UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1
SCIENCES & GÉOGRAPHIE

HABILITATION À DIRIGER LES RECHERCHES

Discipline : Informatique

présentée et soutenue publiquement par

Laurent BESACIER

Le 11 Janvier 2006

Titre :

Transcription enrichie de documents dans un monde multilingue et multimodal

---

JURY

M. Christian Boitet,	**Professeur, Université J. Fourier, Grenoble**	**Président**
M. Jean-Paul Haton,	**Professeur, Institut Universitaire de France**	**Rapporteur**
M. Giuseppe Riccardi,	**Professeur, Université de Trente (Italie)**	**Rapporteur**
M. Liming Chen	**Professeur, Ecole Centrale, Lyon**	**Rapporteur**
M. Jean Caelen,	**Directeur de Recherche, CNRS (lab. CLIPS)**	**Examinateur**
M. Jean-Luc Schwartz	**Directeur de Recherche, CNRS (lab. ICP)**	**Examinateur**

Travaux préparés au sein du laboratoire de Communication Langagière et Interaction Personne-Système,
Fédération IMAG – Université Joseph Fourier – Grenoble I

Avant-propos

Ce mémoire a été rédigé sur une période allant de juillet 2005 à août 2006. Les chapitres 1, 2 et 4 ont été rédigés pendant l'été 2005 tandis que le chapitre 3, qui contient entre autres un court résumé de mes activités à IBM Watson, a été écrit pendant mon séjour aux USA (octobre 2005 – octobre 2006). Pour cette raison, la bibliographie associée aux deux premiers chapitres s'arrête en 2005, tandis que le chapitre suivant qui traite de l'aspect *multilinguisme* est peut-être plus à jour. Pour les mêmes raisons, les chapitres 2 et 4 apparaîtront plus comme une revue générale de mes activités de recherche depuis 1998 (année de ma soutenance de thèse) tandis que le chapitre 3 concerne plus directement les thèmes de recherche que j'aborde aujourd'hui en priorité. Je prie donc le lecteur de m'excuser pour l'éventuel inconfort de lecture provoqué par cette rédaction morcelée…

Table des matières

Chapitre 1 : Introduction

1.1 Contexte scientifique

Les technologies existantes permettent depuis longtemps de numériser l'information avec beaucoup d'avantages : transfert rapide, reproduction illimitée, stockage facile. Pratiquement toute information multimédia se trouve aujourd'hui sous format numérique. Avec la facilité de stockage due en partie à des algorithmes de compression très efficaces, les corpus de documents audio et vidéo ne cessent de croître. La recherche de documents multimédia basée sur le contenu est par exemple devenue une tâche très importante.
Dès l'apparition du numérique dans la représentation de données multimédia, l'effort principal a été consacré à la mise en place d'algorithmes pour la transmission et le stockage, sans envisager la possibilité de manipuler l'information selon le contenu.

1.1.1 Concept de transcription enrichie de documents oraux

Pour pouvoir envisager un accès selon le contenu dans de grandes bases de données multimédia, ces données doivent être annotées. Dans mes activités de recherche, je m'intéresse plus particulièrement aux documents audio ou au canal audio d'un document vidéo. Pour ces signaux, une simple transcription manuelle ou automatique issue d'un moteur de reconnaissance de la parole n'est pas suffisante. Le signal audio, et en particulier le signal de parole, est très *riche* en informations. Il contient notamment des informations de haut niveau que nous pouvons appeler méta-données. Des exemples de telles informations sont : les hésitations, les frontières de phrases, les locuteurs [Reynolds, 2005], etc. Un moteur de reconnaissance de la parole qui fournit une transcription simple (ce qui est dit) peut être complété par un moteur d'extraction de méta-données (Figure 1) pour fournir comme résultat final une *Transcription Enrichie*. Cela a pour but notamment d'améliorer la lisibilité des sorties de transcription pour un humain.

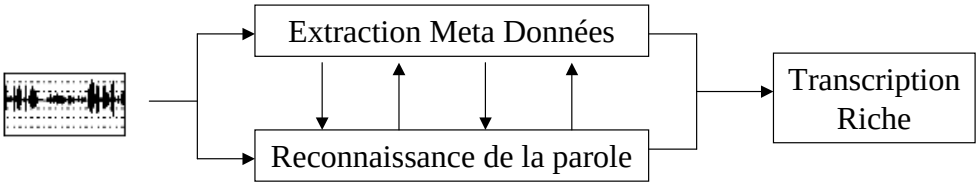


Figure 1 : Architecture d'un moteur de transcription enrichie

1.1.2 Evolution du domaine

Ce concept de transcription enrichie est fortement lié à l'évolution du domaine de la reconnaissance automatique de la parole. Avant les années 90, la recherche dans le domaine s'intéressait principalement aux systèmes de dictée vocale et à la transcription d'enregistrements téléphoniques. Vers la fin des années 90, l'intérêt de la recherche s'est porté vers des données plus riches en information et de moins en moins contrôlées comme les journaux télévisés et radiodiffusés, et les documents issus de réunions enregistrées dans des "environnements perceptifs" équipés de nombreux capteurs. En parallèle, les systèmes de transcription ont évolué depuis des tâches de reconnaissance à vocabulaires limités vers des tâches à très grands vocabulaires dans un contexte de dialogue interactif [Ward-Church 2003]. Cette évolution est illustrée par la figure 2 qui représente

15 ans d'amélioration des performances pour différentes tâches de reconnaissance, dans le contexte des campagnes d'évaluation organisées par l'organisme américain NIST[1] [Pallet 2003].

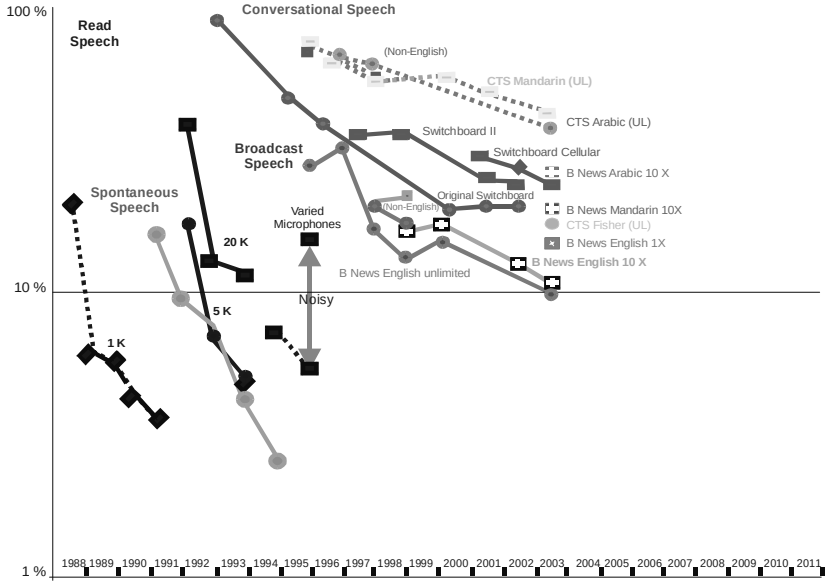
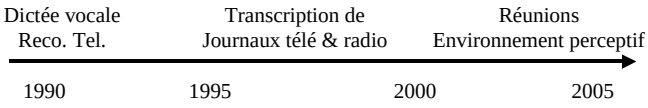


Figure 2 : Evolution des performances (taux d'erreurs) de reconnaissance de la parole (source NIST [Pallet 2003])

En 2003, NIST a mis en œuvre un nouveau programme d'évaluation sur la transcription de documents oraux nommé *EARS* (*Effective, Affordable, Reusable, Speech-to-Text*). Le but de ce programme est d'évaluer les outils de transcription enrichie (*RT : Rich Transcription*). Les efforts du projet EARS sont focalisés sur la transcription de différents types de données présentant chacune des problématiques spécifiques : journaux télévisés (*BN : Broadcast News*), données téléphoniques (*CTS : Conversational Telephone Speech*), et données issues de réunions (*Meetings*).

Ces récentes évolutions amènent de nouvelles difficultés dans le traitement automatique de la parole illustrées par la Figure 3. Par exemple, dans les systèmes de dictée, le locuteur contrôlait lui même le début et la fin de l'enregistrement. Dans les journaux télévisés, le flux audio est continu et hétérogène et c'est le système qui doit segmenter le signal de façon automatique. Un autre nouveau problème est la nécessité de traiter, dans le cas d'enregistrements de réunions par exemple, du signal provenant de capteurs différents (plusieurs microphones, plusieurs cameras vidéo, détecteurs de mouvements, etc.).



Transcription simple	⇨	Transcription riche
Flux audio contrôlé	⇨	Flux audio continu
Un capteur	⇨	Plusieurs capteurs
Monolingue	⇨	Multilingue
Uni-modal	⇨	Multi-modal
(Audio seul)		(Audio, Vidéo, Texte, etc.)

Figure 3 : Ma vision sur l'évolution du domaine en traitement automatique de la parole

[1] http://www.nist.gov/speech/tests

La difficulté des nouvelles tâches apparues est due aussi à la grande quantité de données à traiter. A titre d'exemple, la *Smart Room* NIST [Stanford 2003] qui permet d'enregistrer des réunions génère environ 1 Go de données par minute, issues de multiples capteurs. La nécessité d'indexer une telle quantité de données est alors évidente (cf vision de Kenneth Ward Church[2]).
En résumé, on peut dire que deux des principaux enjeux pour les systèmes futurs sont le *passage à l'échelle* d'une part, et la *portabilité* d'autre part (portabilité vers de nouvelles langues, des nouveaux domaines, etc…).
Plus récemment encore, sont apparues de nouvelles tâches, liées au stockage et à la navigation dans des *bases de données personnelles* [Bell 2004]. En effet, l'enregistrement en continu et la conservation systématique de données multimédia personnelles devient désormais possible avec l'augmentation des capacités de stockage et l'offre actuelle de matériel d'acquisition (appareils photos numériques, enregistreurs, caméras). Le besoin d'outils pour annoter et rechercher des informations dans ces bases de données personnelles devient alors important[3]. Dans ce cas, les données audio personnelles peuvent être collectées dans de multiples situations de la vie courante, dans des environnements très variables, et sont souvent couplées avec d'autres informations, telles que des informations de localisation.

1.1.3 Axes de recherche abordés

Mes activités de recherche au cours des 6 dernières années passées au CLIPS (et de l'année passée comme chercheur invité au centre de recherche *IBM Watson*), se placent dans cette évolution, puisque j'ai essayé d'aborder quelques-uns des points mentionnés dans la figure 3 :
-le *multilinguisme*, autour duquel subsistent un certain nombre de verrous, notamment en ce qui concerne la généricité des systèmes de reconnaissance automatique de la parole, et leur portabilité vers de nouvelles langues ;
-la *multimodalité*, puisque désormais le canal audio n'est plus seul mais le plus souvent accompagné d'autres informations (vidéo et/ou texte) ; ici, les problèmes scientifiques résident dans le traitement conjoint de multiples modalités qui peuvent être asynchrones et dans le choix des outils mathématiques permettant ce traitement.

J'ai abordé les deux aspects du concept de transcription enrichie de la façon suivante :
-comme une suite naturelle de mes travaux de thèse sur la reconnaissance de locuteurs [Besacier, 1998]. J'ai d'une part abordé l'extraction de méta-données (enrichissement) et plus précisément l'extraction *d'éléments non linguistiques* tels que l'identité des locuteurs, les changements de tours de parole et les sons clés, à partir d'un flux de parole ; j'y ai ajouté plus récemment une vision multimodale (synthèse des travaux décrite au *chapitre 2*),
-en fonction du contexte local au CLIPS et des thèmes de recherche qui y étaient développés. J'ai ainsi travaillé dans le domaine de la transcription automatique (reconnaissance automatique de la parole), en insistant notamment sur les aspects multilingues (synthèse des travaux décrite au *chapitre 3*).
Enfin, un cadre applicatif naturel lié à ces problèmes nouveaux, et cohérent avec les problématiques de CLIPS, m'a permis d'évaluer l'apport de mes recherches pour des applications telles que les espaces perceptifs et la recherche d'information multimédia (synthèse des travaux décrite au *chapitre 4*).

[2] *« Search will become more important than coding and dictation », [Kenneth Ward Church, 2003]*.
[3] Voir keynote de G. Bell, Microsoft, lors de la conférence ACM Multimedia 2004 « A New Relevance for Multimedia When We Record Everything Personal », ainsi que le premier workshop « The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences ».

1.2 Plaidoyer

1.2.1 Un monde multilingue

Informatique multilingue

Comme le dit V. Berment dans sa thèse [Berment, 2004], « *le développement des ordinateurs personnels et celui des réseaux font aujourd'hui de l'informatique un instrument pour écrire et communiquer au même titre que le papier et l'imprimerie l'étaient auparavant. Traitements de texte, courriers électroniques, voire des systèmes plus avancés comme la dictée ou la synthèse vocale sont des outils largement répandus. L'idée s'impose alors qu'aux moyens traditionnels doivent s'ajouter les outils informatiques appropriés sans lesquels les buts visés ne peuvent plus être atteints* ». L'informatisation d'une langue occupe ainsi une place essentielle dans ce vaste contexte. Cependant, parmi les 6000 langues parlées dans le monde, seul un tout petit nombre atteint un « niveau d'informatisation » satisfaisant. Pour évaluer de manière quantitative le degré d'informatisation d'une langue, [Berment, 2004] propose dans sa thèse le protocole suivant : à chaque service ou ressource, un groupe d'utilisateurs représentatifs des locuteurs de la langue attribue un niveau de criticité Ck et une note Nk, la moyenne pondérée des notes — appelée indice σ — reflétant leur satisfaction globale. Une langue mal ou peu dotée peut ainsi être définie comme une langue dont l'indice σ n'atteint pas 10/20 et qui est encore insuffisante aux yeux de ses évaluateurs. A titre d'exemple, on peut présenter dans le tableau ci-dessous une évaluation du niveau d'informatisation obtenu ainsi pour la langue khmère, parlée au Cambodge.

	Services / ressources	Criticité (/10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	14	140
	Recherche et remplacement	8	12	48
	Sélection du texte	6	12	72
	Tri lexicographique	5	0	0
	Correction orthographique	2	0	0
	Correction grammaticale	0	0	0
	Correction stylistique	0	0	0
Traitement de l'oral				
	Synthèse vocale	5	0	0
	Reconnaissance de la parole	5	0	0
Traduction				
	Traduction automatisée	8	4	32
ROC				
	Reconnaissance optique de caractères	9	0	0
Ressources				
	Dictionnaire bilingue	10	4	40
	Dictionnaire d'usage	10	0	0
Total				540 / 1760
Moyenne				6,2 / 20

Tableau 1 : Tableau d'évaluation du niveau d'informatisation pour le khmer

Comme nous pouvons le voir sur ce tableau, les services liés au traitement de l'oral sont inexistants pour la langue khmère (synthèse vocale et reconnaissance de la parole). C'est aussi le cas pour une majorité de langues dans le monde, dont certaines sont parlées par plusieurs dizaines de millions de locuteurs (par exemple[4] Bengali : 189 millions, Tamoul : 63 millions), y compris au sein de l'Europe des 25 (lituanien, letton, polonais)!

Multilinguisme dans les échanges et les services

Le multilinguisme est au cœur des enjeux actuels concernant les échanges culturels et économiques qui sont désormais mondialisés. Ainsi, les individus sont de plus en plus amenés à évoluer dans des

[4] source : http://www.populationdata.net

environnements multilingues, comme le montrent certaines tendances récentes du monde et de la société :

-importance croissante d'organisations internationales ou transnationales (organisations multilatérales, Union Européenne, sociétés multinationales, etc…),
-augmentation des échanges culturels et des voyages,
-regain d'intérêt pour les langues régionales qui cohabitent désormais avec les langues nationales.

Ainsi, le développement de services et d'interfaces adaptés à ce contexte peut donner lieu à de nouvelles problématiques dans le domaine du traitement automatique du langage naturel. En ce qui concerne la communication homme / homme médiatisée par la machine, les recherches en traduction automatique de parole sont évidemment centrales [Waibel, 2004] ; pour illustrer cela, on peut notamment citer les projets CSTAR[5] et NESPOLE[6] de traduction automatique dans lesquels le laboratoire CLIPS s'est impliqué. Concernant la transcription automatique, on a vu récemment émerger le thème de *reconnaissance automatique de la parole multilingue*, qui fait désormais l'objet de sessions spéciales dans des conférences telles que ICSLP ou ICASSP. En dehors des travaux pionniers de CMU sur ce thème (voir notamment [Schultz, 2004] [Waibel, 2004] ainsi que les projets GlobalPhone[7] et DARPA Babylon[8]), on peut citer les récentes études suivantes : transcription d'un flux audio bilingue (journaux télévisés avec présentateur galicien et reportages en espagnol standard [Dieguez-Tirado, 2005]), traitement du langage naturel et ressources pour les langues minoritaires (voir action SALTMIL[9] de l'ISCA et le concept d'*E-Inclusion*[10]), travaux sur des langues asiatiques mal dotées (thai [Suebvisai, 2005], indonésien [Martin, 2005]), applications dans des environnements bilingues (travaux d'IBM sur la reconnaissance de noms propres anglais sur un flux audio en français canadien [Lejeune, 2005], reconnaissance bilingue de noms propres anglais / chinois [Ren, 2005]).

Parallèlement à cette évolution, et cela reste pour moi partie intégrante de la thématique scientifique du *multilinguisme*, la domination de certaines langues majoritaires comme l'anglais fait que celles-ci restent très largement utilisées par des locuteurs non natifs dans différents services et contextes. Un autre problème scientifique est donc de développer des systèmes de transcription automatique capables de reconnaître la parole prononcée par des *locuteurs non natifs* : reconnaissance d'*anglais européen* dans le projet CHIL[11], adaptation de modèles acoustiques en anglais à des locuteurs d'origine allemande [Wang, 2003b].

En résumé, il s'agit d'une part d'être capable de traiter rapidement de nouvelles langues, de développer des systèmes capables de commuter d'une langue à l'autre (ce qui peut nécessiter ou pas l'utilisation d'une phase d'identification des langues [Fugen, 2003]), mais aussi d'adapter des systèmes existants pour une langue donnée à des locuteurs non natifs. Tout cela participe pour moi d'un même domaine de recherche que je place sous le terme générique de *reconnaissance automatique de la parole multilingue*.

[5] http://www.c-star.org/
[6] http://nespole.itc.it/
[7] http://www-2.cs.cmu.edu/~tanja/GlobalPhone/index.html
[8] http://darpa-babylon.mitre.org/
[9] Speech And Language Technology for MInority Languages : http://193.2.100.60/SALTMIL/
[10] voir session special Eurospeech 2005 : *"E-Inclusion for Spoken Language Processing"*
[11] http://chil.server.de/servlet/is/101/

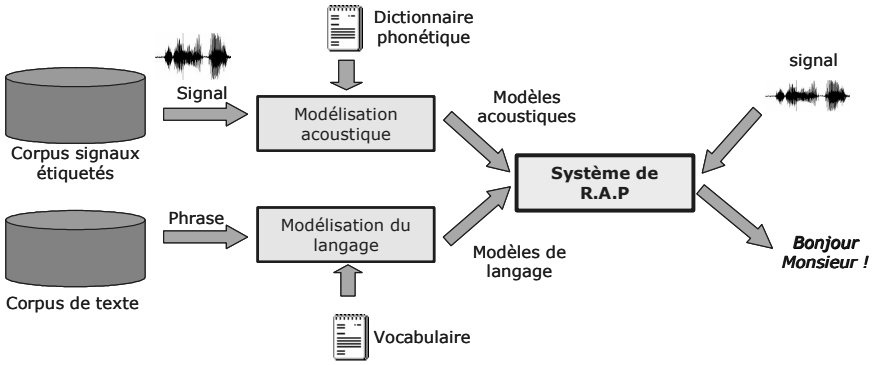


Figure 4 : Schéma d'ensemble de la reconnaissance automatique de la parole

Comme l'illustre la figure 4 ci-dessus, pour développer un système de reconnaissance automatique de la parole continue dans une nouvelle langue, il est souvent nécessaire de rassembler une grande quantité de corpus, contenant à la fois des signaux de parole (pour l'apprentissage des modèles acoustiques du système) et des données textuelles (pour l'apprentissage des modèles de langage du système). De tels corpus et systèmes sont désormais disponibles pour la plupart des langues occidentales (anglais, français, espagnol) et pour quelques langues asiatiques (chinois, japonais). Porter un système de reconnaissance vers une nouvelle langue est cependant une tâche très fastidieuse si aucun corpus de grande envergure n'existe dans la langue cible, puisqu'il faut alors collecter soi-même les ressources nécessaires : signal de parole, lexique, corpus textuels, etc. Précisons aussi qu'étant donné la nature statistique des modèles généralement utilisés en reconnaissance automatique de la parole (modèles acoustiques de phonèmes correspondant à des chaînes de Markov où chaque état est une distribution multigaussienne ; et modèles de langage N-grammes), ces ressources doivent être disponibles en quantité importante.

Les activités de recherche que j'ai abordées et que je décris dans ce document (chapitre 3) sont donc en phase avec l'évolution du domaine de la reconnaissance automatique de la parole de ces vingt dernières années où on est de plus en plus amené à traiter des documents de nature *multilingue*. Dans ce domaine qui, comme nous l'avons vu, commence à être abordé au niveau international, mais encore assez peu au niveau national, il subsiste un certain nombre de verrous, notamment en ce qui concerne la généricité des systèmes de reconnaissance automatique de la parole, et leur portabilité vers de nouvelles langues.
L'originalité de mon approche, par rapport à cet existant, vient de la volonté d'aborder des langues peu dotées, pour lesquelles peu ou pas de corpus sont disponibles, ce qui nécessite des méthodologies innovantes qui vont bien au-delà du simple réapprentissage ou de l'adaptation de modèles. Il s'agit notamment de réduire le fossé actuel qui existe entre les experts techniques et les linguistes, à travers la mise au point d'outils simples et efficaces de collecte de ressources par exemple. Ce problème spécifique (multilinguisme et langues peu dotées) est notamment mis en avant dans un tutoriel[12] d'IBM sur les futurs défis en reconnaissance automatique de la parole, dans lequel les premiers travaux du CLIPS sur le vietnamien (qui seront détaillés au chapitre 3) sont d'ailleurs cités.

1.2.2 Un monde multimodal

Le concept de multimodalité peut prendre différentes formes. En ce qui me concerne, je le présenterai dans ce plaidoyer d'une part du point de vue des *environnements multimodaux* qui

[12] « Portability challenges in developing interactive dialogue systems », Y. Gao, L. Giu, H.-K. Kuo, ICASSP 2005. Philadelphie, USA.

impliquent le traitement d'une information issue de différents flux (capteurs multiples, documents vidéos, interfaces multimodales), et d'autre part du point de vue de la *parole multimodale*, celle-ci étant multisensorielle par essence [Schwartz 2004].

Environnements multimodaux

Les interfaces homme / machine et les systèmes de communication homme / homme médiatisée par la machine sont désormais multimodaux. La modalité *parole* en est un élément central [Flanagan, 2004] et peut se combiner avec d'autres modalités visuelles ou tactiles. Dans d'autres domaines de recherche, la multiplication des canaux d'information, chacun doté d'une pertinence propre, peut permettre d'améliorer la robustesse et l'efficacité des systèmes. C'est par exemple le cas en biométrie, où il s'agit d'identifier un individu à partir de ses caractéristiques physiques ou comportementales. Aucun dispositif d'authentification biométrique monomodal ne satisfait parfaitement l'ensemble des critères nécessaires dans ce domaine (fiabilité, simplicité, accessibilité, complexité,…). Une solution intuitive, qui préfigure les futurs systèmes, consiste alors à rassembler et combiner plusieurs modalités[13] [Jain, 2004].
C'est aussi le cas dans le domaine de la recherche d'information multimodale où les évaluations, TREC Vidéo[14] par exemple, nécessitent des compétences multidisciplinaires pour traiter différents types d'information : audio (transcription automatique de la bande son, détection de catégories non linguistiques), images animées (classification par couleurs, textures, mouvements, ou reconnaissance de formes), texte (extraction des sous-titres ou reconnaissance automatique de textes à partir de la bande image). De nombreuses approches multimodales ont été récemment appliquées dans le cadre expérimental TREC Vidéo sur des tâches de recherche [Christel, 2004], de segmentation de vidéos [Perez-Freire, 2004] [Hsu, 2004] [Ohtsuki, 2003] et d'extraction de plans caractéristiques [Chaisorn, 2003].

La parole multimodale

Redisons-le, la parole est aussi multisensorielle par essence [Schwartz, 2004]. Tout d'abord, une partie du conduit vocal est visible : il s'agit bien sûr des lèvres. Il est désormais acquis que la lecture labiale fait partie intégrante du processus de perception, surtout pour les mal-entendants, mais aussi pour les bien-entendants (voir effet Mac Gurk [Mac Gurk, 1976], ainsi que les études sur la perception en milieu bruité avec ou sans l'information visuelle [Sumby, 1954]). L'effet de la multisensorialité dans le développement du langage est également important. Les modalités qui complètent la parole peuvent s'avérer par ailleurs très importantes lorsque le son est mal perçu : langage parlé complété pour les malentendants [Burger, 2004], débruitage audiovisuel [Girin, 2001] ou reconnaissance automatique de la parole audiovisuelle [Potamianos, 2004].

Les problèmes scientifiques

Ce *monde multimodal* conduit à de nouveaux problèmes scientifiques. En premier lieu, se pose la question du choix d'un mécanisme de fusion permettant de traiter un flux multimodal. Pour cela, on retrouve une large gamme de techniques allant de l'*intégration précoce* des modalités (proche de la fusion de données) jusqu'à l'*intégration tardive* où chaque processus de classification est d'abord appliqué séparément sur les flux monomodaux avant la décision finale [Flanagan, 2004]. On trouve aussi des méthodes plus originales telles que l'utilisation de modèles issus de la psychologie

[13] Voir notamment les projets et actions suivants :
http://www.tele.ucl.ac.be/PROJECTS/M2VTS/
http://www.biosecure.info/
http://www.fub.it/cost275/

[14] http://www-nlpir.nist.gov/projects/trecvid/

expérimentale pour la fusion [Schwartz, 2004]. Une autre difficulté réside dans le fait que les modalités à fusionner peuvent être *asynchrones* : avance du mouvement labial ou du geste sur le signal vocal [Flanagan, 2004] par exemple. Par ailleurs, [Potamianos, 2004] met en évidence la difficulté du déploiement d'approches multimodales telles que la reconnaissance automatique de la parole audiovisuelle dans des contextes applicatifs où la modalité visuelle est mal contrôlée. Le manque de bases de données expérimentales conséquentes est aussi un problème qui conduit certains chercheurs à travailler sur des bases multimodales simulées (bases de *chimères* pour la biométrie [Garcia-Salicetti, 2003] par exemple). Enfin, des questions restent également ouvertes sur la possibilité d'établir des ensembles discrets de symboles, équivalents aux phonèmes de la parole, pour d'autres modalités telles que les gestes (tactèmes), le mouvement des lèvres (visèmes) ou des mains (voir cas particulier du langage parlé complété [Burger, 2004]).

Transcription enrichie dans un monde multimodal

La transcription enrichie, et notamment l'extraction d'informations non linguistiques, peut particulièrement bénéficier d'un contexte multimodal. C'est ce que j'ai essayé d'aborder dans mes récents travaux, qui seront plus précisément décrits dans le chapitre 2 : fusion de modalités audio et vidéo pour la recherche par similarité de vidéos-clips dans une base de vidéos ou pour la segmentation de vidéos en histoires, détection automatique de playback par mesure d'asynchronie lèvre / signal vocal pour des applications biométriques, reconnaissance multimodale de phonèmes en langage parlé complété.

