en permanence, et les observations passées peuvent être reconsidérées sous un jour nouveau par la suite. Il est donc important de ne pas détruire l'information originelle (ce qui est le cas avec les techniques d'apprentissage classique qui extrayent une information utile à un instant donné, puis qui détruisent l'observation), mais de la préserver, au moins lorsqu'elle est porteuse d'une information non prise en compte par ailleurs. Ainsi, la mémorisation d'un grand nombre de fragments de phrases déjà observés me semble une perspective intéressante pour modéliser la structure de l'oral, voire la compréhension du discours.

Conclusion

Ce mémoire a commencé par une discussion sur les sources d'information que constituent les hypothèses de modélisation couramment utilisées dans les méthodes de reconnaissance automatique de la parole robuste au bruit. Ensuite, un bref survol de l'état de l'art du domaine a été réalisé, qui a essentiellement servi à introduire les contributions principales que j'ai pu apportées dans chaque partie. À la fin de chaque chapitre, un résumé de mes contributions principales est donné. Jusqu'au chapitre 4 inclus, seules les informations classiquement utilisées dans un système de reconnaissance sont considérées. Le chapitre 5 est plus exploratoire et propose d'intégrer les systèmes de reconnaissance de la parole avec de nouvelles sources de connaissance dans le cadre des architectures logicielles d'intelligence ambiantes, qui sont amenées à mon avis à remplacer progressivement les systèmes actuels. Pour le moment, seuls les travaux que j'ai réalisés sur l'extraction de ces informations sont mentionnés, car j'envisage seulement aujourd'hui de commencer à travailler sur l'intégration de telles connaissances pour améliorer la reconnaissance de la parole. Deux nouvelles sources d'information sont décrites : les actes de dialogue, et quelques notions de sémantique à travers le thème du discours. À la différence des chapitres précédents, la chapitre 5 ne réalise pas d'état de l'art du domaine mais présente seulement mes contributions, car les domaines abordés sont très vastes et les travaux décrits sont essentiellement exploratoires. Toutefois, le lecteur intéressé peut se reporter aux articles suivants que nous avons publiés et qui présentent respectivement un état de l'art de l'intelligence ambiante [Cerisara et Haradji, 2007], de la reconnaissance des actes de dialogue [Kral, 2007] et de l'extraction d'informations thématiques [Cerisara, 2009].

Dans la mesure du possible, mes contributions présentées dans ce mémoire sont analysées en terme d'information et de contraintes apportées par les hypothèses. Cette analyse est utile car elle modère les conclusions que l'on peut déduire des taux de reconnaissance obtenus sur une tâche donnée, qui sont souvent inversement proportionnels aux caractéristiques robustes d'une méthode lorsque d'autres conditions et environnements que ceux pour lesquels elle a été développée sont utilisés.

Ce compromis qui existe entre performances et robustesse est d'autant plus important qu'il est facile de tirer de fausses conclusions des expérimentations, conclusions qui peuvent considérablement influencer la compréhension que l'on a du domaine dans son ensemble. Ainsi, les taux de reconnaissance très bons affichés dans de nombreuses publications en reconnaissance de la parole peuvent laisser croire à quelqu'un qui n'est pas spécialiste du domaine que la reconnaissance de la parole n'est plus un problème aujourd'hui. Et pourtant, si nous analysons plus en détails les contraintes requises pour obtenir de tels taux de reconnaissance, nous comprenons mieux les limites de tous ces systèmes, et pourquoi la reconnaissance n'est toujours pas prête à être largement exploitée et utilisée en conditions réelles.

Je pense que ce domaine de recherche a aujourd'hui besoin de s'ouvrir à d'autres sources d'information : jusqu'à présent, quasiment seule l'information acoustique a été utilisée. On peut même dire qu'elle a été surexploitée, car d'innombrables algorithmes d'analyse et de modélisation différents ont été proposés pour traiter cette information - HMM, réseaux de neurones, SVM, modèles hybrides, réseaux bayésiens, etc. Il est peu probable d'améliorer encore significativement les performances en continuant à proposer de nouveaux modèles ou de nouveaux algorithmes pour exploiter l'acoustique seule. D'autres sources d'information doivent aujourd'hui être prises en compte, comme le contexte, la syntaxe et la sémantique. Parallèlement à ces efforts, il reste bien entendu d'autres voies de recherche prometteuses, parmi lesquelles, à mon avis, les techniques d'apprentissage non supervisé ou mieux, légèrement supervisé, auront une place de plus en plus importante, car elles permettront aux modèles d'apprendre en permanence - de nouveaux termes, une nouvelle langue, une nouvelle application -, un peu à la manière de l'Homme.

Annexe A

Description des principaux corpus et conditions expérimentales

Au cours des ces dernières années, j'ai utilisé différents corpus dans de nombreuses expérimentations, en fonction du type de bruit que je voulais traiter, du type de tâche, voire de contraintes liées aux projets auxquels je participais à ce moment.

Plutôt que de décrire les conditions expérimentales avant chaque expérience présentée dans le mémoire, ce qui dupliquerait inutilement ces descriptions, j'ai décidé de rassembler dans cette annexe les principales conditions expérimentales utilisées dans ce mémoire. Ainsi, en cas de doute lors de l'interprétation d'une expérience, le lecteur pourra se reporter à cette annexe qui centralise les descriptions des conditions expérimentales.

A.1 Principaux corpus

Les corpus ci-dessous sont décrits par ordre chronologique approximatif.

A.1.1 BREF80

BREF80 [Lamel *et al.*, 1991] est la version française du corpus "Wall Street Journal". Il contient des phrases du journal "Le Monde" lues par des français natifs. Il s'agissait du corpus de référence, au XX^{è} siècle, pour réaliser des tâches de reconnaissance en grand vocabulaire en français.

J'utilisais alors un lexique de 20 000 formes fléchies des mots les plus courant du français. Depuis le XXI^{è} siècle, j'utilise plutôt le corpus ESTER.

A.1.2 OGI

Le corpus OGI est un corpus multilingue de parole enregistrée par téléphone [Muthusamy *et al.*, 1992]. Il contient différents types de parole dans chaque langue : séquences de chiffres, phrases courtes, parole spontanée. La partie que j'ai considérée est celle de parole spontanée, qui est constituée de l'enregistrement des 10 premières secondes de la réponse du locuteur à une requête, comme "Décrivez la pièce dans laquelle vous êtes". Ces parties de parole spontanée ne sont pas transcrites manuellement, mais j'ai utilisé la requête correspondante pour affecter un thème à chacune de ces phrases. Il y a ainsi quatre thèmes de disponibles dans le corpus. Pour le français, chaque thème contient 110 réponses, chacune enregistrée par un locuteur différent.

A.1.3 VODIS

VODIS est un corpus conçu en partie par l'équipe PAROLE en partenariat avec des industriels automobiles [Gassert et Mari, 1998]. Il s'agit de parole enregistrée dans différentes voitures à des vitesses

variables. 140 locuteurs, dont 75 femmes et 65 hommes ont ainsi prononcé des séquences de chiffres et de nombres de taille variable en français. L'enregistrement a été réalisé avec un microphone casque, proche de la bouche et donc avec un rapport signal-sur-bruit très élevé, ainsi qu'avec un microphone positionné au niveau du pare-soleil.

A.1.4 Aurora2

Aurora2 est un corpus standard dédié à la comparaison des méthodes de paramétrisation robuste [Hirsch et Pearce, 2000]. Il comprend des séquences de 11 chiffres en anglais (zero, oh, one, two, three, four, five, six, seven, eight, nine).

Le corpus d'apprentissage est divisé en deux parties, de 8440 phrases chacune : une partie non bruitée, et une partie multi-conditions qui inclut de la parole non bruitée et de la parole bruitée avec 4 bruits différents[24] à 4 rapports signal-sur-bruit chacun (5 dB, 10 dB, 15 dB et 20 dB), ce qui fait au total 17 sous-parties de taille égale. Chaque phrase du corpus multi-conditions correspond à une phrase du corpus non bruité, c'est-à-dire que ces deux phrases sont issues du même signal de parole.

Le corpus de test comprend trois parties : A, B et C. Le test A est bruité avec les 4 bruits utilisés pour l'apprentissage. Le test B est bruité avec 4 autres bruits[25]. Le test C est bruité avec 2 bruits du test B et est de plus filtré pour créer un bruit convolutif. Dans les trois parties, le bruit est ajouté à 7 niveaux de SNR, compris entre -5 dB et 20 dB. Chaque condition de test, bruitée ou non, comprend 1001 phrases.

