Université Pierre et Marie Curie

Indexation et Interface Homme-Machine
Reconnaissance d'un signal vocal

Mémoire d'habilitation à diriger des recherches

soutenu le 14 janvier 1999

par

Claude MONTACIE

Ingénieur de l'Ecole Supérieure d'Electricité
Docteur de l'Ecole Nationale Supérieure des Télécommunications

Devant le Jury composé de :

Daniel	LAZARD	Président
Marc	EL BEZE	Rapporteur
Sadaoki	FURUI	Rapporteur
Jean-Luc	GAUVAIN	Rapporteur
Jean-Pierre	DESCLES	Examinateur
Xavier	RODET	Examinateur
Christel	SORIN	Examinateur

Table des Matières

Table of Contents

Présentation du mémoire

Les travaux présentés dans ce mémoire couvrent la période 1991-1998. Ils s'intègrent dans le thème "Apprentissage et Acquisition des Connaissances" du Laboratoire d'Informatique de l'Université Pierre et Marie Curie (LIP6). Ils se situent dans le domaine du traitement informatique de la parole et des méthodes et architectures logicielles nécessaires pour traiter un tel signal. Cette recherche est fondée sur la coopération de modèles statistiques complexes pour extraire deux informations *a priori* complémentaires : les caractéristiques du locuteur et le contenu du discours. Même si chacun de ces thèmes a sa propre problématique, ils concourent tous à une Communication Parlée Homme/Machine utilisable dans des conditions réelles.

Les recherches en traitement informatique de la parole ont longtemps été freinées par l'absence de bases de test, indispensables à la comparaison des méthodes. La disponibilité récente de telles bases a fait progresser énormément les recherches mais au prix d'une sophistication des méthodes employées. Ces méthodes sont pour l'essentiel basées sur une modélisation statistique, car la taille importante des bases d'apprentissage rend difficile l'utilisation de méthodes orientées " connaissance ". L'utilisation de connaissances est alors reportée dans le choix d'une modélisation statistique adaptée. Cette adaptation va dépendre de la tâche de reconnaissance de la parole ou du locuteur : conditions de bruit, taille du vocabulaire, nombre de locuteurs. Le savoir-faire d'une équipe en traitement automatique de la parole se matérialise par la constitution d'une série de bibliothèques de modules logiciels de modélisations et de traitement permettant de résoudre en grande partie de nouveaux problèmes et guidant les avancées théoriques nécessaires. Ce paradigme d'évolution a permis d'obtenir des technologies vocales applicables dans un futur proche à des domaines aussi divers que le contrôle d'accès ou l'indexation de données multimédia. Avec l'apparition de machines à faible coût suffisamment puissantes pour traiter et reconnaître en temps réel les caractéristiques d'un signal de parole, ces avancées scientifiques vont être mises à la disposition du grand public.

Ce manuscrit comporte quatre parties. La première partie introductive rappelle brièvement l'évolution des différentes problématiques du traitement automatique de la parole. Ce domaine de recherche existe depuis près d'un demi-siècle et compte un certain nombre de congrès spécialisés depuis une trentaine d'années.

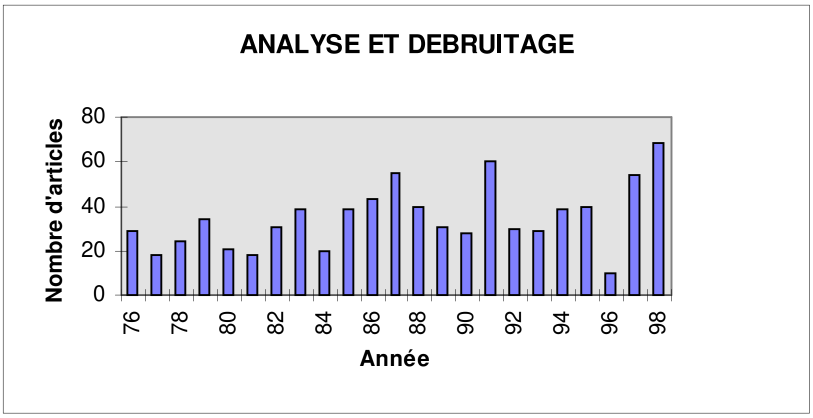
Les deux parties suivantes décrivent les travaux effectués en reconnaissance de la parole et du locuteur ainsi que ma contribution à deux réalisations importantes Orphée, une serrure vocale, et D-DAL, un système de dictée vocale. Dans ces deux parties, l'exposé est organisé de façon chronologique.

La dernière partie traite de mes travaux dans l'étude et la conception de deux systèmes multimédia : l'un consacré à l'interaction homme-machine multimodale, l'autre à la recherche d'informations dans les vidéos.

I Traitement informatique de la Parole

Les applications du traitement informatique de la parole sont l'analyse-débruitage, le codage, la synthèse et la reconnaissance d'un signal vocal. Pour mesurer l'intérêt que porte la communauté scientifique à ces problèmes, je me suis proposé d'analyser la répartition des travaux publiés dans un des congrès phares du domaine " IEEE-ICASSP International Conference on Acoustic, Signal and Speech Processing ". Plus de 3500 contributions sur le traitement de la parole ont été publiées depuis sa création en 1976 et permettent de cerner l'évolution des problématiques de ce domaine. J'ai toujours calculé le pourcentage de contributions par rapport à l'ensemble des publications de ce congrès en traitement de la parole. Plusieurs sous-domaines ont été choisis : analyse et débruitage d'un signal vocal, codage, synthèse, reconnaissance de la parole, reconnaissance de la langue, reconnaissance du locuteur. Les publications d'une session ont été, soit rattachées à un seul de ces sous-domaines si le titre de la session était sans ambiguité, soit examinées une à une. Les deux autres congrès internationaux dans ce domaine sont " EUROSPEECH : European Conference on Speech Communication and Technology" créé en 1989 et " ICSLP : International Conference in Spoken Language Processing " un an plus tard. Ces deux congrès sont bisannuels et alternés. Ils ne sont pas pris en compte dans les histogrammes établis ci-dessous.

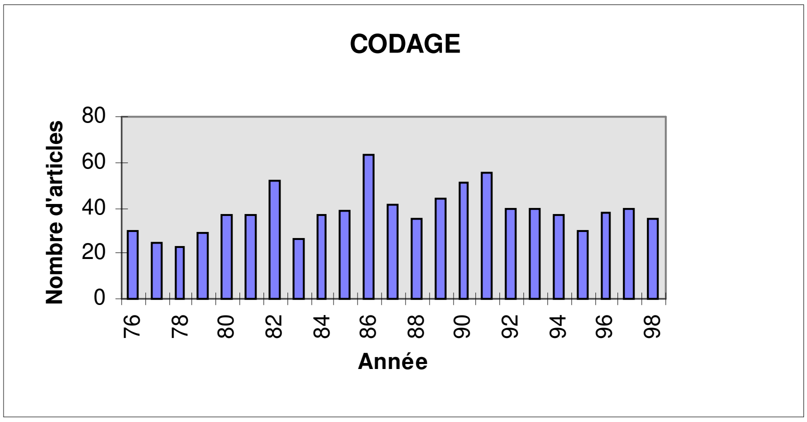
L'analyse et le débruitage d'un signal vocal sont les applications les plus proches du traitement du signal, le débruitage étant considéré comme une analyse robuste au bruit. Elles représentent 18% des contributions et sont au cœur du traitement informatique de la parole.



L'analyse de la parole consiste à extraire du signal de parole des paramètres ou des indices en fonction des contraintes de l'application. Les études de la production et de la perception de la parole sont principalement utilisées pour définir ces analyses. Il s'agit généralement d'analyse à court terme (environ 1/100 de seconde). On peut citer la mélodie, les paramètres provenant d'un modèle source-filtre.
En débruitage de la parole, l'étude des méthodes permet d'effectuer ces analyses même en cas de détérioration du signal de parole, entre son émission et son traitement, par un bruit additif ou convolutif. Ces recherches ont longtemps été très dépendantes du bruit étudié (e.g., téléphonie) et n'étaient pas généralisables. Dans le cas où la source et le type de bruit sont quelconques, la seule solution est d'utiliser simultanément plusieurs microphones (i.e., antenne de microphones). Le coût de ces antennes a été longtemps prohibitif mais la généralisation des techniques de visioconférences les a rendues

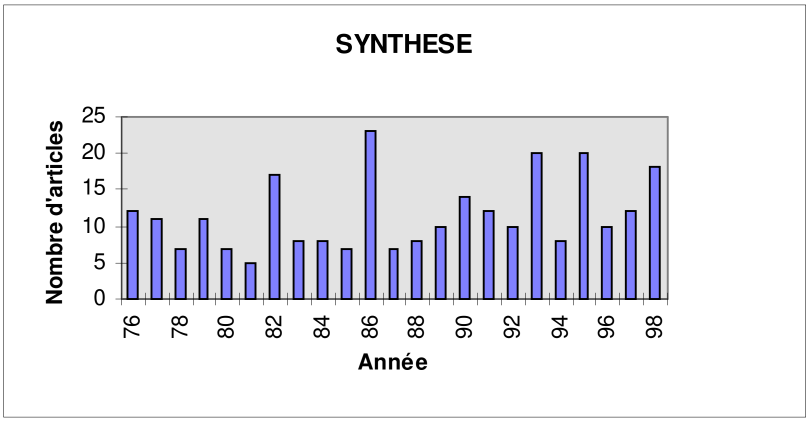
accessibles. Les problématiques du traitement d'antenne sont la topologie de l'antenne, la suppression des échos et l'utilisation des moments d'ordre supérieur.

Le codage d'un signal vocal représente environ 23% des contributions. La principale application est encore la communication par le réseau téléphonique. La contrainte la plus importante est de ne pas introduire un retard préjudiciable à une conversation (environ 1/10 de seconde). La problématique initiale de ce domaine était d'obtenir un algorithme de codage efficace ayant un rapport coût de calcul/facteur de compression le plus faible possible. Avec la diminution des coûts de calcul, cette problématique a en partie perdu de son importance et des algorithmes très coûteux comme les méthodes CELP sont continuellement améliorés.



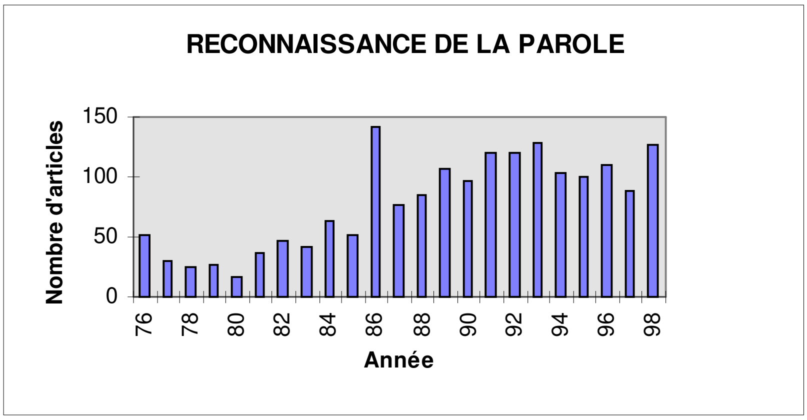
Une application plus récente du codage est le stockage d'un signal audio pour diffusion ultérieure. La principale différence vient de la qualité demandée qui doit être " transparente ", c'est-à-dire perceptivement identique à l'original.
Toutes ces recherches ont donné un certain nombre de normes : GSM en téléphonie, le système MUSICAM en stockage utilisé dans les normes MPEG (couches [1-3]). Ce dernier système devrait être à la base de toutes les futures applications multimédias : il permet actuellement le codage transparent d'un signal Haute-Fidélité stéréo avec un débit de 256 Kbauds.

La synthèse, (7% des contributions), a permis d'obtenir des signaux vocaux synthétisés d'excellente qualité. Les méthodes utilisées sont basées sur la concaténation de segments de bases comme les diphones. Il existe maintenant des outils de qualité disponibles sur le réseau internet comme MBR-PSOLA qui permet une synthèse à partir d'une suite de phonèmes et d'informations prosodiques.



La problématique de ce domaine est d'obtenir une qualité indiscernable de la voix humaine avec la génération automatique d'une meilleure prosodie et le concept de conversion de voix qu'il s'agisse d'imiter une voix bien précise ou les caractéristiques générales d'une population (e.g., accents). On peut remarquer à cet effet l'excellente synthèse d'une voix de castrat réalisée par " morphing " de deux voix : l'une féminine, l'autre masculine (Farinelli - X. Rodet).

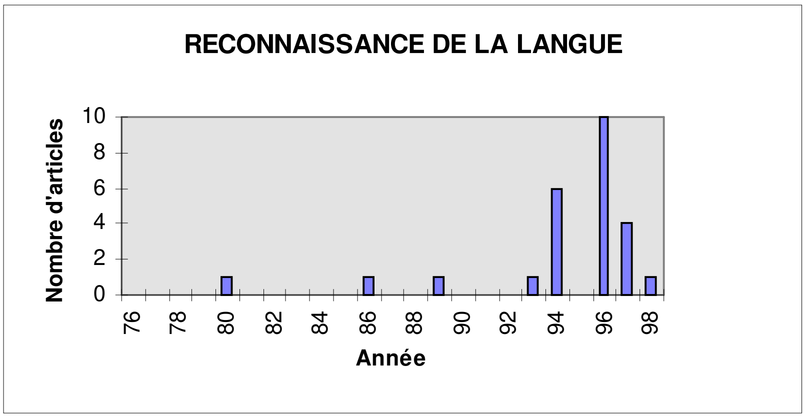
La reconnaissance d'un signal vocal (51% des contributions) recouvre trois applications : la reconnaissance du texte prononcé, de la langue employée et du locuteur. Pour donner une idée de l'ampleur du problème, il y a 6 milliards de locuteurs, 3000 langues et quelques centaines de milliers de mots par langue. La première application représente la majorité des contributions (45%) bien que les deux autres suscitent un intérêt croissant. Les problématiques successives de la reconnaissance du texte prononcé (appelé reconnaissance de parole) sont la reconnaissance de phonèmes (4% des contributions), de mots (7%), de parole continue (8%), la reconnaissance dans le bruit (8%), la modélisation markovienne de la parole (12%). Le reste des contribution traite principalement de la constitution de bases de données de parole ainsi que d'évaluations de systèmes de reconnaissance. On trouve les résultats suivants en calculant ces pourcentages seulement sur les publications de 1998 : reconnaissance de phonèmes (0%), de mots (0%), de parole continue (0%), la reconnaissance dans le bruit (38%), la modélisation markovienne de la parole (39%). Cela montre toute l'évolution qu'a subie ce domaine.



La problématique de la reconnaissance de phonèmes est la localisation et l'identification des phonèmes. La principale application est le décodage acoustico-phonétique (DAP) qui extrait à partir d'une analyse du signal vocal, une hypothèse de treillis des phonèmes. Même si les utilisations de ce treillis sont nombreuses en reconnaissance de parole, le DAP était principalement utilisé comme paradigme d'évaluation dans tous les domaines du traitement de la parole. Mais il n'est pas prouvé que la cause d'une amélioration d'un DAP se répercute à d'autres problématiques. La reconnaissance de mots comporte deux applications : la reconnaissance de suites de mots connectés avec une grammaire limitée (e.g., nom épelé) et la détection d'un nombre limité de mots-clés dans la parole continue. La reconnaissance de la parole continue ou " dictée vocale " consiste à reconnaître tous les mots d'une phrase prononcée de manière continue, c'est-à-dire sans pause artificielle entre les mots. Cette problématique, longtemps considérée comme impossible, est en voie de résolution partielle pour des textes lus, même si les taux de reconnaissance chutent quand l'élocution est plus spontanée. La reconnaissance dans le bruit consiste à prendre en compte la modification possible des conditions entre le test et l'apprentissage dans les techniques de reconnaissance de la parole. Comme les

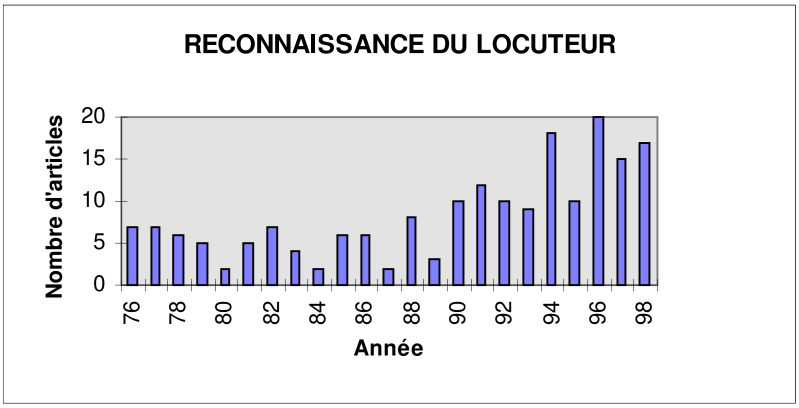
techniques de débruitage ne donnent pas actuellement satisfaction, la solution employée est de compenser ces différences en bruitant les bases d'apprentissage. Cette solution n'est évidemment pas satisfaisante même si elle améliore les performances de reconnaissance.
La modélisation markovienne de la parole est devenue un sujet de recherche en lui-même quand il s'est avéré qu'elle était la seule technique (pour l'instant) d'intégration des informations acoustiques et linguistiques “ utilisable ”. Le mot “ utilisable ” veut dire que l'on dispose de méthodes de modélisation, d'apprentissage et de reconnaissance suffisamment adaptables et automatisables pour prendre en compte rapidement de nouvelles conditions d'utilisation (environnement, locuteur, langue). Les techniques de modélisation markovienne du langage parlé se décomposent en deux parties égales en nombre de contributions : la modélisation acoustique et la modélisation du langage. La modélisation d'une forme acoustique est basée sur un automate probabiliste ou les paramètres d'analyse sont émis par les transitions entre états. Les recherches actuelles portent sur la topologie, l'apprentissage et l'adaptation rapide des paramètres de cet automate. La modélisation linguistique est fondée sur l'utilisation de n-grammes et de n-classes ; c'est-à-dire la probabilité d'une suite de n-mots ou de n-classes (ensemble de mots). La problématique actuelle est de prendre en compte le contexte, de s'adapter à des formes de phrases différentes du langage écrit sans augmenter la complexité de la reconnaissance.

Les contributions sur la reconnaissance de la langue employée sont très peu nombreuses (1% des contributions) puisque ce n'est qu'en 1996 qu'une session entière lui a été consacrée. La première contribution date de 1980, les auteurs arrivaient à un taux d'identification de 80% pour des segments de parole de 4 minutes en cinq langues.



La principale problématique a été de diminuer la durée des segments tout en augmentant le nombre de langues. Les recherches ont pu être grandement facilitées par la disponibilité récente de la base OGI (collectée par Oregon Graduate Institute) dédiée à l'identification des langues comportant des communications téléphoniques exprimées en langage spontané en 22 langues. Deux approches sont étudiées pour cette problématique, l'une est purement acoustique, l'autre plus difficile à mettre en oeuvre est linguistique. L'approche acoustique consiste à représenter une langue par les réalisations acoustiques de ses phones et des règles phonotactiques. Cette approche donne de très bons résultats si les langues ont un domaine acoustique nettement différencié, mais trouve rapidement ses limites quand le nombre de langues augmente. L'approche linguistique est basée sur la reconnaissance de mots spécifiques d'une langue, l'importance des ressources linguistiques à mettre en œuvre (dictionnaire et modèle de langage pour chaque langue à reconnaître) rend cette approche difficile pour l'instant mais plus prometteuse que la précédente. Les applications de cette technique restent cependant peu nombreuses en dehors des services de renseignements.

Même si la reconnaissance du locuteur est une problématique aussi ancienne que la reconnaissance de la parole, elle ne représente que 5% des contributions. Le premier système automatique de reconnaissance de la parole date de 1963, mais ce n'est que tout récemment que ce domaine a éveillé l'intérêt d'un nombre croissant de chercheurs et d'industriels. Cet engouement s'explique par la disponibilité de bases de données comprenant un grand nombre de locuteurs. Les applications sont l'identification du locuteur, la détection de changement de locuteurs et la vérification du locuteur. Au début l'utilisation de mots-clés était nécessaire, mais la plupart des systèmes actuels sont indépendants du texte prononcé.



Les techniques développées utilisent trois types d'approches : statique, dynamique et linguistique. Les systèmes de reconnaissance du locuteur peuvent naturellement combiner plusieurs de ces approches. Les deux premières approches s'appuient uniquement sur des informations purement acoustiques. Elles se différencient par l'utilisation ou non de paramètres représentant l'évolution temporelle du signal vocal. L'approche linguistique est fondée sur l'analyse du lexique employé et des variantes dialectales utilisées, mais il n'existe pas actuellement de méthodes automatiques. Les techniques développées actuellement permettent la vérification de l'identité d'une personne coopérative si les conditions d'acquisition et du test restent identiques. Cependant il ne peut être question de parler d'empreinte vocale par analogie avec les empreintes digitales ou génétiques ; la fiabilité de l'identification vocale n'étant pas du même ordre de grandeur.
La problématique actuelle est de compenser une modification des conditions d'acquisition, de mixer la vérification du locuteur avec celle du texte prononcé, de détecter les changements de locuteurs dans les bases de données multimédia.

Le tableau ci-après récapitule le pourcentage de contributions de chacun des grands axes de la recherche en traitement informatique de la parole.

Nom de domaine	Pourcentage de contributions
Analyse et débruitage	18%
Codage	23%
Synthèse	7%
Décodage acoustico-phonétique	4%
Reconnaissance de mots	7%
Reconnaissance de parole continue	8%
Reconnaissance dans le bruit	8%
Modélisation markovienne acoustique	6%
Modélisation markovienne du langage	6%
Bases de données, évaluations	7%
Reconnaissance de la langue	1%
Reconnaissance du locuteur	5%

Pourcentage de contributions (ICASSP'76-98) en fonction du domaine

Après ce bref exposé des problématiques du traitement informatique de la parole, je vais présenter mes travaux et contributions depuis ma thèse, en 1991. J'ai intégré à cette époque le groupe parole du LAFORIA (Laboratoire des Formes de d'Intelligence Artificielle) qui est devenu une partie du LIP6. J'ai pris avec M.-J. Caraty la responsabilité de ce groupe à la suite du professeur X. Rodet. A part ces deux permanents, ce groupe n'était constitué que de deux ou trois thésards. La plupart de mes travaux pendant cette période ont été faits en collaboration avec M.-J. Caraty. Ils se situent dans le domaine de l'analyse de la parole, de la reconnaissance du locuteur et de reconnaissance de la parole. Ils ont permis la définition de nouvelles approches : multirésolution généralisée pour l'analyse de la parole, modélisation AR-Vectorielle de spectres à très court-terme en reconnaissance du locuteur, nouvel estimateur à base de K-ppv et lexique dynamique pour la reconnaissance de la parole. Ces travaux ont également donné lieu à la définition et la création de ressources (bases de données, lexique, modèle de langage) indispensables au développement de systèmes comme Orphée la serrure vocale et D-DAL la dictée vocale. Ils ont donné l'occasion d'aborder de nouveaux domaines comme la recherche d'informations dans les documents multimédia.

II L’Analyse de la parole

Mes travaux en analyse de parole ont été guidés par l’objectif suivant : introduire les connaissances du domaine et les contraintes d’application directement au niveau de l’analyse. Le but était d’éviter une approche purement “ traitement du signal ” en général déconnectée des problématiques du traitement informatique de la parole, ou d’utiliser des analyses de parole considérées comme normatives.

2.1. La décomposition temporelle généralisée à la mesure de dissimilarité APS

La décomposition temporelle modélise l'évolution spectrale d'un segment de parole, Cette méthode décrit les vecteurs d’analyse par un modèle d'interpolation linéaire. Introduit comme technique de codage [Atal83], sa formulation était motivée par le fait que les paramètres calculés sur des fenêtres successives sont fortement corrélés. Il était donc possible de représenter leur évolution par un nombre plus petit de paramètres. C’est un moyen de détecter les courbures de la trajectoire des vecteurs d’analyse, et de repérer la succession d'attracteurs locaux (ou cibles) par lesquels la trajectoire passe ou semble se diriger successivement. La décomposition temporelle a été développée et utilisée dans différents laboratoires pour de nombreuses applications du traitement de la parole : -le codage, -la synthèse, -le décodage acoustico-phonétique, -la segmentation et -l'étiquetage.
J’ai développé un cadre théorique pour utiliser la décomposition temporelle dans un espace non euclidien en utilisant la notion de géodésique [Montacié91]. Car, la distance euclidienne n'est pas toujours la meilleure mesure de dissimilarité entre vecteurs d’analyse, au sens des performances de reconnaissance. Nous avons choisi pour cette étude la mesure de dissimilarité APS (Appariement de Pics Spectraux) [Caraty87]. L’analyse obtenue permet d’obtenir une ou deux cibles par voyelles sous forme d’ensemble de pics spectraux.

Collaboration : Marie-José Caraty, Claude Barras (thésard)

Bilan sur la période 90-92 : 3 articles dont 1 international [20] et 2 nationaux [38, 39], 1 workshop [53]

Recueil de publications : pp. 101

2.2. Déconvolution homomorphique et paramétrisation

Le modèle source-filtre est l’un des modèles les plus utilisés en traitement de la parole. Sa problématique est qu’une estimation fiable de la partie source et de la partie filtre est nécessaire, car seule la partie filtre est théoriquement utilisable en reconnaissance. La méthode pour estimer la partie filtre est d’utiliser des méthodes de déconvolution homomorphique. Nous avons testé la méthode dite de “ True Enveloppe ” [Imai81] qui permet d’obtenir des enveloppes spectrales plus précises que par la méthode classique du cepstre. Les paramétrisations testées étaient celles les plus utilisées en reconnaissance de la parole : les MFCC (Mel Frequency Cepstral Coefficients) et les

Pics Spectraux. Les résultats de mes travaux ont montré que l'amélioration de la séparation source-filtre n'apportait aucune amélioration substantielle à la paramétrisation MFCC. Une explication possible est que cette paramétrisation est très utilisée, car elle est robuste quand la séparation source-filtre est de mauvaise qualité. En effet, cette enveloppe spectrale est intégrée par des filtres de bandes passantes importantes (e.g., 100 Hz en basse fréquence), ce qui provoque une atténuation des améliorations apportées à l'estimation de l'enveloppe.
Les paramètres issus des analyses décrites dans cette section et dans la section précédente ont été utilisés dans un système d'apprentissage symbolique [Ganascia86] pour la reconnaissance de voyelles [Montacié90].

2.3. La mesure de la stationnarité par la technique de multirésolution généralisée

L'hypothèse de base de l'analyse spectrale est que les signaux analysés sont stationnaires ; c'est-à-dire que leurs propriétés (moyenne , écart-type, moments d'ordres supérieurs) restent identiques pendant la période d'analyse. Ce n'est que dans ces conditions que cette analyse a un sens physique. Dans le cas contraire ce n'est qu'une transformation mathématique qui n'a pas de rapport avec l'activité réelle de l'oreille humaine. Or une partie importante des segments de parole n'est pas stationnaire et une analyse spectrale sans vérification de stationnarité ajoute des artefacts qui théoriquement doivent rendre plus difficiles les traitements ultérieurs comme la reconnaissance. Nous avons, en collaboration avec M.-J. Caraty, introduit la notion de "multirésolution généralisée" qui permet de détecter ces non-stationnarités. Les méthodes développées sont dites généralisées, car elles permettent une rétroaction entre n'importe quelle méthode d'analyse et un processus quelconque de décision qu'il s'agisse de mesures de dissimilarité à minimiser ou de probabilité d'émission à maximiser.

Collaboration : Marie-José Caraty

Bilan sur la période 97-98 : 1 article international [32]

Recueil de publications : pp. 105

Références

[Atal83] B.S. Atal, " Efficient Coding of LPC Parameters by Temporal Decomposition ", IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 81-84, **1983**.

[Caraty87] M.-J. Caraty, " Contribution au décodage acoustico-phonétique. Etudes de distances inter-spectres et reconnaissance de cycles vocaliques ", Thèse de l'Université Paris 6, **1987**.