1.3 Méthodologie

1.3.1 Ligne de conduite

Dans mes activités de recherche, j'essaie de maintenir un équilibre entre les aspects opérationnels et exploratoires. Ainsi, du point de vue *opérationnel*, j'ai fait en sorte de me doter de systèmes ou de plates-formes expérimentales correspondant le plus possible aux performances de l'état de l'art, et de les confronter à la communauté (via une participation soutenue à des campagnes d'évaluation internationales ou nationales) ainsi qu'à des applications réelles (via différents projets). Du point de vue *exploratoire*, comme l'annoncent le titre et le plan de ce mémoire, j'ai essayé d'apporter ma contribution à des problématiques nouvelles (multilinguisme, multimodalité pour la transcription enrichie) et d'aborder de nouveaux verrous scientifiques qui apparaissent dans certains contextes applicatifs (environnements perceptifs, indexation multimédia).

1.3.2 Outils théoriques : modèles probabilistes

Dans le domaine de la transcription enrichie, on est amené à modéliser de nombreuses classes d'objets sonores : type de son (parole / musique / bruit), locuteur, langue, canal de transmission, environnement sonore, phonème, mot, événement sonore (jingle)… La majorité des systèmes actuels repose sur des techniques de modélisation statistique. Nous ne redéfinirons pas ici les notions principales du domaine. Pour cela, le lecteur pourra se reporter à [Jelinek, 1999] [Kay, 1998] et à la bibliographie des thèses du CLIPS suivantes : [Istrate, 2003] [Moraru, 2004a] [Mayorga-Ortiz, 2005]. Précisons tout de même que, pour identifier des objets sonores à partir d'un flux audio (ou signal), on estime le plus souvent une vraisemblance notée $p(y/x)$ où y est une suite de vecteurs de paramètres extraits à partir du flux audio (paramètres MFCC, LPC[15], …), et x est une hypothèse de classe ou d'objet sonore. Par exemple, pour une tâche de transcription, x correspondra

[15] voir à nouveau les bibliographies de [Istrate, 2003] [Moraru, 2004] [Mayorga-Ortiz, 2005] pour plus de détails

à une hypothèse de phonèmes ou de mots, tandis que, pour une tâche de reconnaissance du locuteur, *x* correspondra à une hypothèse d'identité.

Pour estimer ces vraisemblances, les formalismes les plus utilisés sont les Modèles de Markov Cachés (HMM en anglais) dont chaque état correspond à une distribution multigaussienne. En fonction des problèmes à traiter, ces modèles peuvent être simplifiés (1 seul état conduit à une modèle de mélange de gaussiennes très utilisé en reconnaissance du locuteur par exemple), combinés (produits de modèles pour traiter plusieurs flux issus de multiples modalités, par exemple), ou inclus au sein de différents algorithmes (tâches de détection, classification, segmentation ou décodage, comme l'illustre la *figure 5*).

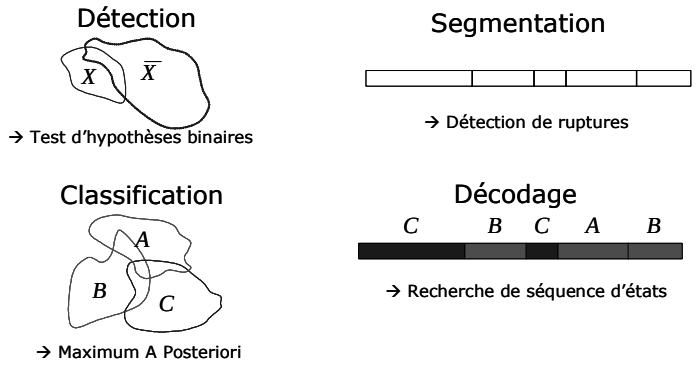


Figure 5 : Différents problèmes pouvant être traités par le formalisme des HMM[16]

Une théorie fondamentale exploitée est également l'approche bayésienne de la décision permettant d'établir des tests d'hypothèses. La formule de Bayes est d'ailleurs fondamentale dans le domaine de la reconnaissance automatique où la classe d'un objet sonore c* reconnue sera telle que :

$$c^* = \arg\max_i p(c_i / y) = \arg\max_i \frac{p(y/c_i).P(c_i)}{p(y)} \cong \arg\max_i p(y/c_i).P(c_i)$$

où *y* est l'observation et *ci* sont les classes candidates.

Par exemple, en reconnaissance automatique de la parole, *p(y/ci)* sera obtenue en utilisant un modèle acoustique, tandis que *P(ci)* (appelée probabilité *a priori*) sera obtenue avec un modèle de langage (mais il peut être intéressant d'utiliser cette probabilité *a priori* pour bien d'autres problèmes, comme proposé dans [Moraru, 2003b]).

Enfin, précisons que, malgré une efficacité démontrée, un certain nombre de verrous sont liés directement à la nature statistique de ces modèles, puisque leur apprentissage nécessite de rassembler des quantités de données importantes, ce qui pose des problèmes, entre autres, de portabilité (à de nouveaux environnements, de nouvelles tâches, de nouvelles langues).

1.3.3 Un cadre expérimental : les campagnes d'évaluation

Afin de conduire des recherches à partir de systèmes de référence représentatifs de l'état de l'art, j'ai essayé de confronter un certain nombre de systèmes développés au CLIPS au cadre expérimental qu'apportent les campagnes d'évaluation organisées par l'institut américain NIST. J'ai notamment participé aux campagnes NIST RT (*Rich Transcription*[17] [Palett, 2003]) proposées chaque année pour encourager la recherche dans le domaine de la transcription enrichie de documents. Les organisateurs proposent un corpus commun et aussi des métriques d'évaluation communes à tous les participants. Les données proposées en quantité suffisamment importante couvrent aussi tous les types d'enregistrements : enregistrements téléphoniques, enregistrements de

[16] dessin issu de la présentation de soutenance HDR de Frédéric Bimbot
[17] http://www.nist.gov/speech/

journaux télévisés et enregistrements de réunions. J'ai aussi participé, en collaboration avec Georges Quenot du CLIPS, aux campagnes d'évaluation TRECVID[18] (*Text REtrieval Conference VIDeo*) 2002, 2003 et 2004 également organisées par NIST. Dernièrement, le CLIPS a aussi participé à l'évaluation française ESTER[19] (Evaluation des Systèmes de Transcription Enrichie des émissions Radiophoniques) [Gravier 2004] qui vise à l'évaluation des performances des systèmes de transcription d'émissions radiophoniques.

Une participation soutenue à de telles campagnes permet de publier des résultats expérimentaux sur des corpus connus de la communauté, apportant ainsi une visibilité nationale ou internationale aux travaux réalisés. Un autre intérêt est la possibilité de participation commune avec d'autres équipes via un travail collaboratif, comme cela fut mon cas pour les évaluations RT (avec le *Laboratoire d'Informatique d'Avignon*) et TREC (avec l'équipe *MRIM* du CLIPS). Ce besoin de travail collaboratif avec d'autres équipes de recherche est aussi dû à la gestion lourde qu'implique une participation à de telles campagnes (quantité de données à traiter, réactivité nécessaire pour renvoyer les résultats dans les délais). Celle-ci est réduite lorsqu'on travaille sous forme de consortium (voir consortium ELISA par exemple [Moraru, 2004b]). Un autre aspect négatif de ces campagnes est qu'on n'a pas toujours le temps d'y proposer des systèmes très exploratoires d'un point de vue des méthodes.

Pour illustrer mon implication au CLIPS dans ces campagnes, le *tableau 2* présente un résumé de celles auxquelles j'ai participé[20] pour des tâches de transcription, de segmentation en locuteur et de recherche d'information (on trouvera plus de détails dans [Quenot, 2002] [Quenot, 2003] [Moraru, 2003a] [Moraru, 2004b] [Meignier, 2004] [Fredouille, 2004] [Besacier, 2004a] [Istrate, 2005a]).

Tâches \ Année	**2002**	**2003**	**2004**	**2005**	**2006**
Segmentation en locuteurs	NIST meeting 1/4 NIST BN 2/4 NIST Tel 3/4	Rich Transcription (RT) BN 2/8*	Rich Transcription (RT) meeting 1/3*	ESTER BN 4/5 Rich Transcription (RT) meeting 2/3*	
Transcription				ESTER BN 6/8	
Recherche d'informations	Extraction de plans Parole 7/13 Monologue 3/9	Extraction de plans Personne X 4/4	Segmentation en histoires 3/6		
Traduction de parole					DARPA/Transtac 1/6**

Tableau 2. Bilan de mes différentes participations à des campagnes d'évaluation au CLIPS (*=collaboration avec le LIA ; **=réalisé lors de mon séjour à IBM ; BN=Broadcast News)

1.3.4 Quelques plates-formes expérimentales

ELISA

Une des plates-formes utilisées est la plate-forme commune de développement conçue par le consortium ELISA [Magrin-Chagnolleau 2001]. Le consortium ELISA a été fondé en 1997 par le LIA (Avignon), l'ENST (Paris) et l'IRISA (Rennes). Il est autofinancé par ses participants et a pour objectif de faciliter les recherches coopératives en reconnaissance du locuteur ainsi que la participation des laboratoires francophones aux campagnes d'évaluation internationales telles que les campagnes NIST. La composition du consortium a évolué au cours des années et le laboratoire CLIPS a rejoint le consortium en 2001. Outre la plate-forme ELISA, le consortium organise des

[18] http://www-nlpir.nist.gov/projects/trecvid/
[19] http://www.afcp-parole.org/ester/index.html
[20] notamment avec D. Moraru, doctorant au CLIPS de 2001 à 2004.

réunions régulières et fournit un support technique et scientifique à ses membres pour les différentes campagnes d'évaluation. Ce consortium a notamment encouragé des travaux collaboratifs entre plusieurs laboratoires dont certaines expérimentations décrites dans ce manuscrit sont un exemple.

Concernant l'extraction de paramètres acoustiques, nous utilisons le module SPRO[21] développé au laboratoire IRISA par Guillaume Gravier. Ce module est capable de calculer plusieurs types de paramètres : coefficients FFT, bancs de filtres, coefficients cepstraux y compris MFCC, énergie ; il peut aussi faire des traitements du signal basiques, comme le filtrage passe-bas par exemple. La partie modélisation de la plate-forme ELISA permet quant à elle d'estimer les paramètres de *modèles de mélanges de gaussiennes* (GMM) avec des matrices diagonales ou pleines. Plusieurs méthodes d'adaptation de modèles par *maximum a posteriori* (MAP) sont également disponibles sur cette plate-forme. Le module de décision permet quant à lui le calcul de vraisemblances associées à chaque modèle pour un corpus de test considéré.

Récemment, la plate-forme a été réécrite sous la forme d'une boîte à outils (en licence GPL) nommée ALIZE[22] et une première version est déjà disponible sous forme de logiciel libre.

JANUS

Tout les systèmes de reconnaissance automatique de la parole du laboratoire ont été développés grâce à la partie *JANUS-Recognition-Toolkit (JRTk)* de la boite à outils Janus III [Zeppenfeld, 1997]. Janus III a été développé dans le laboratoire ISL (Interactive System Laboratory) des universités de Kalsruhe (Allemagne) et Carnegie Mellon (USA). La plate-forme Janus contient tous les composants nécessaires au développement d'un système de reconnaissance phonémique à base de Modèles de Markov Cachés. Janus offre un langage de programmation puissant qui permet de manipuler les structures de données internes et d'écrire des procédures complexes de haut niveau. Le langage de programmation de Janus utilise Tcl/Tk étendu avec des classes d'objets et leurs méthodes.

[21] http://www.irisa.fr/metiss/guig/spro.html
[22] http://www.lia.univ-avignon.fr/heberges/ALIZE

Chapitre 2 : Recherche d'éléments non linguistiques dans un monde multimodal

2.1 L'information « locuteur »

Nous avons vu que le but de la transcription enrichie est de produire un document structuré et très riche en informations à partir d'un signal audio brut. Dans ce contexte, le problème « locuteur » consiste à répondre de façon automatique à la question suivante : "Qui parle et quand sur un document audio?". La réponse à cette question peut être obtenue en plusieurs étapes :
• couper un document audio en segments homogènes appartenant à seulement un locuteur ;
• grouper les segments appartenant à un seul locuteur sur un même document ou sur une grande collection de documents audio ;
• identifier le locuteur auquel appartient une partie d'un document audio ;

La première étape est appelée détection de changements de locuteur (ou de tours de parole) tandis que la deuxième étape est appelée regroupement selon le locuteur. Si nous restons dans le cas où il y a un seul document à traiter, l'ensemble des deux premières étapes définit la tâche de segmentation en locuteurs (souvent appelée *speaker diarization* dans la littérature [Reynolds, 2005]). Si le regroupement est fait sur une grande collection de documents audio, nous pouvons parler d'appariement selon le locuteur [Meignier 2002]. La troisième étape est appelée suivi du locuteur et dans ce cas certains locuteurs appelés locuteurs cibles sont connus à l'avance par le système. Le système dispose alors d'un modèle a priori pour chacun d'entre eux.

Le problème « locuteur » n'est pas spécifique à la transcription enrichie. Dans le domaine du traitement de la parole, de nombreuses recherches ont été menées depuis les années 70 en reconnaissance du locuteur, avec comme but des applications de type biométrique. Les premières tâches dans ce domaine ont été l'identification du locuteur et la vérification du locuteur.

Mes contributions sur ce point portent, d'une part, sur la *biométrie*, et d'autre part, sur la *segmentation en locuteurs*.

2.1.1 Biométrie

J'ai commencé à m'intéresser à l'identification vocale biométrique pendant mes travaux de thèse [Besacier, 1998] où j'ai notamment apporté une contribution sur la répartition temporelle et fréquentielle de l'information *locuteur* dans le signal de parole (synthèse des travaux publiée dans la revue *Speech Communication* [Besacier, 2000]) et son utilisation dans une architecture multi-reconnaisseurs. Par ailleurs, pendant mon post-doctorat dans l'équipe Traitement du Signal de l'Institut de Microtechnique de Neuchâtel (Suisse, 98/99), je me suis intéressé au portage d'algorithmes de vérification vocale d'identité sur une plate-forme utilisable en conditions réelles ainsi qu'à l'évaluation de leurs performances dans ces conditions (projet européen M2VTS d'authentification multimodale de personnes) [Besacier, 1999].

Au CLIPS, ma contribution sur ce thème s'est faite à travers l'action COST275[23] (*Biometric Person Authentication over the Internet*) où j'étais co-animateur du groupe de travail "Evaluation" de 2001 à 2003. Cette action a servi de cadre expérimental à une partie des travaux de thèse de Pedro Mayorga-Ortiz [Mayorga-Ortiz, 2005]. Ce dernier a notamment proposé des protocoles de dégradation de corpus afin de diagnostiquer les performances des systèmes de vérification automatique du locuteur (et de reconnaissance automatique de la parole) dans des conditions de pertes de paquets sur réseaux et de données compressées. Il a observé une dégradation significative

[23] http://www.fub.it/cost275/pages/_home_main/index.htm

de la performance des systèmes de vérification du locuteur due à la compression lors de l'utilisation des codeurs de la parole (codeurs utilisés en voix sur IP ou téléphonie mobile). En revanche, Pedro Mayorga-Ortiz ne constate qu'une faible influence des pertes de paquets sur les performances de vérification du locuteur (contrairement à la tâche de reconnaissance automatique de la parole dont les performances sont dégradées par les pertes de paquets). Une explication de ces résultats est que les modèles acoustiques (GMM) utilisés pour modéliser les caractéristiques du locuteur considèrent chaque trame comme une entité indépendante. Ces modèles (GMM) sont donc moins sensibles aux pertes de paquets qui s'apparentent seulement à une perte de données pour estimer une vraisemblance. Pour ces modèles, l'ordre temporel des données peut donc être perturbé sans conséquence, contrairement au cas des modèles de Markov cachés utilisés en reconnaissance automatique de la parole. Une synthèse de ces travaux a été publiée dans la revue *Applied Signal Processing* [Besacier, 2004c].

2.1.2 Segmentation en locuteurs

Cette partie correspond aux travaux de thèse de Daniel Moraru [Moraru, 2004a] réalisés au CLIPS. Ses principales contributions sont les suivantes.

Evaluation d'un système de segmentation en locuteurs sur le long terme et sur plusieurs types de données.
Au début des travaux de thèse de Daniel Moraru, les principaux problèmes en segmentation en locuteurs étaient : la possibilité d'estimer de façon automatique le nombre de locuteurs présents dans un document (p1), la possibilité de traiter un flux audio hétérogène contenant de la parole de qualité variable, de la musique, du silence, etc. (p2) et la nécessité de traiter de la parole spontanée avec de nombreux recouvrements entre les interventions des locuteurs (p3). Le second problème est spécifique aux données de type journaux télévisés et radiodiffusés tandis que le troisième est lié au traitement d'enregistrements de réunions. Au cours de ses travaux, Daniel Moraru a proposé des solutions en vue de répondre à ces différents problèmes [Moraru, 2003a] [Moraru, 2004b]. Les améliorations obtenues ont pu être validées au fur et à mesure sur différentes campagnes d'évaluation (voir récapitulatif *Tableau 3*).

Données	Perf (%err.) 2002	Perf (%err.) 2003	Perf (%err.) 2004
Données Téléphoniques	16,58 %	-	-
Données Journaux TV	30,33 %	19,25 %	-
Données Réunions	50,20 %	-	22,6 %
Solutions proposées	-	p1 + p2	p1+p3

Tableau 3 Evolution des taux d'erreur de notre système sur 3 ans et sur différentes données. Synthèse de résultats sur les corpus NIST SpRec 2002, NIST RT 2003 et NIST RT 2004 Meeting

Utilisation d'informations a priori pour la segmentation en locuteurs

L'expérience accumulée lors de ces campagnes d'évaluation a permis à D. Moraru d'aborder ensuite une partie plus originale de sa thèse. Par hypothèse, spécifique au cadre de ces évaluations, nous ne disposions d'aucune information concernant les documents analysés pour la segmentation en locuteurs. Cependant, dans certaines applications pratiques et sur certaines données, il est possible de disposer d'informations a priori utiles avant de traiter l'enregistrement.

Nous avons vu notamment (travail en collaboration avec le LIA) qu'une première annotation automatique en classes acoustiques (homme/femme, bande large/étroite, parole/musique) permet de traiter efficacement des documents hétérogènes (problème p2) en conservant un système de segmentation en locuteurs performant [Meignier, 2004].

Dans le même esprit, une présegmentation en locuteurs issue d'un autre système, ou la sortie d'un système de suivi du locuteur cible, peuvent être utilisés avantageusement comme étape préliminaire pour améliorer un système de segmentation [Moraru, 2004c]. Par exemple, l'utilisation en cascade de notre système et de celui développé au LIA nous a permis d'obtenir le deuxième meilleur résultat officiel lors des évaluations NIST RT 2003 parmi huit autres participants (MIT, ICSI, Panasonic, CUED, LIMSI,…).

L'utilisation d'une annotation incomplète, permettant d'avoir des données de référence pour chaque locuteur, conduit également à une réduction importante de l'erreur de segmentation mais aussi du temps d'exécution. L'utilisation de ce type d'information est tout à fait réaliste dans certaines conditions pratiques et permet d'optimiser le résultat d'un système de segmentation classique [Moraru, 2004c].
Le dernier aspect encourageant a été l'utilisation du signal audio issu de plusieurs capteurs, situation spécifique aux enregistrements de réunions (encore en collaboration avec le LIA [Fredouille, 2004]). Sans utiliser de techniques de séparation de sources, nous avons extrait la parole appartenant à chaque locuteur à partir de résultats de segmentation obtenus individuellement sur chaque capteur. Cette approche a donné de bons résultats malgré la difficulté de la tâche, due à la spontanéité de la conversation, et nous a permis de proposer une première piste pour le troisième problème (p3). Le taux d'erreur de segmentation obtenu sur des données de réunion en 2004 est même comparable avec celui obtenu sur des documents de type journaux télévisés.

Une synthèse des travaux conduits en collaboration avec le LIA a été soumise et acceptée dans la revue éditée par Elsevier : *Computer Speech and Language* [Fredouille, 2006].

2.2 Autres Informations

2.2.1 Sons

Un signal audio peut véhiculer d'autres informations que la parole. C'est le cas par exemple pour des enregistrements issus d'espaces perceptifs (ou salles intelligentes, concept plus précisément développé au *chapitre 4*) qui captent en continu l'environnement sonore. La détection et la reconnaissance de signaux sonores a été abordée au CLIPS à travers la thèse de Dan Istrate [Istrate, 2003]. Les sons à reconnaître étaient des sons de la vie courante, en vue d'identifier des alarmes potentielles dans un « appartement intelligent » abritant une personne âgée ou un patient en convalescence.

Une des contributions de cette thèse a été la proposition d'algorithmes de détection d'événements sonores dans le bruit et l'adaptation de méthodes de reconnaissance de la parole et du locuteur à la classification des sons de la vie courante. L'étude s'est notamment concentrée sur la recherche des paramètres acoustiques les mieux adaptés à la reconnaissance de sons de la vie courante. Le système global proposé est constitué de deux parties : la détection des événements sonores et la classification des sons. Les principales contributions du travail de thèse de Dan Istrate sont décrites ci-dessous.

Détection d'événements sonores dans le bruit

Plusieurs algorithmes de détection ont été proposés dans la thèse de Dan Istrate. Un des algorithmes se fonde sur la transformée en ondelettes du signal. Cet algorithme présente de très bonnes performances dans l'environnement bruité réel : un taux d'égal erreur de détection de 5.6% pour un niveau de bruit élevé (RSB=0 dB). Il semble mieux adapté au contexte applicatif où les sons à détecter sont majoritairement impulsionnels (claquements de portes, bris de verres, chutes).

Classification de sons de la vie courante

L'algorithme de classification utilisé est fondé sur les mélanges de gaussiennes (GMM). Comme la paramétrisation du signal est très importante pour une bonne classification des sons, l'étude des paramètres acoustiques adaptés aux sons de la vie courante a constitué l'axe principal de cette recherche. La méthodologie a consisté à évaluer par des méthodes statistiques la pertinence des paramètres utilisés habituellement pour la détection de signaux musicaux (nombre de passages par zéro, roll-off point, centroïde spectral). Étant donné que la transformée en ondelettes est mieux adaptée à l'analyse de signaux impulsionnels, la possibilité d'extraire des paramètres acoustiques à partir de cette transformée a été aussi étudiée. Le meilleur système présenté dans la thèse donne un taux d'erreur de classification de 7.1% pour 7 classes de sons.

Couplage, évaluation et contraintes liées à l'application

Les problèmes liés au couplage entre la détection et la classification, ainsi que le problème de l'évaluation d'un tel système sont aussi abordés dans ce travail. En fin de manuscrit, l'évolution vers un système de reconnaissance de « sons clés », inspirée de la reconnaissance de mots-clés en parole, est ébauchée. Une implémentation en temps réel des algorithmes proposés a été réalisée pour l'application de télésurveillance médicale et a été validée dans l'appartement test disponible pour le projet.

On trouvera une synthèse détaillée de ces travaux dans le journal *IEEE Transactions on Information Technology in Biomedicine* [Istrate, 2006] ainsi que dans [Istrate, 2005b].

2.2.2 Jingles

Sur des documents radiophoniques ou audiovisuels, on trouve ce qu'on peut appeler des «invariants de production » (jingles audio, générique d'une émission, publicités…). Détecter automatiquement sur un document audio ou audiovisuel des « invariants de production » permet d'obtenir de précieux indices aidant à une première macrosegmentation d'un document. Par ailleurs, certains jingles ou génériques apportent une information sémantique forte quant à ce qui va suivre dans le document (ex : générique de météo). Une autre application de cette détection est le « nettoyage » automatique de documents par suppression éventuelle de zones redondantes et non pertinentes pour l'archivage (publicités, génériques, etc…). De récents travaux ont montré l'intérêt de l'utilisation de *signatures audio* fondées sur l'extraction de caractéristiques bas niveau du signal, pour la recherche automatique de génériques sur la bande son d'un document [Allamanche, 2002] [Cano, 2002] [Pinquier, 2004]. Ces techniques, contrairement à des méthodes de reconnaissance de forme classiques, font l'hypothèse d'une quasi non-variabilité du signal à détecter (d'où le concept d'invariant de production).

Ce problème nouveau a été abordé au CLIPS via l'encadrement d'un travail de DEA [Sénéchal, 2004] (en collaboration avec le laboratoire LIS[24]) où une signature audio originale, utilisant la mesure de platitude spectrale (un des descripteurs de bas niveau utilisés dans la norme MPEG-7), a été proposée et évaluée. Dans ce travail, on s'est attaché à retrouver une séquence connue (i.e. dont on connaît le modèle a priori) dans une base de documents. La méthode proposée a été expérimentée sur une partie de la base de vidéos TREC 2003 (34h) et permet de retrouver un ensemble de jingles avec de bonnes performances (précision de 99.5% et rappel de 85%) ; elle est par ailleurs résistant à la compression de signaux au format mp3 (résultats publiés dans la conférence *IEEE ICME 2005* [Senechal, 2005]).

Très récemment, dans une seconde phase plus exploratoire, nous avons également essayé de faire émerger de façon automatique et non supervisée des similarités dans le flux audio (en utilisant la

[24] Laboratoire des Images et des Signaux : http://www.lis.inpg.fr/

technique des dot-plots issue de la génomique), permettant ainsi une première auto-organisation des documents (travail de fin d'étude ingénieur de G. Delafosse [Delafosse, 2005]).

Enfin, il est intéressant de remarquer que de telles techniques « proches des données » (ou empiriques), n'utilisant pas de modèles, se retrouvent dans de récents travaux en traitement de la parole, comme c'est le cas notamment dans [Gillick, 2005] pour la détection de locuteurs où les modèles statistiques de type GMM sont remplacés par des méthodes proches des techniques de signature présentées ci-dessus.

2.2.3 Zones d'intérêt intonatives

Parmi les autres éléments non linguistiques, ne nécessitant pas une phase lourde de transcription automatique, on peut envisager de détecter des « zones d'intérêt » dans le signal pour lesquelles l'intonation est particulière : par exemple des zones de questions ou des zones « chargées » émotionnellement. Cela peut être fait en entraînant des patrons prosodiques, puis en retrouvant des parties de documents où l'intonation y est similaire, avec des outils de type « arbres de décision » par exemple [Shriberg, 2000] [Wang, 2003a].

Ce type d'approche est abordé au CLIPS à travers la thèse de Vu-Minh Quang menée en co-tutelle avec le centre MICA[25] à l'Institut Polytechnique d'Hanoï. Ses premiers travaux consistent à détecter automatiquement des zones où une question est posée, sur un flux de parole. Cela peut être intéressant dans le domaine de la fouille de données audio, si l'on souhaite retrouver des informations pertinentes sur des corpus de grande taille (réunions par exemple) ou dans le domaine du dialogue interactif pour enrichir un résultat d'analyse en détectant les tours de parole qui sont des questions. Pour cela, des paramètres issus de la courbe d'intonation prosodique sont extraits et un arbre de décision permet de classer les tours de parole en deux classes : « question » et « non question ». Les résultats expérimentaux obtenus sur des enregistrements de réunions, découpés manuellement en tours de parole, donnent un taux de bonne classification de 84% [Quang, 2005] mais il s'avère que la courbe de *f0* n'est pas suffisante pour détecter des zones de question, notamment sur des tours de parole courts. Par ailleurs, les premiers arbres de décision obtenus dépendent fortement des données sur lesquelles ils sont appris, et ne généralisent que très peu sur d'autres types de données.