Des scripts d'apprentissage et de reconnaissance utilisant HTK sont fournis avec le corpus.

A.1.5 Aurora4

Aurora4 est une version artificiellement bruitée du corpus anglais "Wall Street Journal", qui contient des phrases du journal de même nom lues par des américains natifs. Comme Aurora2, il contient un corpus d'apprentissage non bruité et un corpus d'apprentissage multi-conditions, chacun d'eux contenant 7138 phrases. Le corpus d'apprentissage multi-conditions contient 893 phrases non bruitées enregistrées, comme le corpus non bruité, avec un microphone Sennheiser, 2676 phrases corrompues artificiellement et aléatoirement par un des 6 bruits parmi un bruit de rue, de gare de train, de voitures, de cocktail-party, de restaurant et d'aéroports, à un SNR également choisi aléatoirement entre 10 et 20 dB, enregistrées également avec le microphone Sennheiser, 893 phrases non bruitées mais enregistrées avec un autre microphone variable, et 2676 autres phrases bruitées comme précédemment et enregistrées avec un autre microphone que le Sennheiser.

Le corpus de test contient 14 conditions de test différentes, chacune composée de 330 phrases pour la version complète, ou de 166 phrases pour la version réduite, qui est celle que j'ai utilisée dans mes recherches. Les tests 1 à 7 sont enregistrés avec le microphone Sennheiser, et les tests 8 à 14 avec un autre microphone. Les tests 1 et 8 n'ont pas de bruit ajouté, tandis que les tests 2 à 7, et 9 à 14 sont artificiellement corrompus par successivement l'un des six bruits cités précédemment, à un SNR aléatoire compris entre 5 et 15 dB.

Un dictionnaire standard de 5000 mots est fourni avec le corpus.

A.1.6 ESTER

ESTER [Gravier *et al.*, 2004] est la version française du corpus américain "Broadcast News". Il a été développé dans le cadre d'une campagne d'évaluation technolangue des systèmes de transcription automatique de journaux radiophoniques des plus grandes radios francophones (France-Info, France-Info, France-Culture, France-Musique, RFI, Radio-Télé-Marocaine, etc.). ESTER comprend un corpus d'apprentissage, un corpus de développement et un corpus de test officiel utilisé à la fin de la campagne d'évaluation. Chacun de ces sous-corpus est composé de fichiers audio monophoniques à 16kHz, chaque fichier correspondant à toute la durée d'un journal, soit généralement une heure, une demi-heure ou vingt minutes. Chaque fichier est également accompagné de sa transcription manuelle.

[24] métro, cocktail-party, voiture, hall
[25] Restaurant, rue, aéroport, gare

Plusieurs phénomènes doivent être pris en compte pour travailler avec ces fichiers : tout d'abord, ces fichiers ne contiennent pas seulement de la parole, mais également des plages musicales, des jingles, des annonces publicitaires, etc. Une variété de locuteurs différents interviennent successivement : le journalistes enregistrés en studio, mais également des hommes politiques dans des meetings, des gens dans la rue. Les conditions d'enregistrement sont donc variables : qualité studio, parole téléphonique, parole plus musique, bruit de fond important, parole spontanée ou au contraire bien structurée, etc.

A.2 Systèmes de reconnaissance

A.2.1 Petit vocabulaire

Toutes les expériences concernant les tâches avec un petit vocabulaire, c'est-à-dire moins de 1000 mots, ont été réalisées avec la boîte à outils HTK [Woodland et Young, 1993]. Cette boîte à outils est devenue un standard en reconnaissance automatique de la parole. Elle propose des outils permettant d'apprendre des modèles acoustiques et des modèles de langage, de construire des arbres de régression pour le regroupement des modèles triphones, d'effectuer une reconnaissance de la parole avec l'algorithme de Viterbi et de calculer des taux de reconnaissance.

Toutefois, dans sa version distribuée librement, cette boîte à outils ne permet pas de faire de la reconnaissance en grand vocabulaire. Mais des décodeurs grand vocabulaire compatibles avec les modèles entraînés avec HTK sont disponibles par ailleurs, et en particulier le décodeur *julius* décrit au paragraphe suivant.

Pour mes expériences, j'ai souvent modifié HTK, par exemple pour la faire fonctionner avec des modèles multi-bandes ou avec les approches de marginalisation de la vraisemblance des observations. J'ai également développé une autre boîte à outil compatible avec HTK en java (JHTK).

A.2.2 Grand vocabulaire

Nous avons développé dans l'équipe une plate-forme de transcription de radio appelée ANTS [52]. Cette plate-forme réalise une succession de traitements sur les fichiers en entrée :

1. Localisation des segments dominés par la parole, la musique et le bruit. Seuls les segments de parole sont passés aux étages suivants.
2. Séparation des segments enregisrés en studio et au téléphone : ces derniers sont codés avec une bande passante limitée à 8kHz.
3. Séparation des segments de parole en homme / femme : un modèle pour chacune de ces quatre conditions (homme/femme, téléphone/studio) est appris.
4. Segmentation par localisation des pauses dans le flux de parole.
5. Sur-segmentation des segments trop longs en segments de 15 secondes recouvrants.
6. Transcription automatique de chacun des segments obtenus avec le décodeur julius (voir ci-dessous).
7. Regroupement des réponses obtenues en recherchant des mots communs dans le cas des segments recouvrants.
8. Normalisation de certaines formes orthographiques (nombres, dates, etc.)
9. Calcul du taux de reconnaissance avec l'outil SCLITE distribué par l'organisme NIST[26].

Julius [Lee *et al.*, 2001] est un décodeur grand vocabulaire compatible avec les modèles acoustiques appris avec HTK et les modèles n-grams au format standard ARPA. Il est implémenté en C et possède plusieurs optimisations qui font de ce décodeur l'un des plus rapides distribués librement aujourd'hui.

Julius transcrit la parole en deux passes. La première passe, de gauche à droite, construit un treillis de reconnaissance avec l'algorithme de passage du jeton, qui est une implémentation de l'algorithme de Viterbi. Elle utilise des modèles triphones approchés et des bigrammes. La seconde passe, de droite à gauche, parcourt le treillis issus de la première passe et en extrait la ou les meilleures solutions par l'intermédiaire d'un algorithme A^* utilisant des modèles triphones exacts et des trigrammes inversés.

[26] `http://www.nist.gov`

Annexe B

Liste complète des publications

Thèses et brevets

[1] C. Cerisara : *Contribution de l'approche Multi-Bandes à la reconnaissance automatique de la parole.* Thèse de doctorat, Institut National Polytechnique de Lorraine, LORIA, Nancy, France, sept. 1999.

[2] C. Cerisara, L. Rigazio, R. Boman et J.-C. Junqua : *Method for Noise Adaptation in Automatic Speech Recognition Using Transformed Matrices.* Brevet. N° de série : 09/551,001. Date de dépôt : 18 avril 2000.

[3] C. Cerisara, L. Rigazio, R. Boman et J.-C. Junqua : *Method for Additive and Convolutional Noise Adaptation in Automatic Speech Recognition Using Transformed Matrices.* Brevet. N° de série : 09/628,376. Date de dépôt : 31 Juillet 2000.

Livres et chapitres de livres

[4] J.-P. Haton, C. Cerisara, D. Fohr, Y. Laprie et K. Smaïli : *Reconnaissance automatique de la parole : du signal à son interprétation.* Dunod, mai 2006.

[5] C. Cerisara et Y. Haradji : *Informatique diffuse*, chap. Nouvelles formes d'interaction homme-machine pour l'informatique diffuse. OFTA, mai 2007.

[6] G. Georgantas, V. Issarny et C. Cerisara : *Ambient Intelligence, Wireless Networking, and Ubiquitous Computing*, chap. Dynamic Synthesis of Natural Human-Machine Interfaces in Ambient Intelligence Environments. Artech House, juil. 2006.

[7] C. Cerisara : *Computational Models of Speech Pattern Processing*, chap. Dealing With Loss of Synchronism in Multi-Band Continuous Speech Recognition Systems. NATO ASI Series F, 1999.

