INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet PAROLE

Analyse, Perception et Reconnaissance de la parole

Lorraine

THÈME 3A

Table des matières

PAROLE est un projet commun à l'INRIA, au CNRS et à l'université Henri Poincaré via le laboratoire LORIA, UMR 7503.

1 Composition de l'équipe

Responsable scientifique

Yves Laprie [Chargé de Recherche, CNRS]

Assistantes de projet

Martine Kuhlmann [CNRS]

Personnel CNRS

Anne Bonneau [Chargée de Recherche]

Christophe Cerisara [Chargé de Recherche, depuis le 1er octobre 2001]

Dominique Fohr [Chargé de Recherche]

Personnel INRIA

Khalid Daoudi [Chargé de Recherche]

Personnel Université

Jean-Paul Haton [Professeur, U. H. Poincaré, Institut Universitaire de France]

Marie-Christine Haton [Professeur, U. H. Poincaré]

Irina Illina [Maître de conférences, I.U.T Charlemagne, U. Nancy 2]

Joseph di Martino [Maître de conférences, U. H. Poincaré, délégué au CNRS depuis le 1er septembre 2000]

Odile Mella [Maître de conférences, U. H. Poincaré]

Nathalie Parlangeau-Vallès [Maître de conférences, I.U.T Charlemagne, U. Nancy 2, depuis le 1er septembre 2000]

Kamel Smaïli [Maître de conférences, U. Nancy 2, délégué au CNRS]

A.T.E.R

Brigitte Bigi [A.T.E.R, U. H. Poincaré, Nancy 1, depuis le 1er octobre 2000]

Vincent Colotte [A.T.E.R, U. H. Poincaré, Nancy 1, depuis le 1er octobre 2001]

David Langlois [A.T.E.R, U. H. Poincaré, Nancy 1, depuis le 1er octobre 2001]

Chercheurs doctorants

Vincent Barreaud [MENRT depuis le 1er octobre 2001]

Yassine Benayed [Bourse tunisienne]

Armelle Brun [MENRT]

Murat Deviren [Bourse INRIA]

Salma Jamoussi [MENRT]

Fabrice Lauri [Bourse CIFRE]

Slim Ouni [Bourse franco-tunisienne]

Ingénieurs sur contrat

Christophe Antoine [projet VODIS jusqu'au 29 février, puis collaborateur extérieur, ingénieur MIC2]

Koray Balci [poste d'accueil INRIA, depuis le 1er octobre 2000]

Michel Pitermann [projet PRIAMM SALSAA]

Collaborateurs extérieurs

François Chaffard [MIC2]
Noëlle Carbonell [Professeur, U. H. Poincaré, Nancy 1]
Virginie Govaere [poste de jeune docteur à l'IUFM jusqu'à fin aout 2001]

2 Présentation et objectifs généraux

PAROLE est un projet commun à l'INRIA, au CNRS et à l'université Henri Poincaré via le laboratoire LORIA, UMR 7503.

L'objectif de notre projet est de traiter automatiquement des signaux de parole pour en comprendre la signification, ou pour analyser et renforcer la structure acoustique. Il s'inscrit dans la perspective de construire des interfaces vocales efficaces et nécessite des travaux en analyse, en perception et en reconnaissance automatique de la parole.

Nos activités se structurent suivant deux thèmes :

Analyse de la parole Nos travaux portent sur l'analyse et la perception des indices acoustiques, l'inversion acoustico-articulatoire et l'analyse de la parole. Ils donnent lieu à un certain nombre d'applications en cours ou à venir : la rééducation vocale, l'amélioration des aides auditives, l'apprentissage des langues.

Modélisation de la parole pour la reconnaissance automatique Nos travaux portent sur les modèles stochastiques (HMM[1], modèles graphiques et trajectoires acoustiques), l'approche multi-bandes, l'adaptation d'un système de reconnaissance à un nouveau locuteur ou au canal de communication et sur les modèles de langage, ce qui donne lieu à un certain nombre d'applications en cours ou à venir : la reconnaissance automatique de la parole, la dictée automatique, l'alignement texte-parole.

Notre culture est pluridisciplinaire et allie des travaux en phonétique et en reconnaissance des formes. Cette pluridisciplinarité se révèle être un atout décisif pour aborder de nouveaux thèmes de recherche, l'apprentissage des langues ou les approches multi-bandes notamment, pour lesquels il faut à la fois disposer de compétences en reconnaissance automatique de la parole et en phonétique.

Notre politique de relations industrielles consiste à favoriser les contrats s'insérant assez précisément dans nos objectifs scientifiques. Nous sommes impliqués dans plusieurs coopérations avec des industriels utilisant la reconnaissance automatique de la parole, notamment MIC2 avec qui nous avons une coopération en cours sous la forme d'un projet RNRT, Syncmagic Procoma avec qui nous avons un contrat PRIAMM sur le Lipsync et Babel Technologies qui commercialise notre logiciel d'analyse de la parole WinSnoori. Par ailleurs, jusqu'en juillet, nous étions impliqués dans le projet européen ISAEUS sur la rééducation vocale. Nous travaillons également avec des enseignants de langue de Nancy dans le cadre d'un projet du Plan État Région.

3 Fondements scientifiques

Mots clés : traitement du signal, phonétique, télécommunications, santé, perception,

[1] Hidden Markov Models

modèles stochastiques, modèles de langage, modèles articulatoires, apprentissage des langues, reconnaissance automatique de la parole, aides auditives, analyse de la parole, indices acoustiques.

Globalement les recherches sur la parole ont donné lieu à deux types d'approches :

- des recherches visant à expliquer comment la parole est produite et perçue, donc incluant des aspects physiologiques (contrôle du conduit vocal), physiques (acoustique de la parole), psychoacoustiques (système auditif périphérique), et cognitifs (construction des phrases),
- des recherches visant à modéliser les observations des phénomènes de la parole (analyse spectrale, modèles stochastiques acoustiques et linguistiques).

Les premières recherches sont motivées par la très grande spécificité de la parole parmi tous les signaux acoustiques : l'appareil de production de la parole est facilement accessible (du moins en première approche), les équations acoustiques relativement abordables d'un point de vue mathématique (au prix de simplifications qui ne sont que modérément restrictives), les phrases produites sont régies par le vocabulaire et la grammaire de la langue étudiée. Cela a conduit les acousticiens à développer des recherches visant à produire un signal de parole artificiel de bonne qualité, les phonéticiens des recherches visant à trouver l'origine de la variabilité des sons de la parole et à expliquer comment les articulateurs sont utilisés, comment les sons d'une langue s'organisent et comment ils s'influencent dans la parole continue. Enfin, cela a conduit les linguistes à mener des recherches pour savoir comment les phrases sont construites. Il est clair que cette approche donne lieu à de nombreux allers et retours entre la théorie et l'expérimentation et qu'il est difficile de maîtriser simultanément tous ces aspects de la parole.

Les résultats disponibles sur la production et la perception de la parole ne permettent cependant pas d'envisager une approche d'analyse par synthèse. La reconnaissance automatique a donc suscité une seconde approche consistant à modéliser les observations des phénomènes de la parole. Les efforts ont porté sur l'élaboration de modèles numériques (d'abord de simples vecteurs de formes spectrales et maintenant des modèles stochastiques ou neuromimétiques) des réalisations acoustiques des phonèmes ou des mots, et sur le développement de modèles de langages statistiques.

Ces deux approches sont complémentaires ; la seconde emprunte à la première les résultats théoriques sur la parole et la première emprunte à la seconde certains outils numériques, les techniques d'analyse spectrale étant sans doute le domaine où les échanges sont les plus marqués. L'existence de ces deux approches est l'une des particularités des recherches en parole menées à Nancy et nous comptons renforcer les échanges entre elles. Ces échanges sont d'ailleurs conduits à se multiplier depuis que les systèmes de reconnaissance automatique (en particulier destinés à la dictée automatique) sont disponibles pour le grand public : il faut augmenter leur robustesse au plan acoustique (robustesse au bruit, adaptation au locuteur) comme au plan linguistique.

Les activités de notre équipe se structurent suivant ces deux approches :

Production et perception Nos recherches portent sur l'analyse et la perception des indices acoustiques, l'inversion acoustico-articulatoire et l'analyse de la parole. Elles donnent et donneront lieu à un certain nombre d'applications : la rééducation vocale, l'amélioration

des aides auditives, l'apprentissage des langues.

Modélisation de la parole pour la reconnaissance automatique Nos recherches portent sur les modèles stochastiques, les modèles de langage et les modèles multi-bandes. Elles donnent et donneront lieu à un certain nombre d'applications : la reconnaissance automatique de la parole, la dictée automatique, l'alignement texte-parole et la classification de signaux différents de la parole.