Ce type d'approche pourrait ensuite être appliqué à d'autres tâches, comme cela est illustré dans les travaux suivants : reconnaissance d'attitudes dans un système de dialogue [Fujie, 2003], détection de zones d'insistance (*emphasis*) sur des documents de réunions [Kennedy, 2003] ou détection de dialecte et d'émotions dans le système HMIHY[26] d'ATT [Shafran, 2003].

2.3 Exploitation et prise en compte de la multimodalité

Bien que spécialiste du traitement automatique de documents audio au départ, j'ai récemment tenté d'élargir mes travaux de recherche à un contexte multimodal où, quel que soit le domaine applicatif, on rencontre des flux de données provenant de multiples modalités (audio, vidéo et autres). Cela s'est déjà concrétisé par une première collaboration avec le laboratoire LIS et l'équipe *Recherche d'Information* du CLIPS dans le domaine de la recherche de documents vidéos (expériences réalisées lors des campagnes d'évaluation TREC Vidéo 2002, 2003 et 2004). Dans ces premiers travaux, nous avons tenté de fusionner des modalités audio et vidéo, tant au niveau des données (intégration précoce, *signatures audiovisuelles pour la détection de génériques*) qu'au niveau des décisions (intégration tardive, *segmentation audiovisuelle de documents*). Enfin, j'ai également proposé une exploitation originale de la nature multimodale de la parole, dans une expérience sur la *mesure de cohérence lèvres / voix pour la biométrie*.

[25] http://www.mica.edu.vn
[26] How May I Help You

Ces trois contributions sont développées dans les trois sous-sections suivantes.

2.3.1 Signatures audiovisuelles pour la détection de génériques

La signature audio présentée précédemment et issue des travaux de DEA de Benjamin Senechal [Senechal, 2004], a été complétée par une signature vidéo (fondée sur l'évolution de barycentres des niveaux de gris de chaque image) pour former ainsi une signature audiovisuelle pour la détection de génériques. La conjugaison des caractéristiques audio et vidéo permet d'obtenir des résultats encore meilleurs et plus résistants aux distorsions. En effet, la méthode audiovisuelle expérimentée sur une partie de la base de vidéos TREC 2003 (34h) permet de retrouver un ensemble de génériques avec une précision de 100% (contre 99.5% pour la signature audio et 98.3% pour la signature video) et un rappel de 94.7% (contre 85% pour la signature audio et 88.7% pour la signature vidéo). Dans ce travail, l'intégration des deux modalités (audio et vidéo) peut être considérée comme précoce puisqu'elle a lieu au niveau des vecteurs de paramètres extraits de façon synchrone à la vitesse image sur les deux flux audio et vidéos. L'association de ces vecteurs audio et vidéo constitue ainsi une signature audiovisuelle.

Des détails expérimentaux sur ces résultats obtenus se trouvent dans l'article [Senechal, 2005] publié à la conférence *IEEE ICME 2005*.

2.3.2 Segmentation audiovisuelle de documents

Cette partie correspond à une autre contribution des travaux de thèse de Daniel Moraru [Moraru, 2004a] où l'utilisation conjointe d'informations issues des canaux audio et vidéo est envisagée pour des tâches de recherche d'information multimédia et de segmentation de vidéos.

Une première tâche abordée, nécessitant la fusion d'informations audio et vidéo, est la détection de plans monologues sur une vidéo qui nécessite l'utilisation conjointe d'un détecteur de visage et d'un système de segmentation en locuteurs (travaux décrits dans [Quenot, 2002]). Cependant, dans ce premier exemple, le signal audio et le signal vidéo sont traités de façon séparée. Or, comme les deux signaux peuvent provenir d'une même source, il peut exister des informations représentées dans les deux flux de données. Un système optimal de traitement de l'information devrait en conséquence exploiter les informations redondantes entre les deux signaux.
Une contribution plus intéressante a donc consisté à proposer une exploitation plus *intégrée* de information multimodale. Par exemple, Daniel Moraru a évalué l'apport de l'information vidéo pour une tâche purement audio comme la segmentation en locuteurs. Le travail présenté dans sa thèse, et publié dans [Moraru, 2005] consiste à détecter les changements de locuteur en utilisant aussi la détection de frontières de plans vidéo. Ainsi, on peut considérer trois différentes détections de changements de locuteur : uniquement audio en utilisant par exemple le critère d'information bayésien (*BIC*[27]), uniquement vidéo en utilisant la détection de frontières de plans vidéo, et enfin en assemblant les deux voies audio et vidéo (BIC + Plans).

Détection de changements de locuteur	Audio (BIC)	Vidéo (Plans)	Audio/Vidéo (BIC + Plans)
Erreur de Segmentation	29,7 %	30,1 %	26,8 %

Tableau 4 : Performances de segmentation en locuteurs sur une partie du corpus TREC 2003 en utilisant différentes stratégies de détection de changements de locuteur [Moraru, 2005]

[27] voir thèse D. Moraru pour plus de détails

Les résultats du tableau 4 montrent l'apport de l'information vidéo sur une tâche purement audio de segmentation en locuteurs, puisqu'on obtient une légère réduction d'environ 3 % en absolu de l'erreur grâce à un système de segmentation en locuteurs "audiovisuel". En fait, l'utilisation des frontières de plans vidéo permet de fournir des segments homogènes issus de plans uni-locuteur et la détection de changements de locuteur basée classiquement sur le canal audio permet de découper en segments homogènes les plans longs contenant la parole de plusieurs personnes.
L'utilisation d'une information multimodale se retrouve également dans la tâche de détection de frontières d'histoires sur des documents audiovisuels. Cette tâche illustre d'ailleurs parfaitement le problème de trou (*gap*) sémantique bien connu en recherche d'information multimédia. L'information que nous désirons retrouver (les frontières d'histoires) est très éloignée de toute information qu'on peut extraire directement à partir du signal audio ou vidéo, comme la détection de plans vidéo ou la détection de changements de locuteur. Une histoire peut contenir plusieurs plans vidéo. Par exemple, le présentateur du journal annonce le sujet et donne la parole à un reporter, plusieurs personnes sont interviewées par le reporter, et le sujet est ensuite clôturé à nouveau dans le studio par le présentateur.
Dans d'autres cas, un plan vidéo peut contenir plusieurs histoires. C'est le cas typique des plans de début de journal ou le présentateur annonce successivement les principaux titres qui constituent autant d'histoires différentes. Une histoire peut également être précédée d'un court générique (jingle) comme la rubrique météo par exemple. Une approche qui tient compte de multiples informations issues de multiples détecteurs est alors indispensable. Pour cela, nous avons fait plusieurs observations empiriques des données de développement TREC 2003. A partir de ces observations, nous pouvons dire que dans certains cas :
a. un long silence peut correspondre à un changement d'histoire ;
b. un changement de plan vidéo peut correspondre à un changement d'histoire ;
c. un changement de locuteur peut correspondre à un changement d'histoire ;
d. le début ou la fin d'une intervention du présentateur du journal peut correspondre à un changement d'histoire ;
e. un jingle (ou générique) peut correspondre à un changement d'histoire ;
f. certains mots-clés détectés à partir de la transcription automatique peuvent correspondre à un changement d'histoire.
Au vu de ces hypothèses, un premier système de segmentation en histoires a été développé au CLIPS et présenté aux évaluations TREC 2004 : il utilise des sorties de détecteurs (intégration tardive) appliqués sur la bande image, la bande son et également le texte obtenu après transcription automatique. Tous les détails techniques sur le système du CLIPS de segmentation en histoires sont présentés dans [Besacier 2004]. Ce premier système permet une détection de frontières d'histoire avec un rappel de 0.6 et une précision de 0.5 (système classé 3ème/6 aux évaluations TREC 2004).

2.3.3 Calcul de cohérence lèvres / voix pour la biométrie

Cette partie correspond à un travail financé par le CNRS sur l'ACI *Sécurité Informatique* (projet BIOMUL[28]). Dans ce contexte, les travaux d'un post-doctorant au CLIPS (Nicolas Eveno, issu du laboratoire LIS/INPG) sur la détection de contours labiaux à partir d'une vidéo, ont été appliqués à la biométrie labiale bimodale.

Choisir les lèvres pour bâtir un système d'identification peut paraître étrange, voire artificiel. En effet, il n'est pas évident que nous utilisions cet indice pour reconnaître nos interlocuteurs. Cependant, des recherches récentes présentent des preuves expérimentales démontrant que la zone de bouche possède un pouvoir discriminant bien plus élevé que n'importe quelle autre partie du visage de même taille. Un autre avantage de l'analyse labiale est la possibilité qu'elle offre d'être combinée à des systèmes d'identification par la parole. De plus le matériel nécessaire (une web-cam et un micro) n'est pas dédié et est relativement courant, ce qui devrait permettre une diffusion très rapide auprès du grand public. Enfin, les mouvements labiaux et la voix produite sont hautement

[28] http://www.lia.univ-avignon.fr/heberges/BIO_MUL/

corrélés car ce sont deux modalités d'un seul et même processus. Ce lien étroit est largement utilisé dans les systèmes de reconnaissance automatique de la parole audiovisuelle [Potamianos, 2004], mais a été un peu moins exploité dans le domaine de la biométrie.

Dans un premier temps, nous avons tenté d'effectuer une détection de playback (ou *liveness*) sur des séquences vidéo. Cela permet de repérer un éventuel falsificateur mimant une phrase d'accès enregistrée sur un magnétophone. En effet, dans le domaine de la biométrie, il est crucial de détecter les imposteurs et de contrer les attaques par enregistrement, dites par playback. Cependant, peu de recherches ont été consacrées à ce type détection, qu'on appelle aussi test de « *liveness* » car il assure que les indices biométriques en cours d'acquisition proviennent réellement d'une personne présente au moment de l'identification. Les systèmes biométriques audiovisuels actuels sont souvent présentés comme des solutions implicites à ce problème car il est extrêmement difficile pour un imposteur de reproduire parfaitement à la fois les indices audio et vidéo. Cependant, dans le cas d'un falsificateur utilisant une phrase d'accès enregistrée sur un magnétophone, l'audio correspond parfaitement à la voix d'un utilisateur autorisé. Le système devient alors unimodal puisque le seul indice disponible réellement discriminant est l'image, ce qui conduit à une chute importante des performances. Dans ces conditions, l'intérêt d'un test de « liveness » précédant le système d'identification proprement dit parait évident.

La méthode de vérification de « liveness » que nous avons développée au CLIPS est basée sur la mesure de cohérence statistique entre les mouvements labiaux et la voix du locuteur (figure 6). Pour effectuer cette mesure, nous avons comparé deux techniques d'analyse de données : l'analyse canonique des correspondances (CANCOR [Hotelling, 1936]) et l'analyse par co-inertie (COIA [Doledec, 1994]). La première est relativement classique, tandis que la seconde, beaucoup plus récente, a été conçue à l'origine pour l'analyse de données en écologie statistique. Malgré sa très grande efficacité, elle est quasiment inconnue dans le domaine du traitement du signal. Nous avons donc tenté d'évaluer son apport au domaine de l'analyse de données audio-vidéo.

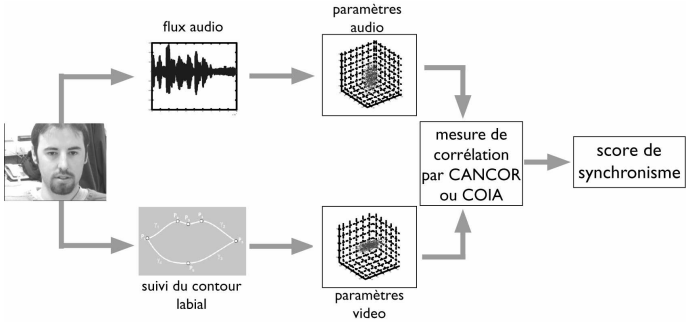


Figure 6 : détection de playback par CANCOR ou COIA

Les analyses COIA et CANCOR permettent de déterminer une valeur de corrélation entre les données audio et vidéo. Cependant, nous avons montré que le seuillage simple de cette corrélation ne permet pas de détecter efficacement une attaque par playback. Nous avons donc introduit un score de *liveness* basé non pas sur une seule valeur de corrélation, mais sur des valeurs successives obtenues en décalant progressivement l'audio par rapport à la bande image. D'après les tests que nous avons menés sur la base XM2VTS, l'analyse de données par COIA conduit à des résultats bien meilleurs que CANCOR. Finalement, la méthode de vérification de *liveness* que nous avons proposée permet pour l'instant de détecter un playback avec un taux d'EER (Equal Error Rate) de 12.5%. De plus, elle est totalement indépendante du locuteur. Ce travail a fait l'objet d'une publication à la conférence Eurospeech 2005 [Eveno, 2005].

Chapitre 3 : Transcription dans un monde multilingue

3.1 Première confrontation avec le multilinguisme : la traduction automatique de parole (projets C-STAR et NESPOLE)

A mon arrivée au CLIPS en 1999, j'ai été impliqué dans deux projets de traduction automatique de parole dont le CLIPS était partenaire : CSTAR[29] (*Consortium for Speech Translation Advanced Research*) et NESPOLE[30] (*Negotiating Through SPOken Language in E-commerce*). Je suis devenu responsable de ces deux projets pour l'équipe GEOD (dont la contribution était surtout de fournir l'étage de reconnaissance automatique de la parole) [Besacier, 2001]. Le but des recherches menées au sein de ces projets était la traduction automatique de parole spontanée avec tous les couples de langues possibles entre les différents partenaires. Dans la phase 3 de CSTAR et dans NESPOLE, la tâche visée était la réservation touristique. Le scénario type est celui d'un client qui communique dans sa langue maternelle, par le biais du système de traduction, avec un agent de voyage qui ne parle pas la même langue que lui. Nous sommes donc dans le cas d'une communication homme/homme médiatisée. L'un des objectifs de NESPOLE était notamment de tester et mettre en œuvre la portabilité et l'extensibilité (*scalability* en anglais) des méthodes de traduction.

Dans un premier temps, mes contributions sur ce thème ont essentiellement consisté à répondre aux objectifs opérationnels fixés dans le projet européen NESPOLE :
-mise au point (avec D. Vaufreydaz) d'un module de reconnaissance automatique de parole continue en français pour une tâche de réservation touristique et intégration de ce module dans la chaîne de traduction complète [Besacier, 2001] [Lavie, 2002],
-organisation de collectes de données liées à la tâche, pour le français [Burger, 2001] [Mana, 2003],
-évaluation de la chaîne complète de traduction, avec des protocoles expérimentaux identiques pour tous les partenaires de NESPOLE [Rossato, 2002] [Blanchon, 2004].

Cependant, vers la fin du projet, j'ai pu identifier un certain nombre de thèmes plus exploratoires parmi lesquels :
-le problème de l'interfaçage efficace entre le module de reconnaissance et le module de traduction (utilisation de treillis ou des N-meilleures hypothèses pour l'analyse, modèle de langage « sémantique »), cela s'est traduit par le co-encadrement, avec H. Blanchon, du DEA de Quang Vu-Minh qui donna lieu à une publication à la conférence TALN 2004 [Vu-Minh, 2004],
-l'adaptation rapide d'un système de reconnaissance vers un nouveau domaine ou l'extension de la couverture d'un domaine [Vaufreydaz, 2001],
-le problème du portage vers une nouvelle langue, qui fait l'objet de la suite de ce chapitre.

Enfin, le projet NESPOLE a été aussi une base d'essais enrichissante pour les expériences menées au CLIPS (thèses de D. Vaufreydaz [Vaufreydaz, 2002] et P. Mayorga-Ortiz [Mayorga-Ortiz, 2005] notamment). L'avantage principal était la possibilité de juger, dans des conditions réelles d'utilisation, de la pertinence de nos modèles dans un cadre dialogique. La mise au point de démonstrateurs et les nombreuses phases de test nous ont aussi permis d'analyser les problèmes et d'améliorer les méthodes et les outils que nous avions développés en transcription automatique.

3.2 Vers une reconnaissance automatique de la parole multilingue : application aux langues peu dotées

3.2.1 Introduction

Cette partie de mes activités de recherche concerne la reconnaissance automatique de parole multilingue. Comme cela est dit dans le chapitre d'introduction, il subsiste un certain nombre de

[29] http://www.c-star.org/
[30] http://nespole.itc.it

verrous en ce qui concerne la généricité des systèmes de reconnaissance automatique de la parole et leur portabilité vers de nouvelles langues. L'originalité de mon approche vient de la volonté d'aborder des langues peu ou pas dotées, pour lesquelles peu ou pas de corpus sont disponibles, ce qui nécessite des méthodologies innovantes qui vont bien au-delà du simple réapprentissage ou de l'adaptation de modèles.
Les langues sur lesquelles j'ai travaillé avec Viet-Bac Le, doctorant au CLIPS de septembre 2002 à juin 2006 (thèse soutenue le 1er juin 2006), sont le vietnamien et le khmer. Le calcul de l'indice σ proposé par Vincent Berment [Berment, 2004] qui mesure le niveau d'informatisation d'une langue me permet de vérifier que ces deux langues font effectivement partie de la classe des langues peu dotées (indice < 10 ; le vietnamien étant cependant à la limite). Le khmer, évalué dans [Berment, 2004] obtient un indice σ d'environ 6/20, tandis que notre évaluation du vietnamien, faite par E. Castelli et Nguyen Quoc Cuong, tous deux chercheurs du centre MICA à Hanoï, donne un indice σ de 10/20 environ (voir tableau 5). Une partie de cette différence s'explique par le fait que le vietnamien, contrairement au khmer, utilise une écriture latine accentuée, rendant plus « simples » certaines tâches comme la reconnaissance automatique de caractères et le tri lexicographique.

	Services / ressources	Criticité (/10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	16	160
	Recherche et remplacement	8	17	136
	Sélection du texte	6	17	102
	Tri lexicographique	5	6	30
	Correction orthographique	2	6	12
	Correction grammaticale	0	0	0
	Correction stylistique	0	0	0
Traitement de l'oral				
	Synthèse vocale	5	0	0
	Reconnaissance de la parole	5	0	0
Traduction				
	Traduction automatisée	8	6	48
ROC				
	Reconnaissance optique de caractères	9	12	108
Ressources				
	Dictionnaire bilingue	10	13	130
	Dictionnaire d'usage	10	0	0
Total				886 / 1760
Moyenne				10 / 20

Tableau 5 : Tableau d'évaluation du niveau d'informatisation pour le vietnamien

Il est important de noter que pour ces deux langues, les services liés au traitement de l'oral sont inexistants. C'est ici que se situe notre problématique. Face à la critique qui pourrait être faite concernant notre choix d'aborder des technologies peut-être moins importantes, en termes de développement, que d'autres liées au traitement de texte et aux dictionnaires, notre argument est que développer des systèmes de traitement de la parole dans une langue « peu dotée » peut permettre de collecter des ressources utiles pouvant être ensuite remises au « pot commun » de cette langue (dictionnaire phonétique, corpus oral, transcriptions de conversations spontanées par exemple).
Les sections suivantes décrivent le contexte de nos travaux, qui a conduit au choix des deux langues abordées dans ce travail, ainsi que la méthodologie qui nous permet de développer et d'adapter le plus rapidement possible un système de reconnaissance automatique dans une nouvelle langue. Nos résultats expérimentaux concernant l'application de cette méthodologie au vietnamien et au khmer sont également présentés.

3.2.2 Contexte

Ce travail a pu se réaliser grâce à la collaboration qui existe entre le CLIPS et le centre MICA[31], créé en 2002 pour participer au développement des technologies de l'information au Vietnam et pour répondre aux préoccupations relatives à leur évolution. Plus récemment, le Département de

[31] http://www.mica.edu.vn

Génie Informatique et Communication de l'Institut de Technologie du Cambodge (ITC[32]) s'est associé à cette collaboration afin de créer un groupe de recherche spécialisé en traitement de la parole en langue khmère. Pour financer ces collaborations, des projets soutenus par l'AUF (Agence Universitaire pour la Francophonie) ou par le MAE (Ministère des Affaires Etrangères) ont été soumis et acceptés[33] [34].

La langue vietnamienne

Elle est parlée par environ 70 millions de personnes dans le monde (source : MSN-Encarta, 2005). Son origine est toujours sujette à débat parmi les linguistes. Il est cependant généralement admis qu'elle a des racines communes et fortes avec le môn-khmer qui fait partie de la branche austro-asiatique et qui comprend le mon parlé en Birmanie, et le khmer, la langue principale du Cambodge appelée aussi cambodgien, aussi bien que d'autres langues parlées par les habitants des îles du nord du Vietnam. Le vietnamien est une langue tonale qui possède six tons. L'orthographe est latine depuis le XVIIè siècle, avec un second niveau de diacritiques, au dessus des accents, pour noter les tons [Nguyen, 2002].

La langue khmère

Le khmer est parlé par une dizaine de millions de personnes dans le monde (source : MSN-Encarta, 2005). Il appartient également au groupe des langues môn-khmères. La langue khmère est une langue atonale – contrairement aux langues chinoises, au thaï et au vietnamien. Cependant, le khmer possède comme ses cousines austro-asiatiques plusieurs registres vocaliques : les voyelles peuvent être longues ou brèves, diphtonguées, reposer sur des consonnes aspirées ou non aspirées. Une différence de ce type entre deux mots en modifie complètement le sens. (Ex. slap = mourir; slaaap = aile d'un oiseau). Cette particularité fait du système vocalique cambodgien un des plus riches au monde. Au niveau de l'écriture, pour adapter les fontes informatiques, il a fallu gérer un ordonnancement, sur plusieurs niveaux, de 33 consonnes, 32 consonnes souscrites, 28 voyelles, 14 voyelles indépendantes et 10 ligatures, sans compter les chiffres et la ponctuation [Berment, 2004].

3.2.3 Méthodologie

Pour développer un système de reconnaissance automatique de la parole continue dans une nouvelle langue, il est souvent nécessaire de rassembler une grande quantité de corpus, contenant non seulement des signaux de parole (pour l'apprentissage des modèles acoustiques du système) mais également des données textuelles (pour l'apprentissage des modèles de langage du système). Pour une langue peu ou pas dotée, il faut alors collecter soi-même les ressources nécessaires : signal de parole, lexique, corpus textuels, etc ou trouver des techniques permettant de se passer de ces grandes quantités de corpus. Les sous-sections suivantes présentent les contributions apportées à ces problèmes par Viet-Bac Le et moi même.

Collecte de ressources

Une première façon d'accélérer le portage des systèmes de reconnaissance automatique de parole continue grand vocabulaire vers une nouvelle langue, est de développer une méthode permettant une collecte rapide et/ou facilitée de ressources textuelles et acoustiques. Cette approche a l'avantage de ne pas modifier fondamentalement le cœur des techniques de reconnaissance utilisées.

[32] http://www.itc.edu.kh/fr/
[33] TALK "Traitement Automatique de la Langue Khmère" projet AUF PCSIU, collaboration etre MICA-Hanoï, l'Institut de la Technologie du Cambodge ITC à Phnom Penh et le CLIPS.
[34] Projet CORUS "Traitement de la parole en langue vietnamienne" collaboration etre MICA-Hanoï et le CLIPS, soutenue par le MAE français.

Recueil de données textuelles

Concernant le recueil de données textuelles en grande quantité (figure 7), une approche intéressante consiste à « aspirer » un grand nombre de sites Web dans la langue donnée et à filtrer les données récupérées pour les rendre exploitables. Ces données textuelles peuvent servir d'une part à calculer des modèles de langages statistiques, et d'autre part à obtenir un corpus pouvant ensuite être prononcé par des locuteurs en vue de la constitution d'une base de signaux conséquente.

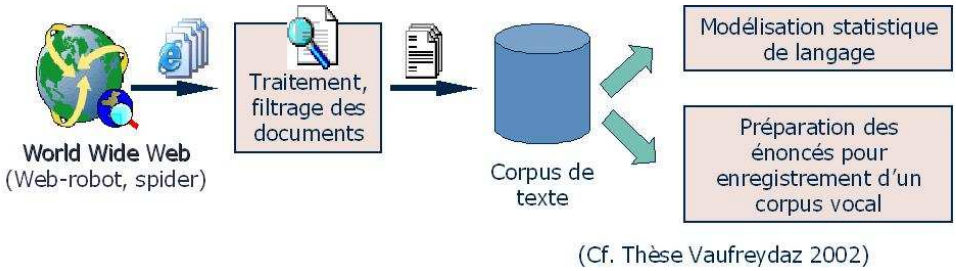


Figure 7 : récupération de données textuelles en utilisant le Web

Une telle approche a déjà été relativement bien validée pour une langue bien dotée telle que le français (co-encadrement de la thèse de D. Vaufreydaz [Vaufreydaz, 2002]). Les problèmes spécifiques pour les langues peu dotées concernent le nombre de sites Web qui peut être peu important, la vitesse de transmission, et la qualité des documents, qui nécessite alors plus d'outils de traitement. On préfèrera par exemple des sites de nouvelles, au fort contenu rédactionnel, tels que VNexpress[35] par exemple, pour le vietnamien.
Afin de rendre les données exploitables, un certain nombre de traitements sont nécessaires tels que : 1) transformation html vers texte, 2) normalisation des balises, 3) conversion des encodages (nous avons choisi de tout convertir vers une représentation interne unique utilisant l'encodage UTF-8 d'Unicode), 4) séparation en phrases et 5) en mots, 6) groupement de mots composés, 7) transcription des symboles et 8) filtrage en fonction d'un vocabulaire donné. Alors que certains traitements peuvent être considérés comme relativement indépendants de la langue cible (1-2-6-8), d'autres doivent être repensés (3-4-5-7) pour chaque nouvelle langue cible : par exemple, la séparation en mots est triviale pour les écritures latines mais problématique pour d'autres systèmes d'écriture comme celui du khmer, surtout si l'on ne dispose pas d'un vocabulaire (i.e. une liste de mots) au départ. Une boîte à outils *open source* rassemblant quelques-uns de ces outils de traitement a été développée au CLIPS[36]. Une description plus détaillée des traitements réalisés et des expérimentations associées pour la modélisation du langage en vietnamien peut être trouvée dans [Le, 2003].

Recueil de signaux de parole

Pour le recueil de signaux de parole, le CLIPS a développé un outil logiciel ne mettant en œuvre que du matériel standard : EMACOP (Environnement Multimédia pour l'Acquisition et la gestion de Corpus Parole) [Vaufreydaz, 98]. La plupart du temps, les campagnes d'enregistrement mobilisent d'importantes ressources humaines pour guider ou assister les locuteurs dans leur tâche de diction, pour organiser l'enregistrement, pour préparer les scénarios et les données, etc. Il faut pouvoir contrôler les différents scénarios pour faire varier les conditions de capture : la lecture d'un texte ou d'une suite de mots ou de mots isolés, la répétition après écoute d'une phrase, le dialogue en réponse à des questions, etc. Les méthodes d'acquisition rigoureusement contrôlées sont donc lourdes et les difficultés sont amplifiées dans le cas des langues mal dotées où les locuteurs ne sont pas forcément rôdés à l'utilisation de moyens informatiques. C'est pourquoi le développement d'un

[35] http://www.vnexpress.net
[36] http://www-clips.imag.fr/geod/User/viet-bac.le/outils/

utilitaire portable de gestion et d'acquisition de grands corpus sur un matériel standard, nous est d'un grand bénéfice. Le logiciel respecte le format SAM (*Speech Assessment Methodologies*) de définition de bases de signaux. Les interfaces ont été adaptées pour manipuler respectivement les caractères vietnamiens et khmers. La figure 8 montre un exemple de l'interface d'EMACOP adaptée à la langue khmère.