[Ganascia86] J.-G. Ganascia, " Learning with Hilbert Cubes ", 2nd European Working session on Machine Learning, Sigma Press, pp.158-171, **1986**.

[Imai81] S. Imai & Y. Abe, " Spectral Envelope Extraction by Improved Cepstral Method. " Electronic and communication, In Japan, Vol 62A, n°4, pp. 10-17, **1979**.

[Montacié90] C. Montacié, M.-J. Caraty & X. Rodet, “ Experiments in the Use of an Automatic Learning System for Acoustic-Phonetic Decoding ”, International Conference on Spoken Language Processing, Kobe, Japan, pp. 357-390, **1990**.

[Montacié91] C. Montacié, “ Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformations spectrales non-linéaires ”, Thèse de l'ENST, **1991**.

III La reconnaissance du locuteur

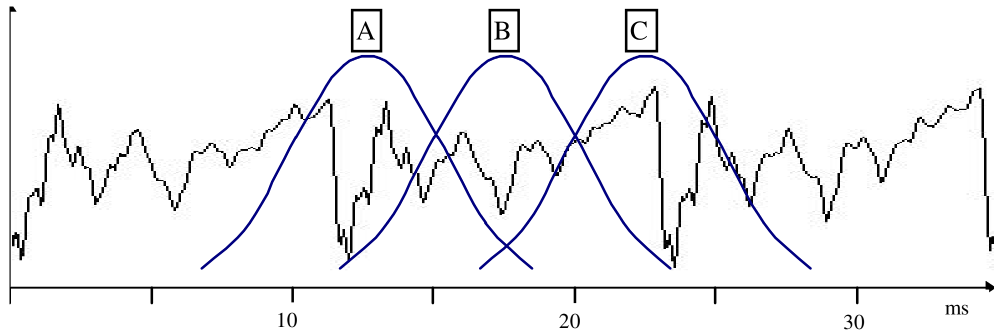
Mes travaux en reconnaissance du locuteur indépendamment du texte sont fondés sur l'intégration des informations sur la source laryngienne et de celles sur le conduit vocal. J'ai utilisé pour cela la modélisation AutoRégressive Vectorielle associée à une analyse originale du signal de parole. Les techniques développées ont fait l'objet d'un dépôt de brevet en 1996, étendue en 1997 aux principaux pays développées. Le rapport de recherche international ne cite que des documents de type A[1], c'est-à-dire non opposable à cette invention. Ce brevet a fait l'objet de licences et est exploité commercialement par diverses sociétés. Après un bref rappel de l'état de l'art des techniques, nous décrirons nos principales résultats en ce domaine.
J'ai classé les méthodes de reconnaissance automatique du locuteur en fonction de l'utilisation ou non de paramètres représentant l'évolution temporelle du signal vocal Les premières que j'appelle " dynamiques " sont généralement dérivées de celles utilisées en reconnaissance de la parole : programmation dynamique [Furui81], modélisation markovienne, réseaux de neurones. Il s'agit des modèles de Markov cachés gauche-droit [Naik89], des modèles de Markov cachés ergodiques [Poritz82], des réseaux de neurones multicouches (Time Delay Neural Network) [Bennani91]. Les méthodes " statiques " font l'hypothèse d'une indépendance entre les vecteurs d'analyse successifs. Elles utilisent des techniques classiques d'analyse de données. Il s'agit de vecteurs prototypes obtenus par quantification vectorielle [Li83], de mesures interlocuteurs sur la matrice de covariance [Bimbot93], de distribution des paramètres modélisée par une somme pondérée de densités de probabilités gaussiennes (Gaussian Mixture Model) [Reynolds94-95]. Cette dernière modélisation qui nous fait revenir aux sources de la Reconnaissance des Formes est paradoxalement la plus utilisée et l'une des plus performantes.
J'ai développé pour ma part une technique originale de reconnaissance du locuteur permettant d'intégrer des informations sur la source laryngienne et celles sur le conduit vocal.

3.1. Intégration des informations sur la source laryngienne et de celles sur le conduit vocal

D'un point de vue physiologique, des informations caractéristiques du locuteur peuvent être trouvées dans l'étude de sa source laryngienne ainsi que dans celle de son conduit vocal. Dans les systèmes de reconnaissance du locuteur traitant ces deux informations, celles-ci sont extraites séparément à partir d'analyses spectrales à court terme puis fusionnées [Matsui90]. Les fenêtres temporelles sont choisies de durée supérieure à la période du fondamental (e.g., 1/100 seconde pour les hommes) pour permettre la séparation des deux informations. Deux séries de paramètres sont alors obtenues : -une estimation de la fréquence fondamentale -une des paramétrisations de la reconnaissance

[1] Le rapport de recherche d'un brevet comporte les documents (articles , conférence orale, brevets, ...) pertinents antérieurs. Ceux-ci peuvent être opposables aux revendications du brevet et les rendre caduques. Les catégories de documents pertinents sont X (particulièrement pertinent à lui-seul), Y (particulièrement pertinent en combinaison avec un autre document de même catégorie), E (document de brevet bénéficiant d'une date antérieure mais publié postérieurement), O (divulgation non-écrite), A (arrière-plan technologique). En conclusion un brevet dont le rapport de recherches comporte des documents X, Y, E ne fait pas preuve d'inventivité.

de la parole (e.g., MFCC, LPCC). Des techniques de fusion sont alors utilisées en fusionnant soit les paramètres, soit les résultats de reconnaissance. Aucune amélioration significative n'a pour l'instant été apportée par l'utilisation de la fréquence fondamentale. Plusieurs explications sont possibles : mauvaise estimation de la fréquence fondamentale, insuffisance de ce paramètre pour représenter la source laryngienne, techniques de fusion inadaptées. Ma démarche a été de ne pas chercher à séparer l'information provenant de la source laryngienne de celle du conduit vocal. La méthode que nous avons retenue a été d'utiliser deux niveaux de modélisation : une analyse à très court terme sensible à la source laryngienne et une modélisation autorégressive vectorielle à partir de ces analyses. J'ai montré expérimentalement que dans ce cas, cette dernière modélisation devenait sensible aux variations de cette source et aux caractéristiques du spectre moyen.
La technique de modélisation autorégressive vectorielle (ou prédiction linéaire vectorielle) est utilisée pour traiter des signaux à plusieurs composantes [Grenier80][Kakusho82]. Le modèle est un ensemble de q matrices de prédiction qui permettent de prédire un échantillon d'un signal vectoriel à partir des q échantillons précédents. Elle n'a été employée que rarement en traitement de la parole. Son principal avantage est la simplicité de sa méthode d'estimation dérivée de celles développées pour la modélisation autorégressive scalaire. Les matrices de prédiction sont estimées par l'algorithme de Levinson-Whittle-Robinson [Whittle63] qui minimise la trace de la matrice covariance des erreurs de prédiction. Le signal vectoriel que nous utilisons est un vecteur d'analyse à très court terme. La principale difficulté a été de trouver le lien entre les paramètres de l'analyse (taille de la fenêtre, pas de l'analyse) avec ceux de la modélisation autorégressive vectorielle (ordre q). J'ai choisi l'heuristique suivante que nous avons ensuite confirmée expérimentalement : le pas de l'analyse multiplié par l'ordre q doit être proche de la taille de la fenêtre d'analyse. Ce choix s'explique dans le cas où la fenêtre d'analyse est égale à la période du fondamental (cf. figure suivante), car dans ce cas chacun des vecteurs d'analyse utilisés pour la prédiction est centré sur une phase différente de l'onde glottique.



Prédiction de la fenêtre C à partir des fenêtres A et B

Les techniques de reconnaissance du locuteur développées à partir de ces principes se sont révélées très efficaces, robustes au bruit, à la dérive temporelle ainsi qu'à un engorgement des voies respiratoires provoqué par un rhume ou une bronchite par exemple. Ces techniques restent cependant sensibles aux variations du style d'élocution (voix lue, préparée, spontanée) ainsi qu'à l'état émotionnel.

Bilan sur la période 91-98 : 1 brevet avec extension internationale [1]

3.2. Développement de distances inter-locuteurs

Une fois la modélisation définie, une mesure de dissimilarité entre un modèle de locuteur et un modèle d'une élocution est indispensable, pour toute application en reconnaissance du locuteur. Mes expériences ont montré que le principe de décision devait être fondé sur la comparaison des erreurs de prédiction produites par ces deux modèles et non sur la trace de la matrice de covariance de ces erreurs. Cela permet de ne pas prendre en compte les erreurs de prédiction dues à des bruits qui ne peuvent être prédits par un modèle autorégressif. Ce principe est très efficace, mais il a un coût de calcul prohibitif pour des applications temps-réel. Dans ce cas, l'approximation que j'ai choisie est d'utiliser des matrices de covariance d'erreurs à la place des erreurs. Nous avons généralisé pour cela les mesures définies pour les modèles de type autorégressif scalaire : les distances d'Itakura-Saito [Itakura75]. Les algorithmes rapides que j'ai développés ont une complexité proportionnelle au carré de l'ordre du modèle multiplié par le cube de la taille des vecteurs d'analyse.
Ces distances interlocuteurs testées sur la base de données TIMIT (base de données de voix lues en anglais américain) [Fisher86] ont permis de reconnaître un locuteur parmi 630 sans erreur avec cinq secondes de parole. Sur cette base, le taux d'erreur en reconnaissance du locuteur était inversement proportionnel à la durée du signal de parole. Cependant, ces résultats ont été obtenus à partir d'un maximum de 12 secondes de parole et ne peuvent *a priori* pas être extrapolés pour une durée supérieure (e.g., 1 heure). Il est certain que ces performances ne peuvent que se dégrader quand le nombre de locuteurs augmente. Cela est dû au fait que les locuteurs sont de plus en plus proches les uns des autres et que les risques de confusion s'accroissent. Des expériences sur des sous-ensembles de TIMIT semblent montrer que ce taux d'erreur est fonction de la racine carrée du nombre de locuteurs [Webb93]. Ces résultats nous ont poussés à développer et à tester un prototype de serrure vocale dans des conditions réelles. Il est décrit en section 3.3 ci-après.

Collaboration : Jean-Luc Le Floch (thésard)

Bilan sur la période 91-92 : 4 articles dont 2 internationaux [18, 19] et 2 nationaux [37, 40]

Recueil de publications : pp. 109

3.3. Le système Orphée et la serrure vocale

La serrure est une application de la vérification du locuteur. L'utilisateur d'une telle serrure décline son identité au système par un moyen quelconque (carte magnétique, code) et prononce une courte phrase. Le système vérifie ensuite la compatibilité des deux informations pour autoriser l'accès.
Orphée, notre serrure vocale modélise un utilisateur par un seul modèle autorégressif vectoriel (7 Ko). Un code temporaire permet une procédure d'apprentissage incrémentale non supervisée. La vérification comporte deux phases : une phase d'identification puis une phase de vérification proprement dite. Une phase d'identification est nécessaire, car plusieurs locuteurs peuvent avoir le même code de vérification. Il s'agit en effet, pour des raisons de mémorisation, du quantième du mois de naissance (code de 1 à 31). Pendant la phase de vérification, la distance interlocuteur entre la phrase de vérification et le modèle de la personne à identifier est comparée à un seuil. Cette distance est normalisée par la méthode des cohortes [Higgins90] [Matsui94]. Cette méthode consiste à mettre en concurrence le modèle de la personne à

identifier avec un ensemble d'autres modèles. Pendant toute notre expérimentation, le seuil ne dépendait pas du locuteur et n'a pas été modifié. Nous avons avec J-L. Le Floch installé et testé Orphée à la porte d'accès de notre laboratoire pendant 20 mois. Située dans un endroit extrêmement bruyant et réverbérant (i.e., une cage d'escalier en béton), cette serrure a été utilisée par 90 membres du laboratoire sans aucune entrée d'imposteur constatée. Les statistiques d'utilisation sont les suivantes : en moyenne, sur 20 tentatives par jour, on relève une tentative déjouée d'imposteur et 12 entrées réussies. L'analyse des 7 autres tentatives montre que ces rejets sont dus dans la majorité des cas à une trop courte durée de parole (i.e., < 2 secondes) ou à de mauvaises conditions d'enregistrement (e.g., bruit, intensité du signal de parole).

Collaborations : Marie-José Caraty, Jean-Luc Le Floch (thésard)

Bilan sur la période 93 : 1 article international [21]

Recueil de publications : pp. 113

3.4. Vérification du locuteur en continu

Le projet AMIBE (Applications Multimodales pour Interfaces et Bornes Evoluées) que je décrirai dans le dernier chapitre comprenait une tâche de reconnaissance du locuteur qui devait être couplée avec de la reconnaissance de visages. L'approche que nous avons développée dans ce projet avait pour but de vérifier en permanence l'identité prétendue de l'utilisateur pour éviter un changement de locuteur au cours d'une transaction. Deux axes de recherche ont été étudiés : le premier pour obtenir une décision de vérification en continu et non à la fin d'une phrase, le second pour déterminer la fiabilité de cette vérification en fonction du type de segments de parole. J.-L. Le Floch a développé une méthode de vérification en continu : la vérification analytique. Dans une vérification "analytique", tout ensemble de vecteurs d'analyse consécutifs permet une vérification du locuteur. La décision est prise en comparant le dernier vecteur avec son estimation par les modèles autorégressifs d'apprentissage à partir des vecteurs précédents. La fiabilité de la vérification analytique a été étudiée en fonction du type de segments phonétiques. Ces expériences de vérification analytique du locuteur montrent que les phonèmes les plus discriminants, par ordre décroissant, sont : les diphtongues, les voyelles, les nasales/liquides, les glissées, les plosives/fricatives/affriquées. Nous avons également constaté que l'élimination des phonèmes les moins discriminants permettait d'améliorer les performances de reconnaissance du locuteur. Ces résultats confirment de précécentes recherches [Sambur75][Atal76] [Rosenberg87] sur la caractérisation du locuteur .

Collaborations : Marie-José Caraty, Jean-Luc Le Floch (thésard)

Bilan sur la période 94 : 2 articles dont 1 international [22] et 1 national [43]

Recueil de publications : pp. 117

3.5. Coopération et Compétition de modèles

Les expériences sur la modélisation du locuteur par une somme pondérée de densités de probabilités gaussiennes (Gaussian Mixture Model) ont montré que les résultats donnés par cette méthode n'étaient pas corrélés avec ceux obtenus par les modèles autorégressifs vectoriels. Cela indiquait qu'il était peut-être possible d'obtenir de meilleures performances en fusionnant les deux méthodes de reconnaissance. En collaboration avec J.-L. Le Floch, nous avons alors défini des procédures de coopération et de compétition de ces modèles. La fusion pouvait se faire soit globalement au niveau de la phrase soit pour chaque vecteur d'analyse. La principale difficulté a été de trouver des méthodes de normalisation qui permettent de rendre comparables les erreurs de prédiction de notre méthode prédictive avec les densités de probabilité des sommes de gaussiennes. Une augmentation des résultats de 10% a été obtenue par la procédure de compétition au niveau de chaque vecteur d'analyse. Les mêmes expériences ont eu lieu en utilisant plusieurs modèles autorégressifs vectoriels par locuteur ; chacun étant appris sur des types de segments phonétiques distincts. Les performances augmentaient également mais dans une moindre proportion. Le coût de calcul important de ces techniques de fusion ne nous a pas permis de les intégrer dans Orphée.

Collaborations : Marie-José Caraty, Jean-Luc Le Floch (thésard)

Bilan sur la période 95-96 : 3 articles dont 2 internationaux [24, 26] et 1 national [45]

Recueil de publications : pp. 121

3.6. Le concours NIST 98

Le National Intitute of Standards and Technology organise chaque année des concours d'évaluation dans différents domaines de l'informatique (réseaux, réalité virtuelle, génie linguistique). Cet organisme spécifie un paradigme d'évaluation, il définit un cahier des charges et fournit des bases de données. Le concours est organisé en deux phases : une phase de développement et une phase de test. La phase de développement dure environ six mois ; c'est un test en blanc qui permet d'optimiser les systèmes à évaluer sur les données de développement. La phase de test a une durée limitée (e.g., 15 jours) : les données de test sont fournies le jour du début du test et les résultats doivent impérativement être renvoyés avant une date limite.

La dernière campagne de tests (09/02/08 au 23/02/98) sur la reconnaissance du locuteur a porté sur la vérification du locuteur au téléphone. Le paradigme d'évaluation imposait de répondre à 350 000 questions du type " Le segment x est-il prononcé par le locuteur y ?". Le fait que le même locuteur pouvait avoir été enregistré sur différents types de microphones (carbone ou électret) et de lignes téléphoniques était la principale difficulté. Une autre difficulté était due à la méthode de construction des segments de parole. Il s'agissait de regrouper de nombreux segments de courte durée provenant de de conversations téléphoniques (base de données SwitchBoard II). La tâche définie n'était plus dans ce cas de la vérification d'un locuteur coopératif mais se rapprochait plus d'une tâche d'identification du locuteur.

Notre entrée tardive dans le concours (2 mois de développement) nous a cependant permis d'acquérir une expérience sur des bases de données difficilement accessibles. Nous avons utilisé le système Orphée en rajoutant seulement un module de filtrage dans la bande téléphonique, sans compenser les différences de microphones et de lignes téléphoniques. Le taux d'égale erreur (EER) obtenu était d'environ 25% ; c'est-à-dire que 25% des réponses positives et 25% des réponses négatives étaient erronées. L'analyse de nos résultats a montré que les erreurs étaient principalement dues aux trois facteurs suivants: -présence de nombreux rires qui perturbent l'estimation des modèles prédictifs, -variation importante de l'état émotionnel d'un même locuteur entre les différents enregistrements, -présence de la voix d'autres locuteurs due à une mauvaise segmentation des conversations. En effet, notre méthode prenant en compte des informations sur la source laryngienne est particulièrement sensible à ces facteurs. Grâce aux enseignements de ce concours, nous avons déjà commencé l'adaptation d'Orphée à la vérification du locuteur au téléphone. Cela nous a également permis d'améliorer nos méthodes d'indexation multimédia que je détaillerai dans le chapitre 5.

Collaboration : Marie-José Caraty
Bilan sur la période 97-98 : 1 concours international d'évaluation [57]

Références

[Atal76] B. S. Atal, " Automatic Recognition of Speakers from Their Voices ", IEEE Proceedings, vol. 64, No. 4, pp. 460-475, **1976**.

[Bennani91] Y. Bennani & P. Gallinari, " On the Use of TDNN extracted features Information for Talker Identification ", IEEE-International Conference on Acoustic Signal and Speech Processing, Toronto, Canada, pp. 385-388, **1991**.

[Fisher86] W. Fisher , V. Zue, J. Bernstein & D. Pallet, " An Acoustic-Phonetic Data Base ", Journal of Acoustic Society of America, Suppl. (A), 81, S92, **1986**.

[Furui81] S. Furui, " Cepstral Analysis Technique for Automatic Speaker Recognition ", IEEE Transactions on Acoustic Signal and Speech Processing, Vol. 29, pp. 254-272, **1981**.

[Grenier80] Y. Grenier, " Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur", 11èmes Journées d'Etude de la Parole, Strasbourg, France, pp. 163-171, **1980**.

[Higgins90] A. Higgins, L. Bahler & J. Porter, " Speaker Verification Using Randomized Phrase Prompting ", Digital Signal Processing, vol. 1, pp. 89-106, **1991**.

[Itakura75] F. Itakura, " Minimum Prediction Residual Principle Applied to Speech Recognition ", IEEE Transactions on Acoustic Signal and Speech Processing, Vol. 23, pp. 67-72, **1975**.

[Kakusho82] O. Kakusho & M. Yanagida, " Hierarchical AR model for Time Varying Speech Signals ", IEEE-International Conference on Acoustic Signal and Speech Processing, Paris, France, pp. 1295-1298, **1982**.

[Li83] K.-P. Li & E.H. Wrench, " An approach to Text-Independant Speaker Recognition with Short Utterances ", IEEE-International Conference on Acoustic Signal and Speech Processing, Tokyo, Japan, pp. 555-558, **1983**.

[Matsui90] T. Matsui & S. Furui, " Text-Independant Speaker Recognition Using Vocal Tract and Pitch Information ", ", International Conference on Spoken Language Processing, Kobe, Japan, pp. 137-140, **1990**.

[Matsui94] T. Matsui & S. Furui, " Similarity Normalization Method for Speaker Verification Based on a Posteriori Probability ", Workshop on Automatic Speaker recognition identification and verification, Martigny, Suisse, pp. 59-62, **1994**.

[Naik89] J.M. Naik, L.P. Netsch & G.R. Doddington, “ Speaker Verification over Long Distance Telephone Lines ”, IEEE-International Conference on Acoustic Signal and Speech Processing ,USA, pp. 524-527, **1989**.

[Poritz82] A.B. Poritz, “ Linear Predictive Hidden Markov Model and the Speech Signal ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Paris, France, pp. 1291-1294, **1982**.

[Reynolds94] A. Reynolds, “ Speaker Identification and Verification using Gaussian Mixture Speaker Models ”, Workshop on Automatic Speaker recognition identification and verification, Martigny, Suisse, pp. 27-30, **1994**.

[Reynolds95] A. Reynolds & C. Rose, “ Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models ”, IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 72-83, **1995**.

[Rosenberg87] A. E. Rosenberg & F. Song, “ Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Dependent Modes ”, Computer Speech & Language, vol. 2, n. 3/4, pp. 143-157, **1987**.

[Sambur75] M. R. Sambur, “ Selection of Acoustic Features for Speaker Identification ”, IEEE Transactions on Speech and Audio Processing, pp. 176-182, **1975**.

[Webb93] J. J. Webb & E. L. Rissanen, “ Speaker Identification Experiments Using HMMs ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Minneapolis, USA, pp. 387-390, **1993**.

[Whittle63] P. Whittle, “ On the Fitting of Multivariate Autoregression and the Approximate Canonical Factorization of a Spectral Density Matrix. ”, Biometrika, Vol. 50, pp. 129-134, **1963**.

IV La reconnaissance de la parole

Mes recherches en reconnaissance de la parole ont porté sur l'adaptation au locuteur des systèmes de reconnaissance et sur une amélioration des densités de probabilité utilisées pour représenter l'espace acoustique. Ces études ont visé également à une meilleure connaissance des systèmes markoviens plutôt que de les utiliser comme des " boîtes noires ", ainsi que cela est souvent le cas.
La méthode choisie pour l'adaptation utilise une topologie des locuteurs définie par nos méthodes de reconnaissance du locuteur. Le remplacement des lois normales par un estimateur non paramétrique comme l'estimateur des K plus proches voisins avait pour but d'améliorer les densités de probabilité employées dans la modélisation markovienne.
Le paradigme d'évaluation utilisé était tout d'abord basé sur un système de décodage acoustico-phonétique qui présentait l'avantage d'être simple à implémenter. Notre équipe en a développé un autre depuis trois ans : un système de dictée vocale [Gauvain94b][Kubala94][Woodland95] indépendant du locuteur sans contrainte de vocabulaire. L'élaboration de ce système nous a permis d'aborder le domaine du génie linguistique, car nous avons dû en grande partie développer les ressources nécessaires à un tel système : dictionnaire phonétisé, modèles de langage, moteur de reconnaissance.
Le Dictaphone-Dactylographe du LIP6 (D-DAL) a participé au premier concours francophone de dictée vocale organisé par l'AUPELF-UREF en 1997.

4.1. Adaptation au locuteur

Malgré l'utilisation de bases de données comportant plusieurs centaines de locuteurs, il est toujours difficile de modéliser toute la variabilité interlocuteur. Cela rend nécessaire, pour l'instant, l'adaptation au locuteur des systèmes de reconnaissance. Cette adaptation peut être supervisée ou non supervisée. Dans le premier cas, le système d'adaptation se focalise sur les erreurs commises par le système de reconnaissance. Cela alourdit l'ergonomie, car la liste des erreurs doit être fournie par l'utilisateur. Dans l'autre cas l'adaptation utilise l'ensemble des paramètres d'analyse de test sans intervention de l'utilisateur. L'adaptation supervisée donne en général les meilleurs résultats mais elle n'est pas toujours applicable. Par exemple elle ne peut être utilisée pour les reconnaissances " off-line " (e.g., indexation). Deux approches sont généralement utilisées pour adapter les systèmes de reconnaissance : -l'estimation d'une transformation spectrale pour projeter au mieux les paramètres d'analyse du locuteur de test sur ceux de l'apprentissage [Shikano86], -l'utilisation de modèles markoviens adaptés au locuteur de test [Kenny90].
L'approche par projection est la plus ancienne, car elle peut être utilisée dans tous les types de systèmes de reconnaissance. Les principales difficultés sont -l'alignement des données de test sur les modèles d'apprentissage, -l'estimation de la fonction de transformation spectrale. Cette fonction est généralement une matrice de projection calculée par des techniques de régression multiple. J'ai utilisé pour ma part des perceptrons multicouches qui permettaient de modéliser des fonctions de transformation non-linéaires.
Dans l'autre approche, la méthode la plus simple et la plus employée consiste à utiliser des modèles markoviens par genre : un modèle appris sur les hommes, un autre sur les femmes. Une autre méthode est d'utiliser la technique dite du " maximum *a*

posteriori (MAP) ” [Gauvain94a]. Il s’agit d’adapter un modèle markovien appris sur un grand nombre de locuteurs en ne modifiant que les densités de probabilité incohérentes avec les données de test. La technique MAP est censée garder la généralité d’un modèle indépendant du locuteur tout en y ajoutant les caractéristiques du nouveau locuteur. Nous avons développé deux méthodes basées sur une topologie des locuteurs d’apprentissage pour adapter les systèmes de reconnaissance. Cette topologie était définie par des distances interlocuteurs basées sur la modélisation autorégressive vectorielle. Dans la première méthode, les modèles markoviens étaient réestimés sur les données provenant des locuteurs les plus proches des données de test, dans la seconde développée par C. Barras, les modèles adaptés étaient obtenus à partir d’une pondération d’un ensemble de modèles préalablement calculés sur des clusters de locuteurs [Barras96]. La première méthode était la plus performante comme nous l’ont montré nos expériences pendant la phase de développement de la campagne de dictée vocale, mais son coût de calcul est prohibitif. La seconde était plus rapide mais l’amélioration par rapport aux modèles par genre n’a pas été significative.

Collaborations : Marie-José Caraty, Claude Barras (thésard), Benjamin Mayrargue (stagiaire DEA),

Bilan sur la période 94-95 : 4 articles dont 1 dans un ouvrage de synthèse [7], 1 international [23] et 2 nationaux [41, 42]

Recueil de publications : pp. 125

4.2. Estimation des probabilités d’émission

Dans les modèles de Markov cachés, la probabilité d’émission d’un état est généralement représentée par une somme de multigaussiennes à matrice de covariance diagonale. Nous avons décidé d’étudier si ce choix était fondé et si une meilleure estimation de ces probabilités pouvait être obtenue par la règle des K plus proches voisins (Kppv) qui est théoriquement l’un des estimateurs les plus performants.
Une étude préliminaire de J. Goût a confirmé que de nombreuses probabilités d’émission étaient mal représentées par une somme de multigaussiennes. Nous avons utilisé pour le montrer un test du chi-carré entre la distribution théorique et la distribution réelle, calculé pour chaque coefficient d’analyse et pour chaque état. Le premier apport de cette étude a été une modification de l’algorithme d’apprentissage des multi-gaussiennes. L’amélioration a consisté à conditionner l’étape de séparation/perturbation d’une multigaussienne par un test de chi-carré multidimensionnel. Ce test a été construit par C. Barras comme la somme des tests de chi-carré sur chacun des coefficients en utilisant l’hypothèse d’indépendance des coefficients. Cette nouvelle méthode d’apprentissage a permis de diminuer de moitié le nombre moyen de multigaussiennes sans baisse de performance. Cependant, ce résultat n’était malgré tout pas totalement satisfaisant et nous avons décidé de lancer une nouvelle étude pour remplacer la modélisation par multigaussienne par un estimateur non paramétrique fondé sur la règle des Kppv. Cette règle consiste à estimer la probabilité d’appartenance d’une observation à une classe par la proportion de ses voisins dans cette classe. Malgré sa simplicité, elle est très efficace et présente les avantages théoriques suivants : l’erreur des 1ppv est toujours inférieure au double de l’erreur minimale (risque de Bayes), de plus cette erreur diminue quand K augmente (pour n suffisamment grand). La règle présente également l’intérêt de retrouver facilement l’origine d’une erreur. Son seul inconvénient est son coût de calcul prohibitif

par rapport à l'utilisation de multigaussiennes. Nous avons résolu ce problème en développant un algorithme rapide de calcul de Kppv que nous décrirons dans la prochaine section. Le gain de temps d'environ 99.8% apporté par cet algorithme rendait possible l'intégration de cet estimateur dans un système de reconnaissance markovien. Ce problème est étudié par F. Lefèvre depuis 3 ans. Il a dû adapter les algorithmes d'apprentissage (Baum-Welch) et de décodage (Viterbi) à ce nouvel estimateur. Les résultats obtenus ont montré les performances n'augmentaient pas sensiblement car le nouvel estimateur modifiait fortement le comportement des modèles markoviens. Si nous sommes certains de l'efficacité de cet estimateur pour une reconnaissance dans R^n (i.e., un seul vecteur d'analyse), cette efficacité n'est pas assurée quand la reconnaissance se passe dans $R^n \times N$ (i.e., un ensemble de vecteurs). Néanmoins, cela nous a donné un nouvel outil pour étudier plus finement le fonctionnement de ces modèles. Nous avons construit pour cela des paradigmes d'évaluation qui devraient nous permettre de mesurer plus exactement le rôle des delta-paramètres et des modèles phonétiques contextuels dans le processus de reconnaissance.