3.1 Analyse de la parole

Participants : Anne Bonneau, Jean-Paul Haton, Marie-Christine Haton, Yves Laprie, Joseph di Martino, Christophe Antoine, Vincent Colotte, Virginie Govaere, Slim Ouni.

3.1.1 Perception

Nous menons des études perceptives afin d'approfondir les connaissances sur les indices essentiels d'identification ainsi que sur les mécanismes de perception des sons de la parole. Les domaines d'application de nos travaux vont de la reconnaissance automatique de la parole aux domaines paramédicaux, comme l'aide aux malentendants, et aux logiciels d'aide à la prononciation.

Nos expériences ont concerné la perception du lieu d'articulation des occlusives sourdes du français, le rôle du contexte vocalique dans leur identification ainsi que l'identification de la voyelle à partir du bruit d'explosion de ces consonnes [12]. Nous avons également étudié l'effet des modifications d'amplitude des formants sur la perception des voyelles. Nous savons que les fréquences formantiques ont un rôle déterminant dans la perception des voyelles et nous avons voulu approfondir le rôle de l'amplitude, un paramètre certes moins important, mais qui, pour certains formants et certaines oppositions vocaliques, peut se révéler également déterminant.

3.1.2 Indices acoustiques

Nous reprenons un travail entrepris sur les indices forts il y a quelques années. Au moment où nous avons introduit le concept d'indice fort, nous désirions pallier une lacune des systèmes de reconnaissance de la parole : l'absence de certitude. En effet, du fait des nombreuses sources de variation qui influencent le signal de parole, les valeurs prises par un indice donné pour deux ou plusieurs unités différentes se recouvrent partiellement. Dans les systèmes de reconnaissance, un coefficient de confiance est attribué en fonction de la valeur de chaque indice considéré et de chaque unité candidate à l'identification. Ainsi l'identification d'un son ou d'un trait repose sur une combinaison de poids complexe. Un tel procédé n'aboutit jamais à une identification certaine. Or certaines configurations acoustiques signalent sans aucune ambiguïté la présence ou l'absence d'un trait ; les jugements définitifs parfois émis par les lecteurs de spectrogrammes nous le confirment. Nous avons donc entrepris de décrire ces formes et nous avons défini deux types d'indices acoustico-phonétiques : des indices « forts », de préférence ou d'exclusion, et des indices « faibles ». Les indices forts de préférence autorisent l'identification immédiate d'un trait, les indices forts d'exclusion éliminent directement un candidat à l'identification. Les indices forts ont donc pour fonction de faire reposer la reconnaissance d'un trait phonétique

sur un certain nombre d'informations présentées comme certaines. L'intérêt de tels indices pour l'analyse lexicale est évident : ils permettent d'élaguer le nombre d'hypothèses de mots, toujours très important dans un système de reconnaissance de la parole à moyen ou grand vocabulaire.

Avec l'apparition de nouvelles technologies qui permettent de renforcer les indices importants, l'intérêt des indices forts ne se limite plus à la reconnaissance de la parole mais trouve des applications dans l'apprentissage des langues et les aides auditives. En effet, un indice fort est un indice très discriminant d'un point de vue phonétique et bien marqué d'un point de vue acoustique. Le renforcement de ce type d'indices doit permettre aux apprenants de mieux assimiler les caractéristiques des sons de la langue qu'ils étudient et aux handicapés de mieux percevoir les sons de parole.

3.1.3 Aides auditives

Dans les aides conventionnelles, le signal est capturé à l'aide d'un microphone, reconditionné par l'aide auditive et diffusé dans l'oreille moyenne. Ces aides utilisent des techniques de filtrage et de contrôle automatique du gain. Le filtrage permet de décomposer le signal en bandes de fréquence traitées en parallèle et le contrôle automatique du gain permet de réduire la dynamique du signal afin d'assurer la perception de l'amplitude et de préserver le confort du patient. La qualité globale d'une aide auditive vient de la stratégie d'utilisation de ces outils de base. L'un des objectifs majeurs de la recherche sur les aides auditives est d'exploiter le mieux possible les spécificités de la parole pour guider les techniques de traitement du signal qui deviennent de plus en plus puissantes. Notre contribution intervient à deux niveaux : celui du diagnostic et celui des stratégies de correction du signal de parole.

En ce qui concerne le diagnostic, il apparaît qu'il faut compléter l'audiogramme tonal actuel mais en évitant de verser dans le développement de tests psycho-acoustiques souvent très lourds à mettre en œuvre et demandant une attention prolongée de la part du patient. Nous utilisons donc des stimuli artificiels mais construits à partir de la parole naturelle.

En ce qui concerne les transformations de la parole, il existe un certain nombre de pistes destinées à compléter les techniques actuelles. Les efforts les plus importants correspondent au renforcement des pics spectraux, le but étant de préserver la perception des pics malgré une perte de sélectivité fréquentielle ou temporelle.

3.1.4 Inversion articulatoire

Les travaux sur l'inversion acoustique articulatoire reposent largement sur une approche d'analyse par synthèse articulatoire qui recouvre trois aspects essentiels :

la résolution des équations de l'acoustique Pour résoudre les équations de l'acoustique adaptées au conduit vocal, on fait l'hypothèse que l'onde sonore est une onde plane dans le conduit vocal et que le conduit peut être redressé. Il existe deux grandes familles de résolutions : **(i)** fréquentielles grâce à l'analogie acoustico-électrique, **(ii)** spatio-temporelles, par la résolution directe des équations aux différences finies issues des équations de Webster.

les mesures du conduit vocal Cet aspect représente un obstacle important car il n'existe pas de méthode fiable pour mesurer le conduit vocal avec précision. L'IRM permet de mesurer le conduit vocal en 3D mais n'est pas assez rapide et les rayons X ne permettent que de récupérer une coupe sagittale du conduit vocal.

la modélisation articulatoire L'un des objectifs de la modélisation articulatoire est de décrire avec un petit nombre de paramètres les formes possibles du conduit vocal tout en préservant les déformations observées sur un conduit réel. Les modèles articulatoires actuels sont souvent le résultat d'analyses statistiques de films ciné-radiographiques, comme par exemple le modèle de Maeda.

L'une des difficultés majeures de l'inversion est qu'une infinité de formes de conduits peuvent donner un même spectre de parole. Les méthodes d'inversion acoustico-articulatoire s'organisent en deux familles :

- les méthodes d'optimisation d'une fonction combinant généralement l'effort articulatoire du locuteur et la distance acoustique entre la parole réelle et la parole synthétisée. Ces méthodes font appel à un certain nombre de contraintes permettant de réduire le nombre de formes de conduits possibles.
- les méthodes par tabulation. Ces méthodes reposent sur un dictionnaire de formes articulatoires indexées acoustiquement (généralement par les fréquences des formants). Après avoir récupéré à chaque instant les formes possibles, une procédure d'optimisation permet de trouver une solution d'inversion sous la forme d'un chemin optimal.

Comme notre contribution ne porte que sur l'inversion, nous avons repris les méthodes de synthèse articulatoire les plus couramment utilisées. Nous utilisons donc le modèle articulatoire de Maeda, l'analogie acoustique électrique pour calculer le spectre de parole et une méthode spatio-temporelle pour produire le signal de parole.

Pour ce qui concerne l'inversion, nous avons choisi d'utiliser le modèle de articulatoire de Maeda pour contraindre les formes de conduit vocal. Ce choix assure que les phénomènes de synergie et de compensation articulatoire sont toujours possibles, ce qui est important pour récupérer des mouvements articulatoires proches de ceux d'un locuteur humain.

3.2 Reconnaissance automatique de la parole

Participants : Dominique Fohr, Jean-Paul Haton, Irina Illina, Odile Mella, Kamel Smaïli, Christophe Antoine, Armelle Brun, Christophe Cerisara, David Langlois, Imed Zitouni, Khalid Daoudi, Michel Pitermann, Yassine Benayed, Murat Deviren, Angel de la Torre Vega, Fabrice Lauri, Vincent Barreaud, Salma Jamoussi.

La reconnaissance automatique de la parole nécessite l'utilisation imbriquée de modèles acoustiques et de modèles de langage. Les modèles acoustiques permettent de prendre en compte des contraintes acoustiques et phonétiques au niveau d'un son ou d'un groupe de sons alors que les modèles de langages définissent les contraintes syntaxiques et sémantiques au sein d'un groupe de mots ou d'une phrase.

Malgré la forte imbrication entre ces deux types de modèles, nous les présentons dans deux paragraphes successifs pour plus de clarté.

3.2.1 Modèles acoustiques

Les techniques stochastiques sont actuellement les plus utilisées pour la modélisation acoustique de la parole. En effet, ce sont celles qui ont permis d'obtenir les meilleurs résultats en reconnaissance de mots isolés, mots enchaînés et parole continue dans des conditions de laboratoire ou en environnement non bruité. En revanche, dans des conditions réelles de traitement de la parole (milieu bruité, parole spontanée, prononciations diverses et variées ...), les performances obtenues par ces techniques sont fortement dégradées ce qui justifie nos recherches actuelles et futures.