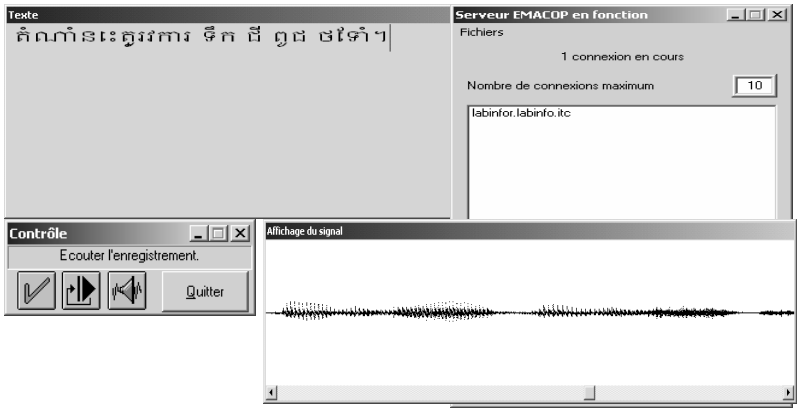


Figure 8 : Interface EMACOP adaptée au khmer

Dictionnaire de prononciation et modèles acoustiques

Dictionnaire de prononciation

Un dictionnaire de prononciation (ou dictionnaire phonétique) est une ressource essentielle pour les tâches de synthèse et de reconnaissance de la parole, ou tout simplement pour enrichir un dictionnaire bilingue, permettant au locuteur étranger de connaître la prononciation du mot en langue cible. Cette tâche est cependant difficile pour des langues peu dotées dont le système phonologique est parfois méconnu, ou sujet à débats (langues peu ou mal décrites). Si nous mettons de côté les méthodes manuelles de phonétisation qui, bien que donnant les dictionnaires de prononciation de meilleure qualité, ne nous semblent pas entrer dans le cadre de notre méthodologie, on peut distinguer trois types d'approche automatique pour constituer un dictionnaire phonétique dans une nouvelle langue :

- Des approches à base de règles, qui nécessitent une bonne connaissance de la langue et de ses règles de phonétisation (qui par ailleurs ne doivent pas connaître trop d'exceptions). Ce type d'approche est assez coûteux en temps (écriture d'un analyseur phonétique), mais donne des dictionnaires de prononciation de qualité très correcte pouvant ensuite être révisés manuellement relativement rapidement.

- Des approches utilisant un système de reconnaissance phonémique appliqué à des enregistrements des mots à phonétiser, permettant un premier étiquetage automatique en phonèmes d'une liste de mots, qui peut être alors révisée par un opérateur humain. L'avantage d'une telle approche est bien sûr sa rapidité. Ses inconvénients sont qu'elle nécessite l'emploi d'un système de reconnaissance automatique de phonèmes qui sera généralement celui d'une langue source bien dotée (par exemple un système de reconnaissance des phonèmes du français) ; l'autre défaut est que les unités phonémiques décrivant les mots en langue cible seront seulement celles pouvant être reconnues par le décodeur en langue source, d'où la nécessité d'employer si possible des décodeurs phonémiques multilingues pour augmenter au maximum la couverture phonémique dans l'alphabet phonétique international (API). La figure 9 illustre ce problème : la langue source utilisée est le français tandis que la langue cible est le vietnamien. Il est évident que la couverture du vietnamien par le français n'est pas du tout optimale (taux de couverture

d'environ 63%). Une telle méthode reste cependant intéressante, notamment lorsqu'on passe d'une langue source à une langue cible qui possède un système phonétique proche.

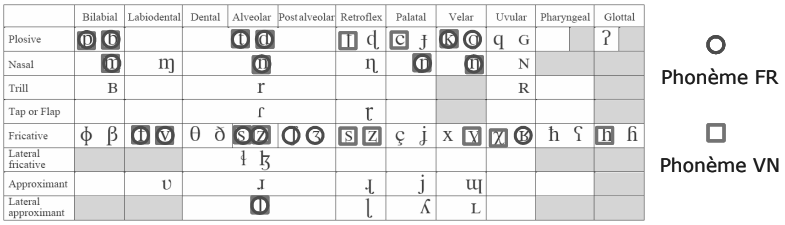


Figure 9 : couverture phonémique du français et du vietnamien pour les consonnes

- Des approches utilisant une représentation orthographique (graphème) d'un mot à la place du phonème comme unité de modélisation acoustique [Killer 2003b], [Stuker 2004]. Pour le système de reconnaissance automatique de la parole à base de graphèmes, la représentation d'une entrée lexicale (un mot) dans le vocabulaire est alors une suite de graphèmes. Par conséquent, la construction automatique d'un dictionnaire de prononciation devient très simple.

Modèles acoustiques

Nous avons vu précédemment qu'il existe des solutions pour collecter rapidement des ressources orales et écrites dans une nouvelle langue. Dans l'idéal, si ces ressources sont en grande quantité, et si un dictionnaire de prononciation est disponible pour la langue cible, l'adaptation du système de reconnaissance peut correspondre alors à un simple réapprentissage des modèles sur ces nouvelles données. Dans la réalité, la quantité de données collectées reste bien souvent inférieure à ce qu'elle est pour les langues bien dotées. La construction d'un système de reconnaissance automatique de la parole nécessite donc également des techniques d'adaptation rapide au niveau des modèles acoustiques comme cela est proposé dans [Schultz, 2006] par exemple.
Une approche possible consiste à obtenir un tableau de correspondances phonémiques (*phone mapping*) entre une ou plusieurs langues sources, et la langue cible. Ensuite, les modèles acoustiques des phonèmes en langue source peuvent être dupliqués pour obtenir des modèles acoustiques initiaux en langue cible. L'avantage d'une telle approche est qu'elle ne nécessite pas ou peu de signaux d'apprentissage en langue cible puisque les modèles acoustiques du système de reconnaissance en langue cible sont en fait ceux d'une autre langue. Cependant, on retrouve dans cette approche les mêmes défauts que ceux mentionnés dans le deuxième point du paragraphe précédent, à savoir le problème de la couverture phonémique (i.e. reconnaître du vietnamien avec des modèles acoustiques appris sur du français !). De tels systèmes peuvent cependant être améliorés en adaptant, par exemple, les modèles acoustiques avec une quantité réduite de signaux en langue cible.
Le problème est aussi d'obtenir le fameux tableau de correspondances phonémiques entre langue cible et langue source. Pour cela, on distingue les méthodes manuelles à base de connaissances (*knowledge-based*), et les méthodes automatiques à base de données (*data-driven*). Les méthodes manuelles consistent à chercher les couples de phonèmes source/cible les plus proches dans le tableau d'API et nécessitent des connaissances acoustiques et phonétiques dans les deux langues (source et cible). Une approche automatique consiste plutôt à disposer d'un corpus vocal en quantité limitée en langue source et étiqueté, puis à utiliser un décodeur phonémique et à calculer la matrice de confusion entre les phonèmes reconnus en langue source et les phonèmes de référence en langue cible.
Une description plus détaillée des traitements réalisés et des expérimentations associées pour l'adaptation rapide de modèles acoustiques au vietnamien indépendants du contexte se trouve dans [Le, 2005]. Récemment, ces travaux ont été enrichis par la proposition de méthodes automatiques d'estimation de similarité (en utilisant le tableau IPA) entre des unités phonémiques différentes

telles que : le phonème (monophone), le groupe de phonèmes, le polyphone, etc. Ces travaux, menés en collaboration T. Schultz du laboratoire ISL/CMU, ont été publies dans [Le, 2006].

3.2.4 Application au vietnamien

Précisons d'abord que la reconnaissance automatique de mots isolés en vietnamien avait déjà été abordée au CLIPS dans le cadre de la thèse de C. Nguyen [Nguyen, 2002] que j'ai co-encadrée avec E. Castelli. Cependant, ce travail de thèse insistait surtout sur la reconnaissance des tons de la langue vietnamienne, tandis que les travaux décrits dans cette section sont, à ma connaissance, les premières expériences en reconnaissance automatique de la parole continue grand vocabulaire en vietnamien.

Ressources collectées

Notre méthodologie a été appliquée au vietnamien. La quantité de pages Web collectées était de 2.5Go. Après filtrage, la quantité de données textuelles pouvant servir à l'apprentissage d'un modèle de langage statistique était d'environ 400Mo (5 millions de phrases). A titre de comparaison, une année complète du journal Le Monde en français correspond à 120Mo en moyenne.
Ensuite, nous avons recueilli un vocabulaire de 6 686 syllabes et un vocabulaire de 40 000 mots à partir des ressources lexicales existantes. Un analyseur phonétique à base de règles a été développé pour obtenir automatiquement un dictionnaire de prononciation vietnamien (VNPhoneAnalyzer, voir [Le, 2004]). Ce dictionnaire phonétique a ensuite été vérifié par des experts de l'Institut Linguistique du Vietnam.
Par ailleurs, un corpus de parole vietnamien est toujours en cours d'enregistrement à MICA. A ce jour, il contient 35 locuteurs, 16 femmes et 19 hommes, venant des régions nord, centre et sud du Vietnam. Chaque locuteur a enregistré environ 1 heure de parole ce qui fait un total de 35 heures. Le corpus contient des séquences de lettres, de nombres et de mots isolés, mais aussi la lecture de phrases complètes et de paragraphes.
Des détails supplémentaires sur les ressources collectées pour le vietnamien se trouvent dans [Le, 2004].

Quelques expériences de reconnaissance automatique du vietnamien

Pour la modélisation acoustique, nous avons utilisé nos méthodes de portage et d'adaptation rapide des modèles acoustiques multilingues vers une langue cible peu dotée. Les résultats obtenus montrent le potentiel des méthodes de portage que nous avons proposées, notamment dans le contexte de langues peu dotées ne possédant pas ou peu de ressources acoustiques.
Pour initialiser le système de reconnaissance du vietnamien, deux modèles en langue source différents ont été utilisés :
-des modèles acoustiques appris sur le français (couverture phonémique français / vietnamien d'environ 63%),
-des modèles acoustiques multilingues (travail en collaboration avec CMU lors du séjour de Viet-Bac Le à Pittsburgh en avril 2005) obtenus à partir de 7 langues : chinois, croate, français, allemand, japonais, espagnol et turc (couverture phonémique multilingue / vietnamien d'environ 87%),
Nous avons testé deux techniques d'obtention du tableau de correspondances phonémiques : à base de connaissances (*IPA*) et à base de données (*data driven*). Les performances du système de reconnaissance automatique de parole continue du vietnamien testé sur un corpus d'une heure de dialogues sont présentées dans le tableau 6 avec respectivement l'utilisation de 1h et 2h de signal en langue vietnamienne pour adapter les modèles acoustiques initiaux.

Système source	Models	Adapt 1h	Adapt 2h
		WA	WA
Français	IPA	60.4	63.6
	Data Driven	61.6	63.8
Multilingue (CMU Global Phone, 7 langues)	IPA	64.6	66.3
	Data Driven	63.8	65.3

Tableau 6 : performances (% taux de reconnaissance de syllabes) de notre système de reconnaissance du vietnamien en fonction de la quantité de signaux d'adaptation utilisée et de la méthode de génération des correspondances phonémiques

Ces résultats montrent le potentiel de l'approche automatique fondées sur les données (*data driven*) pour la génération du tableau de correspondances phonémiques car elle donne des performances équivalentes à celles obtenues avec la méthode manuelle (IPA). Nous voyons également qu'avec une quantité réduite de signaux en langue cible (2h), il est possible d'obtenir des performances acceptables (63.8% de mots correctement reconnus à partir d'un modèle français et 66.3% à partir d'un modèle multilingue).

Des résultats plus détaillés sur la reconnaissance de la parole en vietnamien se trouvent dans la thèse de Viet-Bac Le [Le, 2006b] et dans [Besacier, 2006][37]. Il y est également proposé une modélisation acoustique à base de graphèmes qui montre un potentiel intéressant dans le cas où aucun dictionnaire phonétique n'est disponible ou ne peut être construit.

3.2.5 Application au khmer

Ressources collectées

Des ressources textuelles ont été recueillies et l'enregistrement d'un corpus en langue khmère est en cours à l'ITC. A l'aide d'étudiants cambodgiens, nous avons cherché un nombre réduit de pages Web publiées par le gouvernement cambodgien, par des organisations ou des compagnies. Nous avons d'abord remarqué que beaucoup de sites web hébergés au Cambodge sont en fait écrits en anglais ou en français. Il y a cependant quelques sites écrits en langue khmère. Avec ceux-ci, il y a encore des difficultés de récupération automatique: sites écrits en flash[38], pages encodées dans un système d'encodage spécifique ou privé[39] que nous n'avons pas réussi à convertir en un autre encodage (Unicode). Nous avons trouvé cependant un site des nouvelles en khmer[40].
Ainsi, un corpus de documents html de 174Mo en khmer a été collecté dans un premier temps. Après filtrage et traitement, le corpus de texte obtenu était d'environ 97 Mo, soit 1,1 million de phrases ou 8 millions de mots segmentés. Cela reste bien sûr encore faible par rapport aux 2.5Go de pages web collectées pour le vietnamien et par rapport aux 40Go pour le français (corpus WebFR4) [Vaufreydaz, 2002].
Concernant les modèles de langage, nous avons identifié deux pistes possibles : la première consisterait à construire un système de reconnaissance syllabique où les co-occurrences modélisées seraient des suites de syllabes ; la seconde consisterait à construire un système de reconnaissance de mots où les cooccurrences modélisées seraient des suites de mots. Dans les deux cas, se pose le problème de la segmentation (en syllabes ou en mots) qui est difficile pour les systèmes d'écriture de langues comme le thai et le khmer. Il existe des travaux sur ces langues pour résoudre ce problème : utilisation de grammaires de syllabes [Berment, 2004], méthodes basées sur un vocabulaire, méthodes probabilistes [Meknavin, 1997], etc.

[37] papier invité ICASSP06, session spéciale 'speech to speech translation'
[38] http://www.everyday.com.kh
[39] http://www.seasite.niu.edu/khmer/
[40] www.cambodiacic.org

En appliquant la boîte à outils de traitement de corpus de texte, nous avons obtenu dans un premier temps un corpus de phrases à prononcer afin d'effectuer des enregistrements. Pour cela, nous avons utilisé un vocabulaire de 16000 mots pour filtrer les phrases destinées à être prononcées et enregistrées. Ce vocabulaire a été obtenu à partir du dictionnaire khmer Chuon Nat [41]. Le corpus de parole khmer est en cours d'enregistrement à l'ITC. A ce jour, 3 heures de parole ont été enregistrées par 10 locuteurs phnom-penhois.

Quelques expériences de reconnaissance automatique du khmer
Au moment où nous avons construit le système de reconnaissance automatique de la parole en khmer, aucun dictionnaire phonétique n'existait dans la communauté de traitement de la langue khmère. Ainsi, nous avons choisi le graphème comme unité de modélisation acoustique. Un dictionnaire de prononciation à base de graphèmes a été généré par une procédure de romanisation. Ainsi, un premier système de reconnaissance a été achevé en 4 mois et un taux de reconnaissance de mots d'environ 80% (sur une tâche de parole lue cependant) en mode dépendant du locuteur a été obtenu, ce qui montre l'efficacité de notre méthodologie et des outils développés.

Des résultats plus détaillés sur la reconnaissance de la parole en khmer se trouvent dans la thèse de Viet-Bac Le [Le, 2006b] ainsi que dans [Le, 2006c].

3.3 Résumé des travaux réalisés au centre de recherche d'IBM : traduction de parole irakien - anglais (projet DARPA TRANSTAC)

Cette section présente un résumé des travaux que j'ai réalisés au sein du département « *Speech and Language Technologies* » du centre de recherche d'IBM Watson (Yorktown Heights, NY), dans l'équipe de traduction de parole (MASTOR[42]) dirigée par Y. Gao.

3.3.1 Le projet TRANSTAC

Présentation

Le projet TRANSTAC[43] (*Spoken Language Communication and TRANSlation System for TACtical Use*) est un projet financé par l'agence américaine DARPA. Son but est de développer des technologies permettant à des combattants américains de communiquer avec des locuteurs natifs par le biais d'un système de traduction de parole bidirectionnel. Le projet se consacre notamment aux problèmes posés par le déploiement rapide de systèmes de traduction entre l'anglais et de nouvelles langues cibles peu dotées ou des dialectes. Plus précisément, dans la première phase de TRANSTAC, un système de traduction bidirectionnel anglais – arabe dialectal (le dialecte parlé en Irak) a dû être développé, avec une première évaluation au bout de trois mois, puis deux autres évaluations après six et neuf mois respectivement[44]. Un autre aspect important du projet est l'évaluation de prototypes de traduction sur des terminaux indépendants et légers de type PDA[45].

[41] http://www.khmeros.info/
[42] http://domino.research.ibm.com/comm/research_projects.nsf/pages/mastor.index.html
[43] http://transtac.mitre.org/
[44] Outre la fréquence des évaluations, une autre difficulté dans le projet a été de gérer l'arrivée continue de nouvelles données (parole ou texte) nécessitant une mise à jour régulière des modèles.
[45] Voir par exemple un exemple de terminal sur http://www.phraselator.com/

La problématique de ce projet est donc proche de celle déjà évoquée dans ce manuscrit au sujet des langues peu dotées. En effet, le dialecte parlé en Irak est assez éloigné de l'arabe standard[46] qu'on retrouve dans les journaux télévisés par exemple. Ainsi, pour ce dialecte Irakien essentiellement oral, on trouvera peu de données textuelles en dehors des transcriptions de conversations. C'est donc surtout le manque de données écrites qui fait de ce dialecte une langue « peu dotée » car, en revanche, la quantité de données orales collectées et annotées pendant le projet TRANSTAC est très importante (plus de 200 heures de parole transcrite). Par ailleurs, contrairement aux langues abordées dans les sections précédentes, ce dialecte est lié à une langue bien dotée : l'arabe standard ou MSA, pour lequel il existe une grande quantité de ressources disponibles. Nous verrons par la suite si de telles ressources issues d'une langue bien dotée (l'arabe standard ou MSA) peuvent être ou non avantageusement utilisées pour une langue peu dotée de la même famille (l'arabe dialectal parlé en Irak). On trouvera par ailleurs quelques références sur le traitement automatique de l'arabe dialectal dans [Schultz, 2006] et [Kirchoff, 2004].

Arabe dialectal et arabe standard

Pour illustrer cette différence entre Irakien et arabe standard (MSA), je décris ici quelques expériences de reconnaissance automatique de la parole réalisées au début de mon séjour. A ce moment là, la quantité de données disponibles pour l'apprentissage des modèles de langage était de 7Mo seulement (130k phrases correspondant à des transcriptions de signaux).
Dans un premier temps, j'ai évalué sur un ensemble de test Irakien les performances des modèles acoustiques MSA préexistants (en utilisant pour toutes les expériences un même dictionnaire de prononciation et un même modèle de langage appris sur les données textuelles disponibles). Le taux d'erreur obtenu ainsi est très élevé (62.7%WER), et il demeure important même lorsque les modèles initiaux MSA sont adaptés avec des signaux Irakien (51.3%WER).
A titre de comparaison, plus tard dans le projet, notre meilleur système obtiendra un taux d'erreur d'environ 30% sur le même ensemble de test. Ce premier résultat confirme les différences entre arabe littéral et irakien au niveau acoustique et suggère que des techniques allant au delà du simple réapprentissage ou de l'adaptation doivent être proposées pour obtenir des résultats acceptables.

Dans un second temps, j'ai essayé d'évaluer quel pouvait être l'apport de données textuelles MSA (disponibles en grande quantité) pour la modélisation du langage de l'irakien. Pour cela, j'ai appris deux modèles de langages trigrammes différents : l'un entraîné sur les données textuelles Irakiennes mentionnées ci-dessus (130k phrases), l'autre entraîné sur une grande quantité de données en arabe littéral (environ 50 fois plus importante que les données Irakiennes). La mesure de l'adéquation entre ces modèles et un corpus de test irakien, en utilisant le pourcentage de trigrammes communs (% *trigram hits*), donne respectivement 41% pour le modèle appris sur les données dialectales et 2.7% seulement pour le modèle appris sur les données MSA. Ce dernier chiffre, très faible, semble indiquer que même en très grande quantité, les données MSA ne sont d'aucune utilité pour la modélisation statistique de l'arabe dialectal irakien[47]. Cela est dû, d'une part, aux différences de domaines traités dans les deux types de données (conversation *versus* news), et d'autre part au fait que l'arabe dialectal Irakien est tellement éloigné de l'arabe standard que les deux peuvent être quasiment vus comme des langues différentes du point de vue des modèles statistiques de langage. Cette relative 'inutilité' des données MSA pour la modélisation du langage en arabe dialectal est confirmée dans [Kirchhoff, 2002] pour l'arabe parlé au Moyen-Orient (*levantine Arabic*).

[46] Dans les publications sur le domaine, on utilise l'acronyme MSA (*Modern Standard Arabic*) pour designer l'arabe littéral.
[47] Des mesures de perplexité avec les mêmes données confirment ce résultat et l'interpolation entre les deux modèles littéral (MSA) et dialectal n'apporte rien non plus.

Une idée des performances de traduction obtenues a l'issue de la première phase du projet

Le tableau 7 donne une idée des performances de reconnaissance (WER) et de traduction (BLEU) obtenues en mars 2006 lors de l'évaluation finale de la phase 1 du projet TRANSTAC. Mes contributions sur la reconnaissance de l'Irakien (en gras dans la table) sont décrites plus précisément dans la section suivante. Des détails sur le système complet de traduction anglais – Irakien peuvent être trouvés dans [Gao, 2006].

Participants		**IBM**	SRI	SpeechGear	BBN	Sehda	CMU
ASR WER	E	7.66%	16.72%	20.44%	6.14%	9.01%	5.70%
	I	**13.04%**	28.05%	14.83%	25.70%	38.41%	51.74% (**15.79%)
S2T BLEU	E->I	0.17403	0.08741	0.01480	0.18020	0.12297	0.13277
	I->E	0.41584	0.36114	0.33575	0.36095	0.34410	0.2343 (**0.35649)
T2T BLEU	E->I	0.19410	0.12949	0.03145	0.19524	0.11281	0.14494
	I->E	0.44822	0.43478	0.48465	0.42393	0.43217	0.39091

Tableau 7 : Performances de reconnaissance et de traduction obtenues en mars 2006 lors de l'évaluation finale de la phase 1 du projet TRANSTAC (E=anglais, I=Irakien, T2T=traduction à partir des références, S2T=traduction de parole complète à partir des sorties de reconnaissance)

3.3.2 Mes contributions en reconnaissance automatique de l'arabe dialectal irakien

Modélisation du langage

Comme nous l'avons vu dans la section précédente, la quantité de données disponibles pour la modélisation de l'arabe dialectal irakien est très faible (transcription de données enregistrées seulement) et les données en arabe standard MSA (disponibles, elles, en grande quantité) ne se sont pas avérées utiles. Ce manque de données est amplifié par la morphologie très riche de l'arabe. En effet, en raison du grand nombre d'affixes pouvant être attachés à un même mot, la taille du vocabulaire est généralement plus importante que pour les autres langues si on considère un corpus d'apprentissage de même grandeur. Cela est illustré par la *figure 10* empruntée à [Kirchhoff, 2002]. Cette richesse de la morphologie de l'arabe conduit donc à des taux de mots hors vocabulaire plus élevés et dégrade la qualité de l'estimation des modèles statistiques de langage pour la reconnaissance automatique de la parole ou la traduction.

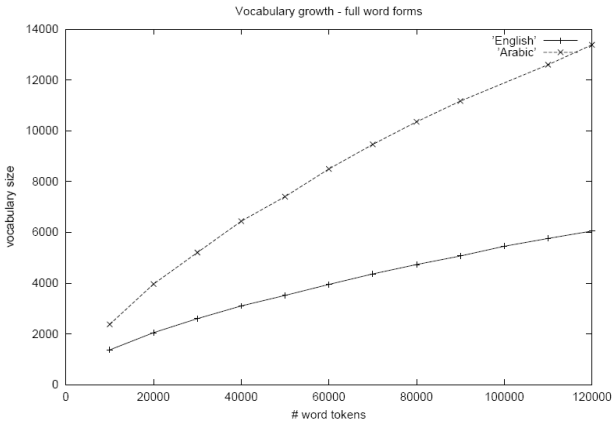


Figure 10 : augmentation de la taille du vocabulaire en fonction de la taille du corpus pour l'arabe et l'anglais

Un moyen de traiter ce problème est d'utiliser un analyseur morphologique qui segmente chaque mot en une séquence du type *préfixe-stem-suffixe*. Cependant, les analyseurs morphologiques de l'arabe standard disponibles, tels que celui de Tim Buckwalter[48], sont peu performants sur le dialecte irakien qui présente des différences dans le choix des affixes notamment. Pour cette raison, j'ai développé un analyseur morphologique qui ne nécessite pas de connaissances préalables sur la langue (liste de préfixes ou suffixes). Le système est entraîné sur des données irakiennes segmentées (ces données ont été distribuées par LDC à tous les participants du projet TRANSTAC), et son implémentation consiste en une cascade de transducteurs à l'aide de la bibliothèque de machines à états finis d'IBM[49].
Cet analyseur morphologique[50] a été ensuite utilisé pour segmenter les données d'apprentissage du modèle de langage Irakien. Cependant, les premiers résultats de reconnaissance de parole obtenus en apprenant un modèle de langage directement sur ces données segmentées montrent une dégradation des résultats par rapport à un modèle non morphologique classique. La raison est que cette segmentation trop « agressive » réduit le contexte pris en compte par les modèles de langage n-grammes. Par exemple, après notre analyse morphologique, un mot comme *AllAEbAn*[51] (« les deux joueurs ») sera décomposé en *Al+lAEb+An*, ainsi la couverture d'un trigramme se limite au mot initial dans ce cas. Ce problème est également mentionné dans [Xiang, 2006] pour l'arabe standard. Pour rendre cette segmentation moins agressive et pour réduire l'influence des erreurs d'analyse, j'ai donc ajouté les règles suivantes :
-les préfixes et les suffixes obtenus après segmentation automatique doivent appartenir à une liste d'affixes prédéfinie au départ,
-la partie centrale (*stem*) doit contenir au moins deux caractères,
-le mot ne peut être décomposé morphologiquement que s'il n'appartient pas à la liste des N mots les plus fréquents.
Cette dernière règle est la plus importante puisqu'elle permet de rendre les mots les plus fréquents non-décomposables, assurant ainsi une meilleure couverture des modèles de langage n-grammes. Par ailleurs, le paramètre N permet de faire varier la taille du vocabulaire et le degré de segmentation. La *figure 11* présente les performances de reconnaissance obtenues sur un corpus d'arabe parlé en utilisant des modèles de langage appris sur un même corpus textuel segmenté avec différentes valeurs de N.

