INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet METISS

Modélisation et Expérimentation pour le Traitement des Informations et des Signaux Sonores

Rennes

THÈME 3A

Table des matières

1. Composition de l'équipe **1**
2. Présentation et objectifs généraux **1**
3. Fondements scientifiques **1**
3.1. Introduction 1
3.2. Approches probabilistes 2
3.2.1. Formalisme et modélisation probabiliste 2
3.2.2. Estimation statistique 3
3.2.3. Algorithmes de calcul de vraisemblance et de décodage 4
3.2.4. Décision Bayésienne 4
3.3. Représentations adaptatives 5
3.3.1. Systèmes redondants et décomposition adaptative 5
3.3.2. Critères de parcimonie 6
3.3.3. Algorithmes de décomposition 6
3.3.4. Construction de dictionnaires 7
3.3.5. Séparation de signaux 7
4. Domaines d'application **8**
4.1. Introduction 8
4.2. Vérification du locuteur 8
4.3. Détection et suivi d'information dans les flux sonores 9
4.3.1. Détection de locuteur 9
4.3.2. Détection et suivi de classes de son 10
4.3.3. Intégration audiovisuelle pour l'indexation 10
4.4. Traitement avancé de signaux sonores 11
4.4.1. Séparation de sources sonores 11
4.4.2. Analyse et décomposition de signaux sonores 11
4.5. Modélisation et décodage de parole 12
5. Logiciels **12**
5.1. Nouvelle version de la plate-forme ELISA 12
5.2. Plate-forme SIROCCO 12
5.3. LastWave 13
6. Résultats nouveaux **13**
6.1. Vérification du locuteur et traitement de la parole 13
6.1.1. Normalisation du rapport de vraisemblance 13
6.1.2. Estimation bayésienne adaptée au locuteur 13
6.1.3. Vérification du locuteur par arbres de décision 14
6.1.4. Incorporation de contraintes phonologiques dans la recherche en faisceaux 15
6.1.5. Détection de classes sonores superposées 15
6.1.6. Fusion d'information sonore et visuelle 15
6.2. Approximation de signaux sonores et séparation de sources 16
6.2.1. Matching Pursuit pour l'analyse de signaux sonores 16
6.2.2. Algorithmes itératifs pour les représentations parcimonieuses 16
6.2.3. Analyse granulaire 16
6.2.4. Approximation non-linéaire 17
6.2.5. Evaluation des algorithmes de séparation de sources sonores 18
6.2.6. Séparation de sources par modèles de signaux sonores monophoniques 18
6.2.7. Séparation de sources par analyse de scènes multicanal 18
7. Contrats industriels **19**

7.1. Conventions de Recherche 19
7.1.1. Contrat CP8 (n° 1 99 C 138 00 31321 01 2) 19
7.2. Actions financées par le RNRT 19
7.2.1. Projet Domus Videum (n° 2 02 C 0100 00 00 MPR 011) 19
7.3. Actions financées par la Commission Européenne 19
7.3.1. Projet BANCA (n° 1 01 C 0296 00 31331 00 5) 19
8. Actions régionales, nationales et internationales 19
8.1. Actions nationales 19
8.1.1. Action Jeunes Chercheurs du GDR ISIS 19
8.2. Actions européennes 20
8.2.1. Consortium ELISA 20
8.2.2. Réseau HASSIP 20
8.3. Participation à des colloques, séminaires, invitations 20
8.4. Participation à des réunions, constructions de groupes de travail 20
8.5. Enseignement 21
10. Bibliographie 21

1. Composition de l'équipe

METISS est un projet commun au CNRS, à l'INRIA, à l'Université de Rennes 1 et à l'INSA.

Responsable scientifique
Frédéric Bimbot [CR CNRS]

Assistante de projet
Marie-Noëlle Georgeault [TR INRIA (avec les projets S4, Sigma2 et Triskell)]

Personnel Inria
Rémi Gribonval [CR]

Personnel CNRS
Guillaume Gravier [CR]

Ingénieurs-Experts
Fabienne Porée [jusqu'au 31 août 2002]
Michaël Betser [depuis le 1er septembre 2002]

Chercheurs doctorants
Mathieu Ben [allocataire MENRT (et moniteur), 2e année]
Laurent Benaroya [allocataire MENRT, 4e année]
Raphaël Blouet [ATER, soutenance prévue le 16 décembre 2002]
Lorcan Mc Donagh [bourse INRIA, 3e année]

2. Présentation et objectifs généraux

Les axes de recherche du projet METISS sont consacrés au traitement de la parole et du signal sonore et comportent trois volets : la caractérisation du locuteur, la détection et le suivi d'information dans les flux audio et le traitement avancé du signal sonore (notamment la séparation de sources). Certains aspects de la reconnaissance de la parole (modélisation et décodage) viennent renforcer ces trois thèmes principaux.

Les principaux secteurs industriels concernés par les thématiques de METISS sont le secteur des télécommunications (notamment l'authentification vocale), celui de l'Internet et du multimédia (en particulier, l'indexation sonore), celui de la production musicale et audiovisuelle, et celui des logiciels éducatifs et des jeux.

Outre la diffusion de nos travaux au moyen de publications dans des conférences et des revues, notre démarche scientifique est accompagnée d'un souci permanent de mesurer nos progrès dans le cadre de campagnes d'évaluation, de diffuser les logiciels (et les ressources) que nous développons et de mutualiser nos efforts avec d'autres laboratoires partenaires.

METISS est, ou a été, récemment impliqué dans plusieurs partenariats bi-latéraux ou multi-latéraux, dans le cadre de groupes de travail (CIDRE), de consortiums et de réseaux de laboratoires (ELISA, HASSIP), d'actions de recherche (AUF-B1, SIROCCO, Action Jeune Chercheur du GDR ISIS), de projets nationaux (AGIR, Domus Videum) ou européens (PICASSO, DiVAN, BANCA) et de contrats industriels (CP8).

3. Fondements scientifiques

3.1. Introduction

Mots clés : *modélisation probabiliste*, *estimation statistique*, *théorie bayésienne de la décision*, *modèle de mélanges de gaussiennes*, *modèle de Markov caché*, *représentation adaptative*, *système redondant*, *décomposition parcimonieuse*, *séparation de sources*.

Les approches probabilistes offrent un cadre théorique général[36] qui a été à l'origine de progrès considérables dans différents domaines de la reconnaissance des formes, et notamment en traitement de la parole[22].

Le cadre probabiliste fournit en effet un formalisme solide qui permet de formuler différents problèmes de segmentation, de détection et de classification. Couplé à des approches statistiques, le paradigme probabiliste permet d'adapter facilement des outils relativement génériques à différents contextes applicatifs, grâce aux techniques d'estimation et d'apprentissage à partir d'exemples.