Collaboration : Marie-José Caraty, Fabrice Lefèvre (thésard), Jérôme Goût (stagiaire DEA), Cécile Lecordier (stagiaire DESS)

Bilan sur la période 97-98 : 3 articles dont 1 dans un ouvrage de synthèse [9] et 2 internationaux [31, 33]

Recueil de publications : pp. 129

4.3. Algorithme rapide des K plus proches voisins et gestion de la mémoire

Un tel algorithme était indispensable, car les bases de données utilisées en traitement de la parole doivent être importantes pour modéliser la variabilité du signal vocal. Nous avons choisi l'une des plus petites, TIMIT, qui correspond déjà à une durée supérieure à 4h. En prenant pour base une analyse centiseconde, cela fait plus d'un million et demi de vecteurs. Dans le cas où les deux tiers sont utilisés à l'apprentissage, une expérience de test demanderait le calcul de plus de 500 milliards de distances entre vecteurs.
Un premier algorithme basé sur les méthodes de Kittler-Richetin [Kittler78] [Richetin80], Friedman [Friedman75] et Fukunaga [Fukunaga75] a été développé. Son principe était d'utiliser un arbre hiérarchique pour partitionner l'espace des vecteurs d'apprentissage. Chaque noeud était caractérisé par son centre, son rayon, et l'ensemble de ses vecteurs. Pour chaque vecteur de test, une suite de parcours en profondeur avec retour arrière contrôlé permettait de sélectionner les noeuds terminaux susceptibles de contenir des Kppv. Seules les distances entre le vecteur de test et les vecteurs des noeuds terminaux sélectionnés étaient alors calculées. Cet algorithme diminuait effectivement le coût de calcul, mais pas la durée du traitement global. La différence venait du nombre important d'entrées-sorties provoquées par l'impossibilité de contenir en mémoire l'ensemble des vecteurs d'apprentissage.
Trois méthodes ont été utilisées avec succès pour résoudre ce problème : meilleure gestion de la mémoire, structuration de l'ensemble des vecteurs de test, ordonnancement de l'arbre de décomposition.
La technique dite " Least Recently Used (LRU) " utilisée pour la gestion de la mémoire a permis effectivement de diminuer le nombre d'entrées-sorties mais pas suffisamment. Le problème venait de la variabilité des vecteurs de test successifs qui oblige à explorer à chaque recherche de Kppv une partie différente de l'arbre de

décomposition. J'ai montré que le traitement d'une centaine de vecteurs peut conduire à une exploration complète de l'arbre de décomposition et donc au chargement en mémoire de tous les vecteurs d'apprentissage.
Pour résoudre ce problème, j'ai pris comme hypothèse que des parties similaires de l'arbre de décomposition étaient explorées pendant la recherche des Kppv de deux vecteurs de test proches. Pour obtenir une faible variabilité entre vecteurs successifs, une solution aurait été de construire un chemin hamiltonien des vecteurs de test, mais elle est inapplicable avec des centaines de milliers de vecteurs. La première méthode employée a été d'ordonner l'ensemble des vecteurs de test en utilisant l'arbre de décomposition. Ce réordonnancement a pour effet de regrouper les vecteurs proches, il permet de diminuer nettement le nombre d'entrées-sorties sans rendre cependant leur influence négligeable dans le temps de traitement. Pour minimiser encore plus les variations entre les vecteurs successifs, j'ai développé des méthodes pour intégrer cette contrainte dès la construction de l'arbre de décomposition. J'ai choisi d'utiliser des arbres de décomposition que j'ai appelés “ ordonnés ”. La technique consistait à modifier l'arbre de décomposition initial en réordonnant pour chaque noeud l'ensemble de ces noeuds-fils. Un chemin hamiltonien entre les centres des noeuds-fils est calculé pour cela.
En conclusion, la prise en compte de la gestion de la mémoire a permis le développement d'un algorithme rapide de Kppv. Cela rend cette technique d'estimation plus utilisable pour de nombreuses applications en Reconnaissance des Formes. Grâce au gain de temps d'environ 99.8%, les cinquante plus proches voisins des quatre cents mille vecteurs de test de TIMIT ont pu être calculés en deux semaines.

Collaboration : Marie-José Caraty

4.4. La contrainte du vocabulaire dans les systèmes de dictée vocale

L'emploi des systèmes de reconnaissance de la parole continue ne pourra être généralisé que s'ils peuvent reconnaître l'ensemble des mots d'une langue. Cela est dû aux problèmes des mots dits “ hors-vocabulaire ”, car les systèmes actuels de reconnaissance ont de grandes difficultés à détecter qu'un segment sonore est un mot inconnu ; en général, ces systèmes reconnaissent, à la place du véritable mot, un autre mot ou une concaténation de mots dont le sens peut être très éloigné. Mais l'utilisation, dans un système de reconnaissance, de tous les mots d'une langue devient aussi un problème, car ce nombre peut dépasser un million si on inclut les noms propres. La principale difficulté se situe au niveau des moteurs de reconnaissance des systèmes de dictée vocale qui procèdent à des calculs détaillés et coûteux de comparaisons entre les modèles acoustiques, lexicaux et linguistiques. La taille mémoire utilisée croît linéairement avec la taille du vocabulaire tandis que l'augmentation du coût de calcul est encore plus rapide (e.g., carré du nombre de mots avec un modèle de langage à base de bigrammes) même si la factorisation de lexique permet de réduire ce coût de calcul [Aubert94][Ortmanns95].
Pour répondre à cette problématique, une solution est de ne plus utiliser un vocabulaire fixé une fois pour toutes, mais de le sélectionner avant la phase de reconnaissance. Cette sélection peut être supervisée par l'utilisateur, par l'application ou par le contexte défini par l'ensemble des phrases précédemment reconnues. Elle peut aussi être basée uniquement sur le signal vocal à décoder. Dans ce cas, la sélection du vocabulaire doit utiliser des techniques de reconnaissance de parole, plus simples naturellement que l'algorithme initial. La solution généralement choisie est d'utiliser des classes de mots

au lieu des mots dans un système simplifié de dictée vocale. Ces classes sont obligatoirement formées par le regroupement des homophones hétérographes. Le vocabulaire sélectionné est alors obtenu à partir de l'ensemble des classes de mots retenu dans le graphe de décodage des classes. Le taux de compression du vocabulaire donné par le rapport du nombre de classes avec le nombre de mots est variable suivant les langues ; nous avons pu l'estimer par exemple à 2.1 en français sans prendre en compte la fréquence des mots, cela est dû au nombre élevé d'homophones [El-Bèze95] dans cette langue. Cependant, si cette méthode permet d'utiliser un vocabulaire plus important, tous les mots d'une langue ne peuvent toujours y être incorporés donc reconnus.

Nous avons développé, pour obtenir un système de dictée vocale sans contrainte de vocabulaire, une méthode originale de lexique dynamique [Fissore89][Waast95][MG96]. Celle-ci est basée sur l'utilisation d'un treillis à N choix de phonèmes obtenu par décodage acoustico-phonétique, algorithme beaucoup moins coûteux qu'une dictée vocale. La construction du lexique dynamique s'effectue en deux étapes. La première étape est le calcul des dissimilarités entre le treillis phonétique décodé d'une phrase (ou d'un paragraphe) et la phonétisation de chacun des mots d'une langue. La seconde étape est l'extraction des N mots les plus proches de ce treillis. Les mesures de dissimilarité acoustico-lexicales sont basées sur un algorithme de programmation dynamique, l'algorithme de Wagner et Fisher [Wagner74], développé à l'origine pour comparer des chaînes de caractères. Il a été modifié pour trouver la position d'une sous-chaîne dans un treillis. Trois opérations peuvent être effectuées sur chaque symbole phonétique de la sous-chaîne : substitution, omission ou insertion dont le coût est estimé sur un ensemble d'apprentissage. Enfin, la dissimilarité obtenue par programmation dynamique est ensuite multipliée par une fonction inverse de la probabilité d'apparition de ce mot. Cela permet de sélectionner des mots “ rares ” mais très proches du treillis tout en n'écartant pas les mots les plus probables.

Cette méthode a été évaluée sur les données de la campagne de dictée vocale en comparant le taux de couverture du lexique dynamique avec celui d'un lexique fixe constitué des mots les plus fréquents. Nous avons pu alors montrer qu'il fallait un lexique fixe d'une taille deux fois plus importante que celle du lexique dynamique pour obtenir le même taux de couverture.

Collaboration : Marie-José Caraty

Bilan sur la période 96-97 : 2 articles dont 1 international [28] et 1 national [47]

Recueil de publications : pp. 133

4.5. Définition d'un modèle de langage

Un système de dictée vocale a besoin d'autres connaissances linguistiques, que la seule liste des mots du vocabulaire, pour pouvoir par exemple reconnaître si une phrase décodée n'est pas qu'une suite de mots sans aucun sens. Comme il est impossible pour l'instant de construire un modèle de langage généraliste, les systèmes de dictée vocale doivent utiliser des modèles statistiques de langage [Bahl83][Jelinek90][Jelinek98] spécialisés dans une tâche donnée (lettres commerciales, diagnostics médicaux). Ces modèles de langage permettent de contraindre la phrase décodée à une compatibilité avec le type de tâche. Ils sont basés le plus souvent sur des n-grammes (i.e., probabilité d'une suite de *n* mots). Ceux-ci ne peuvent être estimés que pour des valeurs faibles de *n* (i.e., 2 ou 3) et ne permettent pas une prise en compte du contexte. Une manière de résoudre ce problème est d'utiliser des n-grammes avec mémoire [Kuhn90] (i.e., probabilité d'une suite de n mots en fonction de la liste des mots des phrases précédentes). Tous ces modèles sont estimés à partir de textes écrits de très grande taille. Une méthode souvent utilisée pour réduire la complexité de cette estimation est de regrouper les mots par classe [Kneser93]. Cela permet d'obtenir des n-classes moins nombreuses que les n-grammes et donc plus faciles à estimer. Cependant l'utilisation simultanée des n-grammes et des n-classes permet d'obtenir les meilleures performances de reconnaissance. Les classes peuvent être construites automatiquement en regroupant les mots ayant le même comportement ; c'est-à-dire ceux qu'on retrouve toujours dans les mêmes contextes. Les classes peuvent être également choisies *a priori* en utilisant par exemple des classes grammaticales. Toutefois, contrairement aux premières méthodes, celles-ci impliquent un recouvrement des classes et non une partition du vocabulaire. Nous avons préféré choisir cette dernière méthode.

Pour construire notre modèle de langage en français, trois opérations ont été nécessaires : le nettoyage des corpus textuels, le marquage grammatical, le calcul des comptes des n-grammes et des n-classes.

Les corpus textuels ont été récupérés à partir des CD contenant les onze dernières années du journal *Le Monde*. Cela fait environ 9 millions de phrases comportant plus de 250 millions de mots. Le nettoyage a consisté à éliminer les balises de mises en forme, à détecter les erreurs lexicales, à les corriger si possible, à homogénéiser les termes employés (e.g., Mr ou Monsieur). Le rôle du nettoyage est primordial, car le texte décodé par le système de dictée vocale utilisera les conventions choisies à ce niveau. Une des principales difficultés rencontrées a été le problème de la décapitalisation des mots des débuts de phrase. Une autre liée à la première est la déambigusation du signe point car ce signe se retrouve en fin de phrase, dans une abréviation ou dans un nombre avec un sens différent à chaque fois.

Après cette étape, il fallait déterminer la classe grammaticale de chacun des mots des textes. J'ai développé pour cela un algorithme de type forward-backward pour le marquage grammatical. Nous avions retenu initialement plus de 400 classes. Mais la trop grande finesse de cette classification ne permettant pas d'obtenir une bonne convergence de l'algorithme de marquage, nous avons alors réduit ce nombre à 74 en regroupant des classes proches (e.g., classes verbales de temps différents et de même personne, différents types de pronom et d'adjectif). Dans ce cas, l'algorithme a un taux de succès qui peut être estimé à 99% de bonnes étiquettes.

La dernière étape est le calcul des comptes des n-grammes. Pour les estimer dans le cas d'un vocabulaire sans contrainte, nous avons défini deux types de structures de données pour les retrouver rapidement : une structure d'arbre pour les n-grammes des 64 K-mots les plus probables, une structure de liste pour les autres. Cela permet d'obtenir rapidement un modèle de langage dynamique adapté à un lexique dynamique tel que défini au paragraphe précédent.

Nous nous sommes limité à deux et trois comme valeurs de n, car il n'était pas possible d'estimer les quadrigammes sur les corpus du journal *Le Monde*. Nous avons trouvé sur ces corpus trois mille biclasses sur environ 5500 possibles et soixante mille triclasses sur environ 400 000 possibles, plus de 10 millions de bigrammes et près de 40 millions de trigrammes. Même si ces nombres sont élevés, ils ne représentent qu'une infime partie des n-grammes possibles. La méthode dite du " backing off " [Katz87] a été utilisée pour le calcul des bigrammes, biclasses, trigrammes et triclasses non-observés.

Collaboration : Marie-José Caraty

4.5. Modélisation acoustique d'un vocabulaire

Dans le cas où le vocabulaire dépasse quelques centaines de mots, il est difficile de faire enregistrer tous ces mots par un nombre suffisant de locuteurs. Cette tâche devient même impossible quand le vocabulaire est sans contrainte. La seule solution est de modéliser ces mots par une concaténation de modèles phonétiques. Le problème de la modélisation acoustique des mots d'une langue devient celui de la recherche d'un dictionnaire phonétisé le plus large possible. Cette ressource est indispensable pour l'apprentissage des modèles, et de sa qualité dépend celle du système de dictée vocale.
Nous avons dû développer un lexique phonétisé devant la difficulté d'en trouver un satisfaisant. Un premier lexique a été construit à partir des 80 Kmots phonétisés du dictionnaire *le Grand Robert*. Ce lexique de base a d'abord été corrigé : en particulier, un module de détection des schwa m'a permis de rajouter plus de 4000 élisions possibles. Ce lexique a ensuite été étendu par l'ajout des formes fléchies (genre, nombre, temps) qui a porté la taille du dictionnaire à plus de 500 000 mots lexicalement différents. Enfin, le cas des liaisons a également été traité. Cela nous a permis d'obtenir plus d'un million de formes phonétiques associées.
Le dernier problème a été la phonétisation d'environ 15 000 noms propres et sigles (sur plusieurs centaine de milliers) . Nous avons utilisé pour les premiers une approche par analogie qui ne s'est pas révélée totalement satisfaisante et dont les résultats ont dû être corrigées par M.-J. Caraty en utilisant le logiciel de synthèse MBR-PSOLA [Dutoit96]. Les principales difficultés se retrouvent dans la prononciation des noms propres d'origine étrangère où au moins deux variantes sont possibles : celle du pays d'origine et celle francisée. L'utilisation des techniques statistiques comme les modèles markoviens n-lettres [Bechet97] devrait résoudre ce problème. La phonétisation des sigles présente elle aussi des difficultés, car la prononciation peut être de type lu ou épelé (e.g., ONU). Nous avons en général préféré garder les deux variantes. Ce dictionnaire doit être continuellement mis à jour pour prendre en compte de nouveaux mots. Nous avons pu estimer, à partir des textes du journal *Le Monde*, qu'il en apparaît à peu près 500 par mois dont 90% de noms propres.
Les textes de la base BREF TOTAL (i.e., 100 heures de parole) [Gauvain90] ont été phonétisés grâce à ce dictionnaire. En éliminant les modèles acoustiques dont le nombre de vecteurs d'apprentissage était trop réduit, cela a permis de définir et d'apprendre la modélisation [Ljole94] acoustique suivante pour le français: 36 monophones, 1100 biphones gauches et droits, 16970 triphones.

Collaborations : Marie-José Caraty, Claude Barras (thésard), Bernard Alary (stagiaire post-DEA), Mathias Heidl (stagiaire MST)

Bilan sur la période 96-97 : 1 article national [29]

4.6. Le moteur de reconnaissance de D-DAL

Le moteur de reconnaissance d'un système de dictée vocale a pour but d'utiliser toutes les ressources précédemment définies pour extraire des hypothèses de phrases à partir du signal de parole. Il procède pour cela à des opérations détaillées de comparaisons entre les modèles acoustiques, lexicaux et linguistiques. Ces opérations demandent des ressources très importantes en taille mémoire et en puissance de calcul. Des compromis et une optimisation poussée sont indispensables pour éviter d'utiliser des machines d'une puissance disproportionnée. Notre but était de pouvoir transcrire, en gardant les

meilleures performances, une heure de parole par semaine en utilisant des machines standards (e.g., PC). Cela correspond à un temps de traitement d'une demi-heure pour une phrase de dix secondes.
Guidé par ces contraintes, le décodage a été classiquement décomposé en deux passes [Schwartz92]. La première est synchrone par rapport aux vecteurs d'analyse et génère un graphe de mots valués en utilisant un modèle de langage basé sur les bigrammes. La seconde passe est asynchrone. Elle utilise un modèle de langage plus complet basé sur les trigrammes et donne l'hypothèse de reconnaissance en réévaluant le graphe de mots. Le principal problème a été d'intégrer efficacement les contraintes linguistiques dans le processus de décodage.
L'utilisation du " Hidden markov ToolKit (HTK) " [Young92] développé à l'université de Cambridge, nous a permis de nous focaliser sur cette problématique sans avoir à réécrire les outils standards de l'analyse markovienne (e.g., calcul des multigaussiennes, algorithme de Baum-Welch, gestion des modèles). Nous avons également gardé, pour implémenter la première passe, l'algorithme de propagation de jetons [Young89]. Cet algorithme de programmation dynamique représente les contraintes sur l'enchaînement des modèles acoustiques par un ensemble de noeuds reliés par des arcs valués. Ces noeuds peuvent représenter l'état d'une machine de Markov ou une fin de mots. Dans le premier cas, la valeur de l'arc représente une densité de probabilité d'émission, dans le second elle doit dépendre du modèle de langage utilisé.
La version originale d'HTK ne permettait pas une dictée vocale grand vocabulaire car seul un modèle de langage à base de paires de mots ou utilisant une matrice de transition était disponible. Nous l'avons modifié pour permettre l'utilisation de modèles de langages plus adaptés. Le graphe de mots généré est plus complet que celui d'origine. Il est obtenu en mémorisant, à chaque vecteur acoustique, les N meilleures fins de mot.
Au cours du décodage, le nombre d'arcs intra-mot est proportionnel à la taille du vocabulaire employé tandis que celui des arcs inter-mots est beaucoup plus élevé, car en général aucune transition n'est interdite. Ce nombre d'arcs inter-mots est égal, pour un modèle de langage de type n-grammes, à la puissance $n^{\text{ième}}$ de la taille du vocabulaire. Pour un grand vocabulaire, il est impossible de garder en mémoire tous ces arcs. Cela rend indispensable leur réestimation à chaque nouveau vecteur acoustique, c'est-à-dire toutes les centisecondes. Dans le cas le plus simple de modèles de langage bigrammes, avec un vocabulaire de 64 K-mots, cela fait plus de 4 milliards d'estimation " backing off " des bigrammes. C'est la raison pour laquelle F. Lefèvre a introduit de nombreuses techniques d'optimisation (recherche en faisceau, utilisation d'heuristiques, h-coding) pour éviter de calculer des arcs inutiles ainsi que pour minimiser le temps de calcul d'une estimation " backing off ". Le gain en temps de calcul est pour l'instant d'un facteur mille. Nous avons comme projet de l'améliorer encore d'un facteur dix.
Les mêmes techniques d'optimisation sont utilisées pour la réévaluation du graphe de mots.

Collaborations : Marie-José Caraty, Fabrice Lefèvre (thésard), Emmanuel Daidone (stagiaire PE)

Bilan sur la période 96-97 : 1 article national [44]

4.7. Evaluation de D-DAL96

Notre système de dictée vocale participe aux deux campagnes d'évaluation de dictée vocale en langue française organisé par l'AUPELF-UREF. La première campagne a eu lieu en 1997 (03/02/97 au 17/03/97). Le test auquel nous avons participé est la dictée vocale indépendante du locuteur sans contrainte de vocabulaire. La tâche de reconnaissance a porté sur 655 phrases lues (structurées en paragraphes) du journal *Le Monde* (novembre 1996), ce qui correspond à plus de deux heures de parole. 70% des mots ont été reconnus, le taux d'insertion était de 8%.
D-DAL 96 était un système 4 passes dont les traitements sont les suivants :
- Détermination du sexe du locuteur par modélisation AR-vectorielle d'un paragraphe.
- Sélection d'un lexique dynamique de 40 k-formes (lexicalement, acoustiquement et grammaticalement différentes) par décodage acoustico-phonétique et analyse d'un treillis phonétique de profondeur 3.
- Une phase synchrone au cours de laquelle est engendré un graphe de mots de profondeur 60 utilisant le modèle de langage des bigrammes-biclasses avec traitement des liaisons et des pauses.
- Une phase synchrone utilisait le graphe de mots et le modèle de langage des trigrammes-triclasses pour donner l'hypothèse textuelle de D-DAL96.

Quatre autres laboratoires participaient à cette campagne d'évaluation mais un seul dans la catégorie sans contrainte de vocabulaire. Tous les systèmes utilisaient une modélisation markovienne de la parole. Notre principale originalité était l'utilisation d'un lexique et d'un modèle de langage dynamique.

Collaborations : Marie-José Caraty, Fabrice Lefèvre (thésard), Action de Recherches Concertées B1

Bilan sur la période 96-97 : 1 campagne d'évaluation internationale [54], 2 rapports de contrat [63, 65]

Références

[Aubert94] X. Aubert, C. Dugast, H. Ney & V. Steinbiss, " Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data ", IEEE-International Conference on Acoustic Signal and Speech Processing, Adelaide, Australie, pp. 129-132, **1994.**

[Bahl83] L. Bahl L., F. Jelinek & R. Mercer, " A Maximum Likelihood Approach to Continuous Speech Recognition ", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol . PAMI-5, pp. 179-190, **1983**.

[Barras96] C. Barras, " Reconnaissance de la parole continue: Adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés ", Thèse de l'Université Pierre et Marie Curie, **1996**.

[Bechet97] F. Béchet, T. Spriet & M. El-Bèze, “ Traitement spécifique des noms propres dans un système de transcription graphème-phonème ”, 1[ères] journées scientifiques et techniques FRANCIL, pp. 469-472, **1997**.

[Dutoit96] T. Dutoit, V. Pagel, N. Pierret, F. Bataille & O. van der Vrecken, “ The MBROLA project : towards a set of high quality speech synthetizers free of use for non commercial purposes ”, International Conference on Spoken Language Processing, Philadelphie, USA, pp. 1393-1396, **1996**.

[El-Bèze95] M. El-Bèze, “ Utilisation des modèles stochastiques de langage ”, Ecole thématique-Fondements et Perspectives en Traitement Automatique de la Parole, Ed. H. Méloni, pp. 129-138, **1995**.

[Fissore89] L. Fissore, P. Laface, G. Micca & R. Pieraccini, “ Lexical Access to Large Vocabularies for Speech Recognition ”, IEEE Transactions on Speech and Audio Processing, vol. 37, n° 8, pp. 1197-1213, **1989**.

[Friedman75] J.H. Friedman, J.L. Bentley & R.A. Finkel, “ An Algorithm for Finding Nearest Neighbours ”, IEEE Transaction on Computer, **1975**.

[Fukunaga75] K. Fukunaga & P.M. Narendra, “ A Branch and Bound Algorithm for Computing K-Nearest Neighbours ”, IEEE Transaction on Computer, **1975**.

[Gauvain90] J.-L. Gauvain, L.-F. Lamel & M. Eskénazi, “ Design Considerations and Text Selection for BREF, a large French read-speech Corpus ”, International Conference on Spoken Language Processing, Kobe, Japan, pp. 2359-2362, **1990**.

[Gauvain94a] J.-L. Gauvain & C.H. Lee, “ Maximum a posteriori Estimation for Multivariate Gaussian Observations of Markov Chain ”, IEEE transactions on Speech and Audio Processing, vol. 2, pp. 291-298, **1994**.

[Gauvain94b] J.-L. Gauvain , L. Lamel, G. Adda & M. Adda-Decker, “ Speaker-independant Speech Dictation ”, Speech Communications, vol. 15, pp. 21-37, **1994**.

[Jelinek90] F. Jelinek, “ Self-Organizing Language Modeling for Speech Recognition ”, in A. Waibel and K. Lee (Eds), Readings in Speech Recognition, pp.450-505, Morgan Kaufman, **1990**.

[Jelinek98] F. Jelinek, “ Statistical Methods for Speech Recognition ”, MIT Press, **1998**.

[Katz87] S.M. Katz, “ Estimation of Probabilities for the language model component of a speech recognizer ”, IEEE transactions on Acoustic Signal and Speech Processing, vol. 3, pp. 400-401, **1987**.

[Kenny90] P. Kenny, M. Lenning & P. Mermelstein, “ Speaker Adaptation in a large-vocabulary gaussian HMM ”, IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 12, n° 9, pp. 917-920, **1990**.

[Kittler78] J. Kittler , “ A Method for Determining K-Nearest Neighbours ”, Kibernetis, vol. 7, pp. 313-315, **1978**.

[Kneser93] R. Kneser & H. Ney, “ Improved Clustering Techniques for Class-based Statistical Language Modelling ”, European Conference of Speech Communication and Technology, Berlin, Allemagne, pp. 973-976, **1993**.

[Kubala94] F. Kubala, A. Anastasakos, J. Makhoul, L. Nguyen, R. Schwartz & G. Zavaliagkos, “ Comparative experiments on Large Vocabulary Speech Recognition ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Adelaide, Australie, pp. 561-564, **1994**.

[Kuhn90] R. Kuhn & R.D. Mori, “ A Cache-based Natural Language Model for Speech Recognition ”, IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 12, n° 6, pp. 570-582, **1990**.

[Ljole94] A. Ljolje, “ High Accuracy Phone Recognition using Context Clustering and Quasi-Triphonic Models ”, Computer Speech and Language, vol. 8, pp.129-151, **1994**.

[MG96] J. Macias-Guarasa, A. Gallardo, J. Feirreiros, J.M. Pardo, “ Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone ”, International Conference on Spoken Language Processing, Philadelphie, USA, pp. 30-33, **1996**.

[Ortmanns95] S. Ortmanns & H. Ney, “ Experimental Analysis of the Search Space for 20 000-Word Speech Recognition ”, European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 901-904, **1995**

[Richetin80] M. Richetin, G. Rives and M. Naranjo, “ Algorithme rapide pour la détermination des k plus proches voisins ”, RAIRO technique et science informatique, vol. 14, n° 4, pp. 369-378, **1980**.