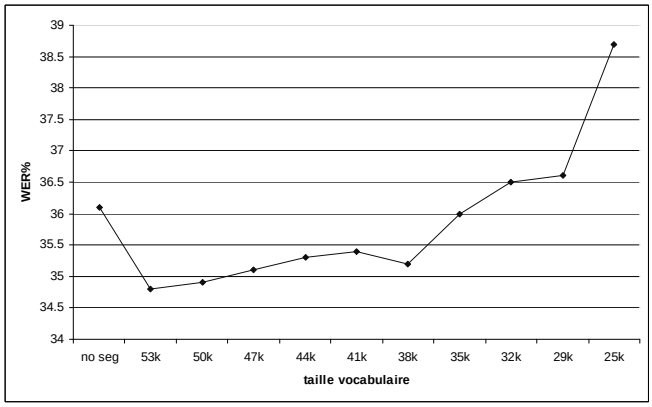


Figure 11 : performance de reconnaissance (WER%) en fonction de la taille du vocabulaire obtenu pour différents degrés d'analyse morphologique (*no seg* = pas de segmentation morphologique)

[48] Voir http://www.qamus.org/morphology.htm
[49] Cette bibliothèque est très proche de la bibliothèque d'ATT présentée dans [Mohri, 1997].
[50] L'évaluation de cet analyseur sur un corpus de 1500 phrases donne un taux d'erreur de mots d'environ 12%
[51] J'utilise ici une translittération de caractères arabes dérivée de Buckwalter (http://www.qamus.org/transliteration.htm)

Cette figure montre que, en utilisant les règles décrites ci-dessus, un modèle de langage « morphologique » améliore les performances de reconnaissance par rapport à un modèle de mots classique. Par ailleurs on peut, par cette méthode, conserver des performances de reconnaissance pratiquement équivalentes tout en réduisant la taille du vocabulaire de 64k à 29k, ce qui est très intéressant pour l'implémentation d'algorithmes sur des terminaux légers type PDA ayant de faibles ressources mémoire.
Cette modélisation « morphologique » du langage ayant donné des résultats prometteurs, elle a été implémentée, sous une forme légèrement différente, dans le prototype de traduction de parole présenté lors de l'évaluation finale de la phase 1 du projet TRANSTAC. Les deux méthodes et les résultats expérimentaux associés sont publiés dans [Afify, 2006].

Modélisation acoustique

Comme nous l'avons vu dans la section 3.3.1, les modèles acoustiques MSA donnent de piètres performances sur l'arabe dialectal irakien. Nous avons donc utilisé les 200 heures de données irakiennes transcrites disponibles pour entraîner de nouveaux modèles acoustiques. Cependant, un des problèmes de l'arabe est que les voyelles courtes, normalement indiquées par des diacritiques, sont le plus souvent omises dans les textes écrits (et par conséquent dans les transcriptions des signaux de parole dont nous disposons). Elles sont par ailleurs fortement dépendantes du contexte lexical, ce qui rend difficile leur inférence automatique. Le problème qui se pose alors pour la reconnaissance automatique de la parole en langue arabe est qu'il est difficile de connaître précisément la séquence de phonèmes correspondant à une transcription, rendant peu précise l'estimation des modèles acoustiques pour les voyelles courtes.
Une première approche (*système graphémique*) consiste à simplement ignorer cette information en utilisant un dictionnaire de prononciation à base de graphèmes comme celui déjà introduit dans la section 3.2.3. Une autre approche (*système voyellé*) consiste à inclure l'information correspondant aux voyelles courtes dans le dictionnaire de prononciation sous la forme de variantes. Une même forme peut cependant avoir un grand nombre de prononciations différentes : par exemple, la racine *ktb* a une vingtaine de diacritiques possibles. Dans ce but, un voyelleur automatique, réentraîné sur des données Irakiennes voyellées, a été utilisé pour générer les variantes de prononciation. Ses performances sont cependant insuffisantes pour apprendre des modèles acoustiques plus efficaces que les modèles à base de graphèmes, comme le montre le tableau 8. En revanche, un apprentissage discriminant de type MPE (*Minimum Phone Error* [Povey, 2002]) a permis également d'améliorer les performances.

Système	%WER
Voyellé	40.7
graphémique	37.9
graphémique + MPE	36.1

Tableau 8 : Performances de reconnaissance pour différentes modélisations acoustiques (apprentissage sur 200 heures d'arabe dialectal) sur une même base de test irakienne

Un autre problème pour l'apprentissage des modèles acoustiques est le fait que les transcriptions manuelles peuvent contenir des erreurs. Ces erreurs peuvent conduire à l'échec du processus d'alignement forcé, rendant inutilisables pour l'apprentissage les signaux mal transcrits. Ce problème est particulièrement critique pour les langues peu dotées pour lesquelles la quantité de données disponibles est plus faible et la qualité plus aléatoire (difficultés pour trouver de bons annotateurs, langues moins décrites et moins normalisées, …). Pour cette raison, j'ai également utilisé une procédure d'apprentissage non supervisé comme cela a déjà été proposé pour la transcription de journaux télévisés [Chan, 2004]. Toutes les données d'apprentissage ont été

transcrites automatiquement avec le meilleur modèle acoustique disponible et un modèle de langage appris sur les transcriptions manuelles. Les transcriptions automatiques ont alors été utilisées pour remplacer les transcriptions manuelles en cas échec de l'alignement forcé. Grâce à cette procédure, le pourcentage de signaux rejetés lors de l'alignement forcé passe de 12.8% à 7.6% et le nouveau modèle acoustique obtenu présente un gain absolu de 0.7% du taux d'erreur de mots.

3.3.3 Traduction de parole pour les langues peu écrites

Ce paragraphe décrit des travaux plus exploratoires, toujours réalisés à IBM, pour la traduction de parole pour des langues peu écrites.
D'après [Schultz, 2006], de nombreuses langues du monde sont essentiellement orales et n'ont pas de forme écrite répandue : par exemple, il est dit que seulement 10% des langages du monde utilisent l'un des 25 systèmes d'écriture connus. Ce grand nombre de langues peu écrites est confirmé par [Nettle, 2000]. Parmi ces langues, on retrouve d'une part des langues indigènes n'ayant pas de tradition écrite, et d'autre part des dialectes utilisés pour la communication orale uniquement. Les langues chinoises et arabes ont un grand nombre de dialectes régionaux peu écrits qui diffèrent significativement de la langue d'origine. Dans ce cas, comme nous avons pu le voir au paragraphe 3.3.1 pour le dialecte parlé en Irak, la collecte de ressources textuelles pour le traitement automatique est plus difficile puisque le seul moyen d'obtenir des données est de transcrire des conversations orales. Cela peut être fait en définissant des règles de transcription du langage oral s'inspirant de la lange d'origine, comme cela a été proposé dans le projet TRANSTAC. Cependant, comment transcrire des dialectes ou des langues qui ne sont liés à aucun standard d'écriture ? Est il impossible de développer des systèmes de traitement du langage naturel pour de telles langues ? Nous pensons plutôt que pour une tâche comme la traduction de parole, la forme écrite peut, sous certaines conditions, être considérée comme secondaire. Les expériences décrites dans cette section constituent notre tentative pour montrer qu'une telle hypothèse est plausible.
Ainsi, nous avons essayé de construire un système de traduction de parole qui utilise un corpus parallèle constitué de signaux en langue source (la langue non écrite depuis laquelle on souhaite traduire) et de leur traduction correspondante en langue anglaise (la langue cible vers laquelle on souhaite traduire). En plus, nous faisons l'hypothèse que les signaux en langue source on été transcrits en phonèmes (en utilisant par exemple l'alphabet phonétique international). Bien sûr, une telle transcription phonétique manuelle est plus longue à produire qu'une transcription en mots (une mesure faite pour le français suggère que la transcription phonétique est environ 3 à 5 fois plus lente que la transcription en mots, pour un locuteur natif) ; le travail d'un annotateur peut cependant être réduit en appliquant par exemple sur les données à transcrire un reconnaisseur phonétique multilingue. Dans un premier temps, nous avons comparé deux approches pour la traduction de parole du dialecte Irakien : une approche *état-de-l'art* pour laquelle un modèle de traduction est appris à partir de données parallèles : mots anglais / mots Irakiens ; et une approche *phonétique* pour laquelle un modèle de traduction est appris à partir de données parallèles : mots anglais / phonèmes Irakiens. Dans les deux cas, la méthode de traduction est une approche statistique fondée sur des séquences (*phrase-based approach* [Koehn, 2003]) qui repose sur l'apprentissage d'une table de traduction (*phrase table*) contenant des séquences de mots en langue cible (*e*) , leur traduction en langue source (*f*) et une valeur de probabilité (*p(f/e)*). La phase de décodage consiste alors à trouver, à partir d'une phrase inconnue *f* , la phrase *e* qui maximise l'expression :

$$argmax(p(e/f))=argmax(p(f/e).p(e))\ .$$

La préparation des données d'apprentissage des modèles de langage et des modèles de traduction pour notre approche phonétique, est décrite dans la *figure 12*.

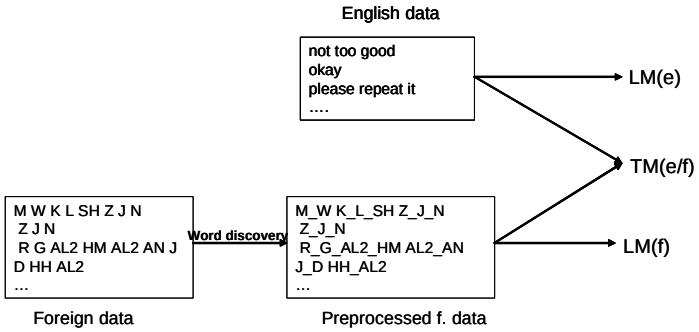


Figure 12 : Préparation des données pour l'apprentissage des modèles de langage et de traduction dans notre approche *phonétique* (LM : modèle de langage ; TM : modèle de traduction)

L'apprentissage direct d'un modèle de traduction entre les phrases en langue cible et les séquences de phonèmes en langue source donne des résultats peu satisfaisants. Pour cette raison, une procédure d'agglomération de séquences phonétiques (ou *word discovery*) non supervisée a été développée et appliquée aux séquences phonétiques avant l'apprentissage des modèles. Elle est plus précisément décrite dans [Besacier, 2006b]. Un signal inconnu est alors décodé en utilisant un modèle acoustique appris grâce aux signaux en langue source et leur transcription phonétique, et un modèle de langage *LM(f)* appris sur les données préparées, comme l'illustre la *figure 12*. Ce modèle de langage permet d'améliorer très significativement les taux d'erreur de phonèmes par rapport à un simple décodeur phonétique de la langue source non écrite (15.1% d'erreur de phonèmes contre 42.6% pour un décodeur phonétique n'utilisant aucun modèle de langage).

Des détails expérimentaux sur cette étude préliminaire sont disponibles dans [Besacier, 2006b]. Le tableau 9 compare les performances de traduction de parole source – cible (irakien – anglais) pour la méthode état-de-l'art et la méthode phonétique sur trois ensembles de test représentant 800 signaux au total (les données d'apprentissage correspondent quant à elles à un corpus parallèle de 366k phrases).

Approche	Test set 1	Test set 2	Test set 3
Etat-de-l'art	0.43583	0.31007	0.31486
Phonétique	0.45178	0.28640	0.23677

Tableau 9 : Performances de traduction de parole (BLEU) obtenues pour deux approches différentes (issu de [Besacier, 2006b]).

Ces premiers résultats montrent la faisabilité d'une approche *phonétique* dont les performances sont juste légèrement inférieures à celles de l'approche *état-de-l'art*. Une évaluation subjective des sorties de traduction confirme la tendance, puisque 58% des phrases sont jugées correctement traduites par la méthode *état–de-l'art* tandis que 54% des phrases sont jugées correctement traduites par la méthode *phonétique*. Il est aussi intéressant de constater que ces méthodes semblent également complémentaires, puisque 64% des phrases sont jugées correctement traduites par au moins l'une des deux approches. Par ailleurs, et c'était le but principal de cette expérience, la méthode phonétique est applicable, en théorie, à n'importe quelle langue peu écrite.

Chapitre 4 : Quelques applications

4.1 La recherche d'information multimédia

4.1.1 Introduction

La recherche d'information (RI) est l'accès par le contenu à des documents satisfaisant les besoins d'information des utilisateurs[52]. Lorsqu'on s'intéresse à la RI multimédia, la difficulté pour des documents audio et vidéo [Berrut 1997], par rapport aux documents textuels, réside dans la distance entre le support de représentation (le signal numérique) et les informations contenues dans le document. Un document textuel est représenté par une chaîne de caractères. Les informations sémantiques sont donc plus facilement extraites à partir des mots et des phrases. Un document audio et vidéo est représenté quant à lui par une chaîne d'octets issus de la numérisation de signaux analogiques, son contenu sémantique est dans ce cas plus difficile à extraire. Malgré les possibilités offertes par le format numérique pour y ajouter des informations sémantiques, la représentation de documents audio et vidéo a été développée, dans un premier temps, sans réellement envisager la possibilité de manipuler l'information selon le contenu.

Je me suis intéressé à l'application de mes travaux en *transcription enrichie* à l'indexation et à la recherche de documents vidéos. Dans ce domaine, les applications concernent l'exploitation efficace de grandes bases de vidéos numériques (archives INA, web, collections cinématographiques privées, vidéo-conférences, archives de l'assemblée nationale...). J'ai commencé à travailler sur ce type de données dans le cadre du projet BQR INPG[53] « vidéo sémantique » dans lequel j'étais impliqué en collaboration avec les laboratoires LIS et LSR[54] de 2002 à 2004.

Jusqu'à maintenant, des solutions partielles ont été proposées pour l'indexation et l'interrogation de grandes bases de vidéos, à partir des modalités séparées [Smeaton 2001].

• audio : transcription par reconnaissance automatique de la parole et utilisation de mots-clés, détection et reconnaissance de catégories non linguistiques (locuteurs, bruits, musique, émotions...) [Nwe, 2005].

• image (animée) : classification par couleurs, textures, mouvement ou reconnaissance de formes [Assfalg, 2002].

• texte : sous-titres ou reconnaissance automatique de textes à partir de la bande image.

Les principales limitations de ces approches, en plus du *fossé sémantique* déjà mentionné pour les modalités audio et vidéo, résident dans l'hétérogénéité des éléments d'indexation, dans les erreurs de reconnaissance ou de classification, et dans les ambiguïtés.

Par ailleurs, le rapprochement entre les domaines du traitement automatique de la parole et de la recherche d'information a donné lieu à de nouveaux travaux tels que :

-résumé de parole [Zechner, 2003] (*speech summarization*), où de nouveaux verrous apparaissent, par rapport au résumé sur des documents textuels, en raison de la nature spontanée (hésitations, disfluences, locuteurs multiples) et continue (pas de séparateurs de phrases clairs) de la parole ; en raison aussi des erreurs sur la transcription automatique,

-fouille de voix [Gazit, 2004] (*voice mining*), où l'on recherche des locuteurs particuliers dans de grandes bases de données téléphoniques.

[52] définition issue du cours de RI de Yves Chiaramella
[53] http://mrim.imag.fr/projets/BQR/
[54] Laboratoire Logiciels Systèmes Réseaux : http://www-lsr.imag.fr

4.1.2 La campagne d'évaluation TREC-VID

La campagne d'évaluation TREC est organisée chaque année depuis 1992 par le NIST. Le but de la campagne est d'encourager les travaux scientifiques en recherche d'information en fournissant de grandes collections de données, des procédures d'évaluation uniformes, et un cadre expérimental pour des laboratoires intéressés à comparer leurs résultats entre eux. A partir de l'année 2001, les données vidéo ont commencé à faire partie des données d'évaluation. Les données sur lesquelles les évaluations ont été faites pendant les années 2001 et 2002 étaient des enregistrements de vieux films. Trois tâches principales ont été définies alors : détection de plans vidéo ; extraction de traits à partir de la vidéo et enfin recherche interactive.

La première tâche consiste évidemment à *détecter les frontières de plans vidéo*. Elle est nécessaire pour réaliser les deux autres tâches. Un plan vidéo est défini comme une séquence continue de trames vidéo similaires d'un point de vue visuel. Les évaluations TREC utilisent le terme "*continuous camera shots*", un plan (shot) étant le résultat des trois actions suivantes : démarrer la caméra vidéo, filmer une séquence vidéo, arrêter la caméra vidéo. Les moments du démarrage et l'arrêt de la caméra vidéo constituent les frontières d'un plan vidéo. Dans le cas de transmissions télévisées, un changement de caméra constitue aussi une frontière. La détection de frontières de plans vidéo est une première structuration simple d'un document vidéo. Même si quelques problèmes demeurent, lors de fondus enchaînés par exemple, la détection de frontières de plans vidéo est un problème quasiment résolu.

L'extraction de traits est toujours évaluée en termes de rappel et précision, calculés à partir du nombre de plans vidéo pertinents. Le plan vidéo est donc l'unité élémentaire pour les évaluations TREC-VID. Le score officiel de la campagne TREC-VID est la moyenne des précisions calculées après chaque document pertinent retrouvé par le système (lorsque celui-ci donne en sortie une liste ordonnée de documents selon un score de confiance). Pour la tâche d'extraction de traits, les participants doivent rechercher des plans vidéo contenant un trait particulier. Selon le support d'information utilisé pour extraire le trait recherché, nous avons trois catégories distinctes :
•traits vidéo : paysage extérieur, paysage intérieur, visage, texte, etc. ;
•traits audio : parole, musique instrumentale ;
•traits audio-vidéo : monologue (présence du visage et de la voix d'une seule personne simultanément).

La recherche interactive est considérée comme la tâche la plus importante pour un utilisateur de moteur de recherche d'informations. Soit la requête suivante : "Je voudrais tous les plans vidéo qui contiennent des images de la ville de New York.". Un opérateur humain transforme ce besoin en une requête exprimée dans le langage du moteur de recherche. Le moteur de recherche sélectionne ensuite des documents pertinents dans une grande collection de données selon des thèmes définis auparavant. Il existe deux possibilités de recherche : manuelle et interactive. La recherche interactive permet à l'opérateur humain de reformuler sa requête en fonction des résultats de la requête initiale si ceux-ci ne le satisfont pas.

A partir de l'année 2003, les données de vieux films ont été remplacées par des données de type journaux télévisés (pour donner une idée de la quantité de données à traiter, 120 heures de journaux télévisés en anglais étaient fournis pour TREC 2003 soit 80 Go de données). Une nouvelle tâche est apparue aussi : la segmentation en histoires (*story segmentation*) d'un document audio-vidéo (voir travaux déjà mentionnés dans la section *exploitation et prise en compte de la multimodalité*).

4.1.3 Application de mes travaux aux évaluations TREC-VID

TREC 2002 (voir [Quenot, 2002])

Une première application naturelle de mes travaux a consisté à utiliser un critère audio pur pour la recherche d'un document vidéo : il s'agissait de la tâche d'extraction de plans contenant de la parole qui correspond à la présence dans un plan vidéo d'une voix humaine prononçant des mots et

reconnaissable comme telle. L'utilisation de modèles GMM de parole et d'un critère simple de maximum de vraisemblance permet d'extraire les plans vidéos contenant de la parole avec une précision (sur 1000 plans) de 99.7% et un rappel de 72% (7ème / 13 participants à la tâche).

Une autre tâche abordée en 2002 était la détection de plans monologues sur une vidéo qui nécessitait l'utilisation conjointe d'un détecteur de visage et de notre système de segmentation en locuteurs (approche multimodale déjà décrite au chapitre 2). Malgré un détecteur de visage peu performant, les résultats obtenus se comparent favorablement aux autres participants à l'évaluation (système classé 3ème / 9 participants à la tâche).

TREC 2003 (voir [Quenot, 2003])

Nous avons participé en 2003 à une nouvelle tâche d'extraction de plans « *Personne X* » qui consiste à retrouver un plan vidéo où une personne donnée est présente. La présence de la personne est considérée seulement du point de vue visuel, mais toute information disponible (audio, vidéo, transcription, etc.) pouvait être utilisée pour détecter les plans pertinents. Lors de TREC-VID 2003, la personne X à trouver était l'ancien secrétaire d'état américaine Madeleine Albright.

Notre participation à cette tâche n'a pas donné de résultats satisfaisants (taux de précision / rappel proches du hasard) car nous avons essayé d'utiliser un système de détection de locuteurs, entraîné sur la voix de la personne cible, qui s'est avéré inefficace du fait que la personne cible parlait seulement sur 5 plans vidéo parmi les 42 plans pertinents (et sur une collection de 35000 plans au total !). On voit donc bien que, pour cette tâche, il aurait fallu utiliser un système de reconnaissance de visage (bien qu'une telle approche ne soit pas non plus très efficace dans un contexte de RI où le nombre de documents pertinents est très inférieur au nombre total de documents dans la collection [Chen, 2004]) et surtout, d'autre part, des approches exploitant du texte issu de la transcription automatique ou de la reconnaissance de caractère sur les images, qui semblent donner les meilleurs résultats sur ce type de tâche [Chen, 2004] [Yang, 2004]. L'intégration de telles informations a été proposée depuis au CLIPS[55] afin de nommer les intervenants d'un journal télévisé.

TREC 2004 (voir [Besacier, 2004a] et [Moraru, 2005])

En plus de la tâche multimodale de segmentation en histoires, déjà mentionnée au chapitre 2 et décrite précisément dans [Besacier, 2004a], nous nous sommes intéressé à l'identification du présentateur d'un journal télévisé sur un document vidéo (sans données a priori disponibles). Cette tâche n'était pas une tâche officielle de la campagne TREC 2004. Cependant, parmi les locuteurs présents dans un document vidéo d'un journal télévisé, il en existe un qui peut être très important pour d'autres tâches comme la segmentation en histoires ou la création d'un résumé automatique ou tout simplement pour mieux structurer le document. Ce locuteur est le présentateur du journal télévisé ou radiodiffusé. Une première possibilité pour extraire les interventions du présentateur est l'utilisation d'un modèle de locuteur pour le présentateur. Un modèle GMM classique peut être appris sur des données audio issues de journaux précédents. On peut facilement obtenir une quantité de données suffisamment grande pour apprendre un modèle fiable. Ce type d'approche (suivi de locuteur) a notamment été abordé pendant la thèse de Daniel Moraru [Moraru, 2004a]. Le principal désavantage d'une telle approche est que le présentateur peut changer pendant une série d'enregistrements du même journal. Dans ce cas, la seule solution est de réapprendre un nouveau modèle du locuteur pour le nouveau présentateur. La portabilité d'une telle approche est donc assez limitée.

Une autre approche consiste à utiliser nos connaissances concernant le tournage d'une émission d'information télévisée ou radiodiffusée. La structure d'une émission d'information peut être décrite et se retrouve parfois, à quelques différences près, d'une chaîne à l'autre, et même d'un pays à

[55] Mbarek Charhad, Daniel Moraru, Stéphane Ayache and Georges Quénot, « Speaker Identity Indexing In Audio-Visual Documents », in Content-Based Multimedia Indexing (CBMI2005), Riga, Latvia, June 21-23, 2005.

l'autre. Un système de segmentation, par contre, ne peut pas donner des identités aux locuteurs détectés. Les locuteurs sont identifiés par des noms génériques tels que L0, L1, L2, etc. La question qui nous intéresse est alors : est il possible de reconnaître parmi les noms génériques le présentateur ? A titre d'illustration, on peut regarder un exemple de résultat de segmentation automatique en locuteurs obtenus sur un document vidéo. Chaque ligne représente un locuteur et chaque zone noire indique une intervention détectée par notre système de segmentation automatique pour le locuteur considéré.

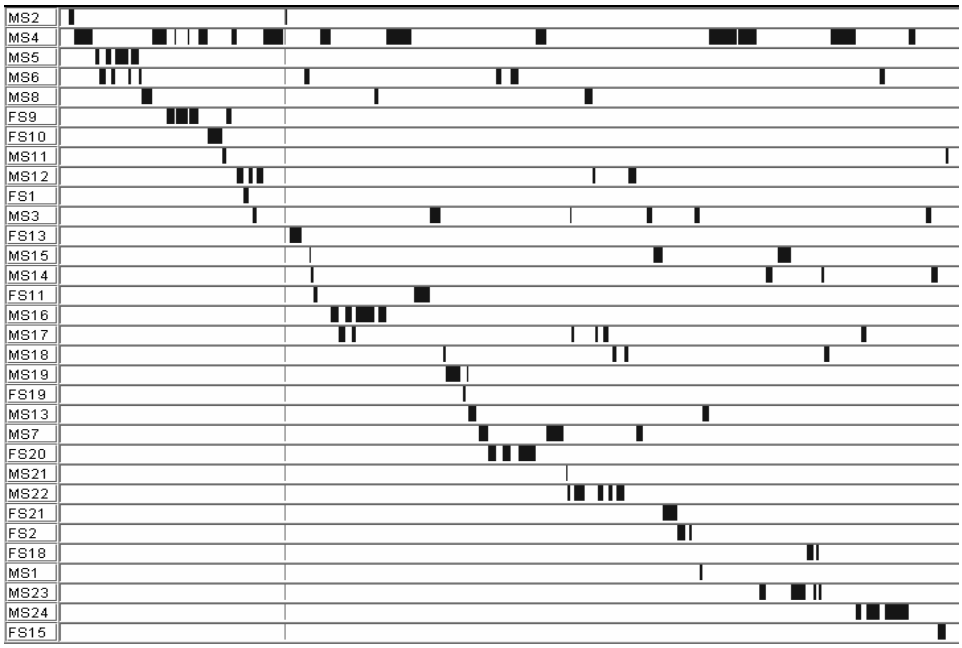


Figure 13 : Sortie d'un système de segmentation en locuteurs sur un journal anglais télévisé de 30 minutes. Le présentateur est le locuteur MS4 (deuxieme ligne)

En regardant les résultats de segmentation présentés dans la figure ci-dessus, nous voyons que la ligne du présentateur semble identifiable car, contrairement aux autres locuteurs qui parlent de façon occasionnelle et pour une période relativement courte, les interventions du présentateur du journal sont réparties plus uniformément sur tout le document. En tenant compte de l'observation précédente, la ligne du présentateur peut être identifiée sans modèle a priori, uniquement à partir de règles empiriques. Une telle approche, permet d'atteindre de bonnes performances de détection de plans où le présentateur parle (rappel de 71% et précision de 93%). Une description plus détaillée de ces résultats est présentée dans [Moraru, 2005].

4.2 Les espaces perceptifs

4.2.1 Introduction

Les espaces perceptifs interviennent dans des contextes applicatifs où il n'est pas possible ou pas souhaitable que l'emplacement du dispositif d'acquisition sonore ou visuel soit visible et situé près des personnes. Ce cas de figure correspond au cas des pièces intelligentes ou *smart rooms* où le dispositif d'acquisition est intégré dans l'espace. Dans ces situations, la personne peut être éloignée du point d'acquisition de la modalité (audio ou image), et pas nécessairement orientée vers celui-ci. De plus, de nombreux événements parasites peuvent être captés en même temps que le signal. Par ailleurs, les environnements perceptifs de ce type introduisent des contraintes supplémentaires :

nécessité de fonctionnement en continu du système (éventuellement 24h/24), besoin éventuel d'optimisations pour le traitement en temps réel de multiples canaux en parallèle. Ce domaine d'application introduit donc de nouveaux problèmes qu'il faut résoudre pour être capable de traiter des documents multimodaux contenant des énoncés de moins en moins contraints.

Une application prédominante dans les espaces perceptifs semble être la mise en oeuvre de ressources facilitant les réunions de travail. Les réunions entre les individus font partie intégrante de la vie courante de tous les groupes de travail. Cependant, la participation sur place aux réunions de travail est souvent rendue difficile à cause du temps requis par les déplacements, de la surcharge des agendas ou d'autres contraintes. Les téléconférences peuvent résoudre ce problème en évitant les déplacements tout en permettant l'interactivité entre participants. Par ailleurs, les enregistrements sonores et vidéo permettent aux personnes ne pouvant se libérer de s'informer, et fournissent des archives.