Les modèles probabilistes auxquels nous nous intéressons sont, pour l'essentiel, des modèles stochastiques de type Modèles de Markov Cachés (dans des formes parfois dégénérées). Le cadre stochastique permet de s'appuyer sur des algorithmes bien connus que ce soit pour l'estimation des paramètres de ces modèles (algorithmes EM, critères MV, MAP, ...) ou pour la recherche du meilleur modèle au sens du maximum de vraisemblance exact ou approché (décodage Viterbi ou recherche en faisceaux, par exemple).

En pratique, cependant, l'utilisation des outils théoriques doit s'accompagner d'un certain nombre d'ajustements pour tenir compte de problèmes survenant dans les contextes d'utilisation réels comme l'inexactitude des modèles, l'insuffisance (voire l'absence) de données d'apprentissage, leur mauvaise représentativité statistique, etc.

Un autre versant des activités de METISS est consacré aux représentations adaptatives de signaux dans des systèmes redondants[40]. L'utilisation de critères de parcimonie ou d'entropie (à la place du critère des moindres carrés) pour contraindre l'unicité de la solution d'un système d'équations sous-déterminé offre la possibilité de rechercher une représentation économique (exacte ou approchée) d'un signal dans un système générateur redondant, mieux à même de rendre compte de la diversité des structures présentes dans un signal. Il en résulte un vaste champ d'investigation scientifique : critères de parcimonie, algorithmes de recherche (poursuite) de la meilleure décomposition, construction du dictionnaire redondant, liens avec la théorie de l'approximation non-linéaire, extensions probabilistes, ...Les débouchés applicatifs potentiels sont nombreux.

Cette section expose brièvement ces différents éléments théoriques qui participent à nos activités.

3.2. Approches probabilistes

Mots clés : *densité de probabilité*, *modèle gaussien*, *modèle de mélange de gaussiennes*, *modèle de Markov caché*, *maximum de vraisemblance*, *maximum a posteriori*, *algorithme EM*, *algorithme de Viterbi*, *recherche en faisceaux*, *classification*, *test d'hypothèses*, *paramétrisation acoustique*.

Depuis près d'une vingtaine d'années, les approches probabilistes sont utilisées avec succès pour différentes tâches de reconnaissance des formes, et plus particulièrement en reconnaissance de la parole, qu'il s'agisse de la reconnaissance de mots isolés, de la retranscription de parole continue, de la vérification du locuteur ou de l'identification de la langue. Les modèles probabilistes permettent en effet de rendre compte efficacement des différents facteurs de variabilité présents dans le signal de parole, tout en se prêtant bien à la définition de mesures de ressemblance entre une observation et le modèle d'une classe de sons (phonème, mot, locuteur, etc.).

3.2.1. Formalisme et modélisation probabiliste

L'approche probabiliste pour la représentation d'une classe X repose sur l'hypothèse d'existence d'une fonction $P(.|X)$ permettant d'associer une densité de probabilité $P(Y|X)$ à toute observation Y.

En traitement de la parole, la classe X peut représenter un phonème, une suite de phonèmes, un mot du vocabulaire, ou bien un locuteur particulier, un type de locuteur, une langue, ...La classe X peut également correspondre à d'autres types d'objets sonores, par exemple une famille de sons (parole, musique, applaudissements), un événement sonore (bruit particulier, jingle), un segment sonore au voisinage d'instants spécifiques (de part et d'autre d'une hypothèse de rupture), etc.

Dans le cas des signaux sonores, les observations Y sont de type acoustique, par exemple des vecteurs issus de l'analyse du spectre à court terme du signal (coefficients de banc de filtres, coefficients cepstraux, composantes principales temps-fréquence, etc.) ou toute autre représentation permettant de rendre compte de l'information nécessaire à la bonne séparation des différentes classes considérées.

Dans la pratique, la fonction de densité de probabilité P n'est pas accessible à la mesure, et l'on a recours à une approximation $\hat{P}$ de cette fonction, que l'on désigne usuellement par fonction de vraisemblance. Celle-ci peut s'exprimer sous la forme d'un modèle paramétrique et les modèles les plus utilisés dans le domaine

du traitement de la parole (et du signal sonore) sont le modèle Gaussien (MG), le Modèle de Mélange de Gaussiennes (MMG) et le Modèle de Markov Caché (MMC).

Dans la suite de ce texte, nous désignerons par Λ l'ensemble des paramètres qui définissent le modèle considéré : une moyenne et une variance pour un MG, p moyennes, variances et poids pour un MMG à p Gaussiennes, q états, q^2 probabilités de transitions et $p \times q$ moyennes, variances et poids, pour un MMC à q états dont les fonctions d'émission sont des MMG à p Gaussiennes. On notera Λ_X le vecteur de paramètres pour la classe X, et l'on écrira, dans ce cas :

$$\hat{P}(Y|X) = P(Y|\Lambda_X).$$

Le choix du type de modèle repose généralement sur un ensemble de considérations faisant intervenir la structure pressentie des données (notamment l'existence ou non d'ordonnancement temporel), des connaissances permettant de fixer les paramètres structurels du modèle (nombre de gaussiennes p, nombre d'états q, etc.), la rapidité de calcul de la fonction de vraisemblance, le nombre de degrés de liberté du modèle par rapport au volume de données d'apprentissage disponibles, etc.

3.2.2. Estimation statistique

La détermination des paramètres du modèle pour une classe X donnée passe le plus souvent par une étape d'estimation statistique consistant à déterminer la valeur « optimale » du vecteur de paramètres Λ, c'est-à-dire celle qui maximise un critère de modélisation pour un ensemble d'apprentissage $\{Y\}_{app}$ constitué d'observations correspondant à la classe X.

Dans certains cas, on utilise le critère du Maximum de Vraisemblance (MV) :

$$\Lambda^*_{MV} = \arg\max_{\Lambda} P(\{Y\}_{app}|\Lambda)$$

Cette approche est généralement satisfaisante dès lors que le nombre de paramètres à estimer est petit devant le nombre d'observations d'apprentissage. Cependant, dans de nombreux contextes applicatifs, on fait appel à d'autres critères d'estimation, plus robustes devant la faible quantité de données d'apprentissage. Citons notamment le critère du Maximum a Posteriori (MAP) :

$$\Lambda^*_{MAP} = \arg\max_{\Lambda} \; P(\{Y\}_{app}|\Lambda) \; . \; p(\Lambda)$$

qui fait intervenir la probabilité a priori $p(\Lambda)$ du vecteur Λ, celle-ci traduisant d'éventuelles connaissances dont on dispose sur la distribution attendue des paramètres pour la classe considérée. Mentionnons également l'apprentissage discriminant comme alternative à ces deux critères, nettement plus complexe à mettre en oeuvre que les critères MV ou MAP.