Aussi notre groupe travaille-t-il sur l'amélioration de la modélisation de parole par des modèles de Markov cachés (Hidden Markov Models ou HMM) et a-t-il développé deux classes de modèles stochastiques originaux pour la reconnaissance automatique de la parole : les réseaux bayésiens et les modèles stochastiques de trajectoires (Stochastic Trajectory Modeling ou STM).

Les **modèles de Markov cachés** nous ont permis de réaliser des systèmes de reconnaissance automatique de lettres épelées, de chiffres connectés ou de parole continue et de tester différents algorithmes de paramétrisation dans le cas de la parole bruitée ou téléphonique.

Les **réseaux bayésiens** consistent à associer un graphe orienté non-cyclique à la distribution jointe d'un ensemble de variables aléatoires donné. Les nœuds de ce graphe représentent les variables, alors que les liens entre les nœuds codent les indépendances conditionnelles qui existent (ou qui sont supposées exister) dans la distribution jointe. Les HMM sont un cas particulier des réseaux bayésiens. Ces derniers nous offrent donc un cadre théorique général qui nous permet de proposer de nouveaux modèles capables de représenter la parole plus fidèlement que les HMM.

Les **modèles stochastiques de trajectoires (STM)** utilisent une approche novatrice pour reconnaître la parole. Plutôt que d'analyser à intervalle de temps fixé le signal de parole, les STM modélisent la trajectoire du signal dans l'espace de représentation (fréquentiel ou cepstral). L'unité à reconnaître – le mot ou le phonème – est retrouvée grâce à une probabilité d'appartenance à une classe qui intègre les informations de durée et d'évolution des paramètres acoustico-phonétiques.

3.2.2 Modèles de langage

Les systèmes de dictée automatique donnent de bons résultats acoustiques ; néanmoins plusieurs problèmes au niveau langagier n'ont toujours pas de solution. La communauté scientifique travaillant sur la reconnaissance automatique de la parole a pris conscience qu'il devient indispensable de fournir plus d'efforts pour concevoir des modèles de langage plus performants et ayant une meilleure interaction avec les niveaux acoustiques. En effet, les modèles de langage d'aujourd'hui sont, dans la plupart des cas, des modèles stochastiques ayant une portée locale ou à court terme (les modèles avec mémoire cache). Même si ces systèmes donnent de bons résultats, ils restent néanmoins limités et ont besoin d'être constamment améliorés pour s'adapter à la complexité de la langue. Afin de maîtriser cette complexité du langage, nous avons continué nos travaux de recherche portant sur l'adaptation dynamique des modèles de langage, par le biais d'études concernant l'identification thématique. Le second axe consiste à essayer de modéliser quelques phénomènes sémantiques de la langue d'une manière statistique

afin de lever certaines ambiguïtés et ainsi améliorer les taux de reconnaissance. Enfin, nous avons décidé de prospecter dans une nouvelle voie de recherche : la compréhension.

4 Domaines d'applications

Les domaines d'application de nos travaux vont de la reconnaissance automatique de la parole aux domaines paramédicaux. Les méthodes d'analyse de la parole contribueront au développement de nouvelles technologies concernant l'aide à la prononciation (par exemple pour les malentendants ou pour l'apprentissage des langues) et les systèmes d'aides auditives.

Par ailleurs, la parole a et aura un rôle de plus en plus important dans les modalités d'interaction homme-machine. En effet, l'expression orale en langue naturelle est un mode de communication susceptible de séduire le grand public, surtout dans un environnement multimodal où l'association à la parole de gestes de désignation sur un écran tactile permet notamment de simplifier l'interprétation des expressions linguistiques de référence spatiale. D'autre part, le recours à la parole s'impose dans de nombreuses applications nouvelles où l'usage du clavier est malaisé, voire impossible : informatique mobile ou embarquée, serveurs vocaux, bornes interactives, informatique domestique, téléphone. Enfin, la multiplication des documents sonores disponibles sur le Web et non répertoriés par les moteurs de recherche classiques comme Google ou Yahoo, ouvre une nouvelle voie d'application. En effet l'indexation automatique de documents sonores ou audiovisuels permettra qu'ils soient référencés et donc exploitables.

Notre intérêt pour l'apprentissage de la voix et de la parole a permis dans le passé le développement d'un ensemble d'outils éducatifs utilisant l'entrée vocale et les techniques d'analyse et de reconnaissance élaborées dans l'équipe (voir le projet européen ISAEUS qui s'est terminé en juillet 2000) [33]. Nous poursuivons dans cette optique des travaux sur le suivi de l'apprenant en situation d'apprentissage d'un cours diffusé sur le web : constitution automatique de « sous-sites », donc de portions de cours, suivi de la navigation pour conseiller l'élève.

5 Logiciels

5.1 Outils logiciels

5.1.1 Snorri

Snorri est le logiciel d'étude de la parole que nous avons développé et amélioré depuis 10 ans. Il est destiné à faciliter le travail du chercheur en reconnaissance de la parole, en phonétique, en perception ou encore en traitement du signal. Les fonctions de base de Snorri permettent de calculer plusieurs types de spectrogrammes et d'éditer le signal de parole de manière très fine (couper, coller, filtrages et atténuations diverses) car le spectrogramme permet de connaître la répercussion acoustique de toutes les modifications. À cela s'ajoute un grand nombre de fonctions destinées à étiqueter phonétiquement ou orthographiquement des signaux de parole, des fonctions destinées à extraire la fréquence fondamentale de la parole, des fonctions destinées à piloter le synthétiseur de Klatt et d'autres à utiliser la synthèse PSOLA.

Snorri a servi de base logicielle pour un grand nombre de travaux dans notre équipe (suivi de formants, identification des occlusives, études perceptives, ...). Étant donné l'intérêt qu'il

représente pour l'étude de la parole nous l'avons diffusé auprès d'une quinzaine d'équipes francophones, dont celle du CNET de Lannion. Initialement développé sous Unix et Motif, nous l'avons porté sous Windows et nous le commercialisons depuis cet automne sous le nom de WinSnoori par l'intermédiaire de Babel Technologies (startup située à Mons en Belgique et vendant des logiciels de synthèse et de reconnaissance automatique de la parole).

5.1.2 Étiquetage de corpus écrits pour la reconnaissance

Nous avons développé un outil d'étiquetage permettant de résoudre syntaxiquement un texte. Il permet d'affecter à chaque mot d'une phrase sa classe syntaxique en fonction du contexte dans lequel celui-ci apparaît. Cet outil d'étiquetage utilise, pour fonctionner, un dictionnaire de 230 000 formes ainsi qu'un jeu de classes syntaxiques comportant 230 étiquettes. Le taux d'erreur de l'étiqueteur est de 1 %.

5.1.3 Classifieur automatique de lexique

Pour adapter nos modèles de langage aux différentes applications de la dictée automatique, nous avons développé un outil permettant, à partir d'un vocabulaire donné et d'un corpus d'apprentissage, de proposer un jeu de classes permettant d'avoir un modèle de langage de perplexité minimale. Cet outil est fondé sur l'algorithme du recuit simulé et comprend plusieurs variantes : classification initiale aléatoire ou fixée, nombre de classes fixé, perplexité fixée, etc.

5.1.4 SALT

SALT (Semi-Automatic Labelling Tool) est un outil d'étiquetage semi-automatique de grands corpus oraux. À partir du texte de la phrase prononcée, d'un dictionnaire phonétique et de règles phonologiques, il génère un graphe des prononciations possibles pour une phrase. Ensuite, il effectue un alignement forcé de ce graphe sur le signal de parole grâce à des modèles de Markov du second ordre (algorithme de Viterbi). L'étiquetage est affiné itérativement à l'aide d'un logiciel de comparaison d'étiquetage.

5.1.5 LIPS

Dans le cadre de la réalisation de dessins animés, il est nécessaire de synchroniser le mouvement des lèvres des personnages avec la phrase prononcée par l'acteur. Cette phase, jusqu'alors réalisée manuellement, peut maintenant être effectuée grâce à notre logiciel LIPS (Logiciel Intégré de Post-Synchronisation) qui permet l'alignement automatique d'un texte anglais ou français avec le signal audio correspondant. Deux versions du logiciel ont été implantées : l'une sous PC-Linux, l'autre sous PC-Windows.

5.1.6 VINICS

L'étude fait partie du projet IMAGIN mené par le CEA dans le domaine des bases de données de centrales nucléaires. Deux aspects de la reconnaissance ont été abordés :

- reconnaissance de la parole continue. Une version de notre système VINICS a été réécrite en C++ par les ingénieurs du CEA.

– Vérification du locuteur. Une interface graphique a été développée pour un de nos systèmes de vérification et l'ensemble a été livré au CEA-Cadarache.