Revues internationales avec comité de lecture

[8] P. Kral et C. Cerisara : Dialogue act recognition approaches. *Computing and Informatics*, 2009. à paraître.

[9] C. Cerisara : Automatic topic discovery and recognition from the speech stream. *Computer Speech and Language*, 23(2):220–239, avr. 2009.

[10] S. Demange, C. Cerisara et J.-P. Haton : Missing data mask estimation with frequency and temporal dependencies. *Computer Speech and Language*, 23(1):25–41, 2009.

[11] C. Cerisara, S. Demange et J.-P. Haton : On noise masking for automatic missing data speech recognition : a survey and discussion. *Computer Speech and Language*, 21(3):443–457, juil. 2007.

[12] P. Král, C. Cerisara et J. Klečková : Lexical structure for dialogue act recognition. *Journal of Multimedia*, 2(3):1–8, juin 2007.

[13] C. Cerisara, L. Rigazio et J.-C. Junqua : α-jacobian environmental adaptation. *Speech Communication*, 42(1):25–41, jan. 2004. Special Issue on Adaptation Methods for Automatic Speech Recognition.

[14] C. Cerisara et D. Fohr : Multi-Band automatic speech recognition. *Computer Speech and Language*, 15(2):151–174, avr. 2001.

Conférences internationales avec comité de lecture

[15] C. Cerisara, O. Mella et D. Fohr : Jtrans, an open-source software for semi-automatic text-to-speech alignment. *In Proc. Interspeech*, Brighton, UK, sept. 2009.

[16] P. Král, T. Pavelka et C. Cerisara : Evaluation of dialogue act recognition approaches. *In Proc. IEEE Workshop on Machine Learning for Signal Processing*, Cancun, Mexico, oct. 2008.

[17] C. Cerisara : Exploiting confidence measures for missing data speech recognition. *In Proc. Acoustics*, Paris, juil. 2008.

[18] P. Král, C. Cerisara et J. Klečková : Importance of prosody for dialogue acts recognition. *In Proc. SPECOM*, Moscow, Russia, oct. 2007.

[19] S. Demange, C. Cerisara et J.-P. Haton : Accurate marginalization range for missing data recognition. *In Proc. Interspeech*, août 2007.

[20] P. Král, C. Cerisara et J. Klečková : Confidence measures for semi-automatic labelling of dialog acts. *In Proc. ICASSP*, p. 153–156, Honolulu, Hawaii, USA, avr. 2007.

[21] S. Demange, C. Cerisara et J.-P. Haton : Missing data mask models with global frequency and temporal constraints. *In Proc. Interspeech*, Pittsburgh, Pennsylvania USA, sept. 2006.

[22] P. Král, J. Klečková et C. Cerisara : Automatic dialog acts recognition based on words clusters. *In 9th Western Pacific Acoustics Conference - WESPAC IX 2006*, p. 6 p. The Acoustical Society of Korea, juin 2006.

[23] C. Cerisara et K. Daoudi : Evaluation of the SPACE denoising algorithm on aurora2. *In Proc. ICASSP*, Toulouse, mai 2006.

[24] S. Demange, C. Cerisara et J.-P. Haton : Mask estimation for missing data recognition using background noise sniffing. *In Proc. ICASSP*, Toulouse, mai 2006.

[25] P. Král, C. Cerisara et J. Klečková : Automatic dialog acts recognition based on sentence structure. *In* IEEE, éd. : *Proc. ICASSP*, p. 61–64. IEEE, mai 2006.

[26] P. Král, C. Cerisara, J. Klečková et T. Pavelka : Sentence structure for dialog act recognition in czech. *In 2nd IEEE International Conference on Information et Communication Technologies : from Theory to Applications - ICTTA'06*. Syrian Computer Society, SCS, avr. 2006.

[27] K. Daoudi et C. Cerisara : An improved version of the SPACE algorithm for noise robust speech recognition. *In Proc. IEEE-EURASIP ISCCSP*, Marrakech, Morroco, mars 2006.

[28] P. Král, C. Cerisara et J. Klečková : Combination of classifiers for automatic recognition of dialog acts. *In Proc. Interspeech*, p. 825–828, 2005.

[29] P. Král, J. Klečková et C. Cerisara : Sentence modality recognition in French based on prosody. *In VI. International Conference on Enformatika, Systems Sciences and Engineering - ESSE 2005*, p. 185–188, 2005.

[30] K. Daoudi et C. Cerisara : The MAP-SPACE denoising algorithm for noise robust speech recognition. *In Proc. IEEE ASRU Workshop*, Cancuun, Mexique, 2005.

[31] P. Král, J. Klečková et C. Cerisara : Analysis of importance of the prosodic features for automatic sentence modality recognition in French in real conditions. *In Proc. ICECS*, vol. 3, p. 1820–1824, Crete, Greece, nov. 2004.

[32] I. Illina, D. Fohr, O. Mella et C. Cerisara : The automatic news transcription system : Ants, some real time experiments. *In Proc. ICSLP*, Jeju island, Korea, oct. 2004.

[33] D. Fohr, O. Mella, I. Illina et C. Cerisara : Experiments on the accuracy of phone models and liaison processing in a French broadcast news transcription system. *In Proc. ICSLP*, Jeju island, Korea, oct. 2004.

[34] C. Cerisara, D. Fohr, O. Mella et I. Illina : Exploiting models intrinsic robustness for noisy speech recognition. *In Proc. ICSLP*, Jeju island, Korea, oct. 2004.

[35] C. Cerisara et I. Illina : Robust speech recognition to non-stationary noise based on model-driven approaches. *In EUROSPEECH'2003*, Geneva, Switzerland, sept. 2003.

[36] C. Cerisara : Towards missing data recognition with cepstral features. *In EUROSPEECH'2003*, Geneva, Switzerland, sept. 2003.

[37] C. Cerisara et D. Fohr : Fast channel and noise compensation in the spectral domain. *In EUSIPCO 2002*, Toulouse, France, sept. 2002.

[38] C. Cerisara, J.-C. Junqua et L. Rigazio : Dynamic estimation of a noise over estimation factor for Jacobian-based adaptation. *In ICASSP 2002*, Orlando, Floride, May 2002.

[39] C. Cerisara, L. Rigazio, R. Boman et J.-C. Junqua : Environmental adaptation based on first order approximation. *In ICASSP 2001, Salt lake City, USA*, mai 2001.

[40] C. Cerisara, D. Fohr, I. Illina, F. Lauri et O. Mella : Comparison of different methods for noise adaptation in a HMM-based speech recognition system . *In International Congress on Acoustics, Italy, Rome*, sept. 2001.

[41] C. Cerisara et K. Daoudi : Modeling dependency between regression classes in MLLR using multiscale autoregressive models. *In Workshop on adaptation methods for speech recognition*, Sophia-Antipolis, France, août 2001.

[42] J.-C. Junqua, C. Cerisara, L. Rigazio et D. Kryze : Environment-adaptive algorithms for robust speech recognition. *In Workshop on Hands-Free Speech Communication*, Kyoto, Japan, avr. 2001.

[43] C. Cerisara, L. Rigazio, R. Boman et J.-C. Junqua : Transformation of Jacobian matrices for noisy speech recognition. *In Proc. ICSLP*, vol. 1, p. 369–372, Beijing, China, oct. 2000.

[44] C. Cerisara, D. Fohr et J.-P. Haton : Asynchrony in Multi-Band Speech Recognition. *In IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP'2000, Istanbul*, juin 2000.

[45] R. Boman, C. Cerisara, L. Rigazio et J.-C. Junqua : Jacobian adaptation and likelihood computation for speech recognition on inexpensive integer processors. *In ICSPAT*, Dallas, Texas, oct. 2000.

[46] J.-P. Haton, C. Cerisara et D. Fohr : Improvement of Multi-Band Speech Recognition. *In SPECOM'99*, 1999.

[47] C. Cerisara, D. Fohr et J.-P. Haton : Robust behavior of multi-band paradigm. *In Robust Methods for Speech Recognition in Adverse Conditions, Tampere*. Nokia, COST249 and IEEE, mai 1999.

[48] C. Cerisara, J.-P. Haton et D. Fohr : Towards a Global Optimization Scheme for Multi-Band Speech Recognition. *In EUROSPEECH'99, Prague*, sept. 1999.