Concernant plus particulièrement les enregistrements audio, la parole présente les spécificités suivantes lorsqu'elle est issue d'une pièce intelligente (*smart room*) :
- la parole de réunion est complètement spontanée, ce qui conduit à la présence d'une grande quantité de phénomènes spécifiques à l'oral comme les hésitations, les faux départs (début d'un mot prononcé une ou plusieurs fois avant le mot complet), etc.
- plusieurs locuteurs parlent souvent en même temps,
- le bruit de fond peut être important (voix de fond, chuchotements, bruit de la ventilation d'un vidéoprojecteur, …),
- il y a non plus un seul, mais potentiellement plusieurs enregistrements issus de multiples microphones disposés dans toute la salle,
- dans le cas où les participants ne disposent pas d'un micro casque ou bouche, les locuteurs peuvent être à une distance importante des microphones placés au plafond ou sur la table ; cela a pour conséquence de diminuer la qualité des signaux enregistrés.

Il y a aujourd'hui de plus en plus de laboratoires de recherche travaillant dans ce contexte applicatif qui sont équipés de leur propre pièce intelligente. On peut notamment citer Microsoft [Cutler, 2003], l'institut NIST [Stanford et al., 2003], le laboratoire IDIAP[56] [McCowan et al., 2003], ICSI[57] [Morgan et al., 2003] et ISL/CMU[58] [Waibel et al., 2003]. Deux projets européens importants traitent également des environnements perceptifs : CHIL[59] et AMI[60].

Mes contributions sur ce thème concernent d'une part la reconnaissance de sons dans un appartement intelligent, et d'autre part la transcription enrichie de réunions ; elles sont décrites dans les deux sections suivantes.

4.2.2 Reconnaissance de sons dans un appartement intelligent

Cet aspect a été abordé au CLIPS dans la thèse de Dan Istrate [Istrate, 2003], déjà mentionnée au chapitre 2, et dans le cadre d'une collaboration (financée par l'IMAG puis par une ACI Santé du CNRS) entre l'équipe GEOD du laboratoire CLIPS et l'équipe AFIRM[61] du laboratoire grenoblois TIMC[62] autour d'un *Habitat Intelligent pour la Santé*[63] *(HIS).* L'objectif général du projet est la conception, la mise au point et l'expérimentation d'un dispositif de télémédecine s'appuyant sur l'utilisation de capteurs de déambulation et d'activité, et l'utilisation d'algorithmes de classification automatique des sons et de la parole pour détecter des situations de détresse. L'originalité de

[56] Institut Dalle Molle pour l'Intelligence Artificielle perceptive,Martigny, Suisse
[57] International Computer Science Institute, Berkeley, USA
[58] Interactive Systems Laboratories, Carnegie Mellon University, USA
[59] http://chil.server.de/servlet/is/101/
[60] http://www.amiproject.org/
[61] Acquisition, Fusion d'Informations et Réseaux pour la Médecine
[62] Techniques de l'Imagerie, de la Modélisation et de la Cognition
[63] http://www-clips.imag.fr/geod/projets/HIS/

l'application se situe dans le remplacement de la surveillance vidéo, mal acceptée par les patients, par une surveillance sonore : analyse des signaux de parole (tels que des gémissements, cris, des appels au secours) et des sons de la vie courante qui viendront compléter les informations données par les capteurs d'activité. Le dispositif prend place au sein d'un habitat intelligent et a pour but, à terme, de contribuer au maintien de patients à leur domicile en transmettant les alarmes vers les centres de médico-surveillance.

Un local du laboratoire TIMC a été entièrement équipé pour en faire un Habitat Intelligent pour la Santé pilote, à des fins d'expérimentation et de simulation. Cette réalisation constitue un prototype d'appartement de type T1 (environ 30 m2), comprenant les zones d'habitat classiques que sont la chambre, le séjour, la cuisine, les toilettes, la douche et un couloir [Rialle et al., 1999]. Le maintien au domicile de personnes dépendantes suppose la détection et l'analyse de situations de détresse. Les capteurs utilisés sont des capteurs d'activité physiologique et des capteurs sonores. Au niveau de l'analyse sonore, en plus de la localisation de la pièce où se trouve la personne, le but est de reconnaître parmi les sons de la vie courante, les éventuels appels au secours ou d'éventuelles alarmes audio (chute, bris de verre, etc.). Pour des contraintes liées à la préservation de la sphère privée de la personne, le son est acquis et analysé en temps réel sans stockage du signal.

Dans ce contexte, la thèse de Dan Istrate [Istrate, 2003] analyse et propose des solutions aux problèmes spécifiques au traitement du son dans les espaces perceptifs. Un problème important des espaces perceptifs a été abordé : c'est la qualité et la quantité des signaux traités. Tous les algorithmes étudiés et proposés dans les travaux de D. Istrate ont notamment été évalués et validés sur des signaux avec un rapport signal sur bruit variant dans une large gamme (de 0 dB à 60 dB). La grande quantité des signaux à traiter en temps réel, due à l'analyse en continu de l'environnement sonore sur plusieurs canaux, impose par ailleurs l'utilisation d'algorithmes rapides mais ayant néanmoins des performances acceptables pour une application de télésurveillance médicale. Ces algorithmes de détection et de reconnaissance d'événements sonores dans le bruit, déjà mentionnés au chapitre 2, ont été récemment complétés par un système de reconnaissance automatique de la parole à vocabulaire limité, adapté à un contexte de télésurveillance où des appels de détresse peuvent survenir.

Ces travaux autour de l'*Habitat Intelligent Santé* continuent au CLIPS[64] et sont désormais dirigés par Michel Vacher, Ingénieur de Recherche au CNRS. Tous les systèmes de traitement du son utilisés (reconnaissance d'appels de détresse, détection et reconnaissance d'alarmes sonores) ont été développés au CLIPS à l'aide des plates-formes décrites au chapitre 1.

4.2.3 Transcription enrichie de réunions

Cet aspect a été abordé grâce à la participation du CLIPS aux campagnes d'évaluation NIST (SpeakerRec 2002, RT 2004 et RT 2005) consacrées aux enregistrements de réunions (*meetings*). Cette participation s'est faite en collaboration avec le LIA et est décrite en détail dans deux articles [Fredouille, 2004] [Istrate, 2005a]. Elle ne concerne que la tâche de segmentation en locuteurs.

Sur ce type de données, la difficulté vient du fait que les interventions des locuteurs présents sont de durée variable, de moins d'une seconde jusqu'à plusieurs minutes. Il arrive aussi très souvent qu'ils se coupent la parole. Si on regarde les données des évaluations RT04-Meeting, pour environ un tiers de la durée de chaque enregistrement, nous avons plusieurs (jusqu'à 5) locuteurs qui parlent en même temps !

Par ailleurs, il existe plusieurs façons d'enregistrer une réunion :
• un seul microphone placé au milieu des participants ;
• un microphone pour chaque participant ;

[64] voir http://www-clips.imag.fr/geod/projets/HIS/SITE_WEB/ pour les récentes publications autour de ce projet

• plusieurs microphones placés dans la salle de réunion (par exemple un dans chaque coin de la salle).

Le premier cas est celui qui donnera, a priori, les plus mauvais résultats, étant donné l'éloignement résultant pour certains locuteurs de la réunion. Ce cas correspond aux évaluations NIST SpeakerRec 2002. Pour le deuxième cas, nous pouvons nous attendre à de meilleurs résultats, étant donné que la segmentation semble réduite à seulement une détection de silence sur chaque microphone. Cependant, l'expérience montre que le microphone de chaque locuteur capte également les interventions des personnes proches de lui ; les choses ne sont donc pas aussi simples. Une somme des signaux de tous les microphones individuels était aussi proposée à la segmentation pour les évaluations NIST SpRec 2002, à titre de comparaison. De façon assez surprenante, notre système a montré peu de différences de performance (environ 1 % en absolu) sur ces deux types de données.

Des réunions enregistrées avec plusieurs microphones placés uniformément dans la salle de réunion ont été proposées pour les évaluations NIST RT-04 et RT-05 Meeting. La difficulté principale dans ce genre de situation est la grande variabilité du signal selon le microphone. Selon la position du microphone, on peut entendre ou non certains locuteurs et la qualité du signal peut aussi varier beaucoup. En effet, selon la position du locuteur qui parle, on peut se retrouver dans une situation favorable dans certains cas : par exemple quand deux locuteurs parlent en même temps, ils peuvent être captés par des microphones différents, pour être ensuite séparés, ce qui n'est pas possible si la réunion est enregistrée avec un seul microphone.

L'utilisation du signal audio issu de plusieurs capteurs différents est donc une situation spécifique aux enregistrements de réunions. Pour cela, nous avons dans un premier temps (RT 04 [Fredouille, 2004]) proposé un système de fusion de décisions (intégration tardive) correspondant aux résultats de segmentation obtenus individuellement sur chaque capteur. Cette approche a donné de bons résultats malgré la difficulté de la tâche due à la spontanéité de la conversation. Le taux d'erreur de segmentation obtenu sur des données de réunion en 2004 est même comparable avec celui obtenu à l'époque sur des journaux télévisés (22.6%). Dans un second temps (RT 05 [Istrate, 2005a]), nous avons proposé une technique consistant à reconstruire un signal « virtuel » à partir des signaux provenant des différents microphones (intégration précoce) en utilisant une estimation du rapport signal sur bruit le long de chaque canal pour pondérer la contribution de chaque source. Cette technique est comparée à celle qui consiste à simplement reconstruire un signal virtuel par addition de chaque canal. Avec ces deux techniques, les performances obtenues sur les données d'évaluation RT05 sont équivalentes sur les enregistrements de réunions (environ 25% d'erreur dans les deux cas), tandis que la reconstruction pondérée donne de meilleurs résultats sur des enregistrements de cours (*lecture*) pour lesquels un microphone (probablement celui situé près de l'orateur) prédomine par rapport aux autres.

4.2.4 Une pièce intelligente au CLIPS ?

Plus localement, le laboratoire CLIPS, en collaboration avec le CNRS, le CEA, France Telecom, ST Microelectronics et IDEAS LAB a participé au projet RNRT COUCOU (Conception participative Orientée Usage de services de Communication et d'objets Ubiquistes) entre 2002 et 2005. Le but de ce projet était, entre autres, de spécifier et commencer à implémenter deux salles de réunions intelligentes (au CEA et au CLIPS) équipées d'outils avancés de prise de notes et de post-traitement des données enregistrées. Mon implication sur ce projet a concerné les aspects audio [Besacier, 2004b] [Jambon, 2003]. Ces spécifications ont permis à la pièce intelligente du CLIPS de voir le jour à partir de septembre 2005 dans les nouveaux locaux du CTL[65] à Grenoble.

[65] Centre des Technologies du Logiciel

Chapitre 5 : Quelques perspectives de recherche et projets à venir

Mes perspectives de recherche s'inscrivent dans la continuité des deux principaux thèmes présentés dans ce document : la *multimodalité* et le *multilinguisme*.

5.1 Multimodalité

Les projets suivants vont me permettre de continuer à développer des activités autour du thème *multimodalité* :

-TELMA (projet ANR RNTS) : ce projet, qui a démarré en 2006, est issu au départ d'un BQR INPG. Il vise à l'étude et au développement algorithmique de fonctionnalités audiovisuelles originales à l'usage des personnes malentendantes, et à l'étude de faisabilité de leur intégration dans un terminal autonome de télécommunication téléphonique. Le projet a pour objectif technique précis d'exploiter la modalité visuelle de la parole, d'une part pour améliorer les techniques de débruitage du son de parole (la minimisation du bruit environnemental permettant une meilleure exploitation des restes auditifs des malentendants), et d'autre part, en mettant en œuvre des techniques d'analyse/synthèse de lecture labiale et de gestes de la Langue Française Parlée Complétée (LPC). Un tel système peut être vu comme une traduction bidirectionnelle entre parole et LPC. La contribution du CLIPS concerne la reconnaissance automatique de la parole et des gestes multimodaux. En liaison avec ce projet, je co-encadre un étudiant en thèse à l'ICP, Nourredine Aboutabit. Ses travaux ont déjà fait l'objet de publications sur la reconnaissance de gestes labiaux et de la main [Aboutabit, 2006a] [Aboutabit, 2006b]. D'un point de vue exploratoire, les problèmes scientifiques qui peuvent encore être développés, en liaison avec ce projet, concernent la fusion multimodale (lèvres + main) pour la reconnaissance automatique du LPC. Si on considère la communication « parole simple vers LPC », une perspective intéressante consiste également à contraindre un système de reconnaissance de parole afin que sa sortie soit une chaîne symbolique directement liée au LPC ; ces contraintes nous amènent à envisager des unités de modélisation originales pour la reconnaissance de la parole (jeu de phonèmes remplacés par un jeu de visèmes ou tactèmes, modélisation acoustique d'unités plus longues telles que la syllabe ou les séquences CV, …).

- WIMUR (*Web Imag Multimedia Document Retrieval System*, Projet IMAG 2006) : ce projet officialise et consolide les collaborations déjà existantes entre le CLIPS, le LIS et le LSR sur la recherche d'information mutlimedia. Il devrait permettre d'approfondir les activités décrites dans la section 4.1 de ce document. Bien que ce projet constitue plus une application de mes recherches au domaine de la recherche d'information, je souhaiterais proposer et expérimenter des paradigmes permettant une approche globale pour la recherche d'information lorsque de multiples sources sont disponibles (texte, audio, vidéo). On pourrait par exemple envisager d'utiliser des outils théoriques tels que les modèles graphiques (par exemple les CRF ou *Conditional Random Fields* [Lafferty, 2001]) qui permettent de modéliser des processus temporels mettant en jeu des dépendances complexes entre paramètres.

- MISTRAL (projet ANR RNTL, accepté, démarrage début 2007) : ce projet s'articule autour du thème de la biométrie multimodale (voir section 2.1 de ce manuscrit) avec des partenaires tels que le LIA, EURECOM, IRIT et LIUM. Dans ce contexte, je souhaiterais développer un thème encore peu abordé : la robustesse aux attaques. En effet, dans le domaine de la biométrie, il est crucial de détecter les imposteurs et de contrer les différents types d'attaque menant à une fausse acceptation. Parmi les différentes attaques possibles, nous pouvons lister les attaques par enregistrement, dites par play-back (déjà décrites au paragraphe 2.3.3), et les attaques par transformation de modalités biométriques (dans ce cas de figure, on suppose que l'imposteur a une connaissance des systèmes biométriques utilisés, et qu'il est capable de transformer les modalités biométriques pour être

identifié à la place du client). La mise en place d'un test de *liveness* exploitant la cohérence lèvres / voix (cf section 2.3.3) peut répondre à ce type d'attaque si la modalité volée ou transformée est reproduite par play-back. Une autre approche intéressante consiste à coupler un système de vérification du locuteur avec un système de reconnaissance automatique de la parole afin de bâtir un protocole d'authentification plus robuste : par exemple, une vérification des informations personnelles sur le locuteur (adresse, numéro de téléphone, etc.) afin de confirmer ou infirmer son identité.

5.2 Multilinguisme

Le thème *multilinguisme* va sans aucun doute devenir le plus important, en raison, d'une part, de ma récente mobilité à IBM qui m'a ouvert des perspectives de recherches nouvelles autour de la traduction de parole, et, d'autre part, en raison de la récente fusion entre les équipes GEOD et GETA (GETALP[66]) qui rend encore plus pertinentes des recherches en reconnaissance automatique de parole multilingue et en traduction automatique de l'oral. J'envisage notamment de soumettre un projet ANR « Jeunes Chercheurs » autour de la traduction automatique de parole pour les langues peu dotées. Un tel projet me permettrait d'étendre ma problématique « langues peu dotées » à la tâche de traduction de parole, avec pour objectif également d'explorer des pistes, ébauchées pendant la fin de mon séjour à IBM, concernant un meilleur interfaçage entre modules de reconnaissance et de traduction. Ce travail va également pouvoir se concrétiser par l'encadrement d'un nouvel étudiant en thèse, d'origine cambodgienne, arrivé au CLIPS à l'automne 2006, et qui a commencé à travailler sur la traduction de parole khmer – français.

Il m'apparaît assez clairement désormais que le principal problème des langues peu dotées, pour la reconnaissance de parole, est le manque de données textuelles permettant l'apprentissage des modèles statistiques de langage. En effet, nous avons vu dans la section 3.2 que le manque de données acoustiques peut être en partie résolu par une méthodologie efficace de collecte de données ou d'adaptation de modèles (méthodes translingues). En revanche, le problème de la faible quantité de données textuelles implique sans doute de réfléchir à des techniques de modélisation sous-lexicale (morphèmes, groupes de phonèmes agglomérés de façon non supervisée), permettant ainsi de réduire la taille du vocabulaire de l'application, tout en gardant théoriquement la même couverture potentielle. Des travaux menés sur l'analyse morphologique non supervisée (voir par exemple [Kurimo, 2006] au sein du réseau d'excellence PASCAL) semblent donc nécessaires afin de bâtir des systèmes de traitement du langage naturel pour des langues peu dotées en ressources textuelles ou même pour des langues non écrites (voir étude préliminaire présentée au paragraphe 3.3.3 pour la traduction de parole et publiée dans [Besacier, 2006b]). Au delà de la reconnaissance automatique de la parole, ce type de modélisation sous-lexicale peut apporter des bénéfices à la traduction automatique ou la recherche d'information multilingue.

Au niveau international, ce travail de recherche pourrait bénéficier des collaborations suivantes :

-prolongement de mes travaux communs avec IBM : dans la seconde phase du projet TRANSTAC, il est prévu d'évaluer les performances des systèmes de traduction de parole sur une *langue surprise* connue seulement quelques mois avant l'évaluation, et pour laquelle une quantité restreinte de ressources serait disponible ; ce cadre expérimental est idéal pour évaluer mes recherches sur les langues mal dotées ;

-lancement d'activités autour de la reconnaissance de la parole au sein du groupe SALTMIL[67], dédié aux langues peu dotées ; les activités d'un tel groupe pourraient même conduire au démarrage d'une action financée par l'Europe du type COST, avec des partenaires représentant des langues peu dotées (Europe de l'est par exemple) et des laboratoires ayant une bonne expérience dans le

[66] Groupe d'Etude sur le Traitement Automatique du Langage et de la Parole
[67] Speech And Language Technology for MInority Languages : http://193.2.100.60/SALTMIL/

domaine de la reconnaissance automatique et de la traduction de parole : Université de Karlsruhe, Université de Trento, …

Par ailleurs, j'aborde également toujours le problème de la reconnaissance de la parole pour des locuteurs non natifs, via la thèse en cours de Tan Tien-Ping. Ses premiers travaux ont porté sur l'analyse automatique des confusions de phonèmes faites par les locuteurs non natifs en utilisant des modèles acoustiques bilingues (langue maternelle ou source et langue seconde ou cible) [Tan, 2006]. L'information sur ces confusions a été ensuite utilisée pour adapter les modèles acoustiques à la parole non native, en utilisant différentes techniques d'interpolation entre modèles acoustiques en langue source et en langue cible [Tan, 2007]. Dans la suite de ces travaux, nous envisageons d'expérimenter différentes méthodes de transformation de dictionnaires de prononciation en utilisant l'information sur les confusions source/cible des locuteurs non natifs. Si on considère la parole spontanée, on peut également envisager de travailler au niveau de la modélisation du langage, car il est très probable que les locuteurs non natifs utilisent des structures de phrases plus simples, voire erronées. Une première idée pourrait être de simuler ces structures en dégradant des corpus textuels de parole spontanée de locuteurs natifs pour apprendre des modèles de langage simulant la parole spontanée de locuteurs non natifs.

Bibliographie

[Aboutabit, 2006a] Noureddine Aboutabit, Denis Beautemps, Laurent Besacier, “Hand and lip desynchronisation analysis in French cued speech: automatic temporal segmentation of hand flow” *IEEE ICASSP 2006*. Toulouse, France. May 2006.

[Aboutabit, 2006b] Noureddine Aboutabit, Denis Beautemps, Laurent Besacier, « Characterisation of cued speech vowels fron the inner lip contour », *Proc ICSLP 2006*, Pittsburgh, USA, September 2006.

[Afify, 2006] Mohamed Afify, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, Laurent Besacier, and Yuqing Gao, “On the use of morphological analysis for dialectal Arabic speech recognition”, *Proc ICSLP 2006*, Pittsburgh, USA, September 2006.

[Allamanche, 2002] E.Allamanche, J.Herre, O.Helmuth, B.Froba, T.Kasten and M.Cremer, “Content-based identification of audio material using Mpeg-7 low level description”, *Proc.of the International Symposium of Music Information Retrieval*, Indiana, USA, Oct 2002.

[Assfalg, 2002] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, “Semantic Annotation of Sports Videos”, *Journal IEEE MultiMedia*, Vol.9, No.2, April-June 2002.

[Bell, 2004] Bell G., “A New Relevance for Multimedia When We Record Everything Personal”, *ACM Multimédia 2004*, NewYork, 10-16 Octobre 2004.

[Berment, 2004] V. Berment, « *Méthodes pour informatiser des langues et des groupes de langues peu dotées*», Doctorat de l’Université Joseph Fourier – Grenoble 1, Mai 2004

[Berrut 1997] C. Berrut, "*Indexation de données multimédia, utilisation dans le cadre d'un système de recherche d'information*", Habilitation à diriger des recherches, Université Joseph Fourier 1997.

[Besacier, 1998] L. Besacier, « *Un modèle parallèle pour la reconnaissance automatique du locuteur* » Thèse de Doctorat, Université d’Avignon, Avril 1998.

[Besacier, 1999] L. Besacier, J. Luettin, G. Maître, E. Meurville "Experimental Evaluation of Text-independent Speaker Verification on Laboratory and Field Test Databases in the M2VTS project" *Eurospeech 99*. Budapest, Hungary. 5-9 September 99.

[Besacier, 2000] L. Besacier, J.F. Bonastre, C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling", *Speech Communication*, n°31 (2000), pp 89-106.

[Besacier, 2001] L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz "Speech Translation for French in the NESPOLE! European Project", *Eurospeech 2001*, Aalborg, Danemark, September 2001.

[Besacier, 2004a] « Video Story Segmentation with Multi-Modal Features: Experiments on TRECvid 2003 », L. Besacier, G. Quenot, S. Ayache, D. Moraru, *6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, October 15-16, 2004, New York, NY USA.

[Besacier, 2004b] L. Besacier, A.C. Descalle, “*Analyse des dialogues et développement de logiciels de traitement*” Rapport Interne 3.4 – SP3.4. Projet RNRT COUCOU. 2004.

[Besacier, 2004c] L. Besacier, A. M. Ariyaeeinia, J. S. Mason, J.-F. Bonastre, P. Mayorga, C. Fredouille, S. Meignier, J. Siau, N. W. D. Evans, R. Auckenthaler, R. Stapert, "Voice biometrics over the Internet in the framework of COST action 275", *EURASIP Journal on Signal Processing, Special issue on biometric signal processing*. n°4, 1 April 2004. p 466-479.

[Besacier 2006] L. Besacier, V-B. Le, C. Boitet, V. Berment, “ASR and translation for under-resourced languages”, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 15-19 May 2006.

[Besacier, 2006b] L. Besacier, B. Zhou, Y. Gao, « Towards speech translation of non written languages », submitted to *IEEE/ACL SLT2006* conference.

[Blanchon, 2004] H. Blanchon, L. Besacier « Traduction de dialogue: résultats du projet NESPOLE! et pistes pour le domaine » *TALN 2004*, Session Poster, Fès, 19-21 avril 2004.

[Burger, 2001] S. Burger, L. Besacier, P. Coletti, F. Metze, C. Morel "The NESPOLE! VoIP Dialogue Database", *Eurospeech 2001*, Aalborg, Danemark, September 2001.

[Burger, 2004] T. Burger « *Caractérisation labiale et classification des phonèmes de la langue française parlée complétée »*, Mémoire de Master 2 Recherche, IMAG, INPG Grenoble, 2004.

[Cano, 2002] P.Cano, E. Battle, E. Gomez, T. Kalker and J. Haitsma, “A review of algorithms for audio fingerprinting”. *In International Workshop on Multimedia Signal Processing*, US Virgin Islands, December 2002.

[Chan, 2004] H.Y. Chan, P.C. Woodland, "Improving Broadcast news transcription by lightly supervised discriminative training", *ICASSP 2004*, Montreal, Canada.

[Chaisorn, 2003] L. Chaisorn, C. Koh, Y. Zhao, H. Xu, T.S. Chua, T. Qi, “Two-level multimodal framework for news story segmentation of large video corpus”, *12th Text Retrieval Conference*, Gaithersburg, MD, USA, 2003.

[Chen, 2004] M-Y. Chen, A. Hauptmann "Searching for a specific person in broadcast news video", *ICASSP 2004*, Montréal, Canada. 2004.

[Christel, 2004] M.G. Christel, C. Huang, N. Moraveji, N. Papernick "Exploiting multiple modalities for interactive video retrieval", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, Montreal, Canada, Mai 2004.

[Cutler, 2003] R. Cutler, "The distributed meetings system". *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 756–759, Hong-Kong. 2003.

[Delafosse, 2005] "*Auto-organisation de documents audiovisuals par recherche d'invariants de production*". Mémoire de projet de fin d'étude ingénieur. ENSERG. INPG. Juin 2005.

[Dieguez-Tirado, 2005] J. Dieguez-Tirado, C. Garcia-Mateo, L. Docio-Fernandez, A. Cardenal-Lopez « Adaptation strategies for the acoustic and language models in bilingual speech transcription ». *IEEE ICASSP*, Philadelphie, USA, 2005.

[Doledec, 1994]S. Doledec, D. Chessel, "Co-inertia analysis: an alternative method for studying species-environment relationships", *Freshwater Biology*, vol. 31, pp. 277–294, 1994

[Eveno, 2005] N. Eveno, L. Besacier «A Speaker independent "Liveness" Test for Audio-Visual Biometrics». *Eurospeech 2005*. Lisbonne, Portugal. Septembre 2005.

[Flanagan, 2004] J.L. Flanagan « Speech-centric multimodal interfaces ». *IEEE Signal Processing Magazine*. November 2004.

[Fredouille, 2004] « The NIST 2004 spring rich transcription evaluation : two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation », C. Fredouille, D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, *RT2004 Spring Meeting Recognition Workshop*, May 17, 2004.

[Fredouille, 2006] C. Fredouille, D. Moraru; S. Meignier, J-F. Bonastre, L. Besacier, "Step-by-step and Integrated approaches in broadcast news speaker diarization". *Computer Speech and Language Journal. Elsevier Ed.* pp303-330, vol 20, Issues 2-3. April-July 2006

[Fügen, 2003] C. Fügen, S. Stüker, H. Soltau, F. Metze, T. Schultz, "Efficient Handling of Multilingual Language Models". *IEEE ASRU 2003*, Virgin Islands, USA, Dec. 2003.

[Fujie, 2003] S. Fujie, Y. Ejiri, Y. Matsusaka, H. Kikuchi, T. Kobayashi "Recognition of paralinguistic information and its application to spoken dialogue system". *ASRU 2003 (Automatic Speech recognition & Understanding)*, Virgin Islands, USA, Dec 2003.

[Gao, 2006] Yuqing Gao, Gu Liang, Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo, Wei-zhong Zhu, Yonggang Deng, Charles Prosser, Wei Zhang and Laurent Besacier « IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-Speech Translator », *First International Workshop on Medical Speech Translation, in conjunction with NAACL/HLT 2006*, New-York, USA. June 2006.