[Schwartz92] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen & P. Placeway, “ New uses for N-best Sentence Hypothesis, within the BYBLOS Speech Recognition System ”, IEEE-International Conference on Acoustic Signal and Speech Processing, vol. 1, pp. 1-4, **1992**.

[Shikano86] K. Shikano, K.F. Lee & R. Reddy, “ Speaker Adaptation through Vector Quantization ”, ”, International Conference on Spoken Language Processing, Tokyo, Japan, pp. 2643-2646, **1993**.

[Wagner74] R. A. Wagner & M. J. Fisher, “ The String to String Correction Problem ”, Journal of the ACM, vol. 21 n°1, pp. 168-173**, 1974**.

[Waast95] C. Waast, L. Bahl & M. El-Bèze, “ Fast Match on Decision Tree ”, European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 909-912, **1995**.

[Woodland95] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, S.J. Young, “ The 1994 HTK LargeVocabulary Speech Recognition System ”, ”, IEEE-International Conference on Acoustic Signal and Speech Processing, vol. 1, pp. 73-76, **1995**.

[Young89] S.J. Young, N.H. Russel, J.H.S. Thornton, “ Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems ”, Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dpt, **1989**.

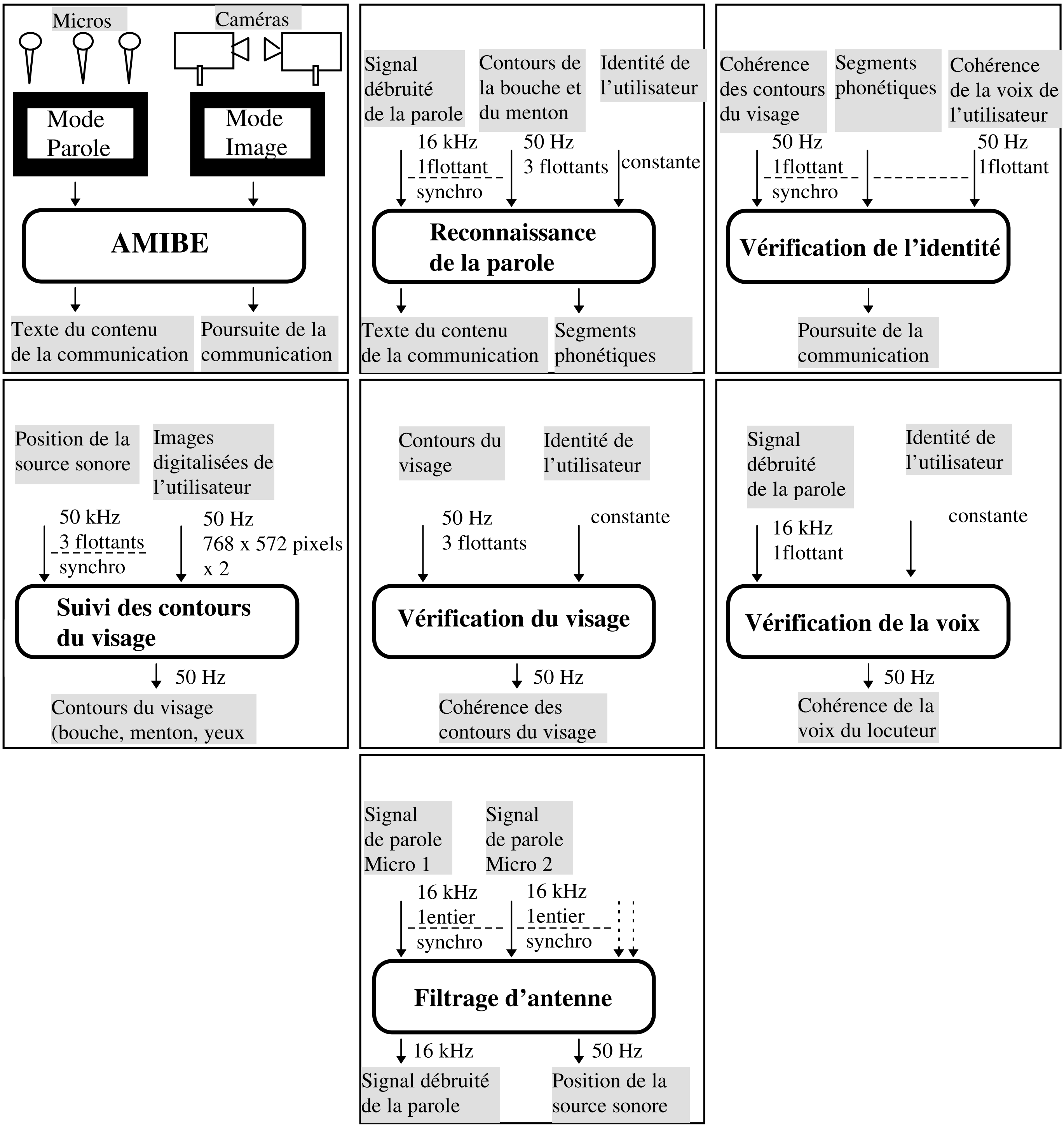
[Young92] S.J. Young, “ HTK Version 1.4: Reference Manual and User Manual ”, Cambridge University Engineering Department - Speech Group, **1992.**

V Systèmes Multimédia

L'augmentation des capacités des autoroutes de l'information permet l'utilisation des données et des techniques multimédia. Grâce à celles-ci, le traitement de l'information s'étend désormais au son et à l'image. Les nouvelles technologies multimédia, qui couvrent l'extraction, la représentation et l'interprétation de ces données, ont déjà un grand nombre de domaines d'application : l'indexation et la gestion des bases de données multimédia, l'accès à l'information par des données biométriques, l'infographie et la communication multimédia. Le traitement des informations multimédia est l'une des hautes technologies qui nous restent à acquérir ou à maîtriser ; il fait appel à la Reconnaissance des Formes et à l'Intelligence Artificielle. Toutes ces composantes permettent réellement une communication Homme-Machine plus évoluée.
Ces travaux ont donné lieu à plusieurs projets comme AMIBE, TRANSDOC ou AVIR.

5.1. AMIBE et l'interaction homme-machine multimodale

Le projet AMIBE (Applications Multimodales pour Interfaces et Bornes Evoluées), dont j'ai été coordonnateur, a débuté en 1993 et a duré 3 ans. Financé par le GDR-PRC Communication Homme Machine, Ce projet comprenait une vingtaine de chercheurs dans six laboratoires (ICP, l'INRIA, l'IRIT, le LAFORIA-IBP, le LIA et le LIUM). Il avait pour but d'augmenter la robustesse des Communications Homme-Machine en fusionnant les modes son et image. Deux modules principaux fondés sur une interaction forte de ces deux modes pouvaient être distingués : la reconnaissance du message oral transmis par l'utilisateur et la vérification de son identité. En effet, à chaque instant (e.g., chaque centiseconde), les deux modes concourent à éliminer les ambiguïtés et à renforcer la fiabilité de l'information transmise (contenu du message oral et identité de l'utilisateur). Trois modules auxiliaires étaient associés aux modules principaux: le suivi du contour des lèvres, la détection du mouvement de la source sonore et la détection des contours du visage.
Les spécifications de ce projet (c.f., figure suivante) étaient les suivantes : la reconnaissance du message oral opérait en synchronisant une “ lecture labiale ” avec un module de reconnaissance des formes acoustiques : la reconnaissance acoustico-labiale [Duchnowski95]. La “ lecture labiale ” consistait en un suivi des lèvres peintes en bleu et une mesure temps réel des paramètres labiaux [Lallouache91]. Le module de détection du mouvement de la source sonore avait deux fonctions pour le décodage du message oral : le débruitage du signal de parole par une technique de filtrage d'antenne et le recalage spatial pour le module du suivi des lèvres.
En dehors de la coordination et de la définition des spécifications, les contributions de notre équipe ont porté plus particulièrement sur les modules vérification du locuteur et traitement d'antenne. Pour ce dernier module, les méthodes classiques de traitement d'antenne ont été adaptées en tenant compte des contraintes de notre projet : il ne s'agit pas de calculer la localisation exacte de la source sonore mais seulement de recaler les modules de suivi des contours du visage et d'extraction du contour des lèvres par différence entre deux localisations successives. Cela permet de résoudre en partie les problèmes posés par la mise au point de la chaîne d'acquisition multi-micros. Les techniques employées [Silverman95] reposaient principalement sur le calcul des intercorrélations dans le domaine fréquentiel et la recherche d'un maximum par une méthode stochastique.



Spécifications logicielles du projet AMIBE

En conclusion, le projet AMIBE a été un pas vers l'intégration des recherches sur la “ parole audible ” et la “ parole visible ”. Elle a montré que la coopération des modes parole et image peut permettre une meilleure communication homme-machine. Les problèmes qui restaient à résoudre pour l'obtention de plates-formes opérationnelles sont de trois ordres. Le premier est le problème de la complexité des algorithmes mis en oeuvre (surtout pour le traitement des images). Il peut être résolu par une optimisation des procédures et la mise au point d'architectures matérielles spécialisées. Le deuxième problème est l'amélioration des performances des différents modules (extraction des paramètres labiaux sans bleu, localisation précise de la source, meilleure reconnaissance acoustico-labiale, ...). Le dernier problème est la définition des interactions entre les différents modules : la difficulté des synchronisations et la présence de boucles de rétroaction nous donnent un système extrêmement complexe à modéliser.
A la suite de l'examen du GDR CHM du 29 Septembre 1994, les commentaires du comité scientifique du GDR sur ce projet ont été les suivants : "le projet AMIBE, qui fusionne signal vocal et image sur un projet très focalisé et illustrant bien certains aspects de la communication homme-machine, a été apprécié du comité scientifique. Son bilan à mi-parcours est prometteur de bons résultats finaux qui pourraient s'intégrer en une démonstration à forte visibilité (en particulier s'il était possible d'étendre le module de reconnaissance bi-sensoriel).

Collaboration : Marie-José Caraty, Nicolas Magdeleine (stagiaire maîtrise), Tristan Behar (stagiaire maîtrise), consortium AMIBE

Bilan sur la période 93-96 : 1 article dans un ouvrage de synthèse [8], 1 rapport de contrat [64]

Recueil de publications : pp. 137

5.2. La recherche d’informations et le multimédia

La recherche d’informations dans les documents sous forme électronique est un domaine en pleine extension [Hartman95]. Cela est dû au nombre sans cesse croissant de documents produits par la généralisation de l’usage de l’informatique. Les premières méthodes de recherche d’informations ont été développées dans le courant des années 60 sur des documents textuels. Ces méthodes sont basées généralement sur les fréquences d’apparition de mots-clés dans le texte recherché [Luh58] et permettent par exemple de représenter un document comme un vecteur de l’espace des mots-clés. Les textes doivent cependant être préalablement normalisés pour minimiser le nombre de termes [SJ71]. Des représentations plus complexes comme des arbres sont également employées mais posent le problème de la métrique utilisée entre les documents. Les problématiques de la recherche d’information textuelle sont principalement le choix de la technique de normalisation et de la liste des mots-clés employés.
La prise en compte des documents multimédia fait actuellement l’objet de nombreuses recherches [Aigrain96]. On peut citer les projets QBIC [Flickner95], VisualSeek, Informedia [Watclar96], OLIVE, DIVAN (DIstributed audio Visual Archives Network), AVIR (Audio Visual Indexing and Retrieval) ainsi que le concours DARPA “ Broadcast News ” [Woodland98]. La recherche d’informations dans les documents multimédia pose deux types de problèmes : l’identification automatique des constituants d’un document composite et de leurs interrelations, la mise au point de techniques d’indexation adaptées aux différents types de données (texte, audio, image fixe, image animée). Le premier problème peut être résolu par un effort de normalisation des

documents multimédia (e.g., normes MPEG). Pour le second problème, deux axes de recherche sont possibles : la définition de métriques entre les données non-textuelles, l'association de descripteurs textuels aux données non-textuelles. Dans le premier cas, cela ouvre la possibilité d'inclure des données non-textuelles dans une requête. Cet enrichissement du langage d'interrogation a cependant plusieurs inconvénients : lourdeur algorithmique du traitement des données non-textuelles, choix des pondérations entre les différentes métriques. Le second axe a pour avantage, la possibilité de réutiliser l'ensemble des techniques d'indexation textuelle mais présente le défaut d'être très dépendant des techniques de reconnaissance des formes employées.
Nous avons choisi le dernier axe en prenant comme domaine d'application l'indexation des documents vidéo. Notre but à long terme est de pouvoir reconstruire le script à partir du document vidéo. L'intérêt d'utiliser un tel type de descripteurs textuels est que son écriture est normalisée et qu'il est riche en informations (décors, protagonistes, actions, dialogues, ...). Nous ne traitons que la partie audio des documents vidéo. D'autres équipes, avec lesquelles nous collaborons dans un certain nombre de projets nationaux ou internationaux, s'occupent de l'indexation des images. La fusion des descriptions provenant des deux sources est toutefois rendue difficile par le faible niveau des informations obtenues à partir des images. Si on peut par exemple détecter les changements de scènes à partir des changements de l'ambiance sonore et des changements de plans [Zhang93], cela est encore impossible pour des informations de plus haut niveau comme l'extraction des protagonistes d'un plan : il est pour l'instant extrêmement difficile de les reconnaître à partir de leur visage [Chellappa95] ou de leur silhouette alors que nos résultats en reconnaissance du locuteur montrent qu'identifier un protagoniste à sa voix semble plus facile.
Mes travaux sur l'indexation du canal audio d'une vidéo ont porté sur l'adaptation de nos systèmes de reconnaissance du locuteur et de dictée vocale, sur la détection des segments silence/bruit/musique/parole ainsi que sur l'alignement du script avec la vidéo.

Deux approches sont possibles pour l'indexation d'un dialogue en terme de locuteurs [Siu92][Yu93][Sugiyama93][Wilcox94][Olsen95] selon que l'on connaît ou non l'identité des acteurs de la vidéo. Le premier cas est *a priori* plus simple mais se révèle difficilement utilisable quand on ne dispose pas de modèle des protagonistes. Cela m'a conduit à choisir la deuxième approche en développant une méthode non-supervisée basée sur la méthode dite de "détection des ruptures" [AO88]. Son principe est de comparer trois modèles autorégressifs vectoriels calculés sur les fenêtres glissantes se recouvrant. Une rupture est détectée quand les modèles divergent trop entre eux. Des expériences préliminaires ont été conduites sur un dialogue comportant plus de 100 locuteurs construit à partir de la base de données TIMIT. Elles ont donné un taux de détection des frontières inter-locuteurs supérieur à 90% (précision de 0.9 avec un rappel de 1). Cependant ces résultats doivent être confirmés sur d'autres bases de données.
L'adaptation de notre système de dictée vocale à la reconnaissance de dialogues d'une vidéo pose deux types de problèmes : multiplicité des conditions d'enregistrement et diversité linguistique des textes prononcés. Une étude d'Informedia [Hauptmann95] a montré que les taux de reconnaissance peuvent varier de 80% pour un présentateur enregistré dans un studio à 25% pour un documentaire. Si le premier problème peut être traité par l'adaptation des modèles acoustiques au canal audio, il est plus difficile d'adapter le modèle de langage au document. Nous avons choisi, comme d'autres équipes, de construire des modèles de langage par type de documents [Seymore97]. Cette approche n'est pas satisfaisante, car elle demande des ressources textuelles très importantes difficiles à rassembler et que le problème de la reconnaissance automatique du type de documents reste à traiter.

Deux méthodes ont été développées successivement pour découper les données audio en silence/bruit/musique/parole. La première était basée sur la méthode de détection des transitions inter-locuteurs développée plus haut. Mais elle s'est révélé inadaptée quand le document sonore comportait une grande variété de bruits comme l'a montré un test sur le film “ Un indien dans la ville ”, produit par Hervé Palud. Pour améliorer la segmentation, nous avons décidé de ne plus utiliser un seul type de modélisation pour décrire tous les types de segments. La nouvelle méthode que nous avons développée en collaboration avec S. Petridis comporte trois étapes. Dans la première étape, un histogramme des énergies à court terme permet une détection silence/non-silence. Ensuite, les segments de non-silence sont étiquetés en bruit/non-bruit en utilisant -l'énergie, -une détection et -un suivi de la fréquence fondamentale -ainsi qu'un modèle de markov du bruit déjà détecté. Enfin les techniques de reconnaissance du locuteur que nous développons, associées à un décodage acoustico-phonétique, permettent de distinguer les segments de parole des segments de musique à partir des segments de non-bruit. Nous avons testé cette méthode sur le film “ Contes de Printemps ” d'Eric Rohmer avec moins de 5% des segments mal classés (précision de 0.96 avec un rappel de 1.).

La méthode d'alignement de script que j'ai développée consiste à associer des étiquettes temporelles à chaque mot du dialogue. Elle est basée sur une méthode d'apprentissage des modèles de Markov cachés phonétiques utilisée quand un étiquetage phonétique des données d'apprentissage n'est pas disponible [Weatley92]. Cette méthode est très employée, car en général seul le texte est fourni. Un réseau des différentes transcriptions phonétiques de ce texte est d'abord construit. Ensuite, l'algorithme de Viterbi aligne ce réseau avec le signal de parole en utilisant un ensemble initial de modèles phonétiques. L'ensemble des segmentations phonétiques des phrases d'apprentissage est alors utilisé pour calculer de nouveaux modèles phonétiques. Cette procédure est réitérée tant que les modèles phonétiques ne se stabilisent pas.
Les principaux problèmes, pour utiliser cette méthode sur un film complet, viennent de la présence de bruit et de musique ainsi que d'une durée du signal audio beaucoup plus grande (e.g., plusieurs heures au lieu de 10 secondes en moyenne pour une phrase). Le premier problème a été traité en partie par la méthode décrite dans la section précédente. Pour résoudre le second problème nous avons défini pour chaque segment phonétique un indice de confiance dépendant de différents paramètres comme sa durée ou sa probabilité d'émission. Seuls les segments phonétiques dont l'indice de confiance est élevé servent à l'apprentissage des nouveaux modèles phonétiques. Cette procédure est réitérée tant que l'indice de confiance moyen augmente. La segmentation en mots est ensuite calculée à partir de la segmentation phonétique.
Nous avons testé cette méthode sur deux documents : un épisode de 20 minutes du dessin animé “ Ivanhoe ” et le film “ Contes de Printemps ” d'une durée de 102 minutes. Dans le premier cas, quatre types d'erreurs étaient définis : décalage d'un phonème, d'un mot, de plusieurs mots, d'une phrase. Plus de 90% des mots étaient parfaitement alignés sans un seul décalage d'une phrase. Dans le second cas, nous avons choisi un paradigme d'évaluation plus simple : un mot-représentant (i.e., le plus long) est choisi pour chaque minute de film. L'erreur d'alignement est donnée par l'erreur moyenne des localisation temporelles des mots représentants. Après segmentation, silence/bruit/musique/parole, cette erreur moyenne était inférieure à une seconde.

Collaborations : Marie-José Caraty, Sergios Petridis (thésard), Pascal Faudemay, consortium TRANSDOC, consortium AVIR

Bilan sur la période 96-98 : 4 articles internationaux [25, 29, 30, 34], 2 workshops [55, 56], 1 rapport de contrat [66]

Recueil de publications : pp. 149, 161

Références

[Aigrain96] P. Aigrain, H.-J. Zhang & D. Petkovic, “ Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review ”, Multimedia Tools and Applications, pp. 179-202, **1996**.

[AO88] R. André-Obrecht, “ A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals ”, IEEE Transactions on Acoustic Signal and Speech Processing, vol. 36, pp. 29-40, **1988**.

[Chellappa95] R. Chellappa, C. L. Wilson & S. Sirohey, « Human and Machine Recognition of Faces: A Survey » IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 83, pp. 705-740, **1995**.

[Duchnowski95] P. Duchnowski, M. Hunke, D. Busching, U. Meir & A. Waibel, « Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition », International Conference on Acoustic Signal and Speech Processing, pp. 109-112, **1995**.

[Flickner95] M. Flickner, H.S. Sawhney & J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, Dr. Petkovic, D. Steele, P. Yanker, “ Query by Image and Video Content: the QBIC System ”, IEEE-Computer, 28, 9, pp. 23-32, **1995**.

[Hartman95] D. Hartmann, “ Overview of the Second Text Retrieval Conference ”, Information Processing and Management, vol. 3, pp. 271-289, **1995**.

[Hauptmann95] A.G. Hauptmann, “ Speech Recognition in the Informedia Digital Video Library : Uses and Limitations ”, IEEE Tools with Artificial Intelligence Conference, 7 pages, 1995.

[Lallouache91] T. Lallouache “ Un poste 'visage-parole' couleur. Acquisition et traitement automatique des lèvres ”, Thèse de Doctorat, INP Grenoble, **1991**.

[Luh58] H.P. Luhn, “ The Automatic Creation of Literature Abstracts ”, IBM Journal of Research and Development ”, vol 2, pp. 159-165, **1958**.

[Olsen95] J. Olsen, “ Separation of speakers in audio data ”, European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 355-358, **1995**.

[SJ71] K. Spark Jones, “ Automatic Keyword Classification for Information Retrieval, Butterworths, **1971**.

[Seymore97] K. Seymore, S. Chen, M. Eskenazi & R. Rosenfeld, “ Language and Prononciation Modeling in the CMU 1996 HUB 4 Evaluation ”, 1997 DARPA Speech Recognition Workshop, 13 pages, **1997**.

[Silverman95] H. F. Silverman & S. E. Kirtman, « A Two-Stage Algorithm for Determining Talker Location from Linear Microphone Array Data » Computer Speech and Language, vol 6, no. 2, pp.129-152, **1995**.

[Siu92] M.-H. Siu, G. Yu & H. Gish, " An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers ", IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 189-192, **1992**.

[Sugiyama93] M. Sugiyama, J. Murakami & H. Watanabe, " Speech Segmentation and Clustering Based on Speaker Features ", International Conference on Acoustic Signal and Speech Processing, pp. 395-398, **1993**.

[Watclar96] H.D. Watclar, T. Kanade, M.A. Smith & S.M. Stevens, " Intelligent Access to Digital Video: the Informedia Project ", Computer, 29, 5, pp. 46-53, **1996**.

[Weatley92] B. Weatley, G. Doddington, C. Hemphill, J. Godfrey, E. Holliman, J. Mac. Daniel & D. Fisher, " Robust Automatic Time Alignement of Ortographic Transcription with Unconstrained Speech ", IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 533-536, **1992**.

[Wilcox94] L. Wilcox, F. Chen, D. Kimber & V. Balasubramanian, " Segmentation of speech using speaker identification ", ", International Conference on Acoustic Signal and Speech Processing, pp. 161-164, **1994**.

[Woodland98] P.C. Woodland, T. Hain., S.E. Johnson, T.R. Niesler, A. Tuerk & S.J. Young, " Experiments in Broadcast News Transcriptions ", IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 909-912, **1998**.

[Yu93] G. Yu & H. Gish, " Identification of speakers engaged in dialog ", IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 383-386, **1993**.

[Zhang93] H.J.Zhang, A.Kankanballi & S.W.Smoliar, " Automatic Partitionning of Full-motion Video ", Multimedia Systems, Vol.1, No.1, pp. 10-28, **1993**.

VI Conclusions et Perspectives

Les travaux résumés tout au long de ce document témoignent de huit années de recherches post-doctorales, de 1991 à 1998. Le domaine de recherche du traitement informatique de la parole a subi pendant ce temps des bouleversements énormes : pour des conditions d'utilisation bien précises, le problème de la reconnaissance de la parole continue est en passe d'être résolu alors qu'il était considéré comme extrêmement difficile il y a moins de 15 ans. On peut faire pratiquement un constat similaire dans le domaine de la reconnaissance du locuteur. Les premiers systèmes de dictée vocale commencent d'ailleurs à être commercialisés pour quelques centaines d'euros même si leurs performances ne sont pas encore celles attendues. Une des explications à ce progrès imprévu, en dehors de l'augmentation de la puissance de calcul et la diminution du coût de la mémoire, est l'organisation de campagnes d'évaluation aux Etats-Unis qui ont permis de faire émerger les techniques les plus prometteuses. Cette explication est cependant insuffisante, car des campagnes similaires avaient été organisées dans les années 70 sans résultats probants.

Les résultats de notre équipe en reconnaissance de la parole et du locuteur nous ont permis d'être au cœur de cette évolution en participant nous-mêmes aux campagnes d'évaluation de reconnaissance du locuteur et de dictée vocale, en apportant au domaine des contributions dont l'importance est attestée par le nombre de publications. Ces activités de recherche ont utilisé des concepts et des méthodes provenant du traitement du signal, de la reconnaissance des formes, du génie linguistique et de l'intelligence artificielle. Elles ont intégré à la fois des aspects expérimentaux, théoriques et techniques, et pris en compte des impératifs industriels. Elles ont nécessité un travail en équipe, indispensable à la réalisation de gros prototypes, et ont donc induit des responsabilités et des tâches administratives, nécessaires à la coordination des travaux et au suivi des contrats qui en assurent le financement.

Les perspectives à moyen terme, dans lesquelles nous allons nous investir, se situent dans le développement de nouvelles technologies d'interprétation et de gestion des données multimédia. Ce domaine va devenir l'un des principaux enjeux de l'informatique de demain que ce soit au niveau des recherches entreprises ou de son impact sur la société civile. Nous avons déjà commencé à l'explorer en participant à des projets sur la communication homme-machine et la recherche d'informations. Notre objectif est de parvenir à un contrôle de la qualité de l'indexation des données multimédia. En effet, nos systèmes de reconnaissance basés sur des approches statistiques n'ont de bonnes performances que pour identifier les objets les plus fréquents : en reconnaissance de la parole, les expressions ou les mots les plus employés. Cela est également dû aux paradigmes d'évaluation utilisés qui par exemple attribuent le même poids à tous les mots. Un taux de 90% de mots reconnus est ainsi considéré comme satisfaisant alors qu'il est presque certain que les 10% restants contiennent beaucoup d'informations pertinentes. Le seul moyen pour parvenir à cette indexation intelligente est d'introduire dans nos systèmes une aptitude à saisir le sens littéral des données textuelles reconnues. Ce domaine de recherche de la compréhension de textes est déjà largement étudié dans le cas de la linguistique écrite : les techniques développées sont utilisées par exemple pour résumer ou traduire automatiquement un texte. La spécificité de la dictée vocale implique que cette composante doit être intégrée, au moins en partie, directement dans le moteur de reconnaissance. Une solution alternative serait de filtrer le graphe de mots créé par le moteur de reconnaissance, mais cette solution est à mon avis sous-optimale dans le cas d'un vocabulaire sans contrainte à cause de la taille limitée du graphe de mots. Plusieurs pistes sont possibles : détection d'indices linguistiques structurants, analyse syntaxique profonde des phrases décodées. Ces recherches ne peuvent être le fruit que d'une collaboration entre la linguistique orale et la linguistique écrite. Celle-ci existe déjà, dans les pays anglo-saxons, où elle a permis l'utilisation commune et le développement de ressources importantes comme les lexiques, les modèles de langage ou les marqueurs grammaticaux. Cela doit également s'accompagner d'une modification des paradigmes d'évaluation qui permette de mieux prendre en compte la destination des résultats de l'indexation.

D'autres perspectives sont également ouvertes dans le domaine de l'analyse sensorielle avec la conception des nez électroniques. Différents acteurs travaillant sur cette application nous ont contactés, car sur certains points nos problématiques sont voisines : Il s'agit de capteurs biologiques dont les propriétés complexes (grande diversité des récepteurs moléculaires) ont été assez bien étudiées, tandis que le traitement de l'information associée est encore mal connu. Différentes hypothèses sur l'inhibition latérale, l'existence d'une image olfactive, l'importance de l'évolution temporelle des données captées ne sont toujours pas validées. L'effort principal porte actuellement sur la fiabilité, la sélectivité et la sensibilité des capteurs, mais la question d'un traitement informatique évolué des données recueillies se pose déjà. Les techniques utilisées actuellement (analyse en composantes principales) relèvent plus de l'analyse de données que de la reconnaissance des formes. Des méthodes simples comme la comparaison dynamique ne sont pas employées : cela permettrait par exemple de vérifier l'hypothèse de l'évolution temporelle. Les nez électroniques sont pour l'instant surtout utilisés pour le contrôle de qualité mais la maîtrise informatique de ce troisième sens après l'ouïe et la vue sera un atout de choix dans le domaine de la réalité virtuelle ou augmentée et plus généralement dans celui de la perception par les ordinateurs de leur environnement.