5.1.7 ESPERE

Nous avons développé un moteur de reconnaissance de parole générique fondé sur les modèles de Markov cachés (HMM). Ce moteur ESPERE (Engine for SPEech REcognition) permet de reconnaître aussi bien des mots isolés que connectés, ou que des mots clefs ou de la parole continue. Entièrement développé en C++, il fonctionne sous UNIX ou sous Windows.

5.2 Corpus

Les recherches menées dans le domaine de la communication parlée ont un point commun : elles nécessitent l'enregistrement, la manipulation et le traitement de corpus de plus en plus importants.

Ainsi, pour mener à bien des études sur les indices phonétiques, il est nécessaire d'enregistrer et d'étiqueter phonétiquement de nombreuses phrases, afin de capturer le maximum d'effets contextuels ; mais ces phrases doivent aussi être prononcées par de nombreux locuteurs, afin cette fois-ci de capturer les variations interlocuteurs. Citons, dans ce cadre, notre participation au projet Européen VODIS pour l'enregistrement, la numérisation et l'étiquetage d'un corpus de plus de 200 automobilistes en conditions réelles.

Depuis plusieurs années déjà, nous avons développé des outils permettant d'éditer, de traiter et d'étiqueter manuellement de telles bases de données de parole, comme Snorri présenté dans le paragraphe 5.1.

Un autre exemple concerne la constitution de grands corpus et leur étiquetage automatique en vue d'entraîner les systèmes de reconnaissance de parole faisant appel à des modèles statistiques, stochastiques ou neuromimétiques. En effet, pour évaluer les paramètres de ces modèles, il faut disposer d'une grande quantité de données d'apprentissage. Les modèles étant de plus en plus précis (contextuels, multigaussiennes,...), le nombre de paramètres libres, donc à apprendre, est devenu de plus en plus grand, ce qui nécessite une augmentation considérable de la taille des corpus étiquetés. A l'heure actuelle, les corpus de parole continue contiennent de nombreuses heures de parole tels ceux du LIMSI (BREF 15 Go) de l'ARPA (Wall Street Journal 20 Go) ou des PTT suisses (Polyphone 10 Go).

De tels corpus, de plusieurs dizaines de milliers de phrases, ne peuvent plus être étiquetés manuellement. Aussi avons nous développé des outils d'étiquetage semi-automatique de grands corpus (cf. paragraphe 5.1.4).

De la même façon, la manipulation de grands corpus de texte est indispensable pour la conception de modèles de langages probabilistes. Ainsi, dans le cadre de la machine à dicter (projet AUPELF-UREF [11]), les modèles bi et trigrammes ont été évalués à partir d'un corpus de 50 millions de mots issus d'articles du journal « Le Monde ».

La taille des corpus disponibles ne cesse d'augmenter. Aux Etats Unis, des corpus de plus de 300 millions de mots sont déjà distribués comme le « North American News Text ». Il sera donc nécessaire d'améliorer continuellement les outils logiciels pour les traiter.

6 Résultats nouveaux

6.1 Analyse de la parole

Mots clés : traitement du signal, phonétique, santé, perception, modèles articulatoires, apprentissage des langues, aides auditives, analyse de la parole, indices acoustiques.

6.1.1 Indices acoustiques

Nous avons élaboré un détecteur du bruit d'explosion des consonnes occlusives[39]. Ce détecteur permet à la fois d'améliorer la détection du bruit d'explosion de ces consonnes et de le segmenter en deux parties : l'attaque et la friction. Le bruit d'explosion est souvent assez bref (inférieur à 30 ms) et parfois faible ; sa détection est donc une opération délicate. Elle est pourtant indispensable à l'identification des occlusives car le bruit contient les indices les plus importants pour le lieu d'articulation. La segmentation en deux parties, attaque et friction, augmente la robustesse de ces indices. En effet, à l'inverse du bruit des fricatives, le bruit d'explosion des occlusives ne correspond pas à la tenue de l'articulation, mais à un mouvement transitoire entre l'articulation de la consonne et celle du son suivant. Par conséquent, le spectre du bruit change très rapidement en fonction du temps et, dans la plupart des cas, les indices consonantiques sont plus discriminants au début du bruit (l'attaque) qu'à la fin de celui-ci. Les deux segments (attaque et friction) sont segmentés par l'emploi d'un critère qui minimise la somme des variances spectrales. Le critère de variance présente l'avantage d'être sensible à la fois aux variations d'énergie et aux variations spectrales. D'autres procédures ont été ajoutées afin d'améliorer la robustesse de la segmentation. Des tests ont montré que l'attaque, segmentée par notre méthode, permet d'obtenir des indices d'identification plus performants que le bruit entier. Nous avons maintenant défini des indices forts pour toutes les classes de voyelles à partir du bruit des occlusives. Nous poursuivons notre définition d'indices pour les transitions formantiques entre les occlusives et les voyelles.

6.1.2 Compréhension orale

Pour améliorer la compréhension orale nous avons développé l'an dernier des outils de transformation de la parole pour ralentir sélectivement le débit de parole et amplifier certains indices acoustiques. L'objectif est d'élaborer des stratégies de transformation qui renforcent l'intelligibilité de la parole. Pour ne pas introduire d'artefacts acoustiques qui risqueraient de détériorer l'identification des sons nous avons adopté une stratégie qui consiste à renforcer seulement les consonnes sourdes et les transitions spectrales rapides. Cette année nous avons conduit une première expérience de perception [22, 23] dans le cadre de l'apprentissage du français langue étrangère. Nous avons utilisé 50 phrases du corpus BDSONS auxquelles nous avons appliqué seulement l'amplification des consonnes sourdes (condition B) ou les transformations complètes qui comprennent l'amplification et le ralentissement sélectif (condition C). La condition A correspond aux phrases sans modifications. Les résultats montrent un gain d'identification de 9% entre la condition A et la condition B (en passant de 72 à 81%) et un gain total de 14% entre la condition A et la condition C (en passant de 72 à 86%). Les exercices

consistaient à compléter des phrases lacunaires et les sujets ont manifestement utilisé une stratégie d'identification de mots. Nous compléterons donc ces expériences au niveau phonétique en utilisant des logatomes sans signification placés dans une phrase porteuse.

Transformations du signal de parole La prosodie couvre les aspects de l'intonation (mesurée par la fréquence fondamentale), de l'énergie et du rythme. Nous avons ajouté cette année la possibilité de manipuler complètement le rythme de la parole. Pour cela nous utilisons les outils de reconnaissance automatique de la parole qui donnent la segmentation du signal en sons. Il suffit alors de jouer avec PSOLA (Pitch Synchronous Overlap and Add) sur la durée de chacun des sons pour modifier le rythme.

Par ailleurs, nous poursuivons nos travaux sur le vocodeur de phase dont l'avantage est de ne nécessiter aucun calcul du fondamental. L'idée est de séparer la contribution de l'amplitude de celle de la phase et de reconstruire un signal qui correspond à la transformation envisagée. L'une des difficultés est l'apparition d'un phénomène acoustique appelé « phasiness » qui donne l'impression d'une voix enregistrée avec un microphone trop éloigné du locuteur. Ce phénomène s'explique par la destruction de la structure des phases du signal initial, et par conséquent de la forme du signal temporel. Pour éliminer cet effet, nous avons conçu une méthode d'optimisation destinée à assurer que la forme du signal temporel, et par conséquent la structure des phases, est bien conservée. Cette étape d'optimisation peut être vue comme une procédure de synchronisation des phases et elle est déclenchée à chaque début de région voisée. Cet algorithme a donné de bons résultats et a été publié à ICASSP-2001 [28]. Au cours de cette année, nous avons travaillé sur la possibilité de déclencher la procédure d'optimisation à tout moment au cours de la zone voisée. Nous y sommes parvenus au moyen d'une technique qui consiste à interpoler linéairement les phases dans une zone de jonction autour de l'instant de synchronisation considéré. Nous avons pu constater que grâce à cette technique et pourvu que le zone de jonction ne soit pas trop étroite temporellement les fichiers sons obtenus, étaient d'excellentes qualité et exempts de « phasiness ».

6.1.3 Inversion articulatoire

L'inversion est effectuée en deux étapes : d'abord retrouver à chaque instant tous les paramètres articulatoires susceptibles d'être à l'origine des paramètres acoustiques observés, ensuite reconstruire une trajectoire articulatoire régulière à partir de ces points. Le modèle articulatoire de Maeda décrit la forme du conduit vocal avec sept paramètres et nous utilisons trois paramètres acoustiques qui sont les trois premiers formants.