[Garcia-Salicetti, 2003] Sonia Garcia-Salicetti, Charles Beumier, Gérard Chollet, Bernadette Dorizzi, Jean Leroux les Jardins, Jan Lunter, Yang Ni, Dijana Petrovska-Delacrétaz, « BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", *Lecture Notes in Computer Science*, Publisher: Springer-Verlag, ISSN: 0302-9743, Volume 2688 / 2003.

[Gazit, 2004] R. Gazit, Y. Metzger, "Voice mining with multiple target speakers". *Proc. Odyssee 2004*. Toledo, Spain. 2004.

[Gillick, 2005] D. Gillick, S. Stafford, B. Peskin "Speaker detection without models", *ICASSP 2005*. Philadelphie, USA.

[Girin, 2001] Girin, L., Schwartz, J.-L. & Feng, G. (2001), "Audio-visual enhancement of speech in noise", *Journal of the Acoustical Society of America*, 109(6), pp. 3007-3020.

[Hotelling, 1936] H. Hotelling, "Relations between two sets of variates", *Biometrika*, 28:321-377, 1936.

[Hsu, 2004] W. Hsu, L. Kennedy, C.W. Huang, S.F. Chang, C.Y. Lin, G. Iyengar, "News video story segmentation using fusion of multi-level multimodal features in trecvid 2003", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, Montreal, Canada, Mai 2004.

[Istrate, 2003] Istrate, Dan. « *Détection et Reconnaissance des Sons pour la Surveillance Médicale* ». Doctorat Spécialité Signal Image Parole Télécom (SIPT), Institut National Polytechnique (INP), Grenoble, 16 Decembre 2003.

[Istrate, 2005a] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J-F Bonastre « NIST RT'05S Evaluation : Pre-Processing Techniques and Speaker Diarization on Multiple Microphone Meetings ». *Proc. RT05S Workshop*. July 2005.

[Istrate, 2005b] D. Istrate, M. Vacher, J. F. Serignat, "Détection et classification des sons : application aux sons de la vie courante et à la parole", *Colloque GRETSI 2005*.

[Istrate, 2006] D. Istrate, E. Castelli, M. Vacher, L. Besacier and J. F. Serignat, "Sound Detection and Recognition for Medical Telemonitoring", *Journal IEEE Transactions on Information Technology in Biomedicine*. Volume 10 Issue: 2 April 2006. pp 264-274.

[Jain, 2004] A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, A. Ross, J.L. Wayman, "Biometrics : A Grand Challenge", *Proc. International Conference on Pattern Recognition (ICPR)*, Cambridge, UK, Vol. 2, pp. 935-942, August 2004.

[Jambon, 2003] F. Jambon, A.C. Descalle, A. Vidal, L. Besacier "*Outils de traçabilité des séances*" Rapport Interne 2.4 – SP2.4. Projet RNRT COUCOU. 2003.

[Jelinek, 1999] F Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Mass., 1999.

[Kay, 1998] S. M. Kay, *Fundamentals of Statistical Signal Processing*, Volume 2: Detection Theory, Prentice Hall PTR. ISBN 013504135X. 1998.

[Kennedy, 2003] L.S. Kennedy, D. P.W. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings". *ASRU 2003 (Automatic Speech recognition & Understanding)*, Virgin Islands, USA, Dec 2003.

[Killer, 2003] M. Killer, S. Stüker, T. Schultz "Grapheme based Speech Recognition", *Eurospeech 2003*, pp3141-3144, Geneva, September, 2003.

[Kirchhoff, 2002] K. Kirchhoff, J. Bilmes, S. Das, M. Egan, G. Ji, F. He, J. Henderson, M. Noamany, P. Schone, R. Schwartz, "Novel speech recognition models for Arabic". *Final report of the 2001 summer workshop on language engineering", Center for Language and Speech Processing, John Hopkins University*, available at http://www.clsp.jhu.edu/ws2002/groups/arabic/arabic-final.pdf

[Kirchhoff, 2004] K. Kirchhoff and D. Vergyri, "Cross-dialectal acoustic data sharing for Arabic speech recognition", *Proceedings of ICASSP 2004*, Montreal, Canada.

[Koehn, 2003] Philipp Koehn, Franz Josef Och, Daniel Marcu, "Statistical Phrase-Based Translation", *Proc HLT NAACL 2003*. pp 127-133.

[Kurimo, 2006] Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E., Saraclar, M., "Unsupervised segmentation of words into morphemes – Morpho Challenge 2005 Application to Automatic Speech Recognition", *Proceedings Interspeech 2006*. Pittsburgh, USA. September 2006.

[Lafferty , 2001] Lafferty, J., McCallum, A., and Pereira, F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" *Proceedings of ICML*, 2001

[Lavie, 2002] A. Lavie & al. "The Nespole Project Consortium "Enhancing the Usability and Performance of NESPOLE! - a Real-World Speech-to-Speech Translation System", *Proc HLT (Human Language Technologies)*, San-Diego, CA. 2002.

[Le, 2003] V.-B. Le, B. Bigi, L. Besacier, E. Castelli, "Using the Web for fast language model construction in minority languages", *Eurospeech '03*, Geneva, Switzerland, September 2003.

[Le, 2004] V.-B. Le, D.-D. Tran, E. Castelli, L. Besacier, J.-F. Serignat, "Spoken and written language resources for Vietnamese", *LREC 2004*, Lisbon, Portugal, 26-28 May 2004.

[Le, 2005] V.-B. Le, L. Besacier (2005), "First steps in fast acoustic modeling for a new target language: application to Vietnamese", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, 19-23 March 2005.

[Le 2006] V-B. Le, L. Besacier, T. Schultz, Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 15-19 May 2006.

[Le 2006b] V-B. Le "*Reconnaissance automatique de la parole pour des langues peu dotées* ». Thèse de l'Université J. Fourier, Grenoble I, soutenue le 1er Juin 2006.

[Le, 2006c] V-B. Le, L. Besacier, « Comparison of Acoustic Modeling Techniques for Vietnamese and Khmer ASR », *Proc ICSLP 2006*, Pittsburgh, USA, September 2006.

[Lejeune, 2005] R. Lejeune, J. Baude, C. Tchong, H. Crepy, C. Waast-Richard "Flavoured acoustic model and combined spelling to sound for asymmetrical bilingual environment". *Eurospeech 2005*. Lisbonne, Portugal.

[Mac Gurk, 1976] H. Mac Gurk and J. Mac Donald. "Hearing lips and seeing voices". *Nature*, 264:746-748, 1976.

[Magrin-Chagnolleau, 2001] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA consortium research activities," *2001 A Speaker Odyssey Workshop*, pp.67–72, Chania, Crete, June 2001.

[Mana, 2003] N. Mana, S. Burger, R. Cattoni, L. Besacier, V. MacLaren, J. McDonough, F. Metze "The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains", *Eurospeech 2003*, Geneva, 1-4 Sept. 2003.

[Martin, 2005] T. Martin, S. Sridharan "Cross-language acoustic model refinement for the Indonesian language", *IEEE ICASSP*, Philadelphie, USA, 2005.

[Mayorga-Ortiz 2005] P. Mayorga-Ortiz, "*Reconnaissance vocale dans un contexte de voix sur IP : diagnostic et propositions*", Doctorat Spécialité Signal Image Parole Télécom (SIPT), Institut National Polytechnique (INP), Grenoble, 177 p., 8 février 2005.

[McCowan et al., 2003] McCowan, I., Bengio, S., Gatica-Perez, D., et al. "Modeling human interaction in meetings". *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 748–751, Hong-Kong. 2003.

[Meignier, 2002] S. Meignier, JF. Bonastre, I. Magrin-Chagnolleau, "Speaker Utterances tying among speaker segmented audio documents using hierarchical classification: towards speaker indexing of audio databases", *Proceedings of International Conference on Spoken Language Processing (ICSLP 2002)*, 2002.

[Meignier, 2004] “Benefit of prior acoustic segmentation for speaker segmentation systems” S. Meignier, D. Moraru, C. Fredouille, L. Besacier, and J.-F. Bonastre, *International Conference on Acoustics Speech & Signal Processing (ICASSP 04)*, Montreal, Canada, May 2004.

[Meknavin, 1997] S. Meknavin, P. Charoenpornsawat, B. Kijsirikul, “Feature-based Thai Word Segmentation”, *In Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 (NLPRS’97)*, Phuket, Thailand.

[Mohri, 1997] M. Mohri, “Finite-State Transducers in Language and Speech Processing,” *Computational Linguistics*, vol. 2, no.23, 1997.

[Moraru, 2003a] "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation", D. Moraru, S. Meignier, L. Besacier, J-F Bonastre, I. Magrin-Chagnolleau. *International Conference on Acoustics Speech & Signal Processing (ICASSP 03)*, Honk-Kong, China, May 2003

[Moraru, 2003b] D. Moraru, L. Besacier "Towards Conversational Model for Speaker Segmentation", *in Speech Technology & Human-Computer Dialogue*, Bucharest, April 10-11, 2003, ISBN 973-27-0963-4.

[Moraru, 2004a] D. Moraru, “*Segmentation en locuteurs de documents audios et audiovisuels : application à la recherche d'information multimedia*”, Doctorat Spécialité Signal Image Parole Télécom (SIPT), Institut National Polytechnique (INP), Grenoble, 217 p., 20 décembre 2004.

[Moraru, 2004b] D. Moraru, L. Besacier, S. Meignier, C. Fredouille, JF Bonastre, « Speaker Diarization in the ELISA Consortium over the last 4 years », *RT2004 Fall Workshop*. November 2004.

[Moraru, 2004c] D. Moraru, L. Besacier, E. Castelli “Using a priori information for speaker diarization”, *Proc. Odyssee 2004, The Speaker and Language Recognition Workshop*, Toledo, Spain, 31 May-4 June, 2004.

[Moraru, 2005] D. Moraru, L. Besacier, G. Quenot, S. Ayache. « Speaker and Story Segmentation Using Audio-Video Information ». *Trends in Speech Technology. Proc. 3d Conference on Speech Technology and Human Computer Dialog*. ISBN 973-27-1178-7. Cluj-Napoca, Roumania, May 13-14, 2005.

[Morgan et al., 2003] Morgan, N., Baron, D., Bhagat, S., et al. “Meetings about meetings : Research at ICSI on speech in multiparty conversations”. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 740–743, Hong-Kong. 2003.

[Nettle, 2000] . Nettle, S. Romaine, *Vanishing Voices, the Extinction of the World’s Languages*, Oxford University Press. 2000.

[New, 2005] T. L. New, H. Li “Broadcast News Segmentation by audio type analysis”, *Proc. ICASSP 2005*. Philadelphie. USA. 2005.

[Nguyen, 2002] C. Nguyen “*Reconnaissance automatique de la parole en langue vietnamienne*”. Doctorat de l’INPG – Grenoble, Juin 2002.

[Ohtsuki, 2003] K. Ohtsuki, K. Bessho, Y. Matsuo, S. Matsunaga, Y. Hayashi, “Automatic indexing of multimedia content by integration of audio, spoken language and visual information”, *ASRU 2003*, Virgin Islands, USA, Dec. 2003.

[Pallett 2003] D.S. Pallett “A look at NIST’s Benchmark ASR tests : past, present and future”. *IEEE ASRU 2003*, Virgin Islands, USA, Dec. 2003.

[Perez-Freire, 2004] L. Perez-Freire, C. Garcia-Mateo “A multimedia approach for audio segmentation in TV broadcast news”, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, Montreal, Canada, Mai 2004.

[Pinquier, 2004] J.Pinquier, R. André-Obrecht , “Jingle detection and identification in audio documents”, *Conference ICASSP 2004*. Montreal, Canada. 2004.

[Potamianos, 2004] G. Potaniamos, C. Neti, J. Huang, J.H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, J. Jiang, « Towards practical deployement of audio-visual speech recognition », *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, Montreal, Canada, Mai 2004.

[Povey, 2002] D. Povey & P.C. Woodland, “Minimum Phone Error and ISmoothing for Improved Discriminative Training,” In *Proc. ICASSP 02*, Orlando, USA, 2002.

[Quang, 2005] V-M Quang, E. Castelli, A. Boucher, L. Besacier "Classification de parole en Question et NonQuestion par arbre de décision", *12-èmes Rencontres de la Société Francophone de Classification*, Montréal, 30 Mai – 1er Juin 2005.

[Quenot, 2002] G. Quénot, D. Moraru, L. Besacier, and P. Mulhem, "CLIPS-IMAG at TREC-11 : Experiments in Video Retrieval", *11th Text Retrieval Conference*, Gaithersburg, MD, USA, 19-22 November, 2002

[Quenot, 2003] G. Quénot, D. Moraru, L. Besacier, “CLIPS at TRECvid: Shot Boundary Detection and Feature Detection”, *12th Text Retrieval Conference*, Gaithersburg, MD, USA, 2003.

[Ren, 2005] X. Ren, X. He, Y. Zhang “Mandarin / English Mixed-lingual Name Recognition for Mobile Phone”. *Eurospeech 2005*. Lisbonne, Portugal.

[Renals and Ellis, 2003] Renals, S. and Ellis, D. “Audio information access from meeting rooms”. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 744–747, Hong-Kong. 2003.

[Reynolds, 2005] D.A. Reynolds, P. Torres-Carrasquillo « Approaches and Applications of Audio Diarization », *IEEE ICASSP*, Philadelphie, USA, 2005.

[Rialle et al., 1999] Rialle, V., Lauvernay, N., Franco, A., Piquard, J. F., and Couturier, P. “A smart room for hospitalized elderly people : Essay of modelling and first steps of an experiment”. *Technology and Health Care*, 7 :343–357.

[Rossato, 2002] S. Rossato, H. Blanchon, L. Besacier “Speech-to-speech translation system evaluation : results for French for the Nespole! Project first showcase”, *Proc ICSLP2002*, Denver, USA, Sept 2002

[Schultz, 2001] T. Schultz, A. Waibel (2001). “Language independent and language adaptive acoustic modeling for speech recognition”. *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.

[Schultz, 2004] T. Schultz, “Towards Rapid Language Portability of Speech Processing Systems”, Invited Paper, *SPLASH-2004, Conference on Speech and Language Systems for Human Communication*, Delhi, India, November 17-19, 2004.

[Schultz, 2006] T. Schultz, K. Kirchhoff, *Multilingual Speech Processing*, Academic Press, Elsevier, 2006

[Schwartz, 2004] J.L. Schwartz, « La parole multisensorielle : plaidoyer, problèmes, perspective ». *XXVèmes Journées d’Etude sur la Parole*, 19-21 avril 2004, Fès, Maroc.

[Senechal, 2004] “*Etude de signatures audio-vidéo pour la recherche d’invariants de production*”, Mémoire de Master Recherche. INPG. 2004.

[Senechal, 2005] B. Senechal, D. Pellerin, L. Besacier, I. Simand, S. Brès "Audio, Video and Audio-Visual Signatures for Short Video Clip Detection: Experiments on Trecvid2003", *IEEE ICME (International Conference on Multimedia and Expo)*. Amsterdam. Holland. July 2005.

[Shafran, 2003] I. Shafran, M. Riley, M. Mohri “Voice signatures”. *ASRU 2003 (Automatic Speech recognition & Understanding)*, Virgin Islands, USA, Dec 2003.

[Shriberg, 2000] E. Shriberg, A. Stolcke et. Al. « Prosody-based Automatic Segmentation of Speech into Sentences and Topics », *Speech Communication*. 32(1-2), Special Issue on Accessing Information in Spoken Audio.2000.

[Smeaton 2003] Smeaton A., Kraaij W., Over P., "TRECVID 2003 – An introduction", *12th Text Retrieval Conference, Gaithersburg*, MD, USA, 2003.

[Standford 2003] Standford V., Garofolo J., Galibert O., Michel M., Laprun C., "The NIST Smart Space and Meeting Room Projects: Signal, Acquisition, Annotation and Metrics", *Proc of ICASSP 2003*, Hong-Kong, China, Mai 2003.

[Stanford et al., 2003] Stanford, V., Garofolo, J., Galibert, O., Michel, M., and Laprun, C. “The NIST smart space and meeting room projects : Signals, acquisition, annotation and metrics”. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 736–739, Hong-Kong. 2003.

[Stuker 2004] S. Stüker, T. Schultz, ‘A Grapheme based Speech Recognition System for Russian’, *SPECOM’04*, St. Petersburgh, Russia, September 2004.

[Suebvisai, 2005] S. Suebvisai, P. Charoenpornsawat, A. Black, M. Woszczyna, and T. Schultz, “Thai Automatic Speech Recognition”, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2005)*, Philadelphia, Pennsylvania, March 2005.

[Sumby, 1954] Sumby, W. H. & Pollack, I. “Visual contributions to speech intelligibility in noise”. *Journal of the Acoustical Society of America*, 26, 212-215. 1954.

[Tan, 2006] Tien-Ping Tan, Laurent Besacier. « A French Non-Native Corpus for Automatic Speech Recognition » Proc *LREC 2006*. Genoa, Italy. May 2006.

[Tan, 2007] Tien-Ping Tan, Laurent Besacier. « Acoustic Model Interpolation for Non-native speech recognition ». Soumis a *ICASSP 2007*.

[Vaufreydaz, 1998] D. Vaufreydaz, M. Akbar, J. Caelen, J.-F. Serignat (1998). “EMACOP Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole”, *JEP'98 (Journées d'Étude sur la Parole)*, Martigny (Switzerland), pp. 175-178, June 1998.

[Vaufreydaz, 2001] D. Vaufreydaz, L. Besacier , C. Bergamini, R. Lamy, “From generic to task-oriented speech recognition: French experience in the NESPOLE! European project", *presented at ITRW Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France 29-30 August, 2001,

[Vaufreydaz, 2002] D. Vaufreydaz, (2002) “*Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*”. Doctorat de l’Université J. Fourier – Grenoble I, Janvier 2002.

[Vu-minh, 2004] V-M Quang, L. Besacier, H. Blanchon, B. Bigi « Modèle de langage sémantique pour la reconnaissance automatique de parole dans un contexte de traduction » *TALN 2004*, Session Poster, Fès, 19-21 avril 2004.

[Waibel et al., 2003] A. Waibel, T. Schultz, M. Bett, et al. “Smart : The smart meeting room task at ISL”. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 752–755, Hong-Kong. 2003.

[Waibel, 2004] A. Waibel, T. Schultz, S. Vogel, C. Fügen, M. Honal, M. Kolss, J. Reichert und S. Stüker, “Towards Language Portability in Statistical Machine Translation” Invited paper, *Special Session on Multilinguality in Speech Processing, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2004)*, Montreal, Canada, Mai 2004.

[Wang, 2003a] D. Wang, L. Lu, H-J. Zhang, « Speech segmentation without speech recognition », *ICASSP 2003*. Honk-Kong. China. 2003.

[Wang, 2003b] Z. Wang, T. Schultz, A. Waibel, ,'Comparison of acoustic model adaptation techniques on non-native speech', *Proc of ICASSP 2003*, Hong-Kong, China, Mai 2003.

[Ward-Church 2003] Ward-Church K., "Speech and Language Processing: Where Have Been and Where Are We Going?", *Eurospeech*, Geneva, Switerland, September 1-4 2003.

[Xiang, 2006] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription" *Proc. ICASSP'06*, Toulouse, France, 2006.

[Yang, 2004] J. Yang, A. Hauptmann "Naming every individual in news video monologues", *6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, October 15-16, 2004, New York, NY USA.

[Zechner, 2003] K. Zechner, "Spoken Language Condensation in the 21st century", *Eurospeech 2003*. Geneva, Switzerland, 2003.

[Zeppenfeld, 1997] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, A. Waibel "Recognition of conversational telephone speech using the Janus speech engine" *IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997.

Annexe : CV Détaillé

Activités d'encadrement scientifique : Doctorants et Masters Recherche

Une grande partie des activités décrites dans ce manuscrit s'est faite à travers l'encadrement d'étudiants de Master ou Doctorat. Qu'ils soient tous remerciés ici, car c'est peut être sur le travail d'encadrement que j'ai le plus appris (et j'espère progressé !) lors de ces six dernières années, notamment au contact des étudiants étrangers (vietnamiens, roumains, mexicain).

Thèses

Encadrement partiel

-C. Nguyen[68] (taux d'encadrement : 30%) : *Reconnaissance automatique de la parole en langue vietnamienne*. Doctorat de l'INPG, école doctorale EEATS Grenoble, **thèse soutenue** en Juin 2002.

-D. Vaufreydaz[69] (taux d'encadrement : 50%) : *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Doctorat de l'Université J. Fourier, école doctorale EDMI Grenoble, **thèse soutenue** en Janvier 2002.

-D. Istrate[70] (taux d'encadrement : 50%) : *Détection et Reconnaissance des Sons pour la Surveillance Médicale*. Doctorat de l'INPG, école doctorale EEATS Grenoble, **thèse soutenue** en Décembre 2003.

-V-B Le (taux d'encadrement : 70%) : *Reconnaissance automatique de la parole pour des langues peu dotées.* Doctorat de l'Université J. Fourier, école doctorale EDMI Grenoble, **thèse soutenue** le 1er Juin 2006.

-Q. Vu-Minh (taux d'encadrement : 50%) : *Détection de zones d'intérêt intonatives dans un flux audio.* Doctorat de l'INPG, école doctorale EEATS Grenoble, soutenance prévue en 2007.

Encadrement total

-D. Moraru[71] : *Segmentation en locuteurs de documents audios et audiovisuels : application à la recherche d'information multimédia*. Doctorat de l'INPG, école doctorale EEATS Grenoble, **thèse soutenue** en Décembre 2004.

-P. Mayorga[72] : *Reconnaissance vocale dans un contexte de voix sur IP : diagnostic et propositions.* Doctorat de l'INPG, école doctorale EEATS Grenoble, **thèse soutenue** en Janvier 2005.

-T.T. Ping : *Les problèmes des locuteurs non natifs en reconnaissance automatique de la parole.* Doctorat de l'Université J. Fourier, école doctorale EDMI Grenoble, soutenance prévue fin 2007.

DEA ou Masters Recherche

-C. Bergamini : *Modèle acoustique dépendant du contexte pour la reconnaissance automatique de la parole.* DEA Informatique Systèmes Communication. Université J. Fourier, Grenoble. Juin 2000. Mention B.

[68] C. Nguyen est enseignant-chercheur à l'Institut Polytechnique de Hanoï (Viet-Nam)
[69] D. Vaufreydaz est enseignant-chercheur à l'Université Pierre Mendes-France (Grenoble II)
[70] D. Istrate est enseignant-chercheur à l'ESIGETEL (Fontainebleau)
[71] D. Moraru est ingenieur de recherche dans une filiale de Motorola
[72] P. Mayorga est enseignant-chercheur à l'Université de Mexicali (Mexique)

-R. Lamy : *Adaptation de modèles acoustiques et traitement des vecteurs acoustiques pour la reconnaissance automatique de la parole téléphonique*. DEA Informatique Systèmes Communication. Université J. Fourier, Grenoble. Juin 2001. Mention B.

-D. Moraru : *Segmentation de signaux en locuteurs*. DEA Signal Image Parole Telecom. INPG, Grenoble. Juin 2001. Mention B.

-V-B. Le : *Reconnaissance automatique de mots clés en anglais en conditions bruitées*. DEA Informatique Systèmes Communication. Université J. Fourier, Grenoble. Juin 2002. Mention AB.

-Q. Vu-Minh : *Meilleur Interfaçage Reconnaissance / Analyse pour la Traduction de Parole*. DEA Informatique Systèmes Communication. Université J. Fourier, Grenoble. Juin 2003. Mention AB.

-B. Senechal[73] : *Etude de signatures audio-vidéo pour la recherche d'invariants de production*. Master-R Signal Image Parole Telecom. INPG, Grenoble. Juin 2004. Mention AB.

Participation à des jurys de thèses :

-examinateur et co-directeur de la thèse de D. Vaufreydaz (CLIPS) soutenue le 7 Janvier 2002,
-examinateur et co-directeur de la thèse de C. Nguyen (CLIPS) soutenue le 19 Juin 2002,
-examinateur et co-directeur de la thèse de D. Istrate (CLIPS) soutenue le 16 Décembre 2003,
-examinateur et co-directeur de la thèse de D. Moraru (CLIPS) soutenue le 20 Décembre 2004,
-examinateur et co-directeur de la thèse de P. Mayorga (CLIPS) soutenue le 19 Janvier 2005,
-examinateur et co-directeur de la thèse de V-B Le (CLIPS) soutenue le 1er Juin 2006,
-examinateur de la thèse de Teva Merlin (LIA / Université d'Avignon) soutenue le 18/11/2004.

Projets industriels et collaborations scientifiques nationales

-3 Contrats : CLIPS/GEOD – Thomson Multimédia 2000-01 ; CLIPS/GEOD – Prosodie 2000-01 ; CLIPS/GEOD – Thalès 2001-02

La nature de ces trois contrats était sensiblement la même. Le projet de convention portait sur une expertise et un développement menés conjointement par le CLIPS et le partenaire industriel (Thomson, Prosodie ou Thalès) dans le but de réaliser un système de reconnaissance automatique de la parole et de le comparer à un système de reconnaissance vocale de référence. La contribution du CLIPS consistait à accompagner le développement du système au niveau de la réalisation d'un état de l'art sur la reconnaissance du point de vue des méthodes et des algorithmes, de l'apport de données d'apprentissage exploitables pour la modélisation acoustique (étiquetage complet d'une base d'apprentissage), de conseils techniques dans le développement (choix des paramètres acoustiques, modèles acoustiques, recherche des hypothèses et modèle de langage), et du suivi scientifique dans le développement du système de reconnaissance en prenant comme point d'appui les résultats du système RAPHAEL du CLIPS pour comparer et valider les résultats à toutes les étapes de la reconnaissance.
Le système propriétaire réalisé par Prosodie, avec l'aide du CLIPS, est actuellement en exploitation dans quelques serveurs vocaux interactifs de la société.

Contrat CLIPS/GEOD – Université de Karlsruhe 2000-01

Dans le cadre d'un contrat de coopération avec ISL (Interactive Systems Laboratories) de l'Université de Karlsruhe pour le développement en commun d'une base de données de parole en français, nous avons enregistré une large base orale (BRAF100 : Base pour la Reconnaissance Automatique du Français avec 100 locuteurs – environ 30h de parole).

[73] co-encadrement avec le laboratoire LIS.

Projet RNRT COUCOU (CLIPS, MSH, CEA, FT-R&D), 2002-05

Le laboratoire CLIPS, en collaboration avec le CNRS, le CEA (IDEAS LAB), France Telecom, et ST Microelectronics a participé au projet RNRT COUCOU (Conception participative Orientée Usage de services de Communication et d'objets Ubiquistes) entre 2002 et 2005. Le but de ce projet était, entre autres, de spécifier et commencer à implémenter deux salles intelligentes de réunions (au CEA et au CLIPS) équipées d'outils avancés de prise de notes et de post-traitement des données enregistrées. Mon implication sur ce projet a concerné les aspects audio (équipement et traitement) d'une salle intelligente.

Projet Technolangue AGILE/ALIZE (CLIPS, LIA, DDL, IRISA, IRIT, ENST), 2003-05

Le projet ALIZE consistait à réaliser une plateforme logiciel libre en vérification automatique du locuteur. L'objectif était de pérenniser le savoir-faire du consortium ELISA, acquis grâce à des participations continues aux campagnes d'évaluation NIST depuis 1998, aux entreprises et laboratoires académiques qui souhaiteraient se lancer dans la vérification automatique du locuteur. La contribution du CLIPS sur ce projet a concerné la segmentation en locuteurs : participation à des campagnes d'évaluation (NIST, ESTER) et re-écriture du système existant au CLIPS en utilisant la plateforme ALIZE.