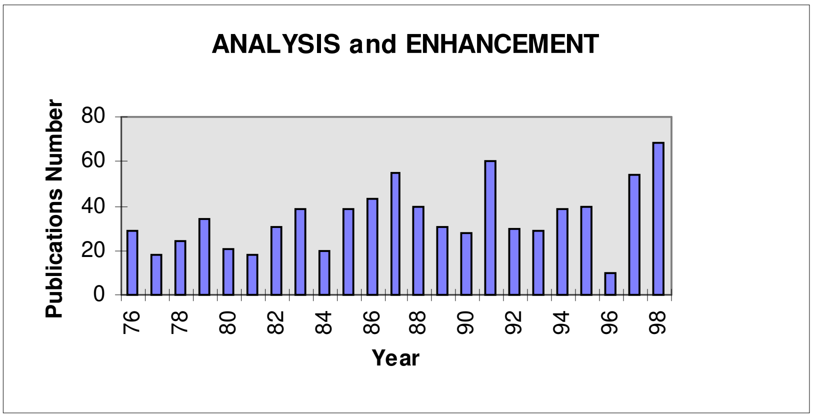
Presentation

The work presented in this study covers the period 1991-1998. It’s a part of the topic “Apprentissage and Knowledge Acquisition” of the Computer Science Laboratory of Pierre and Marie Curie University (LIP6). They are located in the field of the speech processing and the methods and software architectures necessary to process such a signal. This research is based on the cooperation of complex statistical models to extract two a priori additional informations : characteristics of the speaker and contents of the speech. Even if each one of these topics has its own problems, they all contribute to a Spoken Communication Homme/Machine which can be used in real conditions.
Research in speech processing was slowed down a long time by the absence of test databases, essential to the comparison of the methods. The recent availability of such databases made research enormously progress but a consequence was a sophistication of the employed methods. These methods are essentially based on a statistical modeling, for the important size of the bases of training makes difficult the use of methods using “ knowledge ”. The use of knowledge is then in the choice of an adapted statistical modeling. This adaptation will depend on the task of speech or speaker recognition: noise conditions, vocabulary size, numbers of the speakers. The know-how of a speech processing team consists then often of a set of libraries, of software modules, of modelings and processing making it possible to solve mainly new problems and guiding the necessary theoretical axis. This research paradigm made it possible to obtain speech technologies applicable in an immediate future to fields as various as acces control or multimedia data indexation. With the appearance of sufficiently powerful machines at low cost to process and recognize in real time a speech signal, these scientific results will be availability for everybody.
This manuscript comprises four parts. The first introductory part will briefly point out the description of the various problems of the speech processing. This field of research has existed for almost half a century and there were a lot of congresses these last thirty years. The two following parts will describe the work carried out in speech and speaker recognition as well as my contribution to two important achievements Orphée, a vocal lock, and D-DAL, a vocal dictation system. For these two parts the treatment will be chronological.
The last part will treat my work about the study and design of two multimedia systems: the first one is devoted to multimode man-machine communication, the second one is devoted to the information retrieval in videos.

I Evolution of the Speech processing

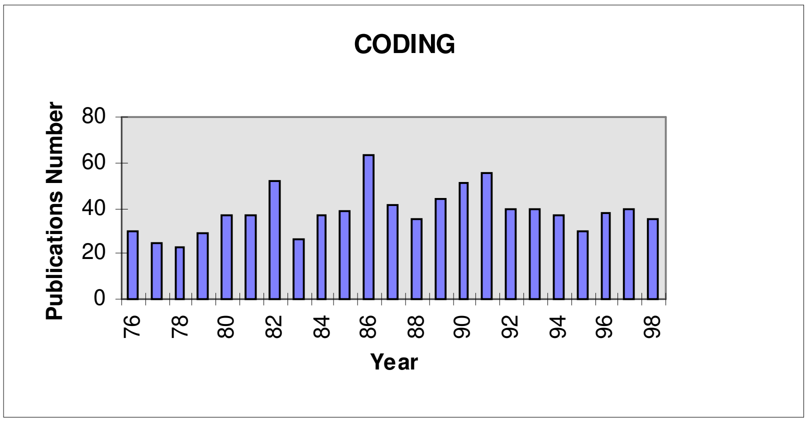
The applications of the speech processing are the analysis- enhancement, the coding, the synthesis and the recognition of a speech signal. To measure the interest which carries the scientific community to these problems, we have decided to st study to one famous of the congresses of the speech processing field “IEEE-ICASSP International Conference one Acoustic, Signal and Speech Processing”. More than 3500 contributions on the speech processing were published since its creation in 1976 and make it possible to understand the description of the problems of this field. The percentage of contributions will be always calculated compared to the bulk of the publications in speech processing since 1976. The two others international congresses in this field are “EUROSPEECH: European Conference one Speech Communication and Technology” creates in 1989 and “ICSLP: International Conference in Spoken Language Processing” one year later. These two congresses are biannual and alternate.

The speech analysis and the speech enhancement are the applications closest to the signal processing, the speech enhancement being regarded as an analysis which resists to the noise. They represent 18% of the contributions and are the core of the speech processing.



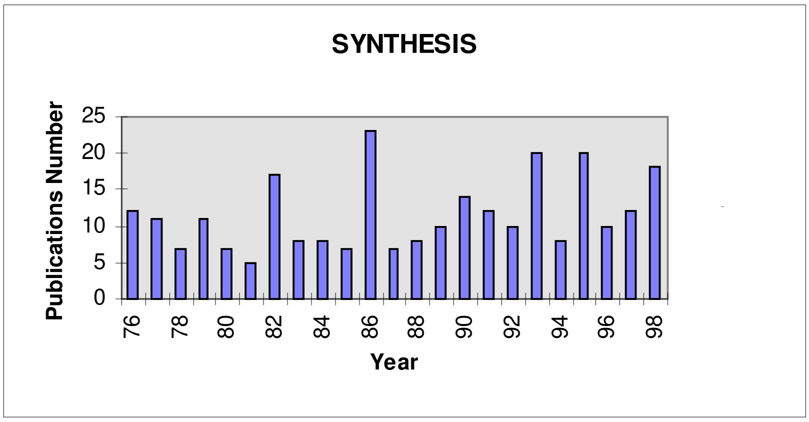
The speech analysis consists in extracting from the speech signal parameters or indices taking into account the application constraints. The studies of the production and the speech perception are mainly used to define these analyses. They usually are short-term analysis (several 1/100 of second). One can quote the melody, the parameters coming from a source-filter model.
The speech enhancement is the study of methods making it possible to obtain these analyses even in the case of deterioration of the speech signal, between its emission and its processingt, by an additive or convolutif noise. This research was a long time very dependent on the studied noise (e.g., telephony) and were not available for other cases. If the source and the type of noise are unspecified, the only solution is the use of microphones array. Their cost was a long time prohibitory but the generalization of the videoconferences techniques made now available. The problems of the array processing are the topology of the microphones array, the suppression of the echoes and the use of high order moments.

Coding a speech signal represent approximately 23% of the contributions. The principal application is still the communication by the telephone network. The most important constraint is not to introduce a delay which makes a trouble in a conversation (several 1/10 of second). The initial problems of this field were to obtain an effective algorithm of coding having the lowest ratio between the computer cost and the compression factor. With the reduction in the costs of calculation, these problems have partly lost their importance and the very expensive algorithms such as methods CELP are continuously improved.



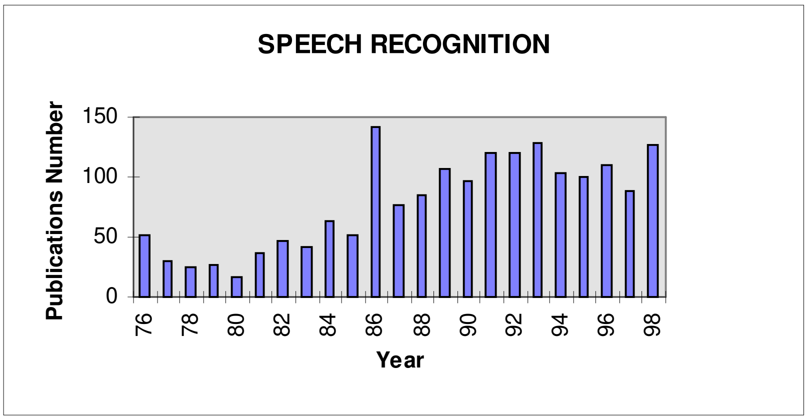
A more recent application of coding is the storage of an audio signal for later diffusion. The principal difference comes from the required quality which must be "transparent", that is to say audibly identical to the original. All this research gave a certain number of standards: GSM in telephony, MUSICAM system in storage used in standards MPEG ([layers 1-3]). This last system should be at the base of all the future multimedia applications: it now allows audibly identical coding of a High-Fidelity stereo signal with a flow of 256 Kbauds.

The synthesis, (7% of the contributions), made it possible to obtain synthesized vocal signals of very good quality. The methods used are based on the concatenation of basic segments such as the diphones. There are now tools of quality available on the network internet as MBR-PSOLA which allows a synthesis using a set of phonemes and prosodic informations.



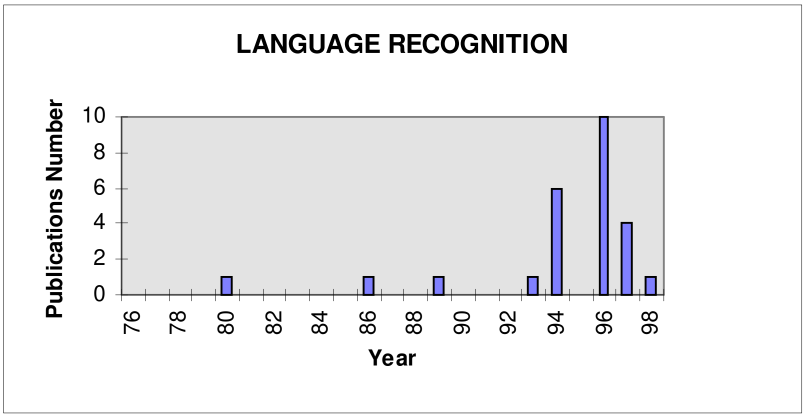
The problems of this field are to obtain an quality indistinguishably of the human voice with the automatic generation of a better prosody and the concept of voice conversion which must be imitated or the general characteristics of a population having an identical accent. One can notice for this purpose the very good synthesis of a castrato voice realized by a mixture of two voices: the first one a female voice, the second one a male voice (Farinelli - X Rodet).

The recognition of a vocal signal (51% of the contributions) covers three applications: the uttered text, the language employed and the speaker recognition. To give an idea of the difficulty of the problem, there are 6 billion speakers, 3000 languages and a few hundreds of thousands of words by language. The first application represents the majority of the contributions (45%) although the two others arise an increasing interest. The successive problems of the uttered text (called speech recognition) are the phonemes recognition (4% of the contributions), the words recognition (7%), the continuous speech recognition (8%), the speech recognition in the noise (8%), the Markovian modeling of the speech (12%). The remainder of the contribution treat mainly constitution of speech databases and recognition systems evaluation. One finds the results following by calculating these percentages only on the publications of 1998: the phonemes recognition (0%), the words recognition (0%) the continuous speech recognition(0%), the recognition in the noise (38%), the Markovian modeling of the speech (39%). That shows all the evolution of the speech processing field.



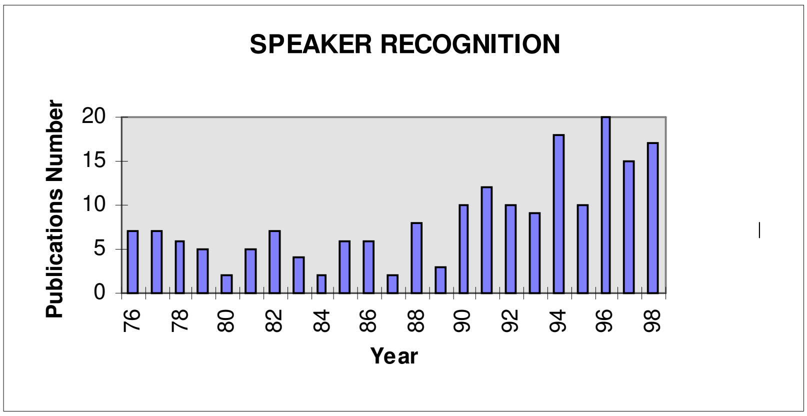
The problems of the phonemes recognition are the localization and the identification of the phonemes. The principal application is the acoustico-phonetic decoding (APD) which extracts from an speech signal analysis, an hypothesis of phonemes lattice. Even if the uses of this lattice are numerous in speech recognition, APD is mainly used such as evaluation paradigm in all the fields of the speech processing. But it's not proved that the cause of an APD improvement has an influence on others speech recognition problems. The words recognition has two applications: the connected words recognition with a limited grammar (e.g., spelled name) and the detection of a limited number of keywords in the continuous speech. The continuous speech recognition or "vocal dictation " consists in recognizing all the words of a uttered sentence in a continuous mode, that is to say without artificial pause between the words. These problems, considered a long time as impossible, were solved partly for reading speech, even if the recognition rates decrease when the voice is more spontaneous. The recognition in the noise consists to adapt the speech recognition techniques to different conditions between the test and the training. As the techniques of speech enhancement currently do not give satisfaction, the solution employed is to compensate for these differences by a noise addition to the training databases. This solution is obviously not satisfactory even if improves the recognition performances. The Markovian modeling of the spoken language became a subject of research in itself when it proved to be the only technique which could be used to integrate acoustic and linguistic information. "which could be used " want to say that we have methods of modeling, of training and recognition which are sufficiently adaptable and automatizable to quickly taking into account new conditions (environment, speaker, language). The techniques of Markovian modeling of the speech can be divided into two equal parts: the acoustic modeling and the language modeling. Acoustic modeling is based on a probabilistic automat in which the analysis parameters are emitted by the transitions between states. Current research relates to topology, training and fast adaptation of the parameters of this automat. Linguistic modeling is founded on the use of N-grams and N-classes; that is to say the probability of continuation of N-words or N-classes (a bulk of words). The current problems are taking into account the context, the adaptation to sentences different from the written language without increasing the complexity of the recognition.

The contributions on the recognition of the language employed are not numerous (1% of the contributions). It's only in 1996 a whole session was devoted to him. The first contribution goes back to 1980, the authors have an identification rate of 80% for speech segments of 4 minutes in five languages.



The principal problems were to decrease the duration of the segments while increasing the number of languages. Research could be largely facilitated by the recent availability of base OGI (collected by Oregon Graduate Institute) dedicated to the languages identification comprising the telephone calls uttered in spontaneous language in 22 languages. The are two approaches of these problems, the one is purely acoustic, the other one is linguistic and more difficult to implement. The acoustic approach consists in representing a language by the acoustic achievements of its phonemes and the phonotactic rules. This approach gives very good results if the languages have clearly differents acoustic spaces, but finds its limits quickly when the number of languages increases. The linguistic approach is based on the recognition of specific words of a language. the importance of the linguistic resources which must be used (dictionary and language model for each language to recognize) makes difficult this approach but more promising than the preceding one. The applications of this technique are not numerous except information services.

Even if the speaker recognition is a problem as old as the speech recognition, it accounts for only 5% of the contributions. The first automatic speaker recognition system goes back to 1963, but it is only very recently that this field woke up the interest of an increasing number of researchers and industrialists. This interest is explained by the availability of databases including a large number of speakers. The applications are the speaker identification, the detection of changes of speakers and the speaker verification. At the beginning the use of keywords was necessary, but the majority of the systems are now text independent.



The developed techniques use three types of approaches: statics, dynamics and linguistics. The speaker recognition systems can naturally combine several of these approaches. The first two approaches are based only on acoustic information. The difference between two approaches proceed from the use of parameters representing the temporal evolution of the speech signal. The linguistic approach is founded on the analysis of the lexicon employed and the dialectal alternatives used, but it do not exist automatic approaches. The techniques which are now developed allow to the verification of the identity of a person who accepts to cooperate if the recording conditions of the test remain identical to those of the training. However there cannot question of vocalprint which could be compared with the fingerprints or geneticprints. the reliability of speech identification is not the same. The actual problems are to compensate a modification of the recording conditions, to mixer the speaker verification with that of the uttered text, to detect the changes of speakers in the multimedia databases.

After this brief exposing of the problems of the speech processing. I will present my work and contributions since my thesis, in 1991. I joined at that time the LAFORIA speech group (Laboratory of the Forms of Artificial Intelligence) which is now a part of the LIP6 laboratory. I am the co-responsible with M.-J. Caraty of this group after the departure of the professor X Rodet. The major part of my work for this period were made in collaboration with other members of the speech group. This work is in the field of the speech analysis, speaker and speech recognition. I have defined and developed new approaches: generalized multiresolution for the speech analysis, AR-Vectorial modeling of very short-term spectra for speaker recognition, new estimator using K-NN and dynamic lexicon for the speech recognition. This work was also the definition and the development of linguistic resources (databases, lexicon, language model) essential to the system such as Orphée the vocal lock and D-DAL the vocal dictation. At last this was an opportunity to approach new fields as the retrieval information in the multimedia documents.

II Speech Analysis

My aim in speech analysis was guided to introduce the knowledge of the speech field and the application constraints on the speech analysis. The goal was to avoid a purely "signal processing" approach which is general disconnected from our problems.

2.1. Generalized temporal decomposition using the dissimilarity measure PSP

The temporal decomposition models the spectral evolution of a speech segment, This method describes the analysis vectors by a linear model of interpolation. Introduced as a coding technique [Atal83], its formulation was justified by the fact that the parameters calculated on successive windows are strongly correlated. It was thus possible to represent their evolution by a smaller number of parameters. It's a means of detecting the curves of the trajectory of the analysis vectors, and of locating the succession of local attractile (or targets) by which the trajectory passes or seems to go on. The temporal decomposition was developed and used in various laboratories for many applications of the speech processing: -the coding, -the synthesis, -the segmentation and the labelling, -the acoustico-phonetic decoding.
We developed a theoretical framework to use the temporal decomposition in a non Euclidean space by using the geodesic concept [Montacié91]. For, the Euclidean distance is not always the best dissimilarity measure between analysis vectors, within the meaning of the recognition performances. We chose for this study the dissimilarity measure PSP (Pairing of Spectral Peaks). [Caraty87] The analysis obtained makes it possible to obtain one or two targets by vowels.

Collaborations : Marie-José Caraty, Claude Barras (Phd thesis)

Results (1990-92) : 3 publications, 1 international [20] and 2 national [38, 39], 1 workshop [53]

Publication selected : pp. 101

2.2. Homomorphic Déconvolution and parameterization

The source-filter modeling is one of the most used modelings in speech processing. Its problems is that a reliable estimate of the part source and of the part filter is necessary, for only the filter part is usable in recognition process. The method to estimate the filter part is to use homomorphic déconvolution methods. We tested the method known as of "True Enveloppe" [Imai81] which makes it possible to obtain spectral envelopes more precise than by the traditional cepstre method. Parameterizations tested were those the most used in speech recognition: the MFCC (Mel Frequency Cepstral Coefficients) and the spectral peaks. The results of our work showed that the improvement of source-filter separation did not get any substantial improvement with MFCC parameterization. A possible explanation is that this parameterization is very much used, for it is robust when separation source-filter is of bad quality. Indeed, this spectral envelope is

integrated by wideband filters (e.g., 100 Hz in low frequency), which causes an attenuation of the improvements of the envelop estimation.
The parameters resulting from the analyses described in this section and the previous section were used in a symbolic learning system [Ganascia86] for the vowels recognition [Montacié90].

2.3. The stationnarity measure by the technique of generalized multiresolution

The usual hypothesis of spectral analysis is that the speech signals are stationary; that is to say that their properties (average, standard deviations, high order moments) remain identical for the analysis period. This is only under these conditions that this analysis keeps a physical sense. In the contrary case this is only a mathematical transformation which has no relationship with real activity of human ear. However many speech segments are not stationary and a spectral analysis without a stationnarity verification add artifacts which must make more difficult the future processing such as the speech recognition. In collaboration with M.-J. Caraty, we have introduced the concept of "generalized multiresolution" which makes it possible to detect these nonstationarinesses. The developed methods are told generalized methods, for they allow a feedback between any analysis method and a process of decision (e.g., minimization of dissimilarity measures or maximisation of the probabilities).

Collaboration : Marie-José Caraty

Results (1997-98) : 1 international publication [32]

Publication selected : pp. 105

References

[Atal83] B.S. Atal, " Efficient Coding of LPC Parameters by Temporal Decomposition ", IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 81-84, **1983**.

[Caraty87] M.-J. Caraty, " Contribution au décodage acoustico-phonétique. Etudes de distances inter-spectres et reconnaissance de cycles vocaliques ", Thèse de l'Université Paris 6, **1987**.

[Ganascia86] J.-G. Ganascia, " Learning with Hilbert Cubes ", 2nd European Working session on Machine Learning, Sigma Press, pp.158-171, **1986**.

[Imai81] S. Imai & Y. Abe, " Spectral Envelope Extraction by Improved Cepstral Method. " Electronic and communication, In Japan, Vol 62A, n°4, pp. 10-17, **1979**.

[Montacié90] C. Montacié, M.-J. Caraty & X. Rodet, " Experiments in the Use of an Automatic Learning System for Acoustic-Phonetic Decoding ", International Conference on Spoken Language Processing, Kobe, Japan, pp. 357-390, **1990**.

[Montacié91] C. Montacié, " Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformations spectrales non-linéaires ", Thèse de l'ENST, **1991**.

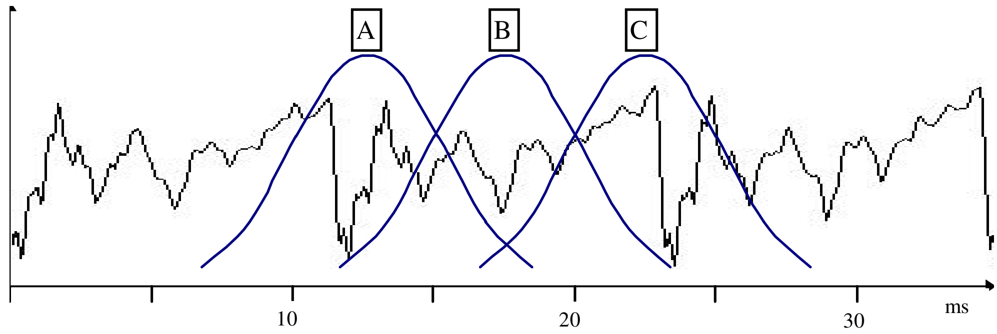
III Speaker recognition

My work in text independant speaker recognition is founded on the integration of informations on the laryngeal source and of those on the vocal tract. I have used an AR-Vectorial modeling associated with an original analysis of the speech signal. The developed techniques were patented in 1996. The international research report only quote type A documents, that is to say nonopposable to this invention. After a short review of the state of art, I shall describe our principal results in this field.
I have classified the speaker recognition methods according to the use of parameters representing the temporal evolution of the speech signal. the first methods that i say "dynamic methods" are generally derived from those used in speech recognition: dynamic programming [Furui81], Markovian modeling and neural networks. They are the left-to-right hidden Markov models [Naik89], the ergodic hidden Markov models [Poritz82], the multi-layer neural network (e.g., TDNN) [Bennani91]. At the base of "the static methods" is the hypothesis of independence between the analysis vectors. These methods are based on usual data analysis techniques such as vectorial quantification [Li83]. They are the interspeakers measure from the covariance matrix [Bimbot93], the parameters distribution modelled by a weighted sum Gaussian densities (Gaussian Model Mixture) [Reynolds94-95]. This last modeling which makes us return to the beginning of the pattern recognition is one of the most performant.
In the next section, I develop an original technique of speaker recognition allowing to integrate informations on the laryngeal source and of those of the vocal tract.

3.1. Integration of informations on the laryngeal source and of those of tract vocal

Physiological speaking, characteristic information of the speaker can be found in the study of his laryngeal source and in his vocal tract. In the speaker recognition systems dealing with these two informations, these informations are extracted separately from short term spectral analysis then they are amalgamated [Matsui90]. To allow the separation of these two informations, the temporal windows are selected of a duration higher than that of the fundamental period (e.g., 1/100 second for the men). Two series of parameters are then obtained: -an estimate of the fundamental frequency -a chosen parameterization (e.g., MFCC, LPCC). Fusion techniques are then used by amalgamating the parameters or the speaker recognition results. No significant improvement can be detected by this use of the fundamental frequency. Several explanations are possible: a bad estimate of the fundamental frequency, an insufficiency of this parameter to represent the laryngeal source, unsuited fusion techniques. My purpose was not to separate the laryngeal source information from the vocal tract information. The method that we retained was to use two types of modelings: a very short term analysis sensitive to the laryngeal source and a AR-Vectorial modeling from these analyses. We showed in experiments that in this case, this last modeling became sensitive to the variations of this source and the characteristics of the average spectrum.
The AR-Vectorial modeling technique is used to process signals with several components [Grenier80][Kakusho82]. The model is a set of q prediction matrices using to predict a vector analysis from the q previous vector analysis. It is seldom employed in speech processing. Its principal advantage is the simplicity of its estimation method derived from those developed for AR-scalar modeling. The matrices of prediction are

estimated by the Levinson-Whittle-Robinson algorithm [Whittle63] which minimizes the trace of the covariance matrix of the prediction errors. The analysis vector that we use is a very short term spectal analysis. The principal difficulty was to find the bond between the analysis parameters (size of the analysis window, analysis step) with the order *q* of the AR-Vectorial modeling.
We chose the following heuristic that we confirmed in experiments: the analysis step multiplied by the order *q* must be close to the size of the window. This choice is available if the window size is equal to the period of fundamental (cf. following figure), for in this case each vector d'analyse used for the prediction is centered on a different phase of the glottal excitation.



Prediction of the window C from the windows A and B

The speaker recognition techniques developed from these principles are very effective, resistant to the noise and temporal drift and also to clogging of the respiratory tracts caused by a clod or a bronchitis for example. However these techniques remain sensitive to the variations of the elocution type (reading, spontaneous or prepared speech) and to those of emotional state.

Results (1991-98) : 1 international patent [1]

3.2. Development of interspeakers measure

Once defined modeling, a dissimilarity measure between a model of speaker and a elocution model is essential, for any application in speaker recognition. My experiments showed that the principle of decision was to be founded on the comparison of the prediction errors produced by these two models and not on the trace of the covariance matrix of the prediction errors. That makes it possible not to take into account the prediction errors due to the noises. Theses noises cannot be predicted by an AR-Vectorial model. This decision principle is very effective, but it's computer cost is prohibitory calculation for an effective application. In this case, the approximation that we chose is to compare covariance matrix of the prediction errors in the place of the errors. We generalized for that the Itakura-Saito dissimilary measures [Itakura75]. The fast algorithms that we developed have a complexity proportional to the square of the model order multiplied by the cube of the vectors analysis size.
These interspeakers mesures tested on the TIMIT database (reading voices databases) [Fisher86] made it possible to recognize a speaker among 630 without error within five seconds. On this database, the error rate in speaker recognition was inversely proportional to the duration of the speech signal. However, these results were obtained from a 12 seconds maximum of speech and cannot be extrapolated a priori for one

higher duration (e.g., 1 hour). It is certain that these performances must be degraded when the number of speakers increases. That is due to the fact that the speakers are increasingly close to each other and that the risks of confusion increase. Experiments on TIMIT database subsets seem to show that this error rate is a function of the square root of the number of speakers [Webb93]. These results are good reasons to develop and test a prototype of vocal lock under real conditions.