Outre le fait que le critère MV n'est qu'un cas particulier du critère MAP (hypothèse d'uniformité de la probabilité a priori de Λ), le critère MAP s'avère expérimentalement mieux adapté aux faibles volumes de données et offre de meilleures capacités de généralisation des modèles estimés (ce qui se mesure par exemple par l'amélioration des performances en classification et en reconnaissance). De plus, le même schéma peut être utilisé pour procéder à l'adaptation incrémentale d'un modèle initial, c'est-à-dire au raffinement des paramètres du modèle à partir de nouvelles données observées ultérieurement (par exemple, en cours d'utilisation du système de reconnaissance). Dans ce cas, la valeur de $p(\Lambda)$ peut être obtenue à partir du modèle avant adaptation et la nouvelle estimation intègre les anciennes données par cet intermédiaire.

Quel que soit le critère considéré (MV ou MAP), l'estimation du vecteur de paramètres Λ s'effectue par l'intermédiaire de l'algorithme EM (Expectation-Maximization), qui fournit une solution correspondant à un des maxima locaux de la fonction de vraisemblance.

3.2.3. Algorithmes de calcul de vraisemblance et de décodage

En phase de reconnaissance, il est nécessaire d'évaluer la fonction de vraisemblance pour les différentes hypothèses de classes X_k. Quand la complexité du modèle est importante - le nombre de classes est élevé et les observations à reconnaître sont multi-dimensionnelles - il est généralement nécessaire de mettre en oeuvre des algorithmes de calcul rapide approché de la fonction de vraisemblance.

Par ailleurs, lorsque le modèle de la classe est un MMC, l'évaluation de la vraisemblance passe par le décodage (implicite ou explicite) de la séquence d'états cachés la plus probable, ce qui nécessite la mise en oeuvre de l'algorithme de Viterbi, outil désormais classique en reconnaissance de la parole.

Si, de plus, les observations sont constituées de segments appartenant à des classes différentes, chaînées par des probabilités de transition entre classes successives et sans que l'on ne connaisse a priori les frontières de segments (ce qui est le cas d'un énoncé en parole continue), il est nécessaire de faire appel à des techniques de recherche en faisceaux (beam-search) pour décoder la séquence d'états (quasi-)optimale à l'échelle de l'énoncé entier.

3.2.4. Décision Bayésienne

Dans les problèmes d'identification en ensemble fermé, où il s'agit d'effectuer la classification d'une observation dans une classe parmi plusieurs (K), le critère de décision usuel est le maximum a posteriori :

$$\hat{X}_k = \arg\max_{X_k} \; p(X_k) \; . \; \hat{P}(Y|X_k)$$

où $\{X_k\}_{1 \leq k \leq K}$ désigne l'ensemble des classes considérées.

Dans d'autres contextes (comme celui de la vérification du locuteur, de la détection de mot ou d'un type de son dans un enregistrement sonore), le problème de la classification se pose sous forme d'un test d'hypothèses binaire, consistant à décider si l'observation doit être considérée comme appartenant à la classe X (hypothèse notée X) ou comme n'y appartenant pas (c'est-à-dire appartenant à la « non-classe », hypothèse notée $\bar{X}$). Dans ce cas, la décision est du type acceptation ou rejet, respectivement notés $\hat{X}$ et $\hat{\bar{X}}$ dans la suite.

Ce second problème peut théoriquement se résoudre dans le cadre de la décision Bayésienne par le calcul du rapport S_X des densités de probabilité pour la classe et la non-classe, et la comparaison de ce rapport à un seuil de décision :

$$S_X(Y) = \frac{P(Y|X)}{P(Y|\bar{X})} \begin{cases} \geq R & \text{hypothèse } \hat{X} \\ < R & \text{hypothèse } \hat{\bar{X}} \end{cases}$$

où le seuil optimal R ne dépend pas de la distribution de la classe X, mais seulement des conditions de fonctionnement du système via le rapport des probabilités a priori des deux hypothèses et le rapport des coûts de fausse acceptation et de faux rejet.

En pratique, cependant, la théorie Bayésienne ne peut pas être appliquée telle quelle, car les quantités fournies par les modèles probabilistes ne sont pas les vraies fonctions de densité de probabilité, mais des valeurs de vraisemblance qui les approchent plus ou moins précisément, selon la qualité du modèle de la classe.

La règle de décision optimale se ré-écrit alors :

$$\hat{S}_X(Y) = \frac{\hat{P}(Y|X)}{\hat{P}(Y|\bar{X})} \begin{cases} \geq \Theta_X(R) & \text{hypothèse } \hat{X} \\ < \Theta_X(R) & \text{hypothèse } \hat{\bar{X}} \end{cases}$$

et le seuil optimal $\Theta_X(R)$ doit être ajusté pour la classe X, en étudiant le comportement du rapport de vraisemblance sur des données dites « de développement ».

Le problème de l'estimation du seuil optimal $\Theta_X(R)$ dans le cas du test de rapport de vraisemblance, peut se formuler de façon équivalente comme une normalisation du rapport de vraisemblance qui ramènerait le

seuil de décision optimal sur le seuil théorique. Plusieurs transformations ont été proposées (dans le cadre de la vérification du locuteur) : z-norm, t-norm, transformation affine, ...

3.3. Représentations adaptatives

Mots clés : *ondelette*, *dictionnaire*, *décomposition adaptative*, *optimisation*, *parcimonie*, *approximation non-linéaire*, *poursuite adaptative*, *algorithme glouton*, *complexité calculatoire*, *atome de Gabor*, *apprentissage à partir des données*, *analyse en composantes principales*, *analyse en composantes indépendantes*.

La famille des signaux sonores comprend une très grande diversité de structures temporelles et fréquentielles, de durées très variables, pouvant aller du régime stationnaire bien entretenu d'une note de violon jusqu'au bref transitoire d'une percussion. La structure du spectre peut être majoritairement harmonique (voyelles) ou nettement bruitée (consonnes fricatives). Plus généralement, la diversité des timbres sonores se traduit par une grande variété des structures fines du signal et de son spectre, ainsi que de son enveloppe temporelle et fréquentielle.

Par ailleurs, la plupart des signaux sonores rencontrés en pratique sont composites, c'est-à-dire qu'ils résultent du mélange de plusieurs sources (voix et musique, mixage de plusieurs pistes, signal utile et bruit de fond). De plus, ils peuvent avoir subi différentes distorsions, dues aussi bien aux conditions de prise de son qu'aux dégradations du support, aux effets du codage et de la transmission, etc.

Ces éléments structurels incitent à employer des techniques de décomposition de signaux sur des systèmes redondants (ou dictionnaires) d'atomes élémentaires correspondant aux différentes structures rencontrées, afin de mieux rendre compte de cette diversité.

3.3.1. Systèmes redondants et décomposition adaptative

Les méthodes classiques de décomposition de signaux s'appuient généralement sur la description du signal dans une base donnée (système libre, générateur et constant pour l'ensemble du signal), sur laquelle la représentation du signal est unique (par exemple, une base de Fourier, de Dirac, d'ondelettes orthogonales, ...). A l'inverse, les représentations adaptatives dans les systèmes redondants reposent sur la décomposition optimale du signal (au sens d'un critère à définir) dans un système générateur (ou dictionnaire) comprenant un nombre d'éléments (très) supérieur à la dimension du signal.