[49] C. Cerisara, J.-P. Haton, J.-F. Mari et D. Fohr : A Recombination Model for Multi-Band Speech Recognition. *In ICASSP'98, Seattle, USA*, mai 1998.

[50] C. Cerisara, J.-P. Haton, J.-F. Mari et D. Fohr : Multi-Band Continuous Speech Recognition. *In EUROSPEECH*, 1997.

Conférences nationales avec comité de lecture

[51] C. Cerisara et C. Gardent : Analyse syntaxique du français parlé. *In Workshop ATALA*, Paris, oct. 2009.

[52] A. Brun, C. Cerisara, D. Fohr, I. Illina, D. Langlois, O. Mella et K. Smaïli : Ants : le système de transcription automatique du loria. *In Journées d'Études sur la Parole*, Fès, Maroc, avr. 2004.

[53] D. Fohr, O. Mella, I. Illina, F. Lauri, C. Cerisara et C. Antoine : Reconnaissance de la parole pour des locuteurs non natifs en présence de bruit. *In Journées d'Études sur la Parole*, Nancy, France, 2002.

[54] C. Cerisara, M. Afify et J.-P. Haton : étude de la recombinaison de plusieurs classifieurs appliquée à deux tâches de reconnaissance de la parole. *In Journées d'Études sur la Parole*, Martigny, Suisse, juin 1998.

[55] C. Cerisara, Y. Gong et J.-P. Haton : Reconnaissance de la parole continue par le modèle STM polynomial. *In Journées d'Études sur la Parole*, Avignon, France, 1996.

Divers

[56] D. Sacchetti, R. Chibout, V. Issarny, C. Cerisara et F. Landragin : Seamless access to mobile services for the mobile user. Demonstration, ASE'04 Conference, sept. 2004.

[57] Y. Laprie et C. Cerisara : Vers le succès en reconnaissance vocale. Inédit, la lettre d'information de l'INRIA, mars 2004. Numéro spécial "Le hasard et l'observation".

[58] D. Sacchetti, A. Talamona, V. Issarny, S. Ben Attalah, C. Cerisara, R. Chibout et W. van Raemdonck : Ozone away environment. Film de démonstration, diffusion INRIA, 2004.

Bibliographie référencée dans le mémoire

[Abad *et al.*, 2003] Abad, A., Nadeu, C., Hernando, J. et Padrell, J. (2003). Jacobian adaptation based on frequency-filtered spectral energies. *In Proc. EUROSPEECH*, pages 1621–1625.

[Acero, 1990] Acero, A. (1990). *Acoustical and environmental robustness in automatic speech recognition.* Thèse de doctorat, Carnegie Mellon University.

[Acero et Stern, 1991] Acero, A. et Stern, R. M. (1991). Robust speech recognition by normalization of the acoustic space. *In Proc. ICASSP*, pages 893–896, Toronto.

[Achlioptas, 2001] Achlioptas, D. (2001). Database-friendly random projections. *In PODS '01 : Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, New York, NY, USA. ACM.

[Adami *et al.*, 2002] Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N. et Sivadas, S. (2002). QUALCOMM-ICSI-OGI features for ASR. *In Proc. ICSLP.*

[Afify et Siohan, 2007] Afify, M. et Siohan, O. (2007). Comments on vocal tract length normalization equals linear transformation in cepstral space. *IEEE Trans. on Speech and Audio Processing*, 15(5): 1731–1732.

[Akbacak et Hansen, 2007] Akbacak, M. et Hansen, J. H. L. (2007). Environmental sniffing : Noise knowledge estimation for robust speech systems. *IEEE Trans. ASSP*, 15(2):465–477.

[Alexandersson *et al.*, 1995] Alexandersson, J., Buschbeck-wolfz, B., Fujinamiz, T., Maier, E., Reithingery, N., Schmitz, B. et Siegel, M. (1995). Dialogue acts in VERBMOBIL-2. Rapport technique. `http://verbmobil.dfki.de/dialog/publications/Alexanderssonetal97a.ps.gz`.

[Allen et Core, 1997] Allen, J. et Core, M. (1997). Draft of Damsl : Dialog Act Markup in Several Layers. *In `http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html`.*

[Allen, 1994] Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Trans. ASSP*, 2:567–576.

[Austin, 1962] Austin, J. L. (1962). How to do Things with Words. *Clarendon Press, Oxford.*

[Bach et Jordan, 2005] Bach, F. et Jordan, M. (2005). Blind one-microphone speech separation : A spectral learning approach. *In Advances in Neural Information Processing Systems (NIPS)*, volume 17, pages 65–72.

[Barker *et al.*, 2005] Barker, J., Cooke, M. et Ellis, D. P. W. (2005). Decoding speech in the presence of other sources. *Speech Communication*, 45(1):5–25.

[Barker *et al.*, 2001] Barker, J., Green, P. et Cooke, M. (2001). Linking auditory scene analysis and robust ASR by missing data techniques. *In Proc. WISP*, pages 295–307, Stratford-upon-Avon, England.

[Bauerecker *et al.*, 2003] Bauerecker, H., Nadeu, C. et Padrell, J. (2003). On the advantages of frequency-filtered features for speech recognition with variable sampling frequencies. experiments with SPEECHDAT car databases. *In Proc. EUROSPEECH*, pages 869–873.

[Besacier, 1998] Besacier, L. (1998). *Un modèle parallèle pour la reconnaissance automatique du locuteur.* Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

[Besacier et Bonastre, 2002] Besacier, L. et Bonastre, J.-F. (2002). Frame pruning for speaker recognition. *In Proc. ICASSP.*

[Bod, 2006] Bod, R. (2006). Unsupervised parsing with U-DOP. *In Proc. CoNLL*, New-York.

[Boll, 1979] Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. ASSP*, 27:113–120.

[Bourlard et Dupont, 1996] Bourlard, H. et Dupont, S. (1996). A new ASR approach based on independent processing an recombination of partial frequency bands. *In Proc. ICSLP*, Philadelphie, USA.

[Bregman, 1990] Bregman, A. S. (1990). *Auditory Scene Analysis.* MIT Press.

[Bronkhorst, 2000] Bronkhorst, A. (2000). The cocktail party phenomenon : a review of research on speech intelligibility in multiple-talker condition. *Acustica*, 86:117–128.

[Brown *et al.*, 2001] Brown, G. J., Barker, J. et Wang, D. L. (2001). A neural oscillator sound separator for missing data speech recognition. *In Proc. IJCNN-01*, volume 4, pages 2907–2912, Washington, DC, USA.

[Brown et Cooke, 1994] Brown, G. J. et Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language*, 8(4):297–336.

[Brown et Wang, 2005] Brown, G. J. et Wang, D. L. (2005). *Speech Enhancement*, chapitre Separation of speech by Computational Auditory Scene Analysis, pages 371–402. Springer, New York, J. Benesty, S. Makino and J. Chen édition.

[Brun, 2003] Brun, A. (2003). *Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole*. Thèse de doctorat, Université Henri Poincaré - Nancy 1, Nancy, France.

[Buera *et al.*, 2007] Buera, L., Lleida, E., Miguel, A., Ortega, A. et Saz, O. (2007). Cesptral vector normalization based on stereo data for robust speech recognition. *IEEE Trans. ASSP*, 15(3):1098–1113.

[Byrne et Gunawardana, 2001] Byrne, W. et Gunawardana, A. (2001). Discounted likelihood linear regression for rapid speaker adaptation. *Computer Speech and Language*, 15(1):15–38.

[Carletta *et al.*, 1997] Carletta, J., Isard, A., Isard, S., Kowtko, J., Newlands, A., Doherty-Sneddon, G. et Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31.

[Cerisara, 1999a] Cerisara, C. (1999a). *Contribution de l'approche Multi-Bandes à la reconnaissance automatique de la parole*. Thèse de doctorat, Institut National Polytechnique de Lorraine, LORIA, Nancy, France.

[Cerisara, 1999b] Cerisara, C. (1999b). Dealing with loss of synchronism in Multi-Band continuous speech recognition systems. *In* Ponting, K. M., éditeur : *Computational Models of Speech Pattern Processing*. NATO ASI Series F.

[Cerisara, 2003] Cerisara, C. (2003). Towards missing data recognition with cepstral features. *In EUROSPEECH'2003*.

[Cerisara, 2008] Cerisara, C. (2008). Exploiting confidence measures for missing data speech recognition. *In Proc. Acoustics*, Paris.