Collaboration : Jean-Luc Le Floch (Phd thesis)

Results (1991-92) : 4 publications, 2 international [18, 19] et 2 national [37, 40]

Publication selected : pp. 109

3.3. The Orphée system and the lock vocal

The vocal lock is an application of the speaker verification. The user of such lock declines his identity with the system by any means (e.g., magnetic card, code) and pronounces a short sentence. Then the system verify the compatibility of two information to authorize the entrance.
Orphée, our lock vocal models a user by only one AR-Vectorial model (7 kB). A temporary code allows an incremental and not supervised training procedure. The verification comprises two steps: an identification step then a verification step. This identification step is necessary, for several speakers can have the same verification code. This verification code is the day of the month of birth. During the verification step, the interspeakers measure between the verification sentence and the model of the person to be identified are compared with a threshold. This distance is standardized by the cohort method [Higgins90]. This method consists in putting in competition the model of the person to be identified with a set of others speakers models. During all our experimentation, the threshold did not depend on the speaker and have not been modified. Orphée was installed and tested by J-L. Floch with the acces door of our laboratory for 20 months. Located in an extremely noisy and reverberating place, this vocal lock was used by 90 members of the laboratory without any entry of an impostor. The statistics of use are the following ones: on average, on 20 attempts per day, one raises a thwarted attempt of an impostor and 12 successful coming in. The analysis of the 7 other attempts shows that these rejects are due in the majority of the cases to a too short speech duration (i.e., < 2 seconds) or to bad recording conditions (e.g., noise, intensity of the speech signal).

Collaborations : Marie-José Caraty, Jean-Luc Le Floch (Phd thesis)

Results (1993) : 1 international publication [21]

Publication selected : pp. 113

3.4. Continuous speaker verification

The project AMIBE (Multimode Applications for Interfaces and Advanced Terminals) which I will describe in the final chapter included a speaker recognition task who was to be used with face recognition. The purpose of the approach that we developed in this project was to permanently verify the alleged identity of the user to avoid a speaker change during a transaction. Two research orientations were studied: the first to obtain a

decision of continuous verification without waiting the end of a sentence, the second to determine the verification reliability according to the type of speech segments. J.-L. Floch developed a method of continuous speaker verification: the “analytical verification”. In an analytical verification, any whole of consecutive analysis vectors allows a speaker verification. The decision is made by comparing the last vector with the estimate by the training AR-vectorial from the preceding vectors. The reliability of the analytical verification was studied according to the phonetic segments. The results of speaker verification experimentsl show that the most discriminating phonemes, by descending order, are: the diphthongues, the vowels, the nasals/liquids, the glides, the stops/fricatives/affricates. We also noticed that the elimination of the least discriminating phonemes allowed to improve the performances of the speaker recognition.
These results confirm previous investigations [Sambur75][Atal76][Rosenberg87] on speaker characterization.

Collaborations : Marie-José Caraty, Jean-Luc Le Floch (thésard)

Results (1994) : 2 publications, 1 international [22] et 1 national [43]

Publication selected : pp. 117

3.5. Co-operation and Competition of models

The experiments on modeling Gaussian Mixture Model showed that the results given by this method were not linked with those obtained by the AR-Vectorial models. That showed that it was perhaps possible to obtain of better performances by amalgamating the two methods of recognition. In collaboration with J.-L. Floch, we then defined procedures of co-operation and competition of theses models. Fusion could be done either for the all sentence or for each analysis vector. The principal difficulty was to find standardization methods which make it possible to make comparable the prediction errors of our predictive method with the probabilit densities of the sums the Gaussian ones. An 10% increase in the results is obtained by the competition procedure. The same experiments took place by using several AR-Vectorialmodels by speaker; each one being trained on distinct phonetic segments. The performances also increased but in a less proportion. The cost of important calculation of these techniques of fusion did not enable us to integrate them in Orphée.

Collaborations : Marie-José Caraty, Jean-Luc Le Floch (Phd thesis)

Results (1995-96) : 3 publications, 2 international [24, 26] et 1 national [45]

Publication selected : pp. 121

3.6. NIST 98 speaker recognition evaluation

Each year, The National Intitute of Standards and Technology organizes evaluations in various fields of computer science (networks, virtual reality, linguistic genius). This organization specifies an evaluation paradigm, it defines a schedule of conditions and provides data bases. The evaluation is organized in two steps: a development test and test step. The development step lasts approximately six months; It is fake test which allows to optimize the systems on the development data. The test step has a limited

duration (e.g., 15 days): The test database is provided on the day of the beginning of the evaluation and the results must imperatively be returned before a deadline.
The last speaker recognition evaluation was on the speaker verification on the telephone. The evaluation paradigm was to answer 350 000 questions “is speech segment x is pronounced by the speaker y”. The fact that the same speaker could be recorded on various types of microphones (carbon or electret) and on several types of telephone lines was the principal difficulty. Another difficulty was due to the method of the construction of a speech segment. This speech segment was composed of several short speech segments uttered in a conversation on telephone. In that case the definite task was not the cooperative speaker verification but was closer to the speaker identification task. However, our late participation in this evaluation (2 months of development) allowed us to have an experiment on such databases. We used the Orphée system by adding only one module of filtering in the telephone band, without compensation of the microphones and telephone lines variability. The equal error rate obtained was about 25%.. The analysis of our results showed that the errors were mainly due to the three following factors: -the presence of many laughter which disturbs the estimation of the predictive models, -the important variation of the emotional state of an speaker between various recordings, -the voices of others speaker due to a bad conversations segmentation. Indeed our method taking into account information on the laryngean source is particularly sensitive to these factors. Thanks to the lesson of this evaluation, we already began the adaptation of the Orphée system with the speaker speaker on the telephone. That also allowed us to improve our multimedia indexation methods which we will detail in the final chapter.

Collaboration : Marie-José Caraty
Results (1997-98) : 1 international evaluation [57]

References

[Atal76] B. S. Atal, “ Automatic Recognition of Speakers from Their Voices ”, IEEE Proceedings, vol. 64, No. 4, pp. 460-475, **1976**.

[Bennani91] Y. Bennani & P. Gallinari, “ On the Use of TDNN extracted features Information for Talker Identification ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Toronto, Canada, pp. 385-388, **1991**.

[Fisher86] W. Fisher , V. Zue, J. Bernstein & D. Pallet, “ An Acoustic-Phonetic Data Base ”, Journal of Acoustic Society of America, Suppl. (A), 81, S92, **1986**.

[Furui81] S. Furui, “ Cepstral Analysis Technique for Automatic Speaker Recognition ”, IEEE Transactions on Acoustic Signal and Speech Processing, Vol. 29, pp. 254-272, **1981**.

[Grenier80] Y. Grenier, “ Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur”, 11èmes Journées d’Etude de la Parole, Strasbourg, France, pp. 163-171, **1980**.

[Higgins90] A. Higgins, L. Bahler & J. Porter, “ Speaker Verification Using Randomized Phrase Prompting ”, Digital Signal Processing, vol. 1, pp. 89-106, **1991**.

[Itakura75] F. Itakura, “ Minimum Prediction Residual Principle Applied to Speech Recognition ”, IEEE Transactions on Acoustic Signal and Speech Processing, Vol. 23, pp. 67-72, **1975**.

[Kakusho82] O. Kakusho & M. Yanagida, “ Hierarchical AR model for Time Varying Speech Signals ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Paris, France, pp. 1295-1298, **1982**.

[Li83] K.-P. Li & E.H. Wrench, “ An approach to Text-Independant Speaker Recognition with Short Utterances ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Tokyo, Japan, pp. 555-558, **1983**.

[Matsui90] T. Matsui & S. Furui, “ Text-Independant Speaker Recognition Using Vocal Tract and Pitch Information ”, ”, International Conference on Spoken Language Processing, Kobe, Japan, pp. 137-140, **1990**.

[Matsui94] T. Matsui & S. Furui, “ Similarity Normalization Method for Speaker Verification Based on a Posteriori Probability ”, Workshop on Automatic Speaker recognition identification and verification, Martigny, Suisse, pp. 59-62, **1994**.

[Naik89] J.M. Naik, L.P. Netsch & G.R. Doddington, “ Speaker Verification over Long Distance Telephone Lines ”, IEEE-International Conference on Acoustic Signal and Speech Processing ,USA, pp. 524-527, **1989**.

[Poritz82] A.B. Poritz, “ Linear Predictive Hidden Markov Model and the Speech Signal ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Paris, France, pp. 1291-1294, **1982**.

[Reynolds94] A. Reynolds, “ Speaker Identification and Verification using Gaussian Mixture Speaker Models ”, Workshop on Automatic Speaker recognition identification and verification, Martigny, Suisse, pp. 27-30, **1994**.

[Reynolds95] A. Reynolds & C. Rose, “ Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models ”, IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 72-83, **1995**.

[Rosenberg87] A. E. Rosenberg & F. Song, “ Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Dependent Modes ”, Computer Speech & Language, vol. 2, n. 3/4, pp. 143-157, **1987**.

[Sambur75] M. R. Sambur, “ Selection of Acoustic Features for Speaker Identification ”, IEEE Transactions on Speech and Audio Processing, pp. 176-182, **1975**.

[Webb93] J. J. Webb & E. L. Rissanen, “ Speaker Identification Experiments Using HMMs ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Minneapolis, USA, pp. 387-390, **1993**.

[Whittle63] P. Whittle, “ On the Fitting of Multivariate Autoregression and the Approximate Canonical Factorization of a Spectral Density Matrix. ”, Biometrika, Vol. 50, pp. 129-134, **1963**.

IV Speech Recognition

My research in speech recognition related on the speaker adaptation of the recognition systems and to an improvement of the densities of probability used to represent acoustic space. These studies also aimed at a better knowledge of the Markovian systems which can often be used without knowing their way of working. The selected method for the adaptation uses a speakers topology which is defined by our speaker recognition methods. The replacement purpose of the gaussian probabilities by a nonparametric estimator such as the K Nearest Neighbors rule was to improve the densities of probability employed in Markovian modeling. First of all, the evaluation paradigm used was based on an acoustico-phonetic decoding system which had the advantage of being simple to implement. Our team has developed another one for three years: a vocal dictation system [Gauvain94b][Kubala94][Woodland95] which is speaker independent without any constraint of vocabulary. The development of this system enabled us to approach the field of the linguistic genius, for we mainly had to develop the resources necessary to such a system: phonetized dictionary, language models, decoding step. The Dictaphone-Typist of LIP6 (D-DAL) took part in the first French-speaking vocal dictation evaluation which was organized by the AUPELF-UREF in 1997.

4.1. Speaker Adaptation

In spite of the use of speech databases which include several hundreds of speakers, it is always difficult to model all speakers variability. That makes necessary, for the moment, the speaker adaptation of the recognition systems. This adaptation can be supervised or not supervised. In the first case, the system of adaptation is focused on the errors made by the recognition system. That makes ergonomics more difficult, for the errors list must be provided by the speaker. In the other case the adaptation uses the whole of the analysis parameters of test without the speaker's intervention. The supervised adaptation gives in general the best results but it is not always applicable: for example in the recognitions "off-line" (e.g., indexing case). Two approaches are generally used to adapt the recognition systems: -the estimate of a spectral transformation to project the analysis parameters of the test speaker on those of the training speaker [Shikano86], -the use of Markovian models adapted to the test speaker [Kenny90].

The approach by projection is the first one which was made, for it can be used in all the types of recognitions systems. The principal difficulties are -the alignment of the test data on the training models, -the estimate of the function of spectral transformation. This function is generally a projection matrix which is calculated by techniques of multiple linear regression. I prefered to use multilayers perceptrons which made it possible to model non-linear transformation functions.

In the other approach, the more simple and most employed method consists in using Markovian models by genre: a model estimated on men, another one on women. Another method is to use the technique which is called the "maximum a posteriori (MAP)" [Gauvain94a]. It is a question of adapting a Markovian model estimated on a large number of speakers by modifying only the probability densities which are incoherent with the test data. The MAP technique is supposed to keep the information of a speaker independant model and at the same time adding the characteristics of the new speaker. We developed two methods which are based on a topology of the training speakers in order to adapt the recognition systems. This topology was defined by interspeakers measures based on AR-Vectorial modeling. In the first method, the Markovian models were estimated once more from the speakers who are closest to the test data. In the second method which was developed by C Barras, the adapted models were obtained from a weighting of a models set which was calculated beforehand on clusters speakers [Barras96]. The first method was most performant. This performance was showed by our experiments during the development step of the vocal dictation evaluation, but its cost of calculation is very high. The second one was faster but the improvement compared to the genre models was not significant.

Collaborations : Marie-José Caraty, Claude Barras (Phd thesis), Benjamin Mayrargue (Master),

Results (1994-95) : 4 publications, dont 2 international [7, 23] and 2 national [41, 42]

Publication selected : pp. 125

4.2. Estimation of emission probabilities

In the hidden Markov models, the emission probabilitiy of a state is generally represented by a sum of gaussian mixtures with matrix of diagonal covariance. We decided to study if this choice were founded or not and if a better estimate could be obtained by the K nearest neighbors rule which is theoretically one of the most effective estimators.
A preliminary study of J. Goût confirmed that many emission probabilities were not represented by a sum of gaussian mixtures. To show this was true, we used to a chi-square test between the theoretical distribution and the data distribution which have been calculated for each analysis coefficient and for each state. The first contribution of this study was a modification of the training algorithm of the gaussian mixtures. The improvement consisted in a use of multidimensional chi-square test in order to split the gaussian mixtures. This chi-square test was developed by C. Barras who used the sum of the chi-square tests on each coefficient. This new method of training made it possible to decrease by half the average number of gaussian mixtures without decrease of performance. However, this result was not completely satisfactory and we decided to launch a new study to replace the gaussian modeling by a nonparametric estimator founded on the KNN rule. In spite of the simplicity of this method, it is very effective and it presents the following theoretical advantages: the error of the 1NN is always lower than the double of the minimal error (i.e., Bayes risk). Moreover this error decreases when K increases. This rule is also the interesting to find the origin of an error more easily. Its only disadvantage is its calculation cost compared to the use of gaussian mixtures. We solved this problem by developing a KNN fast algorithm which we will describe in the next section. The saving of time which is about 99.8% allowed by this algorithm makes it possible the integration of this estimator in a Markovian recognition system. This problem has been studied by F. Lefèvre for 3 years. He had to adapt the training (i.,e., Baum-Welch) and the decoding algorithms (i.e., Viterbi) to this new estimator. The results which were obtained showed that the performances did not increase in a significant way for the new estimator really modified the functionning of the Markovian models. If we can trust of this estimator for its recognition in R^n (i.e., only one analysis vector), we are not assured that the recognition is effective in $R^n \times N$ (i.e., a set of vectors). Nevertheless, that gave us a new tool for a deepest study of the functionning of these models. We developed for that of the evaluation paradigms which

should enable us to measure more exactly the influence in the recognition process of delta-parameters and contextual phonetic models.

Collaboration : Marie-José Caraty, Fabrice Lefèvre (Phd thesis), Jérôme Goût (Master), Cécile Lecordier (Master)

Results (1997-98) : 3 international publications [9, 31, 33]

Publications selected : pp. 129

4.3. K nearest neighbors fast algorithm and memory management

Such an algorithm was essential, for the databases used in speech processing must be important size to model the variability of the speech signal. We chose one of smallest, TIMIT, which already corresponds to a duration longer than 4h. The base being a centiseconde analysis, it gives a than one million and half vectors. If two thirds were used in training, a test experiment would require the calculation of more than 500 billion measures between vectors.

A first algorithm which was based on the Kittler-Richetin methods [Kittler78][Richetin80], the Friedman methods [Friedman75] and the Fukunaga methods [Fukunaga75] was developed. Its principle was to use a hierarchical tree to split the space of the training vectors. Each tree node was characterized by its center, its diameter, and the whole of its vectors. For each vector test, a sucession of in-depth with a backtracking made it possible to select the final nodes likely to contain KNN. Only the measures between the test vector and the vectors of the selected final nodes were then calculated. This algorithm reduced indeed the cost of calculation, but not the duration of the total processing. The difference came from the significant number of input-outputs caused by the impossibility of keeping in memory the whole of the training vectors. Three methods were used to solve this problem successfully: a better management of the memory, a structuring of the whole of the test vectors, scheduling of the hierarchical tree.

The technique called “Least Recently Used (LRU)” used for the management of the memory made it possible to decrease the number of input-outputs but not on a sufficient scale. The problem come from the variability of the successive test vectors which compells to explore the different parts of the hierarchical tree for each KNN search. I showed that the processing of a hundred vectors can lead to a complete exploration of the hierarchical tree and thus to the loading in memory of all the training vectors.

To solve this problem, I chose as a hypothesis: similar parts of the hierarchical tree are explored during the KNN search of two close vectors of test. To obtain a lower variability between successive vectors, a solution would have been to build a hamiltonian way of the test vectors, but it is inapplicable with hundreds of thousands of vectors. The first employed method was to order the whole of the test vectors by using the hierarchical tree. This regrouping allows to gather the close vectors, it makes it possible to decrease in a significant way the number of input-outputs without however making their influence negligible in the duration of the processing. To minimize even more the variations between the successive vectors, I developed methods to integrate this constraint during the construction of the hierarchical tree. I chose to use a hierarchical tree which I called “ordered hierarchical tree”. The technique consisted in modifying the initial hierarchical tree by reordering for each node the whole of these undernodes. A hamiltonian way between the centers of undernodes was calculated for this purpose.

In conclusion, the memory management allows the development of a fast KNN algorithm. That makes this technique of estimate more usable for many applications in pattern recognition. Thanks to the saving of time which is about 99.8%, the fifty nearest neighbors of the four hundred and thousand test vectors of TIMIT were calculated in two weeks.

Collaboration : Marie-José Caraty

4.4. Vocabulary constraint in the vocal dictation systems

It would be difficult to generalize the use of the continuous speech recognition systems without being able to recognize every word of a definite language. That is due to the problems of the words called “out-of-vocabulary”. For the current speech recognition systems have difficulties in detecting that a sound segment is an unknown word; in general, these systems recognize, in the place of the true word, another word or a group of words which meaning can be different. But in a recognition system, the use of all the words of a language becomes also a problem, for this number can exceed a million if proper names are included. The principal difficulty is the cost of calculation of the vocal dictation decoding step. This decoding step compared acoustic, lexical and linguistic models. The used memory size increases in proportion with the size of the vocabulary while the increase in the cost of calculation is even faster (e.g., in proportion with the square of the number of words in a case of a bigram language model) even if the lexicon factorization makes it possible to reduce this cost of calculation [Aubert94][Ortmanns95].

To solve these problems, we have decided not to use one vocabulary fixed but to select it before the recognition step. This selection can depend of the speaker, of the application and also of the context defined by the whole of the sentences which have been previously recognized. It can also be only based on the speech signal to recognizes. In this case, the vocabulary selection must naturally use techniques of speech recognition, which will have to be more simple than the initial decoding step. The generally selected solution is to use words classes instead of the words in a simplified vocal dictation system. In that case, the words of the same classe are prononced in the same way. The selected vocabulary is then obtained from the whole of the classes of words which have being retained in the decoding lattice of the classes. The compression ratio of the vocabulary is given by the ratio beetween the number of classes and the number of words. We could estimate this ratio : for example at 2.1 in French, that is due to the high number of homophons [El-Bèze95] in this language. However, if this method makes it possible to use a more important vocabulary, all the words of a language cannot be used and recognized.

To obtain a vocal dictation system without constraint of vocabulary, we developed an original method of dynamic lexicon [Fissore89][Waast95][MG96]. This method is based on the use of a phonemes lattice which has been obtained by an acoustico-phonetic decoding, for this algorithm is less expensive than a vocal dictation. The construction of the dynamic lexicon is carried out in two steps. The first step is the calculation of dissimilarity measures between the phonemes lattice of a sentence (or of a paragraph) and the phonetisation of each word of a language. The second step is the extraction of the N words which are closest to this lattice. Acoustico-lexical dissimilarity measures are based on a dynamic programming algorithm, the Wagner and Fisher algorithm [Wagner74], which were developed at the beginning to compare character strings. It was modified to find the position of a under-chain in a lattice. Three

operations can be carried out on each phonetic symbol of the under-chain: substitution, omission or insertion, the cost of which was estimated on a training set. Lastly, the dissimilarity measure obtained by dynamic programming is then multiplied by an opposite function of the word frequency. That makes it possible to select “rare word” which are very close to the lattice and at the same time to select words which are frequently used.
This method was evaluated on the database of the vocal dictation evaluation. We compare the cover rate of the dynamic lexicon with that of a fixed lexicon which was made with the most frequent words. So we could show that we need a fixed lexicon of twice more important size than the size of a dynamic lexicon in order to obtain the same cover rate.

Collaboration : Marie-José Caraty

Results (1996-97) : 2 publications, 1 international [28] and 1 national [47]

Publication selected : pp. 133

4.5. Language model definition

The only list of words in a language is not sufficient to recognize if a decoded sentence has a meaning or not. As it is impossible for the moment to build a general language model, a vocal dictation systems must use statistical language models [Bahl83] [Jelinek90][Jelinek98] specialized in a given task (commercial letters, medical diagnoses). These langage models are generally based on N-grams (i.e., probability of a succession of N words). They can only be estimated for low values of N (i.e., 2 or 3) and do not allow a taking into account of a large context. A method of solving this problem is to use N-grams with memory [Kuhn90] (i.e., probability of a succession of N words according to the words list of the preceding sentences). All these models are estimated from very big size written texts. A method often used to reduce the complexity of this estimate is to group the words by class [Kneser93]. That makes it possible to obtain N-class wich are less numerous than the N-grams. So they are easily estimated. However the simultaneous use of the N-grams and the N-class makes it possible to obtain the best recognition performances. The class can be built automatically by gathering the words which are in the same context. Theses class can also be selected a priori by using grammatical class. We preferred to choose this method.
To build our model of language in French, three operations were necessary: the filtering of the textual corpora, grammatical labelling, the calculation of the N-grams and N-classes counts.
The textual corpora were extracted from the newspaper *Le Monde* during these last eleven years. That make approximately 9 million sentences comprising more than 250 million words. Filtering consisted to detect the lexical errors, to correct them if possible, to homogenize the employed terms (e.g., Mr. or Mister). The role of filtering is very important, for the text which has been decoded by the vocal dictation system uses the chosen conventions chosen. One of the principal difficulties was the problem of the word decapitalisation at the beginning of a sentence. Another difficulty which is in a relation with the first one is the meaning of a comma for this sign can be found at the end of the sentence and in an abbreviation and also in a number.
After this first step, it was necessary to determine the grammatical type of each word of the textual corpora. For that purpose, I developed a grammatical labelling algorithm. We

had retained initially more than 400 grammatical class. But the too large number of class did not allow us to obtain a good convergence of the labelling algorithm. So we reduced this number to 74 by gathering close class (e.g.,verbal classes of different tenses and on a same person, various types of pronouns and adjectives). In this case, the algorithm has a success rate which can be estimated at 99%.
The last step is the calculation of the N-grams counts. To estimate them in the case of a vocabulary without constraint, we defined two types of data structures which allow to find them quickly: a tree structure for the most frequent onces, -a list structure for the others. That makes it possible to quickly obtain a dynamic language model adapted to a dynamic lexicon as defined in the preceding section. on these corpora, we found three thousand biclass out of approximately possible 5500 and sixty thousand triclass out of approximately possible 400 000 possible, more than 10 million bigrams and 40 million trigrams. Even if these numbers are high, they represent only one negligible part of the possible N-grams. The method called "backing off" [Katz87] was used for calculation of the bigrams, biclass, trigrams and triclass which have not being observed.

Collaboration : Marie-José Caraty

4.6. Acoustic modeling of a vocabulary

If the vocabulary exceeds a few hundreds of words, it is difficult to make record all these words by a sufficient number speakers. This task becomes even impossible when the vocabulary is without constraint. The only solution is to model these words by a regrouping of phonetic models. The problem of the acoustic modeling of the words becomes the problem of the search for a phonetized dictionary. This resource is essential for the training of the acoustic models, and the performances of the vocal dictation system depend on a quality of this dictionnary.
We had to develop a new phonetic lexicon since no satisfactory one was available. A first lexicon was made with the 80 Kwords phonetized in the *Le Grand Robert* dictionary. This basic lexicon was corrected: in particular, a detection module of schwa added more than 4000 possible pronunciations. This lexicon was then extended by the addition of the derivation forms (genre, plurial, tense) which added more than 400 000 words to existing ones. Lastly, the case of the linking beetween two words was also studied. That enabled us to obtain more than one million phonetic forms.
The last problem was the phonetisation of about 15 000 proper names or initials (on several hundred thousands). For the proper names we used an analogy approach. These was not completely satisfactory and the results had to be corrected by M.-J. Caraty who used the MBR-PSOLA synthesis software [Dutoit96]. The principal difficulty was found in the proper names pronunciation of foreign origin when two choices are possible: the original country pronunciation and the french pronunciation. The use of the statistical techniques such as the N-letters Markovian models [Bechet97] should solve this problem. The phonetisation of the initials presents also difficulties, for the pronunciation can be of type read or spelled (e.g., UNO). In general, we preferred to keep the two pronunciations. This dictionary must be continuously updated to take into account new words. We could estimate, from the texts of the newspaper *Le Monde*, that about 500 new words appear in a month including 90% of proper names.
The texts of the BREF TOTAL speech database (i.e., 100 hours of speech) [Gauvain90] were phonetised thanks to this dictionary. That made it possible to define and train the acoustic models [Ljole94] for the french language: 36 monophones, 1100 biphones (left and right), 16970 triphones.

Collaborations : Marie-José Caraty, Claude Barras (Phd Thesis), Bernard Alary (Master), Mathias Heidl (Master)

Results (1996-97) : 1 national publication [29]

4.7. The decoding step of D-DAL

The purpose of the decoding step of a vocal dictation system is to use all the resources previously definite to extract from the speech signal the sentences hyphothesis. It needs of detailed operations of comparisons between the acoustic, lexical and linguistic models. These operations require very important resources in computer memory and in computing power. Heuristics and a thorough optimization are essential to avoid using computers of a disproportionate cost. Our goal was to be able to transcribe, by keeping the best performances, one hour of speech in a week by using standard machines (e.g., PC). That corresponds to a processing time half an hour for a ten seconds sentence.
Using these constraints, speech decoding is made in two steps [Schwartz92]. The first one is synchronous to the analysis vectors and generates a words lattice by using a bigram language model. The second step is asynchronous. It uses a trigram language model and gives the sentences hypothesis by rescoring the words lattice. The principal problem was to effectively integrate the linguistic constraints in the decoding process.
The use of the “Hidden markov ToolKit (HTK)” [Young92] developed at the Cambridge university, enabled us to focus ourselves on these problems without having to redevelope the Markovian analysis tools (e.g., gaussian mixtures computing, Baum-Welch algorithm, models management). To implement the first step, we have kept the token passing algorithm [Young89]. This dynamic programming algorithm represents the constraints on the sequence of the acoustic models by a set of nodes connected by valued arcs. These nodes represent the state of a Markov model or the end of a word. In the first case, the arc value represents an emission probability density, in the second it must depend on the language model which is used.
The original version of HTK did not allow a vocal dictation using a large vocabulary for only a model of language containing word pairs or using a transition matrix was available. We modified it to allow the use of more sophisticated language models. The lattice of words generated is more complete than that of origin. It is obtained while memorizing the better end of words for each acoustic vector.
During this decoding step, the number of intra-word arcs is proportional to the size of the vocabulary employed while that of the inter-words arcs is much higher, for in general no transition is prohibited. This number of inter-words arcs for a N-grams langage model is equal to the power N of the vocabulary size. For a large vocabulary, it is impossible to keep all these arcs in memory. That makes essential their reappraisal for each new acoustic vector. Using 64 K-words, we should be estimated about 400 billion bigrams per second in the simple case of bigram language model. This is why F. Lefèvre introduced many techniques of optimization (e.g., beamsearch, h-coding) to avoid calculating useless arcs and to minimize the computing time of N-grams. The saving of time of calculation is for the moment of a factor thousand. We have the project of improving it of a factor ten.
The similar optimization techniques are used for the rescoring of the words lattice.