Soit y un signal mono-dimensionnel de longueur T et soit D un dictionnaire redondant composé de $N > T$ vecteurs g_i de dimension T.

$$y = [y(t)]_{1 \leq t \leq T} \qquad D = \{g_i\}_{1 \leq i \leq N} \quad \text{avec} \quad g_i = [g_i(t)]_{1 \leq t \leq T}$$

Si D est un système générateur de R^T, il existe une infinité de représentations exactes de y dans le système redondant D, du type :

$$y(t) = \sum_{1 \leq i \leq N} \alpha_i g_i(t)$$

On notera $\alpha = \{\alpha_i\}_{1 \leq i \leq N}$, les N coefficients de la décomposition.

Le principe de la décomposition adaptative consiste alors à sélectionner, parmi toutes les décompositions possibles, la « meilleure » d'entre elles, c'est-à-dire celle qui satisfait un certain critère (par exemple un critère d'économie de la représentation) pour le signal considéré, d'où le terme de décomposition (ou représentation) adaptative. Dans certains cas, au plus T coefficients seront non nuls dans la décomposition optimale, et l'ensemble des vecteurs de D ainsi sélectionnés sera désigné comme la base adaptée à y. Ce principe peut être étendu à des représentations approchées du type :

$$y(t) = \sum_{1 \leq i \leq M} \alpha_{\phi(i)} g_{\phi(i)}(t) + e(t)$$

avec $M < T$, où ϕ est une fonction injective de $[1, M]$ dans $[1, N]$ et où $e(t)$ correspond à l'erreur d'approximation à M termes de $y(t)$. Dans ce cas, le critère d'optimalité de la décomposition intègre également l'erreur d'approximation.

3.3.2. Critères de parcimonie

L'obtention d'une solution unique pour l'une des équations ci-dessus nécessite l'introduction d'une contrainte sur les coefficients α_i. Celle-ci s'exprime en général sous la forme :

$$\alpha^* = \arg\min_{\alpha} \ F(\alpha)$$

Parmi les fonctions les plus utilisées, citons les différentes fonctions L_γ :

$$L_\gamma(\alpha) = \left[\sum_{1 \leq i \leq N} |\alpha_i|^\gamma \right]^{1/\gamma}$$

Rappelons que pour $0 < \gamma < 1$, la fonction L_γ est une somme de fonctions concaves des coefficients α_i. Par ailleurs, on convient que L_0 correspond au nombre de coefficients non nuls dans la décomposition.

La minimisation de la norme quadratique L_2 des coefficients α_i (qui se résoud de façon exacte par une équation linéaire) a pour effet de disperser les coefficients sur l'ensemble des éléments du dictionnaire. Par contre, en minimisant L_0, on contraint la parcimonie de la représentation adaptative, au sens où la solution obtenue comporte un minimum de termes non nuls. Cependant la minimisation exacte de L_0, est un problème NP-complet.

Une approche intermédiaire consiste à minimiser la norme L_1, c'est-à-dire la somme des valeurs absolues des coefficients de la décomposition. Ceci peut-être réalisé par des techniques de programmation linéaire et on démontre que, sous certaines hypothèses (fortes) la solution trouvée converge vers le même résultat que celui correspondant à la minimisation de L_0. Dans la plupart des cas concrets, cette solution a de bonnes propriétés de parcimonie, sans pour autant égaler les propriétés que l'on obtiendrait avec L_0.

D'autres critères peuvent être pris en compte et dès lors que la fonction F est une somme de fonctions concaves des coefficients α_i, la solution obtenue possède encore de bonnes propriétés de parcimonie. A cet égard, l'entropie de la décomposition est une fonction particulièrement intéressante, compte tenu de ses liens avec la théorie de l'information.

Notons pour terminer, que la théorie de l'approximation non-linéaire est le cadre dans lequel on peut établir des liens entre la parcimonie des décompositions exactes et la qualité des représentations approchées à M termes. Ce type de caractérisation est encore un problème ouvert pour des dictionnaires redondants quelconques.

3.3.3. Algorithmes de décomposition

Trois grandes familles d'approches sont utilisées pour obtenir une décomposition (optimale ou sous-optimale) d'un signal dans un système redondant.

L'approche par « Meilleure Base » (*Best Basis*) consiste à considérer le dictionnaire D comme la réunion de B bases distinctes, puis à rechercher (exhaustivement ou non) parmi toutes ces bases celle qui donne lieu à la décomposition optimale (au sens du critère retenu). Pour des dictionnaires à structure arborescente (paquets d'ondelettes, cosinus locaux), la complexité de l'algorithme est bien inférieure au nombre de bases B, mais le résultat obtenu n'est en général pas optimal pour le dictionnaire D pris dans son ensemble.

L'approche par « Poursuite de Base » (*Basis Pursuit*) minimise la norme L_1 de la décomposition en faisant appel aux techniques de programmation linéaire. L'approche est de complexité importante, mais la solution obtenue possède en général de bonnes propriétés de parcimonie, sans néanmoins atteindre le résultat qui aurait été obtenu par minimisation de L_0.

L'approche « Poursuite Adaptative » (*Matching Pursuit*) consiste à optimiser de façon itérative la décomposition du signal, en recherchant à chaque étape l'élément du dictionnaire qui possède la meilleure corrélation avec le signal à décomposer, puis en soustrayant du signal la contribution de cet élément du dictionnaire. Cette procédure est réitérée sur le résidu ainsi obtenu, jusqu'à ce que le nombre de composantes (linéairement indépendantes) sélectionnées soit égal à la dimension du signal. On peut alors ré-estimer les coefficients α sur la base ainsi obtenue. Cet algorithme de type « glouton » (*greedy*) est sous-optimal mais il possède de bonnes propriétés de décroissance de l'erreur et de souplesse de mise en oeuvre.

Des approches intermédiaires peuvent également être considérées, sur la base d'algorithmes hybrides tentant de rechercher un compromis entre la complexité calculatoire, la qualité de la parcimonie et la facilité de mise en oeuvre.

3.3.4. Construction de dictionnaires

Le choix du dictionnaire D a naturellement une influence importante sur les propriétés de la décomposition obtenue : si le dictionnaire ne contient pas ou peu d'éléments adaptés à la structure du signal, les résultats seront peu satisfaisants car non exploitables.

Le choix du dictionnaire sur lequel s'opère la décomposition adaptative peut provenir de considérations a priori. En premier lieu, on peut viser la simplicité calculatoire, certains systèmes redondants nécessitant moins de calculs que d'autres pour évaluer les projections du signal sur les éléments du dictionnaire. A ce titre, les atomes de Gabor, les paquets d'ondelettes et les cosinus locaux possèdent d'intéressantes propriétés. En second lieu, on peut tenir compte de la structure pressentie des données : toute connaissance sur la répartition et la variation fréquentielle de l'énergie des signaux, sur la position et la durée typique des objets sonores qui le constituent, est de nature à guider le choix du dictionnaire (molécules harmoniques, chirplets, atomes dont la position est pré-déterminée, ...).