La première étape repose sur l'utilisation d'un dictionnaire de formes articulatoires indexées par les formants car une infinité de formes de conduit vocal peuvent donner les mêmes formants. Le dictionnaire de formes articulatoires est structuré sous la forme d'une arborescence d'hypercubes pour accélérer la recherche. La taille des hypercubes est adaptée de manière à ce que la relation articulatoire acoustique soit linéaire à l'intérieur de chaque hypercube. Pour trouver toutes les formes de conduits possibles on commence par récupérer tous les cubes compatibles avec les formants observés. On calcule alors les paramètres articulatoires qui donnent les formants observés pour chacun des cubes candidats. Pour cela il faut résoudre un système sous-déterminé puisqu'il faut trouver sept paramètres articulatoires à partir des trois premiers

formants. La difficulté est donc de trouver les points de l'espace nul à l'intérieur du cube candidat, c'est-à-dire l'intersection de l'espace nul défini par un point et les vecteurs de base avec le cube. Comme il n'est pas possible de résoudre ce problème facilement, nous avons d'abord déterminé les valeurs extrêmes des paramètres articulatoires en résolvant avec l'algorithme du simplexe deux programme linéaires par paramètre articulatoire. Cela permet d'obtenir un polygone qui contient le polygone recherché. Ce polygone est alors échantillonné et nous retenons les points qui appartiennent au cube candidat [40]. De cette manière, nous connaissons pour chaque cube candidat tous les points possibles en fonction d'un pas d'échantillonnage qui dépend de la précision articulatoire souhaitée.

La seconde étape de l'inversion consiste alors à reconstruire les meilleures trajectoires articulatoires possibles en fonction d'un critère qui limite l'effort articulatoire du locuteur et assure une bonne proximité acoustiques avec le signal à inverser.

Nous avons réalisé l'inversion pour des suites de plusieurs voyelles. Faute de données articulatoires l'évaluation a porté sur l'évolution de la forme géométrique du conduit vocal et non pas sur les paramètres articulatoires. Nous avons remarqué en particulier que la minimisation de l'effort articulatoire pénalise les mouvements impliquant deux articulateurs quand il est possible de réaliser la même trajectoire acoustique et géométrique avec un seul paramètre. C'est en particulier le cas pour les trajectoires impliquant la position et la pointe de la langue (lors de l'articulation d'une consonne dentale), pour lesquelles notre méthode a favorisé la déformation de la forme de la langue et dans une moindre mesure la position de la mâchoire.

Par ailleurs, nous avons aussi utilisé le dictionnaire de formes articulatoires pour trouver toutes les formes de conduit correspondant à une voyelle donnée. Cela permet de connaître tous les lieux d'articulation (l'emplacement dans le conduit vocal de l'aire transverse minimale) d'une voyelle [41].

6.1.4 Enseignement des sciences de la parole

Les outils de traitement du signal qui nous ont servi pour la compréhension orale (l'algorithme de détection de la fréquence fondamentale, l'algorithme de marquage des périodes du fondamental et la mise en œuvre de la méthode PSOLA) peuvent être utilisés dans le cadre de l'apprentissage de la prosodie.

L'opportunité de travailler avec des enseignants de langue (français en tant que langue étrangère et anglais) nous a été offerte dans le cadre du Plan État Région. Depuis le mois d'octobre 2000 nous organisons un séminaire commun qui nous a permis de nous fixer comme objectif l'enseignement de la prosodie, sans doute l'un des aspects les moins bien traités par les logiciels éducatifs actuels du domaine.

La construction de turoriaux sur la parole nécessite de pouvoir insérer des outils d'analyse de la parole directement dans des pages web ou des documents Powerpoint. Nous avons donc développé un ensemble de contrôles ActiveX (propres au système Windows) à partir de notre logiciel d'analyse de la parole WinSnoori. WinSnoori est programmé en C++ qui est le langage le plus simple à utiliser pour construire des contrôles ActiveX. Par ailleurs, la plupart des phonéticiens et enseignants des sciences de la parole préfère le système Windows. Ces deux raisons expliquent le choix des contrôles ActiveX. Pour l'instant, les objets et contrôles signal, spectrogramme, annotation, et coupe spectrale sont disponibles avec notamment la possibilité

de modifier tous les paramètres prosodiques.

6.2 Reconnaissance automatique de la parole

Mots clés : télécommunications, modèles stochastiques, modèles acoustiques, modèles de langage, reconnaissance automatique de la parole, apprentissage.

Nos travaux sur la reconnaissance automatique de la parole sont classés suivant le type de modélisation stochastique choisi.

6.2.1 Modèles de Markov Cachés

Moteur de reconnaissance de parole générique Afin de concevoir et de tester de nouveaux algorithmes pour la reconnaissance automatique de la parole, nous avons développé le moteur générique ESPERE basé sur les modèles de Markov cachés (HMM) permettant de définir des modèles, de réaliser leur apprentissage et d'évaluer leurs performances sur de grands corpus de parole continue.

Cette année, nous avons amélioré la phase d'apprentissage des modèles d'une part en implantant un nouvel algorithme de choix des gaussiennes constituant la densité de probabilité des observations et d'autre part en étudiant un nouvel algorithme d'apprentissage discriminant.

Modèles et algorithmes Dans des applications réelles, le système de reconnaissance automatique de la parole doit être capable de s'adapter à un nouveau locuteur et de prendre en compte des mots hors vocabulaire.

L'adaptation au locuteur consiste à modifier les moyennes et les écarts type des gaussiennes des modèles en fonction des premières phrases prononcées par le locuteur. Nous avons débuté une thèse sur l'adaptation et nous nous focalisons sur l'amélioration de la méthode MLLR (Maximum Likelihood Linear Regression) consistant à calculer une ou plusieurs transformations linéaires pour modifier ces moyennes [20].

Toutefois, les méthodes MLLR souffrent souvent du manque de données d'adaptation. En effet, plusieurs phrases doivent être préalablement prononcées, avant que l'adaptation ne puisse être réalisée. Afin de réduire l'ampleur de ce problème, nous avons modélisé, dans MLLR, la dépendance entre les classes de régression regroupant les modèles de Markov, par un processus auto-régressif multi-échelles. Ainsi, grâce à ce modèle, il est possible d'adapter les modèles d'une classe de régression sans que le locuteur n'ait prononcé un seul des phonèmes appartenant à cette classe. Nous poursuivons nos efforts dans cette voie de recherche qui semble prometteuse, mais qui nécessite également beaucoup de travail concernant l'apprentissage du modèle multi-échelles [19] .

Pour certaines applications vocales, la requête d'un locuteur contient des mots indispensables (mots clés) à la compréhension de la requête et des mots outils ou de politesse non nécessaires (mots hors vocabulaires). Une thèse est en cours sur ce sujet, ainsi qu'une collaboration avec la société Yacast.

Détection parole/non-parole L'essor du téléphone et des applications téléphoniques vocales place la reconnaissance de parole téléphonique comme une problématique incontournable

pour notre groupe de recherche. En parole téléphonique, la détection du début et de la fin de la phrase prononcée est primordiale. Nous continuons à tester de nouveaux algorithmes en collaboration avec la société MIC2.

Robustesse au bruit Les voitures sont de plus en plus souvent pourvues d'équipements de haute technologie (systèmes de navigation, téléphones,...) pour lesquels la commande vocale est incontournable. Nous cherchons à augmenter la robustesse de nos modèles acoustiques (HMM) aux bruits habituellement présents dans une voiture ou dans un cockpit d'avion.

L'adaptation à l'environnement permet de transformer les HMM afin que ceux-ci intègrent les nouvelles caractéristiques de l'environnement de test. Dans ce cadre, nous avons :

- amélioré l'algorithme d'adaptation Jacobienne développé l'année précédente, tout d'abord en ajoutant un terme d'adaptation au bruit convolutif [21], puis en incorporant un module permettant d'estimer dynamiquement certains paramètres de l'algorithme d'adaptation. Le but est double : éliminer la nécessité d'utiliser un corpus de développement, et par conséquent, rendre la méthode plus « stable » car moins dépendante des conditions d'apprentissage et de développement.
- proposé et implémenté une méthode concurrente à l'adaptation Jacobienne, qui combine l'algorithme PMC pour l'adaptation au bruit additif, avec l'algorithme CMS pour l'adaptation au bruit convolutif.
- travaillé sur les modèles HMM multi-bandes [13].

Dans ce cadre, nous avons mené une étude comparative de différentes méthodes : MLLR (Maximum Likelihood Linear Regression)et PMC (Parallel Model Combination) qui adaptent les modèles, CMN (Cesptral Mean Normalisation) et la soustraction spectrale qui débruitent le signal [20].

Nous avons également mené une étude en coopération avec l'Université de Granada sur la comparaison de différentes méthodes pour compenser le bruit ambiant dans une voiture [26].