Collaborations : Marie-José Caraty, Fabrice Lefèvre (Phd thesis), Emmanuel Daidone (Master)

Results (1996-97) : 1 national publication [44]

4.8. D-DAL Evaluation

Our vocal dictation system takes part in the two evaluation plans of vocal dictation in French language organized by the AUPELF-UREF. The first evaluation took place in 1997 (03/02/97 to the 17/03/97). The test in which we took part is the speaker independent vocal dictation without constraint of vocabulary. 655 read sentences (structured in paragraphs) in the newspaper *Le Monde* (November 1996) was used for the recognition task. This corresponds to more than two hours of speech. 70% of the words were recognized, the insertion rate was 8%.
D-DAL 96 was a 4 step system whose processing was the following one:
- A determination of the speaker genre using AR-vectorial modeling of the paragraph.
- A selection of a dynamic lexicon of 40 K-forms (lexically, acoustically and grammatically different).
- A synchronous phase during a words lattice of depth 60 was genered. It used a bigram-biclass language model.
- A asynchronous step rescored the words lattice using the the trigrams-triclass language model. It gave the sentence hypothesis of D-DAL96.

Four other laboratories took part in this evaluation but only one in the category without constraint of vocabulary. All the systems used a Markovian modeling of the speech. Our principal originality was the use of a dynamic lexicon and a dynamic language model.

Collaborations : Marie-José Caraty, Fabrice Lefèvre (Phd thesis)

Results (1996-97) : 1 international evaluation [54, 63, 65]

References

[Aubert94] X. Aubert, C. Dugast, H. Ney & V. Steinbiss, “ Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Adelaide, Australie, pp. 129-132, **1994.**

[Bahl83] L. Bahl L., F. Jelinek & R. Mercer, “ A Maximum Likelihood Approach to Continuous Speech Recognition ”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol . PAMI-5, pp. 179-190, **1983**.

[Barras96] C. Barras, “ Reconnaissance de la parole continue: Adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés ”, Thèse de l'Université Pierre et Marie Curie, **1996**.

[Bechet97] F. Béchet, T. Spriet & M. El-Bèze, “ Traitement spécifique des noms propres dans un système de transcription graphème-phonème ”, 1ères journées scientifiques et techniques FRANCIL, pp. 469-472, **1997**.

[Dutoit96] T. Dutoit, V. Pagel, N. Pierret, F. Bataille & O. van der Vrecken, “ The MBROLA project : towards a set of high quality speech synthetizers free of use for non commercial purposes ", International Conference on Spoken Language Processing, Philadelphie, USA, pp. 1393-1396, **1996**.

[El-Bèze95] M. El-Bèze, " Utilisation des modèles stochastiques de langage ", Ecole thématique-Fondements et Perspectives en Traitement Automatique de la Parole, Ed. H. Méloni, pp. 129-138, **1995**.

[Fissore89] L. Fissore, P. Laface, G. Micca & R. Pieraccini, " Lexical Access to Large Vocabularies for Speech Recognition ", IEEE Transactions on Speech and Audio Processing, vol. 37, n° 8, pp. 1197-1213, **1989**.

[Friedman75] J.H. Friedman, F. Baskett & L.J. Shustek, " An Algorithm for Finding Nearest Neighbours ", IEEE Transaction on Computer, vol. C-24, n° 10, pp. 1000-1006, **1975**.

[Fukunaga75] K. Fukunaga & P.M. Narendra, " A Branch and Bound Algorithm for Computing K-Nearest Neighbours ", IEEE Transaction on Computer, vol. C-24, n° 7, pp. 750-753, **1975**.

[Gauvain90] J.-L. Gauvain, L.-F. Lamel & M. Eskénazi, " Design Considerations and Text Selection for BREF, a large French read-speech Corpus ", International Conference on Spoken Language Processing, Kobe, Japan, pp. 2359-2362, **1990**.

[Gauvain94a] J.-L. Gauvain & C.H. Lee, " Maximum a posteriori Estimation for Multivariate Gaussian Observations of Markov Chain ", IEEE transactions on Speech and Audio Processing, vol. 2, pp. 291-298, **1994**.

[Gauvain94b] J.-L. Gauvain , L. Lamel, G. Adda & M. Adda-Decker, " Speaker-independant Speech Dictation ", Speech Communications, vol. 15, pp. 21-37, **1994**.

[Jelinek90] F. Jelinek, " Self-Organizing Language Modeling for Speech Recognition ", in A. Waibel and K. Lee (Eds), Readings in Speech Recognition, pp.450-505, Morgan Kaufman, **1990**.

[Jelinek98] F. Jelinek, " Statistical Methods for Speech Recognition ", MIT Press, **1998**.

[Katz87] S.M. Katz, " Estimation of Probabilities for the language model component of a speech recognizer ", IEEE transactions on Acoustic Signal and Speech Processing, vol. 3, pp. 400-401, **1987**.

[Kenny90] P. Kenny, M. Lenning & P. Mermelstein, " Speaker Adaptation in a large-vocabulary gaussian HMM ", IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 12, n° 9, pp. 917-920, **1990**.

[Kittler78] J. Kittler , " A Method for Determining K-Nearest Neighbours ", Kibernetis, vol. 7, pp. 313-315, **1978**.

[Kneser93] R. Kneser & H. Ney, " Improved Clustering Techniques for Class-based Statistical Language Modelling ", European Conference of Speech Communication and Technology, Berlin, Allemagne, pp. 973-976, **1993**.

[Kubala94] F. Kubala, A. Anastasakos, J. Makhoul, L. Nguyen, R. Schwartz & G. Zavaliagkos, " Comparative experiments on Large Vocabulary Speech Recognition ", IEEE-International Conference on Acoustic Signal and Speech Processing, Adelaide, Australie, pp. 561-564, **1994**.

[Kuhn90] R. Kuhn & R.D. Mori, “ A Cache-based Natural Language Model for Speech Recognition ”, IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 12, n° 6, pp. 570-582, **1990**.

[Ljole94] A. Ljolje, “ High Accuracy Phone Recognition using Context Clustering and Quasi-Triphonic Models ”, Computer Speech and Language, vol. 8, pp.129-151, **1994**.

[MG96] J. Macias-Guarasa, A. Gallardo, J. Feirreiros, J.M. Pardo, “ Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone ”, International Conference on Spoken Language Processing, Philadelphie, USA, pp. 30-33, **1996**.

[Ortmanns95] S. Ortmanns & H. Ney, “ Experimental Analysis of the Search Space for 20 000-Word Speech Recognition ”, European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 901-904, **1995**

[Richetin80] M. Richetin, G. Rives and M. Naranjo, “ Algorithme rapide pour la détermination des k plus proches voisins ”, RAIRO Informatique, vol. 14, n° 4, pp. 369-378, **1980**.

[Schwartz92] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen & P. Placeway, “ New uses for N-best Sentence Hypothesis, within the BYBLOS Speech Recognition System ”, IEEE-International Conference on Acoustic Signal and Speech Processing, vol. 1, pp. 1-4, **1992**.

[Shikano86] K. Shikano, K.F. Lee & R. Reddy, “ Speaker Adaptation through Vector Quantization ”, ”, International Conference on Spoken Language Processing, Tokyo, Japan, pp. 2643-2646, **1993**.

[Wagner74] R. A. Wagner & M. J. Fisher, “ The String to String Correction Problem ”, Journal of the ACM, vol. 21 n°1, pp. 168-173**, 1974**.

[Waast95] C. Waast, L. Bahl & M. El-Bèze, “ Fast Match on Decision Tree ”, European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 909-912, **1995**.

[Woodland95] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, S.J. Young, “ The 1994 HTK LargeVocabulary Speech Recognition System ”, ”, IEEE-International Conference on Acoustic Signal and Speech Processing, vol. 1, pp. 73-76, **1995**.

[Young89] S.J. Young, N.H. Russel, J.H.S. Thornton, “ Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems ”, Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dpt, **1989**.

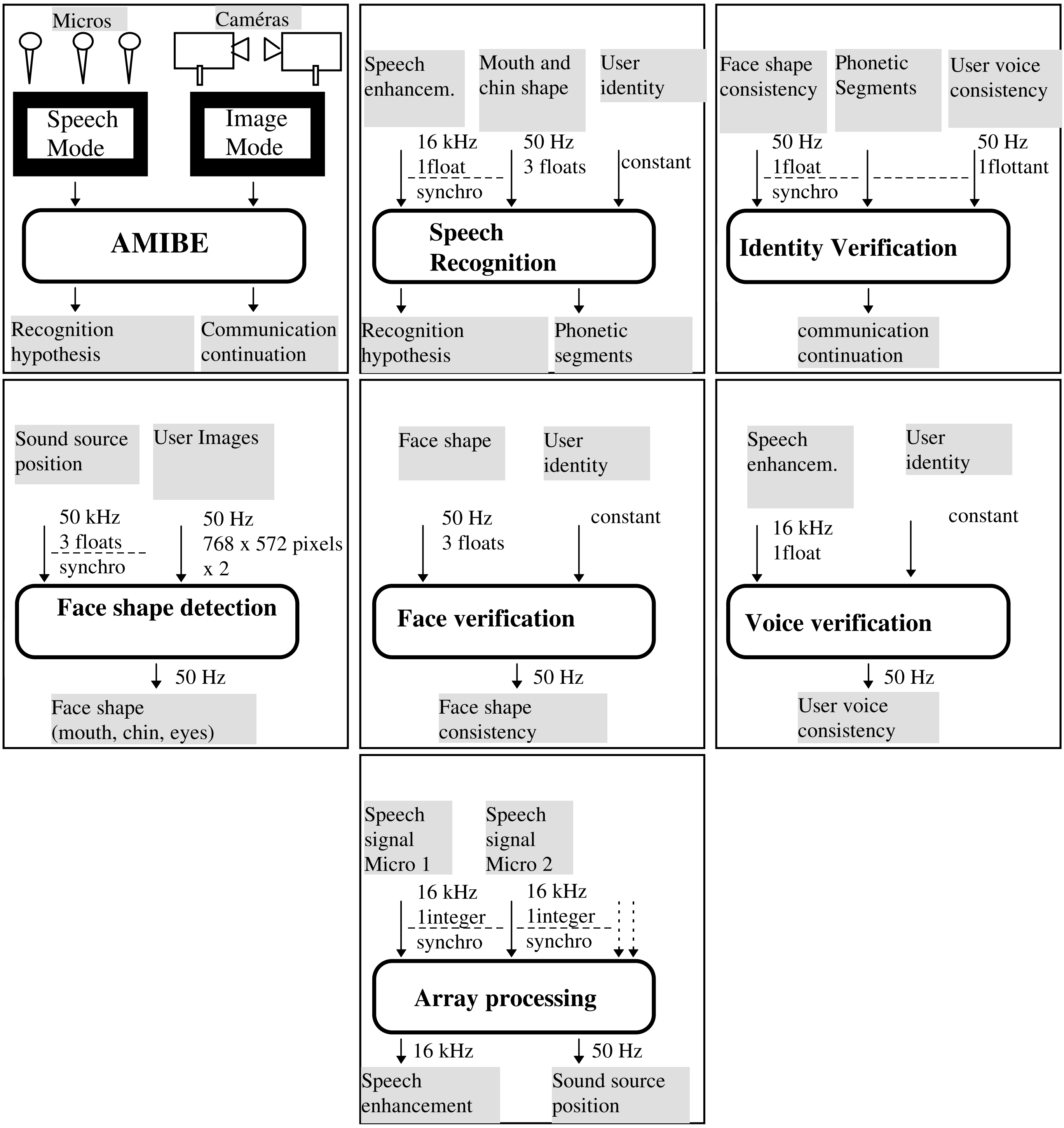
[Young92] S.J. Young, “ HTK Version 1.4: Reference Manual and User Manual ”, Cambridge University Engineering Department - Speech Group, **1992.**

V Multimedia Systems

The increase in the capacities of the internet network allows the use of the multimedia data and techniques. From now, thanks to those, the data processing extends on to the sound and the image. New multimedia technologies, -extraction, -representation and -interpretation of these data, have already a great number of applications: the indexing and the management of the multimedia databases, the access to information by the biometric data verification, the computer graphics and the multimedia communication. The multimedia data processing is one of high technologies which remain us to be acquired or to control; it uses the pattern recognition and the artificial intelligence. All these components allow really a communication more advanced beetwen man and machine.
This work have been used in several projects such as AMIBE, TRANSDOC and AVIR.

5.1. AMIBE project and the man-machine multimode interaction

I was the AMIBE project's coordinator (Multimode Applications for Interfaces and Advanced Terminals), It began in 1993 and lasted for 3 years. This project which was supported by the GDR-PRC Communication Man-Machine included researchers in six laboratories (ICP, the INRIA, the IRIT, the LAFORIA-IBP, LIA and the LIUM). The purpose was to increase the robustness of the Communications Man-Machine by amalgamating the audio and image modes. Two principal modules which were founded on a strong interaction of these two modes have been defined: recognition of the oral message transmitted by the user and the identity verification. At every moment (e.g., each centiseconde), the two modes contributed to eliminate ambiguities and to reinforce the reliability of transmitted information (oral message and identity). Three auxiliary modules were associated with the two principal modules: the lips shape recognition, the detection of the movement of the sound source and the detection of the face shape.
The project specifications (c.f., following figure) were the following ones: the oral message recognition which was operated by synchronizing a "labial reading" with a module of speech recognition: acoustico-labial recognition [Duchnowski95]. The "labial reading" consisted in a recognition of the lips which had being painted in blue and real time measurement of the labial parameters [Lalouache91]. The module of the movement detection of the sound source had two functions for the oral message decoding: speech signal enhancement using a technique of array processing and correction for the module of the lips recognition.
Apart from the coordination and definition of the specifications, our team contributions related more particularly to the speaker verification and array processing modules. For this last module, the usual methods of array processing [Silverman95] were adapted by taking account of project constraints: it was not necessary to calculate the exact localization of the sound source but only to readjust the modules of face shape and lips recognition. These was possible by the study of differences between two successive localizations. That made it possible to partly solve the problems arising to develop the recording multi-microphones system. The techniques that N. Magdeleine and T. Behar. employed were mainly based on the calculation of the intercorrelations in the frequential space and the search for a maximum by a stochastic method.



Spécifications logicielles du projet AMIBE

After the GDR-CHM examination (29/09/1994), the comments of the scientific committee about this project were the following ones: “the AMIBE project, which amalgamates speech signal and image on a very focused project and illustrating some aspects of the man-machine communication, was appreciated by this scientific committee. Its semi-project assessment is promising for good final results which could be integrated into a demonstration (in particular if it were possible to extend the bi-sensory recognition module)”.

This study allowed the integration of the "audible speech" and "visible speech" research. It showed that the cooperation of the speech and image modes allow a better man-machine communication. The problems which remained to be solved in order to obtain effective platforms are of three types. The first is the problem of the complexity of the algorithms (especially for the image processing). It can be solved by a software optimization and the development of specialized architectures. The second problem is the performances improvement of the modules (labial parameters extraction without blue painting lips, more precise sound source localization, better acoustico-labial recognition). The last problem is the definition of the interactions between the various modules: the synchronizations difficulty and the presence of feedback loops give an extremely complex system to model.

Collaboration : Marie-José Caraty, Nicolas Magdeleine (Master), Tristan Behar (Master), AMIBE consortium

Results (1993-96) : 2 national publications [8, 64]

Selected publications : pp. 137

5.2. The multimedia information retrieval

The information retrieval in the electronic documents is a field in full extension [Hartman95]. That is due to the number unceasingly crescent of documents produced by the computer generalization. The first methods of information retrieval were developed since 1950 on textual documents. These methods are generally based on the keywords frequencies in the texts [Luh58]. It allows to represent a document such a vector of the keywords space. However, these texts must be standardized to minimize the number of terms [SJ71]. Representations more complex such as of the trees are also employed but the choice of interdocuments measure is more difficult. The main problems of the text information retrieval are the standardization technique used and the employed keywords list. The taking into account of the multimedia documents is the subject of many researches [Aigrain96]. One can quote projects QBIC [Flickner95], VisualSeek, Informedia [Watclar96], OLIVE, DIVAN (audio DIstributed Visual Archives Network), AVIR (Audio Visual Indexing and Retrieval) as well as contest DARPA "Broadcast News" [Woodland98]. They are two problems for the multimedia information retrieval: the automatic identification of the components of a complex document and their interrelationships, the development of indexing techniques adapted to the various types of data (e.g., text, audio, fixed image, animated image). The first problem can be solved by a standardization effort of the multimedia documents. For the second problem, two research orientations are possible: the definition of a measure between the not-textual data, the association of textual descriptors to the not-textual data. In the first case, that allows to include not-textual data in a request. However, this request language enrichment has several disadvantages: algorithmic heaviness of the not-textual data processing, choices of weights between the different measures ones. The advantage of the second axis is the possibility of reusing the whole of the techniques of textual indexing. Nevertheless its defect is to be very dependent on the pattern recognition techniques which are employed.

We chose this last axis by chosing as application the video documents indexing. Our final goal is to be able to rebuild script from the video document. The interest to use such a type of textual descriptor is that its writing is standardized and that it is rich in

information (decorations, protagonists, actions, dialogues...). We only process the audio part of the video documents. Other teams, with whom we collaborate in a certain number of national and international projects, process image indexing. However, the fusion of descriptions which come from the two sources is made difficult by the low level of the information which was obtained from the images. For example, if one can detect the scenes changes from the sound environment and plans changes [Zhang93], that is still impossible for higher level information such as the extraction of the protagonists of a plan: for the moment it is extremely difficult to recognize them from their faces [Chellappa95] or of their silhouettes whereas the identification of a protagonist from his voice seems easier.
I tried to index the audio channel of a video. I tried to adapt our speaker recognition and vocal dictation system, to split audio data in silence/noise/music/speech segments and to associate a temporal label to each word of the video dialogue (script warping).
Two approaches are possible to index a dialogue beetween several speakers [Siu92] [Yu93][Sugiyama93][Wilcox94][Olsen95] according to whether one knows or not the identity of the actors. If we know the actors identity, a priori it is more simple but it is not easily usable when we have not speaker model of the protagonists. It is the reason why I choose the second approach by developing a not-supervised method based on the method called the "discontinuties dectection" [AO88]. Its principle is to compare three AR-Vectorial models which was calculated on the moving and overlapping windows. A discontinuty is detected when the intermodels measure become higher. Preliminary experiments were made on a dialogue comprising more than 100 speakers. This dialogue was build from the TIMIT database. The detection rate of the interspeakers borders was about 90%. However these results must be confirmed by other databases.
Adaptation of our vocal dictation system to the dialogues recognition has two difficulties: recording conditions variability and linguistic diversity of the uttered texts.
An Informedia study [Hauptmann95] showed that the recognition rates vary from 80% for a presenter who is recorded in a studio to 25% for a documentary. If the first problem can be solved with by the acoustic models adaptation, it is more difficult to adapt the language model to the document. We chose, as other teams did, to build language models per type of documents [Seymore97]. This approach is not satisfactory for two reasons : it requires very important textual resources which are difficult to gather ; the problem of the automatic recognition of the type of documents remains to be studied.
Two methods were developed successively to split the audio data in silence/noise/music/speech segments. The first one was based on the method of speaker change dectection developed in a previous paragraph. But it appeared unsuited when the sound document included a large variety of noises. Theses was showed in a test with the film "Un indien dans la ville", produced by Herve Palud. To improve the segmentation, we decided to use various type of modeling to describe the various types of segments. The new method that we developed in collaboration with S. Petridis has three steps. First one : a short-term energy histogram allows a silence/non-silence dectection. Second one: the segments of not-silence are labelled in noise/non-noise by using –energy coefficent, -a detection of the fundamental frequency -as well as a markov model trained from the already detected noise. And the third ones: The non-noise segments are split into speech and music segments using speaker recognition and acoustico-phonetic decoding techniques. We tested this method on the film "Contes de printemps" of Eric Rohmer with a segmentation error rate lower than 5%.

The method called " script warping " that I developed consists in associating with each word of the dialogue temporal labels. It is based on a training method of the phonetic hidden Markov models when a phonetic labelling of the training data is not available

[Weatley92]. This method is very employed, for in general only the text is provided. A network of the alernative phonetic transcriptions of this text is built. Then, the Viterbi algorithm uses this network to segment the speech signal by the use an initial set of phonetic models. The whole of the phonetic segmentations of the training sentences is then used to calculate new phonetic models. This procedure is reiterated as often as the phonetic models are not stabilized.
To use this method on a complete film, the principal problem araise from the presence of noise and music as well as a duration much larger (e.g., 2 hours instead of 10 seconde). The first problem was partly solved by the method described in the preceding paragraph. To solve the second problem we defined an confidence measure for each phonetic segment. It depends on various parameters such as its duration and its emission probability. Only the phonetic segments the confidence measure of which is good are used for the training of the new phonetic models. This procedure is reiterated as often as the average confidence measure increases. The words segmentation is then calculated from the phonetic segmentation.
We tested this method on two documents: a 20 minutes episode of the cartoon movie "Ivanhoe" and a 102 minutes film "Contes de printemps". In the first case, four types of errors were defined: phoneme shift, word shift, words shift, sentence shift. More than 90% of the words were located perfectly without any sentence shift. In the second case, we chose a more simple evaluation paradigm : A word is selected for each minute of film. The script warping error is given by the average error of the localization temporal of theses chosen words. After a silence/noise/music/speech segmentation, this error average was lower than one second.

Collaborations : Marie-José Caraty, Sergios Petridis (Phd thesis), Pascal Faudemay, TRANSDOC and AVIR consortiums

Results (1996-98) : 4 international publications [25, 29, 30, 34], 2 workshops [55, 56]

Selected publications : pp. 149 161

References

[Aigrain96] P. Aigrain, H.-J. Zhang & D. Petkovic, " Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review ", Multimedia Tools and Applications, pp. 179-202, **1996**.

[AO88] R. André-Obrecht, " A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals ", IEEE Transactions on Acoustic Signal and Speech Processing, vol. 36, pp. 29-40, **1988**.

[Chellappa95] R. Chellappa, C. L. Wilson & S. Sirohey, « Human and Machine Recognition of Faces: A Survey » IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 83, pp. 705-740, **1995**.

[Duchnowski95] P. Duchnowski, M. Hunke, D. Busching, U. Meir & A. Waibel, « Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition », International Conference on Acoustic Signal and Speech Processing, pp. 109-112, **1995**.

[Flickner95] M. Flickner, H.S. Sawhney & J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, Dr. Petkovic, D. Steele, P. Yanker, " Query by Image and Video Content: the QBIC System ", IEEE-Computer, 28, 9, pp. 23-32, **1995**.

[Hartman95] D. Hartmann, “ Overview of the Second Text Retrieval Conference ”, Information Processing and Management, vol. 3, pp. 271-289, **1995**.

[Hauptmann95] A.G. Hauptmann, “ Speech Recognition in the Informedia Digital Video Library : Uses and Limitations ”, IEEE Tools with Artificial Intelligence Conference, 7 pages, 1995.

[Lallouache91] T. Lallouache “ Un poste 'visage-parole' couleur. Acquisition et traitement automatique des lèvres ”, Thèse de Doctorat, INP Grenoble, **1991**.

[Luh58] H.P. Luhn, “ The Automatic Creation of Literature Abstracts ”, IBM Journal of Research and Development ”, vol 2, pp. 159-165, **1958**.

[Olsen95] J. Olsen, “ Separation of speakers in audio data ”, European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 355-358, **1995**.

[Seymore97] K. Seymore, S. Chen, M. Eskenazi & R. Rosenfeld, “ Language and Prononciation Modeling in the CMU 1996 HUB 4 Evaluation ”, 1997 DARPA Speech Recognition Workshop, 13 pages, **1997**.

[Silverman95] H. F. Silverman & S. E. Kirtman, « A Two-Stage Algorithm for Determining Talker Location from Linear Microphone Array Data » Computer Speech and Language, vol 6, no. 2, pp.129-152, **1995**.

[Siu92] M.-H. Siu, G. Yu & H. Gish, “ An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers ”, IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 189-192, **1992**.

[Sugiyama93] M. Sugiyama, J. Murakami & H. Watanabe, “ Speech Segmentation and Clustering Based on Speaker Features ”, International Conference on Acoustic Signal and Speech Processing, pp. 395-398, **1993**.

[SJ71] K. Spark Jones, “ Automatic Keyword Classification for Information Retrieval, Butterworths, **1971**.

[Watclar96] H.D. Watclar, T. Kanade, M.A. Smith & S.M. Stevens, “ Intelligent Access to Digital Video: the Informedia Project ”, Computer, 29, 5, pp. 46-53, **1996**.

[Weatley92] B. Weatley, G. Doddington, C. Hemphill, J. Godfrey, E. Holliman, J. Mac. Daniel & D. Fisher, “ Robust Automatic Time Alignement of Ortographic Transcription with Unconstrained Speech ”, IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 533-536, **1992**.

[Wilcox94] L. Wilcox, F. Chen, D. Kimber & V. Balasubramanian, “ Segmentation of speech using speaker identification ”, ", International Conference on Acoustic Signal and Speech Processing, pp. 161-164, **1994**.

[Woodland98] P.C. Woodland, T. Hain., S.E. Johnson, T.R. Niesler, A. Tuerk & S.J. Young, “ Experiments in Broadcast News Transcriptions ”, IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 909-912, **1998**.

[Yu93] G. Yu & H. Gish, “ Identification of speakers engaged in dialog ”, IEEE-International Conference on Acoustic Signal and Speech Processing, pp. 383-386, **1993**.

[Zhang93] H.J.Zhang, A.Kankanballi & S.W.Smoliar, “ Automatic Partitionning of Full-motion Video ”, Multimedia Systems, Vol.1, No.1, pp. 10-28, **1993**.

VI Conclusions and Perpectives

The work summarized throughout this document testifies eight years of post-doctoral research, from 1991 to 1998. The speech processing research underwent during this time of the enormous modifications: for quite precise conditions of use, the problem of the continuous speech recognition is on the way to be solved whereas it was considered as extremely difficult 15 years ago. One can practically tell the same about the speaker recognition. The first vocal dictation systems start to be sold (a few hundreds euros) even if their performances are not yet those which were excepted. One of the explanations of this unforeseen progress (apart from the increase in computing power and the reduction in the memory cost) is the evaluation plan organization made in the United States. It made it possible to detect the most promising techniques. However, this explanation is insufficient, for a similar evaluation had been made in the seventies without convincing results.

My results in speech recognition and speaker recognition enabled us to participate in this evolution by taking part in the of speaker recognition and vocal dictation evaluations. I brought to the speech processing field contributions the importance of which is attested by a number of publications. These research activities used concepts and methods which use signal processing, artificial intelligence, linguistic genius, and pattern recognition. At the same time, they integrated experimental, theoretical and technical aspects, and took into account industrial requirements. They required a research team which was essential to the making of the prototypes, and thus induced responsibilities and administrative tasks, which was necessary to the research coordination and the contracts management which ensure the financing of the team .
The medium-term prospects, in which we will invest ourselves, are in the development of new technologies to recognize and to manage multimedia data. This research field will become one of the principal stakes of the future computer science in the undertaken researches and in the applications. We already started to explore it in taking part in projects on the man-machine communication and the information retrieval. Our aim is to verify the quality of the multimedia indexing. Indeed, our recognition systems which is based on statistical approaches have only given good performances to identify the most frequent patterns: in speech recognition, they are the most employed expressions or words. That is also due to the evaluation paradigms used in which all the words have the same importance. A word recognition rate of 90% is thus regarded as satisfactory. Still we must noticed it is almost certain that the 10% words which remains contain many relevant informations. The only means to obtain an “intelligent indexing ” is to introduce into our systems a comprehension component. This component must be integrated in the decoding step. It should not be used after the decoding step which is currently done. Theses results can only be obtained by a collaboration between oral linguistics and written linguistics.researches. These one exists in the Anglo-Saxon countries. It allows the common use and the development of important resources such as the lexicons, the language models or the grammatical labellings. That must be also accompanied by a modification of the evaluation paradigms to take a better account of the use of the indexing results.

There are other prospects in the field of the sensory analysis with the study of electronic noses. Various people working on this application kept in touch with us, for some of their problems are ours: They are biological sensors. The complex properties of them (large diversity of the molecular receivers) were well studied, while the associated data processing is badly known. Various hypothesis on lateral inhibition, the existence of an olfactive image, the importance of the temporal evolution of these data are not validated. The principal effort relates to reliability, the selectivity and the sensitivity of the sensors, but the question of an advanced data-processing arises already. The techniques which are used (e.g., principal components analysis) are closer to the date analysis than the pattern recognition. Simple methods such the dynamic comparison are not employed: for example, that would make it possible to verify the hypothesis of the temporal evolution. Actually, the electronic noses are especially used for the quality control. Still the data-processing management of this third sense (after hearing and sight) will be most important in the field of virtual or greater reality and more generally for the perception by the computers of their environment.