A l'inverse, dans d'autres contextes, il peut être souhaitable de construire le dictionnaire grâce à des techniques d'apprentissage à partir des données, celles-ci pouvant être soit les signaux que l'on désire décomposer, soit d'autres exemples de signaux appartenant à la même classe que le signal traité (par exemple, le même locuteur ou le même instrument de musique, ...). A cet égard, l'Analyse en Composantes Indépendantes (ACI) offre des possibilités intéressantes, mais d'autres approches peuvent être considérées (notamment l'optimisation directe de la parcimonie de la décomposition, ou des propriétés de l'erreur d'approximation à M termes) selon l'application visée.

Dans certains cas, l'apprentissage du dictionnaire peut nécessiter la mise en oeuvre d'algorithmes d'optimisation stochastique (de type recuit simulé), mais on peut s'intéresser également aux approches de type EM (Expectation-Maximization) dans les cas où il est possible de formuler la représentation redondante dans un cadre probabiliste.

L'extension des techniques de représentation adaptative peut également s'effectuer par généralisation de l'approche à des dictionnaires probabilistes, c'est-à-dire dont les vecteurs sont des variables aléatoires plutôt que des signaux déterministes. Dans ce cadre, le signal $y(t)$ se conçoit comme la combinaison linéaire d'observations émises par chacun des éléments du dictionnaire, ce qui permet de regrouper dans un même modèle plusieurs réalisations d'un même son (par exemple différentes formes d'onde pour un bruit, si elles sont équivalentes pour l'oreille). Les avancées dans cette direction sont subordonnées à la définition d'un modèle génératif réaliste pour les éléments du dictionnaire et de la mise au point de techniques d'estimation efficaces des coefficients.

3.3.5. Séparation de signaux

METISS s'intéresse à la séparation de sources dans le cas sous-déterminé, c'est-à-dire en présence d'un nombre de sources strictement supérieur au nombre de capteurs.

Dans le cas particulier de deux sources et d'un capteur, le signal mélange (mono-dimensionnel) s'écrit :

$$y = s_1 + s_2 + \epsilon$$

où s_1 et s_2 désignent les sources et ϵ un bruit additif.

Si l'on se place dans le cadre probabiliste et que l'on désigne par θ_1, θ_2 et η les (paramètres des) modèles des sources et du bruit, le problème de la séparation de sources revient à calculer :

$$(\hat{s}_1, \hat{s}_2) = \arg \max_{(s_1,s_2)} [P(s_1, s_2|y, \theta_1, \theta_2)]$$

En appliquant la règle de Bayes et en supposant l'indépendance statistique entre les deux sources, on démontre que l'on obtient le résultat recherché en résolvant :

$$(\hat{s}_1, \hat{s}_2) = \arg \max_{(s_1,s_2)} [P(y|s_1, s_2)P(s_1|\theta_1)P(s_2|\theta_2)]$$

Le premier des trois termes de l'argmax peut s'obtenir grâce au modèle du bruit en s'appuyant sur :

$$P(y|s_1, s_2) \propto P(y - (s_1 + s_2)|\eta) = P(\epsilon|\eta)$$

Les deux autres termes sont obtenus à partir de fonctions de vraisemblances correspondant à des modèles estimés des sources à partir d'exemples (ou de connaissances), par exemple, des modèles Laplaciens, des Mélanges de Gaussiennes ou des Modèles de Markov Cachés.

Ces modèles peuvent porter sur la distribution des coefficients de représentation dans un système redondant réunissant plusieurs bases adaptées aux sources présentes dans le mélange.

4. Domaines d'application

4.1. Introduction

Les principaux domaines d'application de METISS sont centrés autour de l'authentification du locuteur, la détection et le suivi d'information dans les flux audio et la séparation de sources. Certains aspects de la reconnaissance de la parole (modélisation et décodage) viennent accessoirement renforcer ces 3 thèmes principaux.

4.2. Vérification du locuteur

Mots clés : *vérification d'identité*, *sécurisation*, *personnalisation*.

Participants : Frédéric Bimbot, Fabienne Porée, Mathieu Ben, Raphaël Blouet.

Un message parlé ne véhicule par seulement le sens de ce que veut exprimer l'individu qui l'émet. Il porte également des informations sur l'individu lui-même et en premier lieu sur des éléments de son identité. L'étude de cette variabilité inter-individuelle de la voix est désignée par le terme de caractérisation du locuteur.

L'un des débouchés naturels des travaux en caractérisation du locuteur est celui de la reconnaissance automatique du locuteur, dans ses différentes variantes (identification, vérification, détection, suivi). L'essentiel des activités actuelles du groupe METISS en caractérisation du locuteur portent sur la vérification du locuteur.

La vérification (automatique) du locuteur est la tâche qui consiste à décider, à partir d'un enregistrement sonore, si celui-ci a été prononcé par un locuteur particulier (dit locuteur proclamé). Pour ce faire, on dispose d'un ou de plusieurs exemples de parole du locuteur proclamé, à partir desquels on a préalablement construit un modèle de sa voix. L'étape de vérification consiste alors à effectuer un test d'hypothèses à choix binaire, visant à déterminer si l'enregistrement de test est issu ou non du modèle du locuteur proclamé. Un débouché industriel de ces travaux est celui de l'authentification de l'utilisateur (ou du client) lors d'une transaction vocale (téléphonique ou sur l'Internet).

L'état-de-l'art dans le domaine repose sur l'utilisation de modèles probabilistes de la distribution du spectre à court terme du signal de parole (observations acoustiques vectorielles sous forme de coefficients cepstraux, par exemple) : Modèles de Markov Cachés (MMC), lorsque le contenu phonétique de l'énoncé est prédéterminé (mode *dépendant du texte*) ou Modèles de Mélanges de Gaussiennes (MMG) en mode dit *indépendant du texte*. La décision s'appuie alors sur le calcul d'un rapport de vraisemblance pour l'énoncé de test.

Comme nous l'avons évoqué plus haut, plusieurs difficultés nuisent à l'efficacité immédiate de cette approche, notamment :

- l'existence d'importants phénomènes de *variabilité intra-locuteur*, liés à l'imprécision motrice du locuteur, son état de santé, son état psychique, le style de parole qu'il utilise, son intention ou non d'être reconnu, etc.
- les problèmes de *robustesse* aux changements des conditions d'utilisation (notamment de prise de son) ;
- la mauvaise qualité de l'*estimation* du modèle du locuteur, en raison de l'insuffisance et de la faible représentativité des données d'apprentissage imposées, dans le cadre applicatif, par des considérations ergonomiques ;
- la déviation du seuil de décision optimal par rapport à sa valeur théorique en raison des imprécisions des estimateurs de probabilité, qu'il faut compenser par un *ajustement* du rapport de vraisemblance.