BQR INPG 2003 « Vidéo-Sémantique » (CLIPS, LIS, LSR)

Ce projet financé par l'INPG avait pour but de fédérer les collaborations entre trois laboratoires grenoblois dans le domaine de l'indexation multimédia. Ma contribution, au sein du CLIPS, a concerné le traitement automatique de la bande son de documents vidéo et la participation annuelle aux campagnes d'évaluation TREC-Vidéo.

BQR INPG 2004 « TELMA » (CLIPS, LIS, ICP)

Ce projet financé par l'INPG vise à l'étude et au développement algorithmique de fonctionnalités audiovisuelles originales à l'usage des personnes malentendantes, et à l'étude de faisabilité de leur intégration dans un terminal autonome de télécommunication téléphonique. Le projet a pour objectif technique précis d'exploiter la modalité visuelle de la parole, d'une part pour améliorer les techniques de débruitage du son de parole (la minimisation du bruit environnemental permettant une meilleure exploitation des restes auditifs des malentendants), et d'autre part, en mettant en œuvre des techniques d'analyse/synthèse de lecture labiale et de gestes de la Langue Française Parlée Complétée (LPC). Une suite de TELMA, financée par l'ANR vient de démarrer en 2006.

Projet ACI-SI BIOMUL (CLIPS, EURECOM, LIA, INT), 2004-06

Ce projet est financé par le CNRS sur l'ACI *Sécurité Informatique* (BIOMUL : Biométrie et Multimodalités). Dans ce cadre, j'ai pu accueillir un post-doctorant au CLIPS (Nicolas Eveno, issu du laboratoire LIS/INPG) pour travailler sur la biométrie labiale bimodale. Nous avons également abordé au cours de ce projet la problématique de segmentation en locuteurs dans des environnements perceptifs.

Collaborations internationales

Projet Européen NESPOLE & Consortium C-STAR

A mon arrivée au CLIPS en 1999, j'ai été impliqué dans deux projets de traduction automatique de parole dont le CLIPS était partenaire : CSTAR (*Consortium for Speech Translation Advanced Research*) et NESPOLE (*Negotiating Through SPOken Language in E-commerce*). Je suis devenu responsable de ces deux projets pour l'équipe GEOD (dont la contribution était surtout de fournir l'étage de reconnaissance automatique de la parole). Le but des recherches menées au sein de ces projets était la traduction automatique de parole spontanée avec tous les couples de langues possibles entre les différents partenaires.

Action COST275 (Biometric Person Authentication over the Internet)

J'ai été impliqué, entre 2001 et 2004, dans l'action européenne COST 275 traitant de la biométrie. J'ai notamment contribué au sous-groupe « évaluation » de l'action que j'ai co-animé avec J-F Bonastre.

Groupe SPLC (Special Interest Group on Speaker and Language Characterization) de l'ISCA (International Speech Communication Association)

Je suis membre du groupe de travail SPLC de l'ISCA traitant de la reconnaissance et de la caractérisation du locuteur. Je fais notamment partie du comité scientifique de la conférence *Speaker Odysee*[74] sur le domaine.

Projets TALK (Traitement Automatique de la Langue Khmère) et CORUS (Traitement de la parole en langue vietnamienne)

Ces deux projets, soutenus par l'AUF (Agence Universitaire pour la Francophonie) ou par le MAE (Ministère des Affaires Etrangères), concrétisent une collaboration qui existe entre le CLIPS, le centre MICA à Hanoï, et (plus récemment) l'Institut de Technologie du Cambodge (ITC) sur le domaine du traitement automatique du vietnamien et du khmer. Mon implication dans ces projets concerne la reconnaissance automatique de la parole pour le vietnamien et le khmer.

Projets DARPA TRANSTAC (pendant mon séjour à IBM de Octobre 2005 à Octobre 2006)

TRANSTAC (*Translation System for Tactical Use*) est un projet financé par le département de défense américain (DARPA). Ce projet évalue notamment les technologies permettant le déploiement rapide de systèmes de traduction automatique pour de nouveaux langages ou dialectes, ayant peu de ressources. Le but du projet était de développer un système de traduction de parole anglais-arabe dialectal (le dialecte concerné étant celui parlé en Irak). Ma contribution sur ce projet a concerné surtout la reconnaissance automatique de l'arabe Irakien.

Fonctions d'intérêt collectif

-Membre élu du bureau de l'AFCP (Association Francophone de la Communication Parlée) depuis 2001,
-Relecteur pour les revues *Computer Speech and Language*, *IEEE Transactions on Speech and Audio Processing*, *IEEE Signal Processing Letters*, *IEEE Transactions on Signal Processing*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Information Forensics and Security*, *Pattern Recognition Letters*, *Traitement du Signal*, *Acta Acustica*,

[74] http://www.speakerodyssey.com/

-Relecteur pour les conférences : *ICSLP 2006, Eurospeech 2005, RECITAL 2005, Speaker Odyssee 2004 et 2006, Journées d'Etude sur la Parole 2002 2004 et 2006, ECCTD 2001*,
-Organisation d'une session spéciale Biométrie[75] à la conférence *ISPA 2005*.
-Membre des commissions de spécialistes 27ème section de l'Université J. Fourier (depuis 2003) et de l'Université d'Avignon (2001-2004).
-Membre du conseil scientifique de l'université J. Fourier (depuis Janvier 2005).
-Evaluateur de projets pour l'ACI Masse de données 2005 et 2006.

Enseignement en DEA et Master-R

-Enseignement du module « Traitement de l'Oral » du DEA Informatique Système & Communication (ISC) de l'Université Joseph Fourier de Grenoble : années 99/00, 00/01, 01/02, 02/03,
-Enseignement du module « Traitement de l'Oral » commun aux deux Master-R ICPS (Ingénierie de la Communication Personne-Système) et 3I (Intelligence, Interaction, et Information) des universités grenobloises depuis 2004.

Liste complète des publications

Thèse de doctorat

"Un modèle parallèle pour la reconnaissance automatique du locuteur" PhD, University of Avignon, April 1998.

Revues internationales

"Time-frequency analysis of circumferential wave energy distribution for spherical shells. Application to sonar target recognition" P.Chevret, F.Magand, **L. Besacier**. *Applied Signal Processing*, Springer Verlag, (3). pp 136-142. **1996**.

"Subband approach for automatic-speaker recognition" **L. Besacier**, J.F. Bonastre. European Journal Signal Processing, n°80 (**2000**), Elsevier. Special Issue on Emerging Techniques for Communication Terminals. pp 1245-1259.

"Localization and selection of speaker-specific information with statistical modeling" **L. Besacier**, J.F. Bonastre, C. Fredouille, Speech Communication, n°31 (**2000**), pp 89-106.

"Overview of compression and packet loss effects in speech biometrics" **L. Besacier,** J.-F. Bonastre, P. Mayorga, C. Fredouille, S. Meignier, 2003 , IEE Proceedings Vision, Image & Signal Processing - Special issue on Biometrics on the Internet . Vol. 150, n°6, December **2003**.

"Voice biometrics over the Internet in the framework of COST action 275", **L. Besacier**, A. M. Ariyaeeinia, J. S. Mason, J.-F. Bonastre, P. Mayorga, C. Fredouille, S. Meignier, J. Siau, N. W. D. Evans, R. Auckenthaler, R. Stapert, EURASIP Journal on Signal Processing, Special issue on biometric signal processing. n°4, 1 April **2004**. p 466-479.

"Système de télésurveillance sonore pour la détection des situations de détresse" Dan Istrate, M. Vacher, J. F. Serignat, **L. Besacier**, E. Castelli. ITBM-RBM (Elsevier) Revue Européenne de Technologie Biomédicale. **2006**.

« Information Extraction From Sound for Medical Telemonitoring » D. Istrate, D.; Castelli, E.; Vacher, M.; **Besacier, L**.; Serignat, J.-F.. IEEE Transactions on Information Technology in Biomedicine. Volume: 10 Issue: 2 Date: April **2006**. pp 264-274.

« Step-by-step and Integrated approaches in broadcast news speaker diarization » Corinne Fredouille, Daniel Moraru; Sylvain Meignier, Jean-Francois Bonastre, **Laurent Besacier**. Computer Speech and Language Journal pp303-330, vol 20, Issues 2-3. April-July **2006**. (Elsevier).

[75] http://www.isispa.org/sssipb.html

Chapitre de livre

"Subband approach for automatic-speaker recognition : optimal division of the frequency domain" **L. Besacier**, J.F. Bonastre. *Audio and Video Biometric Person Authentication.* Bigun & al. eds. Lectures Notes in Computer Science 1206 (Springer Verlag), pp 195-202 / ISBN 3-540-62660-3.

Congrès internationaux

2006

« Towards speech translation of non written languages» **Laurent Besacier,** Bowen Zhou, Yuqing Gao. IEEE / ACL SLT 2006. Aruba, December 2006.

« IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-Speech Translator » Yuqing Gao, Gu Liang, Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo, Wei-zhong Zhu, Yonggang Deng, Charles Prosser, Wei Zhang and **Laurent Besacier**, First International Workshop on Medical Speech Translation, in conjunction with NAACL/HLT 2006, New-York, USA. June 2006.

« Comparison of Acoustic Modeling Techniques for Vietnamese and Khmer ASR », Viet Bac Le, **Laurent Besacier,** Proc ICSLP 2006, Pittsburgh, USA, September 2006.

« ON THE USE OF MORPHOLOGICAL ANALYSIS FOR DIALECTAL ARABIC SPEECH RECOGNITION », Mohamed Afify, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, **Laurent Besacier**, and Yuqing Gao, Proc ICSLP 2006, Pittsburgh, USA, September 2006.

« Characterisation of cued speech vowels fron the inner lip contour », N. Aboutabit, D. Beautemps, L. Besacier, Proc ICSLP 2006, Pittsburgh, USA, September 2006.

« ASR AND TRANSLATION FOR UNDER-RESOURCED LANGUAGES » **L. Besacier**, V-B. Le, C. Boitet, V. Berment. Proceedings IEEE ICASSP 2006. Toulouse, France. May 2006.

« ACOUSTIC-PHONETIC UNIT SIMILARITIES FOR CONTEXT DEPENDENT ACOUSTIC MODEL PORTABILITY » Viet Bac Le, **Laurent Besacier**, Tanja Schultz. IEEE ICASSP 2006. Toulouse, France. May 2006.

« HAND AND LIP DESYNCHRONIZATION ANALYSIS IN FRENCH CUED SPEECH: AUTOMATIC TEMPORAL SEGMENTATION OF HAND FLOW » Noureddine Aboutabit, Denis Beautemps, **Laurent Besacier.** IEEE ICASSP 2006. Toulouse, France. May 2006.

« A French Non-Native Corpus for Automatic Speech Recognition » Tien-Ping Tan, **Laurent Besacier** Proc LREC 2006. Genoa, Italy. May 2006.

2005

"First steps in fast acoustic modeling for a new target language. Application to Vietnamese" Viet-Bac Le, **Laurent Besacier**. Proceedings IEEE ICASSP 2005. Philadelphia, USA. April 2005.

"Audio, Video and Audio-Visual Signatures for Short Video Clip Detection: Experiments on Trecvid2003", Benjamin Senechal, Denis Pellerin, **Laurent Besacier**, Isabelle Simand, Stéphane Brès . Accepté à IEEE ICME (International Conference on Multimedia and Expo). Amsterdam. Holand. July 2005.

« A Speaker independent "Liveness" Test for Audio-Visual Biometrics », Nicolas Eveno, **Laurent Besacier**. Accepté à Eurospeech 2005. Lisbonne, Portugal. Septembre 2005.

« Speaker and Story Segmentation Using Audio-Video Information » D. Moraru, **L. Besacier**, G. Quenot, S. Ayache. Trends in Speech Technology. Proc. 3d Conference on Speech Technology and Human Computer Dialog. ISBN 973-27-1178-7. Cluj-Napoca, Roumania, May 13-14, 2005.

« NIST RT05S Evaluation : Pre-Processing Techniques and Speaker Diarization on Multiple Microphone Meetings » Dan Istrate, Corinne Fredouille, Sylvain Meignier, **Laurent Besacier**, and Jean Francois Bonastre. Proc. RT05S Workshop. July 2005.

2004

-« Speaker Diarization in the ELISA Consortium over the last 4 years », D. Moraru, **L. Besacier,** S. Meignier, C. Fredouille, JF Bonastre, RT2004 Fall Workshop. November 2004.

« Video Story Segmentation with Multi-Modal Features: Experiments on TRECvid 2003 », **L. Besacier**, G. Quenot, S. Ayache, D. Moraru, 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, October 15-16, 2004, New York, NY USA.

« The NIST 2004 spring rich transcription evaluation : two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation », C. Fredouille, D. Moraru, S. Meignier, **L. Besacier**, J.-F. Bonastre, RT2004 Spring Meeting Recognition Workshop, May 17, 2004.

"Benefit of prior acoustic segmentation for speaker segmentation systems" S. Meignier, D. Moraru, C. Fredouille, **L. Besacier**, and J.-F. Bonastre, International Conference on Acoustics Speech & Signal Processing (ICASSP), Montreal, Canada, May 2004.

"The ELISA consortium approaches in Broadcast News speaker segmentation during the NIST 2003 Rich Transcription evaluation". D. Moraru, S. Meignier, C. Fredouille, **L. Besacier**, and J.-F. Bonastre, International Conference on Acoustics Speech & Signal Processing (ICASSP), Montreal, Canada, May 2004.

"Using a priori information for speaker diarization", Daniel Moraru, **Laurent Besacier**, Eric Castelli Proc. Odyssee 2004, The Speaker and Language Recognition Workshop, Toledo, Spain, 31 May-4 June, 2004.

"ELISA Nist RT03 Broadcast News Speaker Diarization Experiments" Daniel Moraru, Sylvain Meignier, Corinne Fredouille, **Laurent Besacier**, Jean-François Bonastre, Proc. Odyssee 2004, The Speaker and Language Recognition Workshop, Toledo, Spain, 31 May-4 June, 2004.

"SPOKEN AND WRITTEN LANGUAGE RESOURCES FOR VIETNAMESE" Viet-Bac Le , Do-Dat Tran, Eric Castelli, Laurent Besacier, Jean-François Serignat, Proc. LREC2004, Lisbonne, Portugal. 2004

« Spoken Dialogue Translation Systems Evaluation: Results, New Trends, Problems and Proposals » Hervé Blanchon, Christian Boitet, Laurent Besacier. Proc. IWSLT 2004 (ICLSP 2004 Satellite Workshop). Kyoto, Japan. September 30 - October 1, 2004. vol. 1/1: pp. 95-102.

2003

"AUDIO PACKET LOSS OVER IP AND SPEECH RECOGNITION ", Pedro Mayorga, **Laurent Besacier**, Richard Lamy and Jean-Francois Serignat, ASRU 2003 (Automatic Speech recognition & Understanding), Virgin Islands, USA, Dec 2003.

"Using the Web for fast language model construction in minority languages" Viet Bac LE, Brigitte BIGI, **Laurent BESACIER**, Eric CASTELLI, Eurospeech 2003, Geneva, 1-4 Sept. 2003.

"The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains" Nadia Mana, Susanne Burger, Roldano Cattoni,

Laurent Besacier, Victoria MacLaren, John McDonough, Florian Metze, Eurospeech 2003, Geneva, 1-4 Sept. 2003.

"Smart Audio Sensor for Telemedicine" Michel Vacher, Dan Istrate, **Laurent Besacier**, Eric Castelli, Jean-Francois Serignat, Smarts Objects Conference (SOC) 2003, 15-17 May, Grenoble, France.

-"The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation", D. Moraru, S. Meignier, **L. Besacier**, J-F Bonastre, I. Magrin-Chagnolleau. Accepted to International Conference on Acoustics Speech & Signal Processing (ICASSP), Honk-Kong, China, May 2003.

-"Towards Conversational Model for Speaker Segmentation", D. Moraru, **L. Besacier**, "Speech Technology & Human-Computer Dialogue", Bucharest, April 10-11, 2003, ISBN 973-27-0963-4.<zip>

"HABITAT TELEMONITORING SYSTEM BASED ON THE SOUND SURVEILLANCE" Eric Castelli, Michel Vacher, Dan Istrate, **Laurent Besacier**, Jean-Francois Serignat, ICICTH (International Conference on Information Communication Technologies in Health), 11-13 July 2003, Samos Island, Greece.

"Life Sounds Extraction and Classification in Noisy Environment" M. Vacher and D. Istrate and **L. Besacier** and J.F.Serignat and E. Castelli, IASTED International Conference on Signal & Image Processing, 12-14 August 2003, Kauai, 2003.

"Non-linear acoustical pre-processing for multiple sampling rates ASR and ASR in noisy condition", Richard LAMY, **Laurent BESACIER**, worshop NOLISP 03, Le Croisic, France, 20-23 mai 2003.

"CLIPS at TRECvid: Shot Boundary Detection and Feature Detection", Georges M. Quénot, Daniel Moraru, **Laurent Besacier**, 12th Text Retrieval Conference, Gaithersburg, MD, USA, 2003.

2002

METHODOLOGY FOR EVALUATING SPEAKER VERIFICATION ROBUSTNESS OVER IP NETWORKS **L. Besacier**, P. Mayorga, J.F. Bonastre, C. Fredouille, Proceedings of the COST275 Workshop on The Advent of Biometrics on the Internet, Rome, Nov 2002, ISBN 92-894-4848-2

"CLIPS-IMAG at TREC-11 : Experiments in Video Retrieval", Georges M. Quénot, Daniel Moraru, **Laurent Besacier**, and Philippe Mulhem , 11th Text Retrieval Conference, Gaithersburg, MD, USA, 19-22 November, 2002.

S. Rossato, H. Blanchon, **L. Besacier** "Speech-to-speech translation system evaluation : results for French for the Nespole! Project first showcase", Proc ICSLP2002, Denver, USA, Sept 2002.

P. Mayorga-Ortiz, R. Lamy, **L. Besacier** "Recovering of packet loss for distributed speech recognition", Proc. Eusipco 2002, Toulouse, France, Sept. 2002.

The Nespole Project Consortium : A. Lavie, F. Metze, R. Cattoni, E. Costantini, S. Burger, D. Gates, C. Langley, K.Laskowski, L. Levin, K. Peterson, T. Schultz, A. Waibel, D. Wallace, J. MacDonough, H. Soltau, G. Lazzari, N. Mana, F. Pianesi, E. Pianta, **L. Besacier**, H. Blanchon, D. Vaufreydaz "A Multi-Perspective Evaluation of the Nespole! Speech-to-Speech Translation System", Proc. ACL2002 Workshop on Speech-to-Speech Translation : Algorithms and Systems, Philadelphia, PA, July 7-12 2002

The Nespole Project Consortium "The NESPOLE! Speech-to-Speech Translation System", Proc HLT (Human Language Technologies) 2002, San-Diego, CA

The Nespole Project Consortium "Enhancing the Usability and Performance of NESPOLE! - a Real-World Speech-to-Speech Translation System", Proc HLT (Human Language Technologies) 2002, San-Diego, CA

2001

D. Vaufreydaz, **L. Besacier** , C. Bergamini, R. Lamy, "From generic to task-oriented speech recognition: French experience in the NESPOLE! European project", presented at ITRW Workshop on Adaptation Methods for Speech Recognition, Sophia-Antipolis, France 29-30 August, 2001,

L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz "Speech Translation for French in the NESPOLE! European Project", Eurospeech 2001, Aalborg, Danemark, September 2001.

S. Burger, **L. Besacier**, P. Coletti, F. Metze, C. Morel "The NESPOLE! VoIP Dialogue Database", Eurospeech 2001, Aalborg, Danemark, September 2001.

L. Besacier , C. Bergamini, D. Vaufreydaz, E. Castelli "THE EFFECT OF SPEECH AND AUDIO COMPRESSION ON SPEECH RECOGNITION PERFORMANCE " IEEE Multimedia Signal Processing Workshop, Cannes, France, October 2001.

2000

D. Vaufreydaz, C. Bergamini, J. F. Serignat, **L. Besacier** and M. Akbar, "A New Methodology for Speech Corpora Definition from Internet Documents," presented at LREC'2000, 2nd International Conference on Language Ressources and Evaluation, Athens, Greece, 31 May-2 June, 2000, I, pp.423-426.

"GSM Speech Coding and Speaker Recognition," **L. Besacier**, S. Grassi, A. Dufaux, M. Ansorge and F. Pellandini, presented at ICASSP 2000, Istambul, Turkey, 5-9 june, 2000 .

"INFLUENCE OF GSM SPEECH CODING ON THE PERFORMANCE OF TEXT-INDEPENDENT SPEAKER RECOGNITION" S. Grassi, **L. Besacier**, A. Dufaux, M. Ansorge, and F. Pellandini.EUSIPCO 2000, Tampere, Finland, Sept. 4-8, 2000

"AUTOMATIC SOUND DETECTION AND RECOGNITION FOR NOISY ENVIRONMENT", Alain Dufaux, **Laurent Besacier**, Michael Ansorge, and Fausto Pellandini, EUSIPCO 2000, Tampere, Finland, Sept. 4-8, 2000

"SPEAKER RECOGNITION ON COMPRESSED SPEECH" S. Grassi , A. Dufaux , **L. Besacier** , M. Ansorge , F. Pellandini, Workshop on friendly exchanging through the net, Bordeaux (France), March 22-24, 2000.

1999

"Experimental Evaluation of Text-independent Speaker Verification on Laboratory and Field Test Databases in the M2VTS project" L. Besacier, J. Luettin, G. Maître, E. Meurville. Eurospeech 99. Budapest, Hungary. 5-9 September 99.

"Automatic Sound Recognition relying on statistical methods, with application to telesurveillance". L. Besacier, A. Dufaux, M. Ansorge, and F. Pellandini. International Workshop on Intelligent Communication Technologies and Applications, with emphasis on mobile communications. Neuchâtel, Switzerland. May, 5-7, 1999.

"Influence of GSM speech coding algorithms on text-independent speaker identification performance". S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, F. Pellandini. International Workshop on Intelligent Communication Technologies and Applications, with emphasis on mobile communications. Neuchâtel, Switzerland. May, 5-7, 1999.

"Multi Modal Verification for Teleservices and Security Applications (M2VTS)" G. Richard, Y. Menguy, I. Guis, N. Suaudeau, J. Boudy, P. Lockwood, C. Fernandez, F. Fernández, C. Kotropoulos, I. Pitas, R. Heimgartner, P. Ryser, C. Beumier, S. Pigeon, G. Matas, J. Kittler, J. Bigün, Y. Abdeljaoued, E. Meurville, L. Besacier , G.Maitre, J. Luettin, S. Ben-Yacoub B. Ruiz. In Proc. IEEE Conference on Multimedia Computing and Systems'99. Florence, Italy, 7-11 June 1999.

"Automatic Detection and Classification of Wideband Acoustic Signals" A. Dufaux, L. Besacier, M. Ansorge, F. Pellandini. Joint 137th meeting of the Acoustical Society of America and Forum Acusticum 99. Berlin, Germany. 14-19 March, 1999.

1998

"Frame Pruning for Speaker Recognition". **L. Besacier**, J.F. Bonastre. Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 12-15 May 1998. Seattle (USA).

"Time and frequency pruning for speaker identification". **L. Besacier**. Proc 14th International Conference on Pattern Recognition (ICPR), 16-20 August 1998. Brisbane (Australia).

"Frame Pruning for Speaker Recognition". **L. Besacier**, J.F. Bonastre. Proc. Eusipco, 8-11 September 1998. Rhodes (Greece).

"Time and frequency pruning for speaker identification" **L. Besacier**, J.F. Bonastre. Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), 20-23 April, 1998. Avignon (France).

1997

"Independent processing and recombination of partial frequency bands for automatic speaker recognition" **L. Besacier**, J.F. Bonastre. *Fourteenth Internationnal Conference on Speech Processing*. IEEE Korea Council, IEEE Korea Signal Processing Society, Seoul, Korea, August 26-28, 1997.

"Subband architecture for automatic speaker recognition on partially corrupted speech" **L. Besacier**, J.F. Bonastre. C*OST 254 Workshop on emerging techniques for communication terminals.* Toulouse, France. 7-9 July 1997.

1995

"Time frequency analysis of Stoneley wave energy distribution for spherical and cylindrical shells. Application to sonar target recognition." P.Chevret, F.Magand, **L. Besacier**, . UK Symposium on applications of time-frequency and time-scale methods. pp 233-240. Coventry, UK, 30-31 Aout 1995.

Congres nationaux

« Reconnaissance Automatique de la Parole pour des Langues peu Dotées : Application au Vietnamien et au Khmer » **L. Besacier** , V.-B. Le , E. Castelli, S. Sethserey , L. Protin , TALN 2005 – Atelier TALN et langues mal dotées. Dourdan, France. Juin 2005.

"Premiers pas du CLIPS sur les données d'évaluation ESTER", R. Lamy, D. Moraru, B. Bigi, **L. Besacier**, JEP 2004, Fès, Maroc, Avril 2004

"Segmentation selon le locuteur: les activités du Consortium ELISA dans le cadre de Nist RT03", D. Moraru, S. Meignier, C. Fredouille, **L. Besacier**, J-F Bonastre, JEP 2004, Fès, Maroc, Avril 2004

Vu Minh, Q., **Besacier, L**., Castelli, E., Bigi, B., and Blanchon, H.. (2004). Interchange format-based language model for automatic speech recognintion in speech-to-speech translation. Proc. RIVF'04 (Recherche Informatique Vietnam-Francophonie). To be published in a special issue of Studia Informatica Universalis [Suger Editor]. February 2-5, 2004. vol. 1/1: pp. 47-50.

Hervé Blanchon, **Laurent Besacier** « Traduction de dialogue: résultats du projet NESPOLE! et pistes pour le domaine » TALN 2004, Session Poster, Fès, 19-21 avril 2004.

Quang Vu-minh, **Laurent Besacier**, Hervé Blanchon, Brigitte Bigi « Modèle de langage sémantique pour la reconnaissance automatique de parole dans un contexte de traduction » TALN 2004, Session Poster, Fès, 19-21 avril 2004.

D. Moraru, **L. Besacier** « Segmentation en locuteurs de conversations sur IP », XXIVèmes Journées d'Etude sur la Parole, Nancy, Juin 2002.

R. Lamy, **L. Besacier** "Adaptation spectrale par quantification vectorielle : exemple de la RAP à fréquences d'échantillonnage multiples", XXIVèmes Journées d'Etude sur la Parole, Nancy, Juin 2002.

S. Rossato, H. Blanchon, **L. Besacier** "Évaluation du premier démonstrateur de traduction de parole dans le cadre du projet NESPOLE!", Congrès TALN (Traitement Automatique du Langage Naturel), Nancy, Juin 2002.

"Système d'élagage temps-fréquence pour l'identification du locuteur" **L. Besacier**, J.F. Bonastre. *22èmes Journées d'Etude sur la Parole.* Martigny, Suisse. 15-19 Juin 1998.

"Architecture en sous-bandes pour la reconnaissance automatique du locuteur en milieu bruite" **L. Besacier**, J.F. Bonastre, C. Fredouille. *Proc. RFIA 98.* Clermont-Ferrand, France. 20-22 Janvier 1998.

" Traitement indépendant de sous-bandes fréquentielles par des méthodes statistiques du second ordre pour la reconnaissance automatique du locuteur. " **L. Besacier**, J.F. Bonastre. 4th French Congress on Acoustics. Marseille, France, 14-18 April 1997.

"Caractéristiques individuelles de la durée vocalique intrinsèque en français lu; une étude pilote. " D. Duez, **L. Besacier.** 4th French Congress on Acoustics. Marseille, France, 14-18 April 1997.