6.2.2 Réseaux Bayesiens

Notre stratégie de recherche dans ce domaine consiste à concevoir de nouveaux systèmes de reconnaissance dont la robustesse est liée à la fidélité de la modélisation de la parole plutôt qu'à des améliorations de notre système à base de HMM. Ceci est motivé par le fait que les HMM ne modélisent que la dynamique temporelle du signal de parole alors que sa dynamique fréquentielle, très informative d'un point de vue phonétique, n'est pas prise en compte. Pour avoir une modélisation plus fidèle, il est donc naturel de penser à des modèles qui capturent les caractéristiques à la fois temporelles et fréquentielles de la parole.

Pour ce faire, nous nous inspirons de l'approche multi-bandes, mais au lieu d'utiliser des HMM indépendants pour chacune des bandes spectrales, nous lions les états correspondants aux différentes bandes. De cette façon, nous obtenons un RB plus complexe que les HMM mais qui permet de mieux représenter les aspects temporels et fréquentiels de la parole.

Nous avons en effet développé un RB multi-bandes qui permet de prendre en compte l'asynchronisme et la dépendance entre les bandes de fréquence. Nous avons ensuite adapté un algorithme d'inférence (junction tree algorithm) à notre RB. Puis, nous avons développé un algorithme d'apprentissage pour ce modèle. Cette année, nous avons modifié l'algorithme

de reconnaissance afin de pouvoir traiter de la parole continue. En milieu bruité, nous avons effectué plusieurs expérimentations en ajoutant, aux données de test, différents bruits colorés à bandes limitées. Nous avons comparé notre RB à un HMM, à un modèle multi-bandes classique et à un RB multi-bandes synchrone. Dans tous les cas de figure, les performances de notre modèle surpassent largement celles des autres modèles. Tous ces résultats montrent d'une part que notre modèle capture mieux les caractéristiques dynamiques du signal et d'autre part qu'il est bien adapté pour la reconnaissance de la parole corrompue par un bruit à bande limitée [24].

De plus, nous avons développé un algorithme de modélisation de la parole qui ne fait aucune hypothèse de dépendance *a priori* entre les variables observées et cachées, mais plutôt qui apprend les dépendances à partir des données d'apprentissage (*l'apprentissage structurel*). Il garantit de meilleures performances de reconnaissance que les HMM tout en permettant un contrôle absolu à l'utilisateur sur la complexité souhaité du moteur de reconnaissance [27].

6.2.3 Modèles de langage

L'adaptation des modèles de langage constitue l'une de nos préoccupations majeures. Nous abordons le problème par le biais de l'identification thématique qui a pour objet d'assigner un label thématique à un segment de texte parmi un ensemble prédéfini de labels possibles. Dans le cadre de l'application à la reconnaissance automatique de la parole, le thème est identifié avec les premiers mots dictés puis le modèle de langage approprié intervient dynamiquement dans le processus de reconnaissance. Cette année, nous avons développé différentes méthodes d'identification thématique et avons effectué une étude sur leur complémentarité. Nous avons ainsi proposé un système de vote qui détermine le modèle le plus pertinent devant chaque nouvelle situation, et nous avons implémenté une combinaison linéaire de ces modèles qui permet d'améliorer sensiblement la qualité de l'identification [16]. Une autre application de ces techniques de classification est le routage automatique de courriers électroniques sur lequel nous avons travaillé avec l'entreprise MIC2. Cette tâche consiste à attacher un courrier à un dossier thématique dès son arrivée, ce qui s'apparente au problème de la classification thématique. Cependant, cette tâche est rendue très ardue car les données sont bruitées, un taux important de recouvrement entre les thèmes et un manque récurrent de données pertinentes. Malgré ces difficultés, la méthode développée permet un taux de classement intéressant [15]. Depuis peu, nous nous intéressons à la classification dans une hiérarchie de thèmes, appliquée aux forums de discussion. Dans ce cas, la tâche est rendue très difficile par la mauvaise qualité des données, qui, comme les courriers électroniques, sont entachées de fautes d'orthographe ou de grammaire, et de plus contiennent des fichiers attachés, des images, etc. Les premiers résultats nous permettent d'atteindre un score encourageant d'identification [17].

Un autre volet de notre travail concerne les relations syntaxiques ou sémantiques entre les composantes d'une phrase ou d'un texte qui sont en grande partie distantes. Pour tenir compte de cette propriété, dans un premier temps, nous avions utilisé une combinaison linéaire de modèles n-grammes distants et non distants. Nous avions ainsi obtenu de premiers résultats encourageants. Toujours dans ce cadre, nous avons généralisé le travail des années précédentes sur les modèles n-grammes distants. Nous avons ainsi mis en œuvre des techniques de combinaison linéaire des modèles distants ou non via diverses classifications des historiques ; ceci

a permis d'accroître les performances d'une combinaison linéaire indépendante de l'historique. L'année passée, les travaux sur les modèles distants ont montré l'importance d'une combinaison fondée sur une sélection dynamique du modèle de langage en fonction de l'historique. Cette méthode de sélection [38] a le même potentiel qu'une combinaison linéaire dépendante de l'historique tout en requérant moins de paramètres. Mais surtout, elle a mis en lumière des séquences de mots pouvant être considérées comme des unités lexicales. Ces unités ont été intégrées avec profit dans un modèle de langage simple [37].

Depuis deux ans maintenant, nous introduisons dans les modèdes statistiques de langage une connaissance supplémentaire : la notion d'événements impossibles de la langue, non considérée dans les modèles classiques. Leur prise en compte pourrait permettre de modéliser de manière plus précise les contraintes langagières. La première phase de ce travail consiste à recenser un ensemble d'événements impossibles. Nous avons complété ce recensement cette année en utilisant diverses sources de connaissance supplémentaires. De plus, les travaux de l'année passée ont montré la nécessité d'une méthode de report efficace de la masse de probabilités initialement allouée à ces événements impossibles, vers les événements possibles. Nous avons déterminé que cette méthode doit être adaptée aux données de test afin que la modification globale des distributions de probabilités ait une incidence sur les performances du modèle final [18].

Enfin, l'axe récent que nous avons pris concerne la compréhension. Le but de la compréhension de la parole est de percevoir la signification - le sens - de la phrase et de traduire cette signification dans un langage interprétable par une machine ; ce peut être par exemple, un langage dont chacune des phrases est constituée d'un ensemble de concepts. Le premier travail réalisé pour répondre à cet objectif concerne la collecte de corpus. Puis, nous avons choisi d'utiliser un Perceptron Multi-Couche (PMC) pour faire la transformation d'une phrase en langage de concepts. Les résultats obtenus par notre réseau sont encourageants et ont démontré la fiabilité de l'approche. En terme de «concept», nous avons obtenu un taux d'erreur de 14,2 %, c'est-à-dire qu'on arrive à trouver les bons concepts dans 85,8 % des cas, ce qui montre que nous sommes sur la bonne piste.

6.3 PROCOMA

La société Procoma développe un progiciel permettant de concevoir et réaliser des dessins animés. Nous intervenons dans ce progiciel en fournissant un outil de post-synchronisation en français et en anglais (cf. paragraphe 7.2.2)

6.4 MIC2

Nous continuons notre collaboration avec la société MIC2 et cette année nous avons développé un nouveau moteur de reconnaissance nommé MICLOR qui nous permet de tester de nouveaux algorithmes pour l'adaptation au locuteur ou la robustesse de bruit.

7 Actions régionales, nationales et internationales

7.1 Actions régionales

7.1.1 Action « Assistance à l'apprentissage des langues » (thème Téléopérations et assistants intelligents du Pôle Intelligence Logicielle du Plan État Région)

Depuis le mois d'octobre 2000 nous travaillons avec des enseignants de français langue étrangère et d'anglais dans l'objectif de mieux exploiter les outils d'analyse et de reconnaissance automatique en apprentissage des langues. Outre un séminaire mensuel (les trois derniers ont porté sur la revue des articles récents sur l'apprentissage des langues, la présentation de nos résultats de perception sur l'amélioration de la compréhension orale et la présentation des contrôles ActiveX pour l'analyse de la parole) nous définissons actuellement nos objectifs en termes d'apprentissage de la prosodie et nous avons développé un ensemble de contrôles ActiveX qui peuvent être utilisés pour créer des tutoriels.

7.2 Actions nationales

7.2.1 Projet RNRT IVOMOB

Mise au point d'un moteur de reconnaissance automatique de la parole pour l'interface vocale des télé services accessibles depuis un véhicule automobile en mouvement

Projet RNRT MIC2, Mémodata, Technium et LORIA

Les interfaces des serveurs vocaux interactifs exploitent actuellement la technique DTMF robuste mais d'une trop grande rusticité pour constituer un véritable navigateur. La mise au point d'une véritable interface vocale en langue française est de nature à développer de nombreux services accessibles par le public : commerce électronique associant internet et téléphone, recherche vocale de site web, exécution par commande vocale de calcul de performances d'actions et demande de transmission du résultat par e-mail ou télécopie, etc.. Pour maîtriser un tel « navigateur » vocal, il est indispensable d'associer plusieurs disciplines : Spécialistes du traitement du signal et de la reconnaissance automatique de la parole, linguistes spécialisés dans le traitement automatique de la langue française, ergonomes spécialisés dans les interfaces vocales, spécialistes du télémarketing, etc.