Notons qu'en raison de ces nombreux facteurs de variabilité, les meilleurs systèmes de reconnaissance du locuteur fournissent actuellement des performances rarement en-dessous que quelques pour cent de taux d'erreur, ce qui a des implications sur le profil des applications dans lesquels ils s'intègrent.

Le groupe METISS s'intéresse également à la vérification du locuteur indépendamment du texte, que ce soit à travers le téléphone (consortium ELISA pour les évaluations NIST) ou directement à partir d'un terminal dédié (convention de recherche avec Bull). Dans ce contexte, nos efforts portent sur l'amélioration de l'estimation du modèle du locuteur en utilisant des techniques d'adaptation d'un modèle indépendant du locuteur (adaptation Bayésienne, approche MAP, etc.), ainsi que sur la mise en œuvre d'algorithmes distribués sur des architectures disposant de faibles ressources de calcul et de mémoire.

4.3. Détection et suivi d'information dans les flux sonores

L'accroissement constant de la masse de documents sonores (enregistrements radiophoniques, bandes sonores de programmes télévisés, messages parlés, etc.) rend indispensable le développement d'outils automatiques de repérage et de navigation dans ces enregistrements. La définition de descripteurs sonores et leur extraction automatique a pour objectif de donner une représentation plus structurée du matériau audio, pour en faciliter l'accès par le contenu ou selon des critères de similarité.

4.3.1. Détection de locuteur

Mots clés : *flux sonore*, *détection*, *suivi*, *classe sonore*.

Participants : Mathieu Ben, Guillaume Gravier, Frédéric Bimbot.

Les caractéristiques d'un locuteur (genre, tranche d'âge, accent, identité, ...) constituent des descripteurs de première importance pour l'indexation d'enregistrements de parole, ainsi que toute information indiquant la présence d'un locuteur particulier dans un document sonore, les changements de locuteur, la présence de plusieurs locuteurs simultanés, etc. Plus précisément, on peut identifier au moins trois tâches d'intérêt :

- la détection de présence d'un locuteur dans un enregistrement sonore (classification) ;
- la localisation d'un locuteur dans un enregistrement sonore (marquage temporel) ;
- la segmentation en locuteurs d'un enregistrement sonore (détection de changements).

Ces problématiques possèdent naturellement de nombreux points communs avec la vérification du locuteur, avec laquelle elles partagent des aspects théoriques et pratiques - notamment l'utilisation d'un test statistique, que ce soit à partir d'un modèle de locuteur connu au préalable (détection de présence et localisation d'un locuteur), ou de modèles estimés au vol, à partir de l'enregistrement proprement dit (segmentation en locuteurs). Néanmoins, les particularités de la tâche nécessitent la mise en oeuvre de solutions pour neutraliser les facteurs de variabilité spécifiques au problème traité.

4.3.2. Détection et suivi de classes de son

Mots clés : *flux sonore*, *détection*, *suivi*, *classe sonore*, *indexation audio*.

Participants : Guillaume Gravier, Michaël Betser, Frédéric Bimbot, Rémi Gribonval.

Dans le cadre de l'annotation automatique de bandes sonores (programmes de radio et de télévision, archives audiovisuelles, etc.), il est utile de repérer différents sons ou classes de sons comme le silence, la parole, la musique, les applaudissements, ou encore certains évenements caractéristiques (jinggle, bruit de balle, etc). Ces différents éléments sont en effet des points de repère essentiels dans une émission ou une série d'émissions et leur localisation automatique permet de focaliser immédiatement les recherches manuelles ou automatiques sur les plages sonores d'intérêt.

L'approche par rapport de vraisemblance classiquement utilisée en reconnaissance du locuteur se généralise immédiatement à ces types de problèmes, moyennant l'apprentissage préalable d'un modèle statistique de la classe de signaux à détecter et/ou à localiser. Ainsi, on détectera des plages de parole en mettant en compétition un modèle (probabiliste) de parole et un modèle de non-parole, de même pour la musique et les autres classes de son. Ces modèles sont typiquement des modèles de distribution de la densité spectrale de puissance, plus ou moins contraints dans leur structure temporelle (selon qu'il s'agisse de classes générales, comme parole, musique, etc. ou d'événements particuliers, comme un jingle). Une autre approche possible consiste à utiliser des modèles de Markov cachés où chaque état représente l'une des classes de sons à détecter, permettant ainsi une détection conjointe à la segmentation.

Dans ce domaine, les principaux problèmes à résoudre sont la détection de classes superposées dans les scènes sonores complexes (par exemple les publicités lors desquelles différentes classes de signaux sont couramment mélangées), la détection d'évenements caractéristiques souvent courts et pour lesquels on ne dispose que d'un faible volume de données d'apprentissage, et l'apprentissage automatique de modèles à partir de peu de données afin de limiter l'étape d'annotation manuelle de corpus.

4.3.3. Intégration audiovisuelle pour l'indexation

Mots clés : *indexation*, *multimédia*, *fusion d'informations*.

Participants : Guillaume Gravier, Frédéric Bimbot.

Dans le cadre de l'indexation de documents audiovisuels, l'indexation de la bande audio en classes de sons telle qu'exposée au paragraphe précédent permet d'extraire une information partielle sur la structure du document audiovisuel. Cette information partielle doit ensuite être combinée à l'information issue du traitement de la vidéo afin de permettre une structuration du document servant de point de repère à la recherche dans les documents, ou encore à la génération d'un résumé automatique.

La combinaison des informations sonores et visuelles est possible à plusieurs niveaux. Les approches émergentes se basent, la plupart du temps, sur une détection indépendante des événements sonores et visuels, les deux flux d'informations étant ensuite combinés pour structurer le document. La théorie bayésienne de la décision fournit un cadre théorique puissant pour la combinaison d'informations dès lors que l'extraction de l'information se base sur une approche statistique. Il est également possible d'utiliser le résultat de la détection d´événements dans l'une des modalités pour aider l'indexation de la seconde modalité. Dans ce cas, les modèles de Markov cachés multi-flux, permettant l'intégration de flux de données multiples, offrent un grand potentiel.

Les thèmes de recherche dans ce domaine sont la définition de modèles de combinaison d'informations adaptés aux problèmes traités et la définition d'approches permettant une fusion des informations le plus tôt possible afin d'éviter l'extraction d'informations indépendantes pour chacune des modalités.

4.4. Traitement avancé de signaux sonores

Mots clés : *séparation de sources*, *événements sonores*, *indexation*, *son multicanal*, *modèles granulaires*.

Dans de nombreux contextes applicatifs, le signal de parole est présent à côté d'autres signaux sonores ou mélangés avec eux, notamment des signaux musicaux et des bruits. De plus, les signaux traités sont bien souvent composites, c'est-à-dire qu'ils résultent de la superposition de plusieurs sources via le mixage de plusieurs pistes (ou voies). Ils sont également soumis à toutes sortes de distorsions, qu'elles soient dues aux conditions de prise de son ou au canal de transmission.