[Cerisara, 2009] Cerisara, C. (2009). Automatic topic discovery and recognition from the speech stream. *Computer Speech and Language*, 23(2):220–239.

[Cerisara et Daoudi, 2001] Cerisara, C. et Daoudi, K. (2001). Modeling dependency between regression classes in MLLR using multiscale autoregressive models. *In Workshop on adaptation methods for speech recognition*, Sophia-Antipolis, France.

[Cerisara et Daoudi, 2006] Cerisara, C. et Daoudi, K. (2006). Evaluation of the SPACE denoising algorithm on Aurora2. *In Proc. ICASSP*, Toulouse.

[Cerisara *et al.*, 2007] Cerisara, C., Demange, S. et Haton, J.-P. (2007). On noise masking for automatic missing data speech recognition : a survey and discussion. *Computer Speech and Language*, 21(3):443–457.

[Cerisara et Fohr, 2001] Cerisara, C. et Fohr, D. (2001). Multi-Band automatic speech recognition. *Computer Speech and Language*, 15(2):151–174.

[Cerisara *et al.*, 2001] Cerisara, C., Fohr, D., Illina, I., Lauri, F. et Mella, O. (2001). Comparison of different methods for noise adaptation in a HMM-based speech recognition system. *In Proc. International Congress on Acoustics, Italy, Rome.*

[Cerisara et Haradji, 2007] Cerisara, C. et Haradji, Y. (2007). *Informatique diffuse*, chapitre Nouvelles formes d'interaction homme-machine pour l'informatique diffuse. OFTA.

[Cerisara *et al.*, 1999] Cerisara, C., Haton, J.-P. et Fohr, D. (1999). Towards a global optimization scheme for Multi-Band speech recognition. *In Proc. EUROSPEECH*, Prague.

[Cerisara *et al.*, 1998] Cerisara, C., Haton, J.-P., Mari, J.-F. et Fohr, D. (1998). A recombination model for Multi-Band speech recognition. *In Proc. ICASSP*, Seattle, USA.

[Cerisara et Illina, 2003] Cerisara, C. et Illina, I. (2003). Robust speech recognition to non-stationary noise based on model-driven approaches. *In EUROSPEECH'2003*.

[Cerisara *et al.*, 2004] Cerisara, C., Rigazio, L. et Junqua, J.-C. (2004). α-Jacobian environmental adaptation. *Speech Communication*, 42(1):25–41.

[Chen *et al.*, 2006] Chen, J., Benesty, J., Yiteng, H. et Doclo, S. (2006). New insights into the noise reduction Wiener filter. *IEEE Trans. ASSP*, 14(4):1218–1234.

[Chen *et al.*, 2000] Chen, K. T., Liau, W. W., Wang, H. M. et Lee, L. S. (2000). Fast speaker adaptation using eigenspace-based maximum-likelihood linear regression. *In Proc. ICSLP*, volume 3, pages 742–745.

[Chesta *et al.*, 1999] Chesta, C., Siohan, O. et Lee, C. H. (1999). Maximum a posteriori linear regression for hidden Markov model adaptation. *In Proc. EUROSPEECH*, pages 211–214, Budapest.

[Chou, 1999] Chou, W. (1999). Maximum a posteriori linear regression with elliptically symmetric matrix priors. *In Proc. EUROSPEECH*, pages 1–4, Budapest.

[Chou et He, 2003] Chou, W. et He, X. (2003). Maximum a posteriori linear regression variance adaptation for continuous density HMM. *In Proc. EUROSPEECH*, pages 1513–1516, Geneva.

[Christensen *et al.*, 2000] Christensen, H., Lindberg, B. et Andersen, O. (2000). Employing heterogeneous information in a multi-stream framework. *In Proc. ICASSP*, volume 3, pages 1571–1574, Istanbul, Turquie.

[Christensen *et al.*, 2007] Christensen, H., Ma, N., Wrigley, S. et Barker, J. (2007). Integrating pitch and localisation cues at a speech fragment level. *In INTERSPEECH*, Antwerp.

[Clark, 2001] Clark, A. (2001). *Unsupervised language acquisition : theory and practice*. Thèse de doctorat, Univ. of Sessex.

[Comsky, 1991] Comsky, N. (1991). *The Chomskyan turn*, chapitre Linguistics and cognitive science : Problems and mysteries. Wiley-Blackwell.

[Cooke, 1993] Cooke, M. (1993). *Modelling auditory processing and organization*. Cambridge University Press.

[Cooke, 2005] Cooke, M. (2005). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Amer.*, 119(3):1562–1573.

[Cooke *et al.*, 2001] Cooke, M., P., G., Josifovski, L. et Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34.

[Cox et Dasmahapatra, 2002] Cox, S. et Dasmahapatra, S. (2002). High-level approaches to confidence estimation in speech recognition. *IEEE Trans. on Speech and Audio Processing*, 10(7):460–471.

[Coy et Barker, 2007] Coy, A. et Barker, J. (2007). An automatic speech recognition system based on the scene analysis account of auditory perception. *Speech Communication*, 49:384–401.

[Daoudi et Cerisara, 2005] Daoudi, K. et Cerisara, C. (2005). The MAP-SPACE denoising algorithm for noise robust speech recognition. *In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Cancuun, Mexique.

[Daoudi et Cerisara, 2006] Daoudi, K. et Cerisara, C. (2006). An improved version of the SPACE algorithm for noise robust speech recognition. *In Proc. IEEE-EURASIP ISCCSP*, Marrakech, Morroco.

[Daoudi *et al.*, 2003] Daoudi, K., Fohr, D. et Antoine, C. (2003). Dynamic Bayesian networks for multi-band automatic speech recognition. *Computer Speech and Language*, 17:263–285.

[Demange, 2007] Demange, S. (2007). *Contributions à la reconnaissance automatique de la parole avec données manquantes*. Thèse de doctorat, Université Henri Poincaré.

[Deng *et al.*, 2003] Deng, L., Droppo, J. et Acero, A. (2003). Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Trans. on Speech and Audio Processing*, 11(6):568–580.

[Deng *et al.*, 2004] Deng, L., Droppo, J. et Acero, A. (2004). Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features. *IEEE Trans. on Speech and Audio Processing*, 12(3):218–233.

[Deng *et al.*, 2005] Deng, L., Wu, J., Droppo, J. et Acero, A. (2005). Analysis and comparison of two speech feature extraction/compensation algorithms. *IEEE Signal Processing Letters*, 12(5):477–480.

[Deviren, 2004] Deviren, M. (2004). *Systèmes de reconnaissance de la parole revisités : réseaux bayésiens dynamiques et nouveaux paradigmes.* Thèse de doctorat, Université Henri Poincaré, Nancy I.

[Dharanipragada *et al.*, 2007] Dharanipragada, S., Yapanel, U. H. et Rao, B. D. (2007). Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method. *IEEE Trans. ASSP*, 15(1):224–234.

[Dhillon *et al.*, 2004] Dhillon, R., S., B., Carvey, H. et E., S. (2004). Meeting Recorder Project : Dialog Act Labeling Guide. Rapport technique TR-04-002, International Computer Science Institute.

[Digalakis *et al.*, 1995] Digalakis, V., Ritchev, D. et Neumeyer, L. (1995). Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Trans. ASSP*, 3:357–366.

[Droppo *et al.*, 2002] Droppo, J., Deng, L. et Acero, A. (2002). Evaluation of SPLICE on the Aurora2 and 3 tasks. *In Proc. ICSLP*, volume 1, pages 29–32, Denver, Colorado.

[Dupont et Ris, 2001] Dupont, S. et Ris, C. (2001). Assessing local noise level estimation methods : application to noise robust ASR. *Speech Communication*, 34(1-2):141–158.

[Ellis, 1996] Ellis, D. P. W. (1996). *Prediction-driven computational auditory scene analysis.* Thèse de doctorat, MIT.

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

[Gales, 2000] Gales, M. (2000). Cluster adaptive training of hidden Markov models. *IEEE Trans. on Speech and Audio Processing*, 8:417–428.

[Gales et Woodland, 1996] Gales, M. et Woodland, P. C. (1996). Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264.

[Gales et Young, 1993] Gales, M. et Young, S. J. (1993). HMM recognition in noise using parallel model combination. *In Proc. EUROSPEECH*, pages 837–840, Berlin.