Ce projet est constitué des 6 sous-projets suivants :

- réalisation de module de pré-traitement des signaux vocaux,
- constitution de corpus acoustico-phonétique et textuel,
- conception d'un système hybride permettant de faire coopérer plusieurs techniques de RAP,
- conception d'une méthodologie de développement des applications vocales,
- réalisation d'une librairie d'automates vocaux,
- réalisation d'une maquette validant l'interopérabilité de l'ensemble des composants.

Nous commençons actuellement l'enregistrement du corpus dans une voiture.

7.2.2 Projet PRIAMM SAALSA

Dans le cadre du programme PRIAM (Programme pour la Recherche et l'Innovation dans l'Audiovisuel et le Multimédia), l'objectif du projet SAALSA (Système Auto Adaptatif de Lip Sync Automatique) est de développer de nouveaux outils de production pour l'automatisation de la phase de synchronisation de la voix et des mouvements de bouche en animation 2D/3D en appliquant les résultats des travaux de recherche en traitement automatique de la parole effectués dans les laboratoires INRIA-LORIA et ENST.

Traditionnellement, cette phase de synchronisation est réalisée manuellement par un opérateur expérimenté qui annote le signal sonore en l'écoutant. Cette tâche est longue, rébarbative et coûteuse. Sur la base d'un premier logiciel développé par INRIA-LORIA en collaboration avec la société PROCOMA qui automatise une partie de cette phase de synchronisation en utilisant des algorithmes de reconnaissance automatique de la parole, nous proposons de développer un prototype logiciel mieux adapté à la réalité du marché en constante évolution.

Dans ce cadre, les travaux actuels ont consisté en l'amélioration de la phonétisation (transformation du texte en suite de phonèmes) et l'implantation de méthodes d'adaptation au locuteur pour mieux tenir compte des voix utilisées dans le monde de l'animation qui sont souvent caricaturales ou extrêmes (voix enfantine, chuchotée, criée, très aiguë ou très grave, imitations d'animaux ou d'accents).

7.3 Actions européennes

7.3.1 Projet Européen COST 278

Nous sommes membre du projet Européen COST 278 : "Spoken Language Interaction in Telecommunication" qui regroupe des chercheurs de différents pays européen (Allemagne, Norvège, Hollande, France, Slovaquie, Gréce, Espagne, Slovenie, Suisse et Belgique). Nous avons participé à la réunion inaugurale qui s'est tenue à Bruxelles.

7.3.2 OZONE

Projet IST, OZONE (New technologies and services for emerging nomadic societies) n°IST-2000-30026 avec Philips, Interuniversitair Micro-Electronica Centrum, Laboratoires d'Electronique Philips, EPICTOID, Eindhoven University of Technology, THOMSON multimedia R&D France

Avec plusieurs autres projets de l'INRIA (dont le projet Langue et Dialogue et l'action MAIA à Nancy) nous participons à ce projet où notre rôle concerne le développement d'une interface multimodale généraliste destinée à utiliser des services nomades. L'objectif général du projet OZONE est de développer un cadre logiciel qui puisse s'adapter très souplement aux différents types de matériels et de situations rencontrés dans le cadre de l'informatique nomade.

L'interface multimodale reposera sur la parole et le geste et devra être capable de modéliser la situation dans laquelle se trouve l'utilisateur et les services nomades auxquels l'utilisateur a accès. Trois applications sont envisagées pour évaluer cette architecture :

- la réservation d'un cybercar (véhicule autonome intelligent) pour un utilisateur souhaitant se déplacer dans une ville,
- la recherche et la sélection de documents multimédia récréatifs sans aucune contrainte ni sur le format des documents, ni sur leur mode de stockage,
- la domotique nomade qui offre à l'utilisateur la possibilité d'interagir à distance avec sa maison.

7.4 Visites, et invitations de chercheurs

- Louis-Jean Boë, Institut de la Communication Parlée à Grenoble, séminaire intitulé « Croissance du conduit vocal et stratégies articulatoires en production vocalique »,
- Florian Gallwitz, Université d'Erlangen-Nürnberg, intitulé « Integrated Recognition of Words and Phrase Boundaries »,
- Gérard Chollet, École Nationale Supérieure des Télécommunications, Paris, séminaire intitulé « Codage de la parole à très bas débit »,
- Marc Sigelle, École Nationale Supérieure des Télécommunications, Paris, séminaire intitulé « Modélisation par champ de Markov du signal de parole et application à la reconnaissance vocale ».
- Philippe Martin, Professeur à l'université de Toronto, a fait un seminaire sur la modelisation de l'intonation.

8 Diffusion de résultats

8.1 Animation de la Communauté scientifique

Relectures pour les journaux IEEE Transactions on speech and audio processing, IEEE Transaction in Information Theory, Speech communication, Langue, Journal of Phonetics, JASA.

Co-responsabilité du thème Télé-opérations et assistants intelligents dans le cadre du pôle Intelligence logicielle du Plan État Région (Yves Laprie).

Co-responsabilité de l'action « Assistance à l'apprentissage des langues » dans le cadre du thème Télé-opérations et assistants intelligents du Plan État Région (Anne Bonneau).

Membre élu du bureau du G.F.C.P, groupe francophone de la communication parlée, (Yves Laprie).

Membre de IASTED Technical Committee on Pattern Recognition (K. Daoudi).

Membre du comité de programmes du European Symposium of Young Researchers in Artificial Intelligence (K. Daoudi).

8.2 Enseignement universitaire

- Forte participation à divers enseignements dans les établissements lorrains (Université de Nancy 1 et II, INPL) : Maîtrise et DEA d'Informatique, IUT, MIAGE, DESS Informatique, DESS Information Scientifique et Technique, DEA de Chimie Informatique et Théorique ;

- Responsabilité du DESS IST de l'UHP (M. C. Haton) ;
- Responsabilité du DEA d'Informatique de Nancy (J.-P. Haton) ;
- Responsabilité du DESS Informatique de l'UHP (O. Mella) ;

8.3 Participation à des colloques, séminaires, invitations

- Séjour post-doctoral de Imed Zitouni aux Bell Laboratories de Lucent Technologies
- Participation à des jurys de thèses de doctorat D. Fohr, J.-P. Haton, M.-C. Haton, Y. Laprie, K. Smaïli ;
- Exposé de Michel Pitermann sur les têtes parlantes biomécaniques lors du séminaire de l'INRIA au National Science Council de Taiwan en mars,
- Exposé de Yves Laprie sur l'analyse de la parole au laboratoire DDL à Lyon en décembre.
- Exposé de Khalid Daoudi sur les modèles probabilistes graphiques pour la modélisation statistique de la parole lors des journées « statistiques » à l'IRISA en novembre.
- Formation pour les ingénieurs de la société MIC2 sur la conception d'un système de reconnaissance automatique de la parole faite par K. Daoudi, D. Fohr, J.-P. Haton et K. Smaili.
- On se reportera à la bibliographie pour la liste des conférences et *workshops* auxquels les membres de l'action ont participé.

9 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] A. Bonneau, F. Charpillet, S. Coste-Marquis, J.-P. Haton, Y. Laprie, P. Marquis, « Towards a Multilevel Model for Hyptothetical Reasoning in Continuous Speech Recognition », *in : Levels in Speech Communication : Relations and Interactions*, C. Sorin, J. Mariani, et H. Méloni (éditeurs), Elsevier, 1994.

[2] A. Bonneau, L. Djezzar, Y. Laprie, « Perception of the Place of Articulation of French Stop Bursts », *Journal of the Acoustical Society of America 100*, 1, 1996, p. 555–564.

[3] J. Di Martino, « Dynamic time warping algorithms for isolated and connected word recognition », *in : New Systems and Architectures for Automatic Speech Recognition and Synthesis, Nato Asi Series, vol F16*, R. de Mori et C. Y. Suen (éditeurs), Springer-Verlag, Berlin, 1984.

[4] D. Fohr, J.-P. Haton, Y. Laprie, « Knowledge-based Techniques in Acoustic-Phonetic Decoding of Speech : Interest and Limitations », *International Journal of Pattern Recognition and Artificial Intelligence 8*, 1, 1994, p. 133–153.

[5] M.-C. Haton, « Issues in Using Models for Self Evaluation and Correction of Speech », *in : Computational Models of Speech Pattern Processing*, M. Ponting, K. (éditeur), *Computer and Systems Sciences*, Springer-Verlag, Berlin, 1998.

[6] i. Illina, M. Afify, Y. Gong, « Environment Normalization Training and Environment Adaptation Using Mixture Stochastic Trajectory Model », *Speech Communication 24*, 1998.