Les progrès récents dans le domaine des technologies vocales (reconnaissance de la parole et du locuteur) incitent à étudier l'utilisation et l'adaptation de ces techniques à des classes plus larges de signaux, notamment les signaux musicaux.

Ainsi, nous nous intéressons aux thèmes de la séparation de sources et de la représentation de signaux sonores.

4.4.1. Séparation de sources sonores

Participants : Laurent Benaroya, Rémi Gribonval, Frédéric Bimbot.

En toute généralité, le problème de la séparation de sources consiste à décomposer un signal sous forme d'une somme de deux termes ou plus. Dans le cas de la *séparation de locuteurs*, le problème consiste à séparer deux signaux de parole superposés prononcés par des locuteurs distincts. Cette problématique peut s'étendre à la *séparation de voix*, consistant à isoler les différentes contributions simultanées dans un enregistrement sonore (parole, musique, chant, instruments, etc.). Dans le cas du *débruitage*, il s'agit de séparer le signal « utile » (c'est-à-dire portant l'information) du bruit perturbateur. Il est même judicieux de considérer la compression sonore comme un cas particulier de séparation de sources, l'un des signaux étant le signal comprimé, l'autre le résidu de compression. Ainsi, le problème de la séparation de sources recouvre en fait une grande diversité de problématiques et de débouchés.

Alors que dans certains contextes, comme celui de la prise de son, le problème de la séparation de sources peut se poser sous l'hypothèse d'un nombre de capteurs supérieur ou égal au nombre de sources, les travaux du projet METISS se placent dans le cas sous-déterminé, et plus précisément dans le cas d'un seul capteur (enregistrement mono) pour 2 sources, ou dans celui de 2 capteurs (stéréophonie) pour $n > 2$ sources.

4.4.2. Analyse et décomposition de signaux sonores

Participants : Lorcan Mc Donagh, Rémi Gribonval, Frédéric Bimbot.

Les normes de la famille MPEG (et notamment MPEG-4) définissent des formats de description et de transmission de signaux sonores sous forme d'une « partition » (description de haut-niveau de type MIDI) et d'« instruments » (décrivant des textures sonores). Ces formats promettent des codages à très bas débit et des facilités d'indexation et de navigation. Cependant les méthodes pour transformer un enregistrement sonore existant en une représentation de ce type restent à mettre au point.

Les techniques de décomposition de signaux par addition d'atomes élémentaires (parfois désignées par méthodes *granulaires*), qui font l'objet d'un intérêt croissant pour la synthèse sonore, peuvent être vues comme une première étape où les instruments sont les éléments du dictionnaire. Dans le modèle classique, les « grains sonores » sont des fonctions déterministes (sinusoïdes modulées, chirps, molécules harmoniques, voire formes d'ondes prétabulées, etc.). Le signal reconstruit $y(t)$ apparaît alors comme l'approximation adaptative à M termes du signal original dans un dictionnaire D. La théorie de l'approximation non-linéaire et les méthodes de décomposition de type Matching Pursuit fournissent un cadre et des outils puissants pour aborder ce type de problème.

Les techniques *granulaires* décomposent les signaux sonores en un grand nombre de signaux « élémentaires » de courte durée. Les méthodes d'analyse inspirées de l'idée d'*atome sonore* de Gabor utilisent des signaux de type cosinus locaux, modulés ou non en fréquence. Les techniques de synthèse granulaire permettent quant à elles de créer des textures sonores d'une grande complexité, mais posent le problème du contrôle par l'utilisateur du résultat sonore. Nous travaillons sur une méthode d'analyse

adaptative qui utilise des signaux non-déterministes (*prototypes* ou modèles), équivalents aux vecteurs de base utilisés en analyse fonctionelle. Ces prototypes sont obtenus à partir du signal, il peuvent ensuite être utilisés pour reconstruire partiellement le signal d'origine (compression), segmenter les notes d'une mélodie, re-synthétiser le son avec des paramètres facilement accessibles à l'utilisateur.

4.5. Modélisation et décodage de parole

Mots clés : *modèles de Markov cachés*, *algorithme de Viterbi*, *recherche en faisceaux*, *beam-search*, *reconnaissance de parole*.

Participants : Guillaume Gravier, Frédéric Bimbot.

Le projet METISS consacre une partie de ses efforts à des sujets tels que la modélisation et le décodage acoustique pour la reconnaissance automatique de la parole. En effet, ces thématiques apportent des compléments indispensables pour améliorer l'impact des applications dans certains domaines et pour valider nos approches au sein d'un système complet. En particulier, cette activité est fortement complémentaire de nos activités en indexation sonore et en sécurisation de transactions. Dans le premier cas, les algorithmes d'indexation sonore sont souvent utilisés en amont d'un système de reconnaissance de la parole afin de détecter les plages de parole à retranscrir ainsi que les changements de locuteurs, ces derniers étant utilisés d'une part pour l'adaptation au locuteur du système de transcription et d'autre part pour l'enrichissement de la transcription. Dans le cas de la sécurisation de transaction, la reconnaissance de la parole peut permettre de vérifier le contenu linguistique du message prononcé (reconnaissance d'un mot de passe, caractérisation du contenu linguistique pour aider à la vérification, etc).

5. Logiciels

5.1. Nouvelle version de la plate-forme ELISA

Participants : Mathieu Ben, Raphaël Blouet, Frédéric Bimbot.

Une nouvelle version de la plate-forme ELISA, pour la participation à la campagne d'évaluation NIST 2002, a été mise en place, sur la base des modules développés à l'IRISA et dans les laboratoires partenaires du Consortium.

Il s'agit d'un système modulaire de vérification du locuteur utilisable à des fins expérimentales. Il est développé en langage C. L'IRISA et le LIA ont été les principaux contributeurs à la version 2002.

L'approche utilisée est la modélisation probabiliste par mélange de gaussiennes (modèle GMM) avec apprentissage par critère MAP (Maximum A Posteriori) et normalisation du rapport de vraisemblance.

Les efforts de l'IRISA cette année ont principalement porté sur la mise en place de nouvelles normalisations des scores de vérification sur la plateforme ELISA : la d-norm [13] et la dt-norm. L'implémentation de ces normalisations a été faite en collaboration avec le LIA et la participation de l'IRISA aux évaluations NIST 2002, dans le cadre du consortium ELISA, a permis de valider ces approches.

5.2. Plate-forme SIROCCO

Participants : Guillaume Gravier, Frédéric Bimbot.

Suite à l'Action de Recherche Concertée SIROCCO de l'INRIA, qui a permis le développement et la distribution sous license libre (GNU General Public License) d'une plate-forme de reconnaissance de la parole grand-vocabulaire, METISS poursuit une activité de maintenance et de valorisation de la plate-forme Sirocco. Cette dernière est maintenant diffusée en dehors des partenaires du projet et a été présentée à la communauté francophone lors des Journées d'Etudes sur la Parole 2002 [17].