[Gassert et Mari, 1998] Gassert, C. et Mari, J.-F. (1998). Spécification, réalisation et validation d'un corpus oral pour la reconnaissance de la parole dans une voiture. *In Proc. JEP*, pages 171–174.

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

[Ge et Song, 2004] Ge, Y. et Song, J. (2004). Confidence weighting missing feature approach for robust speech recognition. *In Proc. EUSIPCO*, pages 337–340.

[Ghanbaria et Karimi-Mollaei, 2006] Ghanbaria, Y. et Karimi-Mollaei, M. R. (2006). A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets. *Speech Communication*, 48(8):927–940.

[Glotin, 2001a] Glotin, H. (2001a). *Elaboration & études comparatives de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole : incorporation d'indices de voisement et de localisation.* Thèse de doctorat, ICP/Inst. National Polytechnique de Grenoble & IDIAP/EPF Lausanne.

[Glotin et Berthommier, 2000] Glotin, H. et Berthommier, F. (2000). Test of several external posterior weighting functions for multi-band full combination asr. *In Proc. ICSLP*, Beijing.

[Glotin *et al.*, 1998] Glotin, H., Berthommier, F., Tessier, E. et Bourlard, H. (1998). Interfacing of casa and multistream recognition. *In int. conf. Text, Speech & Dialog (TSD), pp207–12, Sojka Ed., Masaryk Univ. (www.fi.muni.cz/tsd98/), Brno-Czech Republic.*

[Godsmark et Brown, 1999] Godsmark, D. et Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366.

[Goel *et al.*, 2002] Goel, V., Visweswariah, K. et Gopinath, R. (2002). Rapid adaptation with linear combinations of rank-one matrices. *In Proc. ICASSP*, pages 581–584, Orlando.

[Gowdy et Tufekci, 2000] Gowdy, J. N. et Tufekci, Z. (2000). Mel-scaled discrete wavelet coefficients for speech recognition. *In Proc. ICASSP.*

[Gravier *et al.*, 2004] Gravier, G., Bonastre, J.-F., Galliano, S., Geoffrois, E., Tait, K. M. et Choukri, K. (2004). ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques. *In Proc. JEP*, Fez.

[Gupta et Gilbert, 2001] Gupta, M. et Gilbert, A. (2001). Robust speech recognition using wavelet coefficients features. *In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 50–61.

[Haeb-Umbach *et al.*, 1995] Haeb-Umbach, R., Beyerlein, P. et Geller, D. (1995). Speech recognition algorithms for voice control interfaces. *Philips Journal of Research*, 49:381–397.

[Hagen et Morris, 2000] Hagen, A. et Morris, A. (2000). Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust ASR. Rapport technique, IDIAP.

[Häkkinen et Haverinen, 2001] Häkkinen, J. et Haverinen, H. (2001). On the use of missing feature theory with cepstral features. *In CRAC Workshop*, Aalborg, Danemark.

[Händel, 2007] Händel, P. (2007). Power spectral density error analysis of spectral subtraction type of speech enhancement methods. *EURASIP Journal on Advances in Signal Processing*, 2007:Article ID 96384, 9 pages. doi :10.1155/2007/96384.

[Handel, 1986] Handel, S. (1986). *Listening : An Introduction to the Perception of Auditory Events*, chapitre Breaking the acoustic wave into events : Stream segregation, pages 185–217. MIT Press, London.

[Harris, 1985] Harris, Z. (1985). *Distributional Structure*. Oxford University Press, New York.

[Haton *et al.*, 2006] Haton, J.-P., Cerisara, C., Fohr, D., Laprie, Y. et Smaïli, K. (2006). *Reconnaissance automatique de la parole : du signal à son interprétation*. Dunod.

[Haykin et Chen, 2005] Haykin, S. et Chen, Z. (2005). The cocktail party problem. *Neural Computation*, 17:1875–1902.

[Hermansky et Morgan, 1994] Hermansky, H. et Morgan, N. (1994). RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2:578–589.

[Hermansky *et al.*, 1991] Hermansky, H., Morgan, N., Bayya, A. et Kohn, P. (1991). Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP). *In Proc. EUROSPEECH*, pages 1367–1370.

[Hilger et Ney, 2006] Hilger, F. et Ney, H. (2006). Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Trans. on Speech and Audio Processing*, 14(3):845–854.

[Hirsch et Pearce, 2000] Hirsch, H. G. et Pearce, D. (2000). The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. *In ISCA ITRW ASR2000*, Paris.

[Hu et Wang, 2004] Hu, G. et Wang, D. L. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. on Neural Networks*, 15(5):1135–1150.

[Ishizuka *et al.*, 2007] Ishizuka, K., Nakatani, T., Fujimoto, M. et Miyazaki, N. (2007). Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio. *In Proc. Interspeech*.

[Ishizuka *et al.*, 2006] Ishizuka, K., Nakatani, T. et Minami, Y. (2006). Speech feature extraction method using subband-based periodicity and nonperiodicity decomposition. *J. Acoust. Soc. Amer.*, 120(1):443–452.

[Jansen et van Hamme, 2005] Jansen, W. et van Hamme, H. (2005). PROSPECT features and their application to missing data techniques for vocal tract length normalization. *In Proc. Interspeech*.

[Jurafsky *et al.*, 1997] Jurafsky, D., Shriberg, E. et Biasca, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation (Coders Manual, Draft 13). Rapport technique 97-01, University of Colorado, Institute of Cognitive Science.

[Kanerva *et al.*, 2000] Kanerva, P., Kristoferson, J. et Holst, A. (2000). Random indexing of text samples for latent semantic analysis. *In* Gleitman, L. et Josh, A., éditeurs : *Proc. Conference of the Cognitive Science Society*, Pennsylvania.

[Kannan, 1997] Kannan, A. (1997). *Adaptation of spectral trajectory models for large vocabulary speech recognition*. Thèse de doctorat, Boston University.

[Kaplan *et al.*, 2008] Kaplan, F., Oudeyer, P.-Y. et Bergen, B. (2008). *Infant and Child Development*, chapitre Computational Models in the Debate over Language Learnability.

[Kim et Ruwisch, 2002] Kim, H.-G. et Ruwisch, D. (2002). Speech enhancement in non-stationary noise environments. *In Proc. ICSLP*, Denver, USA.

[Kim et Chung, 2003] Kim, J. et Chung, J. (2003). Reduction of dimension of HMM parameters using ICA and PCA in MLLR framework for speaker adaptation. *In Proc. EUROSPEECH*, pages 1461–1464, Geneva.

[Kim *et al.*, 2000] Kim, K., Youn, D. H. et Lee, C. (2000). Evaluation of wavelet filters for speech recognition. *In Proc. IEEE International Conference on Systems, Man, and Cybernetics*.

[Kim *et al.*, 2005a] Kim, N. S., Lim, W. et Stern, R. (2005a). Feature compensation based on switching linear dynamic model. *IEEE Trans. ASSP*, 12(6):473–476.

[Kim et Stern, 2006] Kim, W. et Stern, R. M. (2006). Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise. *In Proc. ICASSP*, volume 1, pages 305–308, Toulouse, France.

[Kim *et al.*, 2005b] Kim, W., Stern, R. M. et Ko, H. (2005b). Environment-independent mask estimation for missing-feature reconstruction. *In Proc. Interspeech*, pages 2637–2640, Lisbon, Portugal.

[Klein, 2005] Klein, D. (2005). *The unsupervised learning of natural language structure*. Thèse de doctorat, Stanford Univ.

[Kotnik et Kačič, 2007] Kotnik, B. et Kačič, Z. (2007). A comprehensive noise robust speech parameterization algorithm using wavelet packet decomposition-based denoising and speech feature representation techniques. *EURASIP Journal on Advances in Signal Processing*, 2007.

[Kral, 2007] Kral, P. (2007). *Reconnaissance automatique des actes de dialogue*. Thèse de doctorat, Université Henri Poincaré.

[Král *et al.*, 2005] Král, P., Cerisara, C. et Klečková, J. (2005). Combination of Classifiers for Automatic Recognition of Dialog Acts. *In Interspeech'2005*, pages 825–828, Lisboa, Portugal. ISCA.

[Král *et al.*, 2006] Král, P., Cerisara, C. et Klečková, J. (2006). Automatic Dialog Acts Recognition based on Sentence Structure. *In ICASSP'06*, pages 61–64, Toulouse, France.