[7] J.-C. Junqua, J.-P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer Academic, 1996.

[8] Y. Laprie, M.-O. Berger, « Cooperation of Regularization and Speech Heuristics to Control Automatic Formant Tracking », *Speech Communication 19*, 4, octobre 1996, p. 23.

[9] O. Mella, D. Fohr, « TwoTools for Semi-automatic Phonetic Labelling of Large Corpora », *in : First International Conference on Language Resources and Evaluation, Grenade, Espagne*, mai 1998, `http://www.loria.fr/publications/1998/98-R-028/98-R-028.ps`.

[10] K. Smaïli, I. Zitouni, F. Charpillet, J. P. Haton, « An Hybrid Language Model For a Continuous Dictation Prototype », *in : 5th European Conference On Speech Communication And Technology - Eurospeech'97, Rhodes, Greece*, p. 2723–2726, sep 1997.

Articles et chapitres de livre

[11] F. Bimbot, M. El Beze, S. Igounet, M. Jardino, K. Smaili, I. Zitouni, « An alternative scheme for perplexity estimation and its assessment for the evaluation of language models », *Computer Speech and Language*, 2000, à paraître.

[12] A. Bonneau, « Identification of vocalic features from French stop bursts », *Journal of Phonetics*, 2001, à paraître.

[13] C. Cerisara, D. Fohr, « Multi-band automatic speech recognition », *Computer Speech and Language 15*, 2, avril 2001, p. 151–174.

[14] M. Grice, M. D'Imperio, M. Savino, C. Avesani, « Towards a Strategy for ToBI labelling varieties of Italian », *in : Prosodic Typology and Transcription : A Unified Approach*, S.-A. Jun (éditeur), Oxford University Press, Oxford, England, 2001.

Communications à des congrès, colloques, etc.

[15] B. Bigi, A. Brun, J.-P. Haton, K. Smaïli, I. Zitouni, « A comparative study of Topic Identification on Newspaper and E-mail », *in : Proceedings of the 8th International Symposium on String Processing and Information Retrieval (SPIRE'01), Laguna de San Rafael, Chili*, p. 238–241, novembre 2001.

[16] B. Bigi, A. Brun, J.-P. Haton, K. Smaïli, I. Zitouni, « Dynamic Topic Identification : Towards Combination of Methods », *in : Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria*, Galia Angelova, Kalima Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov, p. 255–257, septembre 2001.

[17] B. Bigi, A. Brun, K. Smaïli, J.-P. Haton, « A Hierarchical Approach for Topic Identification », *in : Proceedings of the international workshop Speech and Computer (SPECOM'01), Moscow, Russia*, novembre 2001.

[18] A. Brun, D. Langlois, K. Smaïli, J.-P. Haton, « Improving Statistical Language Models by Removing Impossible Events », *in : Proceedings of the International Workshop "Speech and Computer" (SPECOM2001), Moscow, Russia*, 2001.

[19] C. Cerisara, K. Daoudi, « Modeling dependency between regression classes in MLLR using multiscale autoregressive models », *in : Adaptation methods for speech recognition*, août 2001.

[20] C. Cerisara, D. Fohr, I. Illina, F. Lauri, O. Mella, « A comparison of different methods for noise adaptation in a HMM-based speech recognition system », *in : International Congress on Acoustics, Italy, Rome*, septembre 2001.

[21] C. Cerisara, L. Rigazio, R. Boman, J.-C. Junqua, « Environmental adaptation based on first order approximation », *in : International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2001, Salt lake City, USA*, mai 2001.

[22] V. Colotte, Y. Laprie, A. Bonneau, « Perceptual experiments on enhanced and slowed down speech sentences for second language acquisition », *in : European Conference on Speech Communication and Technology, Aalborg, Denmark*, septembre 2001.

[23] V. Colotte, Y. Laprie, A. Bonneau, « Signal transformation strategies to improve speech intelligibility for second language acquisition », *in : 17th International Congress on Acoustics, Rome, Italy*, septembre 2001.

[24] K. Daoudi, D. Fohr, C. Antoine, « A Bayesian network for time-frequency speech modeling and recognition », *in : International Conference on Artificial Intelligence and Soft Computing, Cancun, Mexico*, mai 2001.

[25] K. Daoudi, D. Fohr, C. Antoine, « Continuous Multi-Band Speech Recognition using Bayesian Networks », *in : IEEE ASRU, Terento, Italy*, IEEE, décembre 2001.

[26] A. de la Torre, D. Fohr, J.-P. Haton, « On the comparison of front-ends for robust speech recognition in car environments », *in : ISCA ITR Workshop : adaptation methods for speech recognition, Sophia-Antipolis France*, ISCA (éditeur), p. 105–108, août 2001.

[27] M. Deviren, K. Daoudi, « Structural Learning of Dynamic Bayesian Networks in Speech Recognition », *in : EUROSPEECH, Alborg, Denemark*, septembre 2001.

[28] J. Di Martino, Y. Laprie, « Suppression of Phasiness for Time-Scale Modifications of Speech Signals Based on a Shape Invariance Property », *in : International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2001, Salt Lake City, USA*, IEEE, mai 2001.

[29] M. D'Imperio, « Language-specific knowledge and syllable structure effects in the perception of tonal contrast », *in : Invited Talk, Max Plank Institut for Psycholinguistics*, 2001.

[30] M. D'Imperio, « Language-specific knowledge and the perception of tonal contrasts in Italian and English », *in : 141st Meeting of the Acoustical Society of America, Chicago, Illinois, USA*, *Journal of the Acoustical Society of America*, *109*, 5, p. p. 2475, juin 2001.

[31] M. D'Imperio, « The Role of Perception in Defining Tonal Targets and their Alignment », *in : Invited Talk, Department of Linguistics, University of Saarbruecken, Germany*, Martine Grice, avril 2001.

[32] M. D'Imperio, « Tonal alignment, scaling and slope in Italian question and statement tunes », *in : EUROSPEECH '01, Aalborg, Denmark*, septembre 2001.

[33] M.-C. Haton, J.-P. Haton, « Word recognition for all : application to speech training », *in : Universal Access in Human-Computer Interaction, New-Orleans, Louisiana, USA*, C. Stephanidis (éditeur), *3*, Lawrence Erlbaum Associates, Publishers, p. 329–333, 10, Industrial Avenue, Mahwah, New Jersey 07430, USA, août 2001.

[34] I. Illina, D. Mostefa, « Structural Maximum a Posteriori Adaptation for Mixture Stochastic Trajectory Framework », *in : WorkShop International on Adaptation Methods for Automatic Speech Recognition, Sophia Antipolis, France*, *1*, 1, Eurecom, août 2001.

[35] S. Jamoussi, F. Alexandre, « Implantation d'une carte associative pour l'orientation d'un robot autonome à l'aide d'une image vidéo », *in : Traitement et Analyse d'Images Méthodes et Applications - TAIMA'01, Hammamet, Tunisie*, juin 2001.

[36] s. Jamoussi, K. Smaïli, J.-P. Haton, « Contribution à la compréhension de la parole par des réseaux neuronaux », *in : Quatrièmes Rencontres des Jeunes Chercheurs en Parole - RJC 2001, Mons, Belgique*, septembre 2001.

[37] D. Langlois, K. Smaili, J.-P. Haton, « Efficient Language Models Combination : Application to Phrase Finding », *in : Proceedings of the International Workshop "Speech and Computer" (SPECOM2001), Moscow, Russia*, 2001.

[38] D. Langlois, K. Smaïli, J.-P. Haton, « A New Method Based on Context for Combining Statistical Language Models », *in : Third International Conference on Modeling and Using Context - CONTEXT 01, Dundee, Scotland*, Varol Akman, Paolo Bouquet, Richmond Thomason, Roger A. Young (éditeur), *Lecture Notes in Artificial Intelligence*, *2116*, Springer, p. 235–247, juillet 2001.

[39] Y. LAPRIE, A. BONNEAU, « Burst segmentation and evaluation of acoustic cues », *in : Eurospeech'01, Aalborg, Danemark*, *1*, septembre 2001.

[40] S. OUNI, Y. LAPRIE, « Exploring the Null Space of the Acoustic-to-Articulatory Inversion Using a Hypercube Codebook », *in : Eurospeech, Aalborg, Danemark*, *1*, p. 277–280, septembre 2001.

[41] S. OUNI, Y. LAPRIE, « Studying articulatory effects through hypercube sampling of the articulatory space », *in : 17th International Congress on Acoustics , Rome, Italy*, *4*, septembre 2001.

[42] M. PITERMANN, K. G. MUNHALL, « A face-to-muscle inversion of a biomechanical face model for audiovisual and motor control research », *in : Eurospeech'01, Aalborg, Denmark*, ISCA, septembre 2001.