Curriculum Vitae

Claude Montacié

Etat-civil :	né le 6 août 1961 à NICE célibataire
Adresse personnelle :	10, rue de Vintimille 75009 Paris tel. (33) 1 42 82 16 53
Adresse professionnelle	Laboratoire d'Informatique de Paris 6 Université Pierre et Marie Curie 4, place Jussieu - BP. 169 75232 Paris Cedex 05 Tel : (33) 1 44 27 62 81 E-mail : Claude.Montacie@lip6.fr
Situation actuelle :	Maître de conférences à l'Université Pierre et Marie Curie

Diplômes et titres

Licence de Mathématiques, Université Denis Diderot (Paris 7), juillet 1983

Licence d'Informatique, Université Pierre et Marie Curie (Paris 6), juillet 1983

Maîtrise d'Informatique, Université Pierre et Marie Curie (Paris 6), juillet 1984

Diplôme d'Ingénieur de l'Ecole Supérieure d'Electricité, juillet 1986

Docteur de l'Ecole Nationale Supérieure des Télécommunications, février 1991

Titre : *Décodage acoustico-phonétique : Apport de la décomposition temporelle généralisée et de transformations spectrales non-linéaires.*

Jury : J. Mariani (président), L. Miclet (rapporteur), C. Wellekens (rapporteur), X. Rodet (directeur de thèse), G. Chollet, H. Méloni

Activités professionnelles

Depuis 1991	Maître de Conférences en Informatique à l'Université Paris 6 1995, promotion 1ère classe des Maîtres de Conférences Depuis 1994, titulaire d'un contrat d'encadrement doctoral et de recherche (renouvelé en 1998)
1990-1991	Attaché temporaire d'Enseignement et de Recherche en Informatique à l'Université Paris 6

Activités administratives

De 1992 à 1996 : membre élu au Conseil de l'UFR d'Informatique

De 1993 à 1996 : membre du conseil de direction du GDR-PRC CHM

Depuis 1998 : membre élu (titulaire) à la commission de spécialistes de l'Université Paris 6

membre élu du conseil de laboratoire du LIP6

membre (suppléant) de la commission de spécialistes de l'Université Paris 11

Activités scientifiques

Organisation de manifestations scientifiques

Participation au comité scientifique des JEP 94, JEP 96, JEP 98

Responsabilités scientifiques dans des projets français

Intégration et modélisations complémentaires du signal vocal : projet du GDR-PRC Communication Homme Machine (1990-1992)
Coresponsable du projet pour l'Université Pierre et Marie Curie

Architectures hybrides pour la reconnaissance de la parole : projet Cognisciences " réseaux de neurones formels " (1991-1993)
Coresponsable du projet pour l'Université Pierre et Marie Curie

Détection de mots dans la parole continue : projet du GDR-PRC Communication Homme Machine (1992-1993)
Coresponsable du projet pour l'Université Pierre et Marie Curie

AMIBE (Applications Multimédia pour Interfaces et Bornes Evoluées) : projet du GDR-PRC Communication Homme Machine (1993-1996)
Coordinateur du projet (six laboratoires, vingt personnes)

TRANSDOC : projet du ministère de l'Industrie (1996-1998)
Responsable de l'indexation audio

Responsabilités scientifiques dans des projets internationaux :

ARC-B1 : projet de l'AUPELF-UREF (1996-2000)
Participation aux Evaluations des systèmes de dictée vocale

SRE : projet du National Institute of Standard and Technology (1998)
Participation aux Evaluations des systèmes de vérification du locuteur

AVIR : projet ESPRIT 28798 (1998-2000)
Conception et réalisation d'un serveur d'indexation audio

Comité de lecture :

journaux (Speech Communication, Traitement du Signal) depuis 1994

Annexe A
Encadrements de travaux de recherche

Encadrement de thèses (toutes ces thèses sont co-encadrées par M.-J. Caraty)

Claude Barras (40%): Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans le modèles de Markov cachés
Diplôme de Docteur de l'Université Pierre et Marie Curie, 29 mai 1996
Jury : Pt R : R. De Mori, R : R. André-Obrecht, M.-J. Caraty, M. El-Bèze, P. Gallinari, C. Montacié, X. Rodet

Claude Barras a développé et testé différentes approches pour l'apprentissage (classes de locuteurs, test du khi-carré multidimensionel) des modèles de Markov et le décodage (rupture de modèles). Il les a implémenté et testé à l'aide du logiciel HTK sur la base de données TIMIT.

Jean-Luc Le Floch (60%):. Identification du locuteur en milieu bruité indépendamment du texte
Diplôme de Docteur de l'Université Pierre et Marie Curie

Jean Luc Le Floch a développé des méthodes de reconnaissance du locuteur par coopération/compétition de modèles AR-vectoriels et de sommes pondérées de gaussiennes. Il a testé ces méthodes sur la base de données N-TIMIT. Il a arrêté sa thèse pour l'instant pour développer une serrure vocale chez un industriel à partir de notre brevet.

Fabrice Lefèvre (60%): Etude et amélioration des modèles de Markov cachés discriminants par des méthodes non paramétriques.
Diplôme de Docteur de l'Université Pierre et Marie Curie
date de soutenance prévue : début 1999

Fabrice Lefèvre a développé les méthodes d'apprentissage des modèles de Markov cachés en utilisant un estimateur non-paramétrique. Il a également participé à la première campagne de dictée vocale en implémentant le moteur de reconnaissance de notre système.

Sergios Pétridis (40%): Indexation sonore : parole et musique
Diplôme de Docteur de l'Université Pierre et Marie Curie
date de soutenance prévue : 2000
Sergios Pétridis est en thèse depuis un an. Il a développé un algorithme de segmentation silence/bruit/parole.

Encadrement de stages

Younes Bennani (100%) : Décodage acoustico-phonétiques et réseaux de neurones, 01/04/87-01/10/87, DEA Paris 5

Patrick Chaze (40%) : Apprentissage automatique d'un système de règles, 01/04/88-01/10/88, DEA IARFA Paris 6.

Jean-Luc Docci (50%) : Contrôle de la qualité d'une base de données de parole, 01/04/89-01/10/89, DEA IARFA Paris 6.

Luc Julia (10%) : Vers une interface multimodale, 01/04/92-01/10/92, DEA Paris 7.

Alain Rousseau (100%) : Aide à la prononciation des langues étrangères, 01/10/92-01/10/93,scientifique du contingent.

Jérome Goût (50%) : Modélisation non paramétrique des modèles de Markov cachés, 01/04/93-01/10/93, DEA IARFA Paris 6.
coencadrant Marie-José Caraty

Benjamin Mayrargue (30%) : Adaptation au locuteur par modèles AR-vectoriel, 01/04/94-01/10/94, DEA IARFA Paris 6.

Bernard Alary (50%) : Modèle de langage pour la reconnaissance de la parole, 01/10/94-01/10/95, Post-DEA (congé de formation).

Fabrice Lefèvre (100%) : Modèles de langage et N-best, 01/04/95-01/10/95, DEA IARFA Paris 6.

Cécile Lecordier (40%) : Etude d'un modèle spécifique de reconnaissance des voyelles, 15/04/96-15/09/96, DESS Paris 5.

Emmanuel Daidone (50%) : Optimisation de la gestion mémoire d'un moteur de reconnaissance, 01/05/96-01/08/96, PE (Programmeur d'Etude) Paris 6.

Nicolas Magdeleine (100%) : Traitement d'antenne, 01/06/97-01/09/97, Maîtrise Paris 6.

Tristan Behard (100%) : Localisation par Traitement d'antenne, 01/06/97-01/09/97 Maîtrise Paris 6.

Sergios Pétridis (70%) : Reconnaissance de la langue, 01/06/97-01/09/97, DEA IARFA Paris 6.

DEA IARFA (Intelligence Artificielle, Reconnaissance des Formes et Applications)

Participation à des jurys de thèses

Thierry Rouge : Optimisation de l'architecture des serveurs vocaux.
Diplôme de Docteur de l'Université Pierre et Marie Curie, 29 octobre 1993.
Jury : Pt : C. Girault, R : B. Cousin, R : G. Hébuterne, E. Horlait, C. Gagnoulet, C. Montacié.

Delphine Charlet : Authentification vocale par téléphone en mode dépendant du texte.
Diplôme de Docteur de l'Ecole Nationale Supérieure des Télécommunications, 24 Novembre 1997.
Jury : Pt : J.-P. Tubach, R : G. Chollet, R : C. Montacié, J.-F. Bonastre, S. Furui, M.-C. Haton, D. Jouvet.

Driss Matrouf : Adaptation de Modèles Acoustiques pour la reconnaissance de la Parole bruitée.
Diplôme de Docteur de l'Université d'Orsay, 17 décembre 1997.
Jury : Pt : D. Etiemble, R : R. André-Obrecht, R : J.-P. Haton, M. Adda-Decker, J.-L. Gauvain, C. Montacié.

Annexe B
Travaux et Publications

Brevets

1. **C. Montacié**, M.-J. Caraty & J.-L. Le Floch, “ Procédé de reconnaissance vocale d'un locuteur mettant en oeuvre un modèle prédictif, notamment pour les applications de contrôle d'accés. ” , Brevet français 9605566, déposé par l’Université Pierre et Marie Curie, le 3 mai **1996.** Extension internationale PCT/FR 9700753 déposée le 25 Avril **1997.**

Revues Spécialisées avec Actes et avec Comité de Lecture

2. K. Choukri, S. Soudoplatoff, A. Wallin, F. Bimbot, G. Chollet, **C. Montacié**, P. Gallinari, Y. Bennani, A. Varga & M. Immendorfer, “ SPRINT : Speech Recognition Using Integrated Neuro-Computing Techniques ” *in* International Journal of Neuro-Computing, Elsevier, Vol. 2, pp. 71-98, **1990.**

3. J.-P. Tubach, G. Chollet, K. Choukri, **C. Montacié**, C. Mokbel & H. Valbret, “ Adaptation au locuteur de systèmes de reconnaissance. Régression linéaire multiple et perceptrons multicouches ” *in* Traitement du Signal, Vol. 7, n° 4, pp. 285-292, **1991.**

Ouvrages de Synthèse

4. G. Chollet & **C. Montacié**, “ Reference Systems for Speech Recognition Research, Development and Evaluation ” *in* H. Niemann (Editor) *Recent Advances in Speech Understanding and Dialog Systems*, NATO ASI Series, Vol. F 46, Springer-Verlag, Berlin, pp. 345-348, **1988**.

5. K. Choukri, G. Chollet & **C. Montacié**, “ Speaker Adaptation Using Multi-Layer Feed-Forward Automata and Canonical Analysis ” *in* F. Fogelman Soulié (Editor) *Neurocomputing Algorithms, Architectures and Applications*, NATO ASI Series, Vol. F 68, Springer-Verlag, Berlin, pp. 261-264, **1990**.

6. M.-J. Caraty, **C. Montacié** & X. Rodet, “ Automatic Learning of a Production Rule System for Acoustic-Phonetic Decoding ” *in* P. Laface and R. De Mori (Editor) *Speech Recognition and Understanding, Recents Advances, Trends and Applications*, NATO ASI Series, Vol F 75, Springer-Verlag, Berlin, pp. 311-316, **1992**.

7. C. Barras, M.-J. Caraty & **C. Montacié**, “ Temporal Control and Learning Set Choice Applied to Hidden Markov Models ” *in* A. Ayuso (Editor) *Speech Recognition and Coding, New Advances and Trends*, NATO ASI Series, Vol F147, Springer-Verlag, Berlin, pp. 72-75, **1995.**

8. **C. Montacié,** M.-J. Caraty, R. Andre-Obrecht, L.-J. Boë, P. Deléglise, M. El-Bèze, I. Herlin, P. Jourlin, T. Lallouache, B. Leroy. & H. Méloni, “ Applications Multimodales pour Interfaces et Bornes évoluées” , Ecole d'été "*Fondements et perspectives en traitement de la parole*", pp. 155-164, **1995.**

9. F. Lefèvre, **C. Montacié** & M.-J. Caraty, “ K-Nearest Neighbours Estimator in a HMM-Based System ” *in* K.M. Ponting (Editor) *Computational Models of Speech Patterns Processing*, NATO ASI Series, Springer-Verlag, Berlin, **1998** (à paraître).

Congrès Internationaux avec Actes et avec Comité de Lecture

10. F. Bimbot, P. Deléglise, G. Chollet & **C. Montacié**, “ Temporal Decomposition and Acoustico-Phonetic Decoding of Speech ”, IEEE-International Conference on Acoustic Signal and Speech Processing, New York, USA, pp. 445-448, **1988**.

11. P. Deléglise, F. Bimbot, **C. Montacié** & G. Chollet, “ Temporal Decomposition and Acoustico-Phonetic Decoding for the Automatic Recognition of continuous Speech ” International Conference on Pattern Recognition, Rome, Italie, pp. 839-841, **1988**.

12. G. Chollet, K. Choukri & **C. Montacié**, “ Test Workstation for the Evaluation of Speech Recognition Algorithms, Applications and Data Bases ”, I7th FASE Symposium, Edinburgh-Ecosse, pp. 145-151, **1988**.

13. **C. Montacié**, K. Choukri & G. Chollet, “ Speech Recognition Using Temporal Decomposition and Multi-Layer Feed-Forward Automata ”, IEEE-International Conference on Acoustic Signal and Speech Processing, Glasgow, Angleterre, pp. 409-412, **1989**.

14. M.-J. Caraty, **C. Montacié**, P. Chazé & X. Rodet, “ Application of an Automatic Learning System for Acoustic-Phonetic Decoding ”, Third Cognitiva Symposium, Madrid **,** Espagne, pp. 847-850, **1990**.

15. J.-P. Tubach, G. Chollet, F. Bimbot, P. Deléglise & **C. Montacié**, “ Temporal Decomposition of Speech : Developments and Experiments at E.N.S.T. in Speech Analysis, Recognition and Synthesis ”, International Conference on Signal Processing, Beijing, Chine, **1990**.

16. K. Choukri, S. Soudoplatoff, A. Wallin, F. Bimbot, G. Chollet, **C. Montacié**, P. Gallinari, Y. Bennani, A. Varga & M. Immendorfer, “ Speech Recognition Using Integrated Neuro-Computing Techniques ”**,** International Neural Network Conference, Paris, **1990**.

17. **C. Montacié**, M.-J. Caraty & X. Rodet, “ Experiments in the Use of an Automatic Learning System for Acoustic-Phonetic Decoding ”, International Conference on Spoken Language Processing, Kobe, Japon, pp. 357-360, **1990**.

18. **C. Montacié**, P. Deléglise, F. Bimbot & M.-J. Caraty, “ Kinematic Techniques for Speech Processing : Temporal Decomposition and Multi-variate Linear

Prediction ", IEEE-International Conference on Acoustic Signal and Speech Processing, San Francisco, USA, pp. 153-156, **1992**.

19. **C. Montacié** & J.-L. Le Floch, " AR-Vector Models for Free-Text Speaker Recognition ", International Conference on Spoken Language Processing, Banff, Canada, pp. 611-614, **1992**.

20. M.-J. Caraty, **C. Montacié** & C. Barras, " Integration of Frequential and Temporal Structurations in a Symbolic Learning System ", International Conference on Spoken Language Processing, Banff, Canada, pp. 475-478, **1992**.

21. **C. Montacié** & J.-L. Le Floch, " Discriminant AR-Vector Models for Free-Text Speaker Verification ", European Conference of Speech Communication and Technology, Berlin, Allemagne, pp. 161-164, **1993**.

22. J.-L. Le Floch**,** **C. Montacié** & M.-J. Caraty, " Investigations on Speaker Characterization from Orphée System Technics ", IEEE-International Conference on Acoustic Signal and Speech Processing, Adelaide, Australie, pp. 149-152, **1994**.

23. C. Barras, M.-J. Caraty & **C. Montacié**, " Temporal Control and Training Selection for HMM-Based System ", European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 27-30, **1995**.

24. J.-L. Le Floch, **C. Montacié** & M.-J. Caraty, " Speaker Recognition Experiments on the NTIMIT Database ", European Conference of Speech Communication and Technology, Madrid, Espagne, pp. 379-382, **1995**.

25. P. Faudemay, L. Cheng, **C. Montacié**, M.-J. Caraty, C. Maloigne, X. Tu, M. Ardebilian & J.-L. Le Floch, " Multi-channel Video Segmentation ", International Conference on Multimedia Storage and Archiving Systems, SPIE Symposium, Boston, USA, pp. 252-264, **1996**.

26. J.-L. Le Floch, **C. Montacié** & M.-J. Caraty, " GMM and ARVM Cooperation and Competition for Text-Independent Speaker Recognition on Phone Speech ", International Conference on Spoken Language Processing, Philadelphie, USA, pp. 2411-2414, **1996**.

27. **C. Montacié**, M.-J. Caraty & C. Barras, " Mixture Splitting Technic and Temporal Control in a HMM-Based Recognition System ", International Conference on Spoken Language Processing, Philadelphie, USA, pp. 977-980, **1996**.

28. M.-J. Caraty, **C. Montacié** & F. Lefèvre, " Dynamic Lexicon for a very large vocabulary vocal dictation ", European Conference of Speech Communication and Technology, Rhodes, Grèce, pp. 2691-2694, **1997**.

29. P. Faudemay, **C. Montacié** and M.-J. Caraty, " Video Indexing Based on Image and Sound ", International Conference on Multimedia Storage and Archiving System, SPIE Symposium, Dallas, USA, pp. 57-69, **1997**.

30. **C. Montacié** & M.-J. Caraty, “ Sound Channel Video Indexing ”, European Conference of Speech Communication and Technology, Rhodes, Grèce, pp. 2359-2362, **1997**.

31. **C. Montacié**, M.-J. Caraty & F. Lefèvre, “ K-NN versus Gaussian in a HMM based Recognition System ”, European Conference of Speech Communication and Technology, Rhodes, Grèce, pp. 529-532, **1997.**

32. M.-J. Caraty & **C. Montacié**, “ Multiresolution Analysis for Speech Recognition, International Conference on Spoken Language Processing ”, Sydney, Australie, **1998**, (à paraître).

33. F. Lefèvre, **C. Montacié** & M.-J. Caraty, “ Influence of the Probability Estimator in a HMM-based System ”, Australian Conference on Speech Science and Technology, Sydney, Australie, **1998**, (à paraître).

34. **C. Montacié** & M.-J. Caraty, “ A Silence/Noise/Music/Speech Splitting Algorithm ”, International Conference on Spoken Language Processing, Sydney, Australie, **1998**, (à paraître).

Congrès National avec Actes et Comité de Lecture

35. **C. Montacié** & G. Chollet, “ Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance automatique de la parole ”, 16èmes Journées d'Etudes de la parole, Hammamet, Tunisie, pp. 323-326, **1987**.

36. P. Deléglise, **C. Montacié** & F. Bimbot, “ La décomposition temporelle, une technique cinématique de segmentation et de décodage acoustico-phonétique ”, 18èmes journées d'Etudes de la parole, Montréal, Canada, pp. 347-352, **1990**.

37. T. Artières, Y. Bennani, P. Gallinari & **C. Montacié**, “ Connectionist and Conventional Models for Free-Text Talker Identification Tasks ”, NeuroNîmes, Nîmes, France, pp. 189-202, **1991**.

38. C. Barras, M.-J. Caraty, P. Deléglise, **C. Montacié**, R. André-Obrecht & X. Rodet, “ Décomposition temporelle et ruptures de modèles pour le décodage acoustico-phonétique ”, 19èmes Journées d'Etudes de la Parole, pp. 335-340, Bruxelles, Belgique, **1992**.

39. M.-J. Caraty & **C. Montacié**, “ Intégration de la décomposition temporelle généralisée dans un système d'apprentissage symbolique. Application à la reconnaissance des voyelles ”, 19èmes Journées d'Etudes de la Parole, pp. 387-392, Bruxelles, Belgique, **1992**.

40. **C. Montacié** & J.-L. Le Floch, “ Modèles autorégressifs vectoriels et reconnaissance du locuteur ”, 19èmes Journées d'Etudes de la Parole, pp. 439-443, Bruxelles, Belgique, **1992**.

41. C. Barras, M.-J. Caraty & **C. Montacié**, " Contrôle temporel et sélection de l'apprentissage appliqués aux modèles de Markov cachés ", RFIA, Reconnaissance des Formes et Intelligence Artificielle, Paris, France, pp. 391-396, **1994**.

42. C. Barras, M.-J. Caraty & **C. Montacié**, " Topologies des locuteurs et pré-segmentation pour les modèles de Markov cachés ", 20èmes Journées d'Etudes de la Parole, Lannion, France, pp. 491-495, **1994**.

43. J.-L. Le Floch, **C. Montacié** & M.-J. Caraty, " Investigation de protocoles pour la reconnaissance des locuteurs fondée sur le système Orphée ", 20èmes Journées d'Etudes de la Parole, Lannion, France, pp. 255-258, **1994**.

44. M.-J. Caraty, C. Barras, F. Lefèvre & **C. Montacié**, " D-DAL : un système de dictée vocale développé sous l'environnement HTK ", 21èmes Journées d'Etudes de la Parole, Avignon, France, pp. 289-292, **1996**.

45. J.-L. Le Floch, **C. Montacié**, M.-J. Caraty, " Coopération et compétition de modèles en reconnaissance du locuteur ", 21èmes Journées d'Etudes de la Parole, Avignon, France, pp. 395-398, **1996**.

46. P. Faudemay, L. Chen, **C. Montacié**, M.-J. Caraty, X. Tu, " Segmentation multi-canaux de vidéos en séquences ", 3ème Journées d'Etudes et d'Echanges "COmpresion et REprésentation des Signaux Audiovisuels", Issy les Moulineaux, France, **1997**.

47. **C. Montacié,** M.-J. Caraty & F. Lefèvre, " Lexique dynamique dans un système de reconnaissance de la parole ", Congrès français d'acoustique, Marseille, France, pp. 413-416, **1997**.

Workshops

48. F. Bimbot, G. Chollet, P. Deléglise & **C. Montacié**, " Décomposition temporelle et décodage acoustico-phonétique ", 1ères Journées du GRECO-PRC Communication Homme-Machine, Nancy, France, **1988**.

49. **C. Montacié** & G. Chollet, " Recognition of Continuous Speech Using Temporal Decomposition "**,** Second Franco-British Meeting on Speech, Sussex, Grande-Bretagne, **1988**.

50. K. Choukri, **C. Montacié**, F. Bimbot & G. Chollet, " Transformations spectrales par corrélations canoniques ou automates en couches ", Séminaire Variabilité et Spécificité du locuteur, Marseille, France, pp. 192-195, **1989**.

51. K. Choukri, F. Bimbot, **C. Montacié** & G. Chollet, " Voice Transformation Using Neural Networks for Speech Recognition and Synthesis "**,** Eurasip Workshop on Neural Networks, Sesimbra, Portugal, **1990.**

52. M.-J. Caraty, **C. Montacié**, P. Chazé & X. Rodet, " Application d'un système de règles au décodage acoustico-phonétique ", 2èmes Journées du GRECO-PRC Communication Homme-Machine, Paris, France, pp. 255-265, **1991**.

53. C. Barras, M.-J. Caraty, **C. Montacié**, P. Deléglise, R. André-Obrech, F. Bimbot & J.-L. Le Floch, “ Approche événementielle pour la reconnaissance de la parole ”, 3ème Journées du GRECO-PRC Communication Homme-Machine, Nancy, France, pp. 24-34, **1994**.

54. M.-J. Caraty, **C. Montacié** & F. Lefèvre, “ D-DAL le Dictaphone-Dactylographe du LIP6 ”, Atelier concours de dictée vocale de l’AUPELF-UREF, Avignon, France, **1997**.

55. **C. Montacié**, “ Caractérisation du locuteur ”, Action Indexation Multimedia, Paris, France, **1997**.

56. P. Faudemay, M.-J. Caraty & **C. Montacié**, “ Benchmarking Methods for Audio ”, MPEG Requirements Group, MPEG Meeting, Tokyo, Japon, Document ISO/MPEG N3313, **1998**.

57. **C. Montacié** & M.-J. Caraty, “ AR-Vector Models for Speaker Recognition ”, Speaker Recognition Workshop, National Institute of Standart and Technology, Maryland, USA, **1998**.

Thèse

58. **C. Montacié,** “ Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformations spectrales non linéaires ”, Thèse d’Université, Ecole Nationale Supérieure des Télécommunications, Paris, France, 170 pages, **1991**.

Rapports de Fin de Contrat

59. G. Chollet & **C. Montacié**, “ Reference system and speech recognizer assessment ”. Contributions à WP2, contrat ESPRIT1 : Speech Assesment Methods, **1988**.

60. **C. Montacié** & H. Barral, “ Réseaux neuromimétiques ”. Rapport de contrat EDF, **1989**.

61. Y. Bennani, K. Choukri, P. Gallinari, **C. Montacié**, O. Segard, H. Valbret, A. Varga, A. Wallin, “ SPRINT, Esprit Project - BRA 3228, Deliverable of Task2 ”, Edited by K. Choukri, **1990**.

62. C. Barras, M.-J. Caraty, O. Deste, J.-G. Ganascia & **C. Montacié**, “ Acquisition de Connaissance en Traitement de la Parole ”, Actes de la Journée IBP, pp. 32-37, Paris, **1992**.

63. M.-J. Caraty, **C. Montacié**, Rapport d’Activité 1995 (arrêté à la date du 31 mars 1996) ARC “ Linguistique, Informatique et corpus oraux : Dictée vocale ”, “ ILOR B1 ”, Convention ARC n° X/7.10.04/ILOR.B1t.rec/ Paris6.LAFORIA.MJC, **1996**.

64. **C. Montacié**, M.-J Caraty, T. Lallouache, P. Deléglise, P. Jourlin, R. André-Obrecht, I. Herlin, T. Guiard, M. Alissali, M. El-Bèze, B. Jacob, B. Leroy, A. Adjoudani, A. Rogozan, H. Méloni, N. Parlangeau, F. Pellegrino, C. Cénac,

“ Projet du GDR-PRC Communication Homme Machine AMIBE (Applications Multimodales pour Interfaces et Bornes évoluées) ”, 90 pages, **1996**.

65. M.-J. Caraty, **C. Montacié**, Rapport Biennal d’Activité Scientifique 1995-1996 (arrêté à la date du 25 mars 1997) ARC “ Linguistique, Informatique et corpus oraux : Dictée vocale ”, “ ILOR B1 ”, Convention ARC n° X/7.10.04/ILOR.B1t.rec/ Paris6.LAFORIA.MJC, **1997**.

66. **C. Montacié** & M.-J. Caraty, “ Reconnaissance du Locuteur en Indexation ”, Rapport d’Activité Scientifique 1996 (arrêté à la date du 21 mai 1997), projet du ministère de l’Industrie TRANSDOC, **1997**.

Annexe C
Recueil de publications

L'analyse de la parole

La reconnaissance du locuteur

La reconnaissance de la parole

Systèmes multimédia

C. Montacié, M.-J. Caraty, R. Andre-Obrecht, L.-J. Boë, P. Deléglise, M. El-Bèze, I. Herlin, P. Jourlin, T. Lallouache, B. Leroy. & H. Méloni, " Applications Multimodales pour Interfaces et Bornes évoluées" , Ecole d'été "*Fondements et perspectives en traitement de la parole*", pp. 155-164, **1995.** 137

P. Faudemay, L. Cheng, **C. Montacié**, M.-J. Caraty, C. Maloigne, X. Tu, M. Ardebilian & J.-L. Le Floch, " Multi-channel Video Segmentation ", International Conference on Multimedia Storage and Archiving Systems, SPIE Symposium, Boston, USA, pp. 252-264, **1996**. 149

C. Montacié & M.-J. Caraty, " A Silence/Noise/Music/Speech Splitting Algorithm ", International Conference on Spoken Language Processing, Sydney, Australie, **1998**, (à paraître). 161