[Král *et al.*, 2007a] Král, P., Cerisara, C. et Klečková, J. (2007a). Confidence Measures for Semi-automatic Labeling of Dialog Acts. *In ICASSP'07*, pages 153–156, Honolulu, Hawaii, USA.

[Král *et al.*, 2007b] Král, P., Cerisara, C. et Klečková, J. (2007b). Lexical Structure for Dialogue Act Recognition. *Journal of Multimedia (JMM)*, 2(3):1–8.

[Kristjansson *et al.*, 2004] Kristjansson, T., Attias, H. et Hershey, J. (2004). Single microphone source separation using high resolution signal reconstruction. *In Proc. ICASSP*, pages 817–820.

[Kryze *et al.*, 1999] Kryze, D., Rigazio, L., Appelbaum, T. et Junqua, J.-C. (1999). A new noise robust subband front-end and its comparison to PLP. *In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado.

[Kuhn *et al.*, 2000] Kuhn, R., Junqua, J.-C., Nguyen, P. et Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Trans. on Speech and Audio Processing*, 8(6):695–707.

[Lai, 2003] Lai, Y. P. (2003). Maximum likelihood normalization for robust speech recognition. Mémoire de D.E.A., The Hong Kong University of Science and Technology.

[Lamel *et al.*, 1991] Lamel, L., Gauvain, J.-L. et Eskenazi, M. (1991). BREF, a large vocabulary spoken corpus for French. *In Proc. EUROSPEECH*.

[Lee *et al.*, 2001] Lee, A., Kawahara, T. et Shikano, K. (2001). Julius - an open source real-time large vocabulary recognition engine. *In Proc. EUROSPEECH*, pages 1691–1694.

[Lee *et al.*, 1991] Lee, C. H., Lin, C. H. et Juang, B. H. (1991). A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Trans. on Signal Processing*, 39:806–814.

[Lee *et al.*, 2006] Lee, M., Wang, W. et Yu, H. (2006). Exploring supervised and unsupervised methods to detect topics in biomedical text. *BMC Bioinformatics*, 7(140).

[Leggetter et Woodland, 1995a] Leggetter, C. J. et Woodland, P. C. (1995a). Flexible speaker adaptation using maximum likelihood linear regression. *In Proc. EUROSPEECH*, pages 1155–1158, Madrid.

[Leggetter et Woodland, 1995b] Leggetter, C. J. et Woodland, P. C. (1995b). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–186.

[Levit *et al.*, 2002] Levit, M., Gorin, A. et Nöth, E. (2002). Using EM-trained string-edit distances for approximate matching of acoustic morphemes. *In Proc. ICSLP*, pages 1157–1160.

[Li *et al.*, 2006] Li, P., Xu, B. et Liu, W. (2006). Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. *IEEE Trans. ASSP*, 14(6):2014–2023.

[Li *et al.*, 2007] Li, W., Takeda, K. et Itakura, F. (2007). Robust in-car speech recognition based on nonlinear multiple regressions. *EURASIP Journal on Advances in Signal Processing*, (1).

[LingPipe, 2007] LingPipe (2007). LingPipe 3.2.0. `http://www.alias-i.com/lingpipe/`.

[Lippmann *et al.*, 1987] Lippmann, R. P., Martin, E. et Paul, D. B. (1987). Multi-style training for robust isolated-word speech recognition. *In Proc. ICASSP*, pages 705–708.

[Luettgen *et al.*, 1993] Luettgen, M. R., Karl, W. C., Willsky, A. S. et Teney, R. R. (1993). Multiscale representation of Markov random fields. *IEEE Trans. on Signal Processing*, 41(12):3377–3396.

[M. Cooke et Ellis, 2001] M. Cooke, M. et Ellis, D. P. W. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35:141–177.

[Ma *et al.*, 2006] Ma, N., Bouchard, M. et Goubran, R. A. (2006). Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementation. *IEEE Trans. on Speech and Audio Processing*, 14:19–32.

[Mak et Hsiao, 2007] Mak, B. K.-W. et Hsiao, R. W.-H. (2007). Kernerl eigenspace-based MLLR adaptation. *IEEE Trans. ASSP*, 15(3):784–795.

[Masuda-Katsuse et Kawahara, 1999] Masuda-Katsuse, I. et Kawahara, H. (1999). Dynamic sound stream formation based on continuity of spectral change. *Speech Communication*, 27:253–259.

[McCourt *et al.*, 2000] McCourt, P. M., Vaseghi, S. V. et Doherty, B. (2000). Multiresolution subband features and models for HMM-based phonetic modelling. *Computer Speech and Language*, 14(3).

[Miller et Heise, 1950] Miller, G. A. et Heise, G. A. (1950). The trill threshold. *J. Acoust. Soc. Amer.*, 22:637–638.

[Ming et Smith, 2002] Ming, J. et Smith, J. (2002). Robust speech recognition using probabilistic union models. *IEEE Trans. on Speech and Audio Processing*, 10(6):403–414.

[Molla et Hirose, 2007] Molla, M. K. I. et Hirose, K. (2007). Single-mixture audio source separation by subspace decomposition of Hilbert spectrum. *IEEE Trans. ASSP*, 15(3):893–900.

[Moore, 1982] Moore, B. C. J. (1982). *An Introduction to the Psychology of Hearing.* Academic Press, 24/28 Oval Road, London NW1.

[Moreno *et al.*, 1996] Moreno, P. J., Raj, B. et Stern, R. M. (1996). A vector Taylor series approach for environment independent speech recognition. *In Proc. ICASSP*, pages 733–736, Munich.

[Morgan *et al.*, 2005] Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cetin, O., H., B. et Athineos, M. (2005). Pushing the enveloppe - aside. *IEEE signal processing magazine*, 81.

[Morris, 2001] Morris, A. (2001). Data utility modelling for mismatch reduction. *In Proc. CRAC (workshop on Consistent & Reliable Acoustic Cues for sound analysis)*, Aalborg, Denmark.

[Morris *et al.*, 2001a] Morris, A., Barker, J. et Bourlard, H. (2001a). From missing data to maybe useful data : Soft data modelling for noise robust ASR. *In Proc. WISP*, Stratford-upon-Avon, England.

[Morris *et al.*, 1998] Morris, A., Cooke, M. et P., G. (1998). Some solutions to the missing feature problem in data classification, with applications to noise robust ASR. *In Proc. ICASSP*, pages 737–740, Seattle, USA.

[Muthusamy *et al.*, 1992] Muthusamy, Y. K., Cole, R. A. et Oshika, B. T. (1992). The OGI multi-language telephone speech corpus. *In Proc. ICSLP*, pages 895–898.

[Myers et Rabiner, 1981] Myers, C. S. et Rabiner, L. R. (1981). Connected digit recognition using a level-building DTW algorithm. *IEEE Trans. ASSP*, 29(3):351.

[Nadeu *et al.*, 1995] Nadeu, C., Hernando, J. et Gorricho, M. (1995). On the decorrelation of filter-band energies in speech recognition. *In Proc. EUROSPEECH*, pages 1381–1384, Madrid.

[Nadeu *et al.*, 2001] Nadeu, C., Macho, D. et Hernando, J. (2001). Time and frequency filtering of filter-band energies for robust HMM speech recognition. *Speech Communication*, 34:93–114.

[Nakatani, 2002] Nakatani, T. (2002). *Computational Auditory Scene Analysis based on residue-driven architecture and its application to mixed speech recognition.* Thèse de doctorat, Kyoto University.

[Neumeyer et Weintraub, 1994] Neumeyer, L. et Weintraub, M. (1994). Probabilistic optimum filtering for robust speech recognition. *In Proc. ICASSP*, volume 1, pages 417–420, Adelaide.

[Nguyen *et al.*, 1999] Nguyen, P., Wellekens, C. et Junqua, J.-C. (1999). Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments. *In Proc. EUROSPEECH*, pages 2519–2522.

[NOISE-ROM-0, 1990] NOISE-ROM-0 (1990). NATO : AC243/(Panel 3)/RSG-10. Projet ESPRIT 2589-SAM. Institute for Perception-TNO, The Netherlands. Speech Research Unit, RSRE, United Kingdom.

[Pitel et Grefenstette, 2008] Pitel, G. et Grefenstette, G. (2008). Semi-automatic building method for a multidimensional affect dictionary for a new language. *In Proc. LREC*, Marrakech, Morocco.

[Plapous *et al.*, 2006] Plapous, C., Marro, C. et Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Trans. ASSP*, 14(6):2098–2108.

[Poh et Bengio, 2004] Poh, N. et Bengio, S. (2004). Noise-robust multi-stream fusion for text-independent speaker authentication. *In Proc. ISCA Odyssey, The Speaker and Recognition Workshop*, pages 199–206, Toledo, Espagne. `ftp://ftp.idiap.ch/pub/norman/reports/norman-2004-odyssey.pdf`.

[Potamitis *et al.*, 2000] Potamitis, I., Fakotakis, N. et Kokkinakis, G. (2000). Reliable ASR based on unreliable features. *In Workshop ISCA ITRW ASR2000, Automatic Speech Recognition : Challenges for the new Millennium*, volume 1, pages 53–57.

[Potamitis *et al.*, 2001] Potamitis, I., Fakotakis, N. et Kokkinakis, G. (2001). Bayesian independent component analysis as applied to one-channel speech enhancement. *In Proc. Intl. Conf. Artif. Neural Networks*, volume 2130, pages 593–600.

[Quine, 1960] Quine (1960). *Word and object.* MIT Press, Cambridge.

[Raj, 2000] Raj, B. (2000). *Reconstruction of incomplete spectrograms for robust speech recognition.* Thèse de doctorat, Carnegie Mellon University.

[Razik, 2007] Razik, J. (2007). *Mesures de confiance trame-synchrones et locales en reconnaissance automatique de la parole.* Thèse de doctorat, Université Henri Poincaré.

[Renevey, 2001] Renevey, P. (2001). *Speech recognition in noisy conditions using missing feature approach.* Thèse de doctorat, Ecole Polytechnique Fédérale de Lausanne.

[Reyes-Gomez *et al.*, 2004] Reyes-Gomez, M., Ellis, D. P. W. et Jojic, B. (2004). Multiband audio modeling for single channel acoustic source separation. *In Proc. ICASSP*, volume 5, pages 641–644, Montreal.

[Reynar, 1998] Reynar, J. C. (1998). *Topic segmentation : Algorithms and applications.* Thèse de doctorat, Computer and Information Science, University of Pennsylvania.

[Roweis, 2003] Roweis, S. T. (2003). Factorial models and refiltering for speech separation and denoising. *In Proc. EUROSPEECH*, pages 1009–1012, Geneva, Switzerland.

[Sagayama *et al.*, 1997] Sagayama, S., Yamaguchi, Y., Takahashi, S. et Takahashi, J. (1997). Jacobian approach to fast acoustic model adaptation. *In Proc. ICASSP*, pages 835–838, Munich.

[Sagi *et al.*, 2001] Sagi, S., Nemat-Nasser, S. C., Kerr, R., Hayek, R., Downing, C. et Hecht-Nielsen, R. (2001). A biologically motivated solution to the cocktail party problem. *Neural Computation*, 13:1575–1602.

[Sahlgren, 2005] Sahlgren, M. (2005). An introduction to random indexing. *In Proc. Methods and Applications of Semantic Indexing Workshop at the 7th Conf. on Terminology and Knowledge Engineering*, Copenhagen, Denmark.

[Sarikaya et Hansen, 2000] Sarikaya, R. et Hansen, J. H. L. (2000). High resolution speech feature parameterization for monophone-based stressed speech recognition. *IEEE Signal Processing Letters*, 7(7).

[Seltzer, 2000] Seltzer, M. L. (2000). Automatic detection of corrupt spectrographic features for robust speech recognition. Mémoire de D.E.A., Departement of Electrical and Computer Engineering, Carnegie Mellon University.

[Seltzer *et al.*, 2004] Seltzer, M. L., Raj, B. et Stern, R. M. (2004). A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43:379–393.

[Shinoda et Lee, 2001] Shinoda, K. et Lee, C. H. (2001). A structural Bayes approach to speaker adaptation. *IEEE Trans. on Speech and Audio Processing*, 9:276–287.

[Sim et Wright, 2005] Sim, J. et Wright, C. C. (2005). The Kappa statistic in reliability studies : Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268.

[Siohan *et al.*, 2000] Siohan, O., Myrvoll, T. A. et Lee, C. H. (2000). Structural maximum a posteriori linear regression for fast HMM adaptation. *In Workshop on Automatic Speech Recognition : Challenges for the new Millenium*, pages 120–127, Paris.

[Siu et Chan, 2006] Siu, M. et Chan, A. (2006). A robust Viterbi algorithm against impulsive noise with application to speech recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 14(6):2122–2133.

[Smaragdis, 2007] Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. *IEEE Trans. ASSP*, 15(1):1–12.

[Smith et Gasser, 2005] Smith, L. et Gasser, M. (2005). The development of embodied cognition : six lessons from babies. *Artificial Life*, 11:13–30.

[Solan, 2006] Solan, Z. (2006). *Unsupervised learning of natural languages*. Thèse de doctorat, Tel Aviv Univ.

[Srinivasan *et al.*, 2007] Srinivasan, S., Samuelsson, J. et Kleijn, W. B. (2007). Codebook-based Bayesian speech enhancement for nonstationary environments. *IEEE Trans. ASSP*, 15(2):441–452.

[Srinivasan et Wang, 2007] Srinivasan, S. et Wang, D. (2007). Transforming binary uncertainties for robust speech recognition. *IEEE Trans. ASSP*, 15(7):2130–2140.

[Tchorz et Kollmeier, 2002] Tchorz, J. et Kollmeier, B. (2002). Estimation of the signal-to-noise ratio with amplitude modulation spectrograms. *Speech Communication*, 38(1):1–17.

[Tibrewala et Hermansky, 1997] Tibrewala, S. et Hermansky, H. (1997). Sub-band based recognition of noisy speech. *In Proc. ICASSP*, pages 1255–1258, Münich, Allemagne.

[van der Kouwe *et al.*, 2001] van der Kouwe, A. J. W., Wang, D. L. et Brown, G. J. (2001). A comparison of auditory and blind separation techniques for speech segregation. *IEEE Trans. on Speech and Audio Processing*, 9:189–195.

[van Hamme, 2003] van Hamme, H. (2003). Robust speech recognition using missing feature theory in the cepstral or LDA domain. *In Proc. European Conference on Speech Communication and Technology*, pages 3089–3092, Geneva.

[van Hamme, 2004a] van Hamme, H. (2004a). PROSPECT features and their application to missing data techniques for robust speech recognition. *In Proc. ICSLP*, volume 1, pages 101–104, Jeju Island, Korea.

[van Hamme, 2004b] van Hamme, H. (2004b). Robust speech recognition using cepstral domain missing data techniques and noisy masks. *In Proc. ICASSP*, volume 1, pages 213–216, Montreal, Quebec, Canada.

[Virtanen, 2007] Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. ASSP*, 15(3):1066–1074.

[Wang et Brown, 1999] Wang, D. L. et Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Networks*, 10:684–697.

[Wang, 2004] Wang, K. (2004). A detection-based approach to robust speech understanding. *In Proc. ICASSP*, volume 1, pages 413–417.

[Weintraub, 1985] Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation.* Thèse de doctorat, EE dept., Standford.

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

[Wiener, 1949] Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series : With Engineering Applications.* Principles of Electrical Engineering. MIT Press.

[Woodland et Young, 1993] Woodland, P. C. et Young, S. J. (1993). The HTK continuous speech recogniser. *In Proc. EUROSPEECH*, pages 2207–2219, Berlin.

[Wu et Manber, 1992] Wu, S. et Manber, U. (1992). Fast text searching : allowing errors. *Commun. ACM*, 35(10):83–91.

[Yang *et al.*, 2007] Yang, C., Soong, F. K. et Lee, T. (2007). Static and dynamic spectral features : Their noise robustness and optimal weights for ASR. *IEEE Trans. ASSP*, 15(3):1087–1097.

[Yantorno *et al.*, 2003] Yantorno, R. E., Smolenski, B. Y. et Chandra, N. (2003). Usable speech measures and their fusion. *In Proc. ISCAS.*

[Zhao et Kleijn, 2007] Zhao, D. Y. et Kleijn, W. B. (2007). HMM-based gain modeling for enhancement of speech in noise. *IEEE Trans. ASSP*, 15(3):882–892.