La publication des résultats de nos travaux sur l'introduction de contraintes phonologiques dans l'algorithme de décodage de la plate-forme Sirocco (cf. résultats ci-après) sous la forme d'un chapitre de livre [4] a également permis de présenter la plate-forme Sirocco à la communauté internationale.

http://www.irisa.fr/sirocco

5.3. LastWave

Participants : Rémi Gribonval, Lorcan McDonagh.

METISS contribue de façon régulière au développement du logiciel de traitement du signal *LastWave*, dont le noyau est développé par Emmanuel Bacry, du Centre de Mathématiques Appliquées de l'Ecole Polytechnique. *LastWave* est diffusé sous licence libre (GNU General Public License), fonctionne sous MacOS et Unix, et compte près de 300 utilisateurs enregistrés.

LastWave est un programme modulaire de traitement du signal orienté objet, et METISS contribue principalement au développement, à la maintenance et à la valorisation des modules de *Matching Pursuit* et d'Analyse de Fourier à Court Terme. Ces modules ont également été repris (indépendamment de *LastWave*) dans le logiciel *Guimauve* de Fabien Brachere, du Laboratoire d'astrophysique/Observatoire Midi-Pyrénées de Toulouse. Les efforts de METISS cette année ont notamment consisté à développer et porter le noyau de LastWave sur différentes plate-formes grâce à Java, ainsi qu'à implémenter et optimiser les nouvelles méthodes de décomposition de signaux sonores multicanal connues sous le nom de *Fast Matching Pursuit* avec des *chirplets* et des *molécules harmoniques*.
http://webast.ast.obs-mip.fr/people/fbracher/
http://software.linux.com/projects/guimauve/.

6. Résultats nouveaux

6.1. Vérification du locuteur et traitement de la parole

Mots clés : *vérification du locuteur*, *normalisation de test statistique*, *distance de Kullback-Leibler*, *méthode de Monte-Carlo*, *arbres de décision*, *recherche en faisceaux*, *contraintes phonologiques*.

6.1.1. Normalisation du rapport de vraisemblance

Participants : Mathieu Ben, Raphaël Blouet, Frédéric Bimbot.

Les travaux menés l'an dernier sur la d-norm [13] ont été complétés cette année par de nouvelles expériences effectuées lors des évaluations NIST 2002 [47]. Cette nouvelle technique de normalisation des scores en vérification automatique du locuteur diffère des approches précédentes (z-norm, t-norm) [19] par le fait qu'elle ne nécessite pas de données de parole supplémentaires ni de populations de locuteurs externes, et peut donc toujours être appliquée, même lorsque l'on ne dispose pas de données additionnelles. Les expériences menées précédemment ont montré que les performances de la d-norm sont comparables à celles de la z-norm et que la première peut ainsi remplacer avantageusement la seconde en allégeant considérablement la procédure de normalisation.

Lors des évaluations NIST 2002 en vérification du locuteur, les performances de la d-norm ont été comparées à celles de la t-norm et à celles de la dt-norm qui est une association des deux normalisations précédentes. Les résultats ont montré que la d-norm apporte une amélioration significative des performances pour les points de fonctionnement du système à faible taux de faux rejet et n'apporte pas d'amélioration pour les points de fonctionnement à faible taux de fausse acceptation. A contrario, la t-norm n'apporte pas d'amélioration pour les points de fonctionnement à faible taux de faux rejet mais apporte une amélioration significative pour les points de fonctionnement à faible taux de fausse acceptation. En termes de taux d'erreur EER (*Equal Error Rate*), ces deux normalisations apportent une amélioration significative et donnent des résultats comparables. Les expériences menées avec la dt-norm montrent que cette normalisation, qui associe la d-norm et la t-norm, conserve les propriétés avantageuses de ces deux normalisations, en apportant une amélioration significative pour tous les points de fonctionnement du système.

6.1.2. Estimation bayésienne adaptée au locuteur

Participants : Mathieu Ben, Frédéric Bimbot.

Les recherches menées sur la d-norm et les distances de Kullback ont montré que les scores imposteurs renvoyés par un modèle de locuteur sont fortement corrélés à la distance de Kullback entre ce modèle locuteur et le modèle non-locuteur. A partir de ces constatations, nous avons mis au point une procédure de normalisation au niveau même de la modélisation, afin d'amener chacun des modèles locuteurs d'une base de données à une même distance du modèle non-locuteur.

Lors de l'estimation bayésienne d'un modèle locuteur avec le critère du Maximun a Posteriori (MAP) [44], le choix du coefficient de pondération α entre les données d'apprentissage du locuteur considéré et les données a priori (données du monde) détermine également la distance entre le modèle locuteur estimé et le modèle non-locuteur, celui-ci servant de modèle a priori pour l'estimation. Ainsi, plus la part des données du locuteur est importante par rapport aux données a priori, plus la distance entre le modèle locuteur et le modèle non-locuteur est grande. L'estimation des modèles se fait en utilisant un algorithme de type « Expectation-Maximization » avec un critère MAP contraint par une distance de référence. A partir de cette distance de référence, nous utilisons une procédure itérative afin de déterminer pour chacun des locuteurs, le coefficient α adéquat qui aménera chaque modèle locuteur à la distance de référence du modèle non-locuteur. Cette procédure permet donc d'obtenir un coefficient α *adapté au locuteur* contrairement au système de base qui utilise un coefficient α constant quel que soit le locuteur.

Des tests menés sur ces modèles ont montré qu'ils réagissent tous de façon similaire à un énoncé imposteur, en fournissant, en moyenne, des scores imposteurs de valeurs comparables. Il en résulte que la distribution des scores imposteurs lors d'une évaluation complète est plus concentrée que celle obtenue avec le système de base. La séparation des scores imposteurs et des scores clients s'avère alors plus facile et les performances obtenues sont améliorées. D'autre part, des expériences plus complètes ont montré que cette procédure de sélection individuelle du coefficient α pour chaque locuteur donne de meilleures performances qu'une procédure d'optimisation globale de ce coefficient (c.a.d à α constant) sur toute la base de locuteurs. Enfin, nous avons observé que les résultats obtenus avec ces modèles normalisés, sans aucune normalisation des scores, sont identiques à ceux que l'on obtient avec le système de base suivi d'une normalisation des scores de type d-norm ou z-norm. La normalisation des modèles permet donc un léger gain de temps et de place mémoire lors des tests car elle dispense d'appliquer aux scores de vérification une normalisation comme la z-norm ou la d-norm, et de stocker les paramètres correspondants. Cependant, du fait de la procédure itérative lors de l'apprentissage des modèles, l'estimation des modèles normalisés est moins rapide que dans le cas classique.

6.1.3. Vérification du locuteur par arbres de décision

Participants : Frédéric Bimbot, Raphaël Blouet.

METISS s'intéresse aux problèmes relatifs à la vérification du locuteur sous de fortes contraintes de calcul et de mémoire, dans des architectures distribuées. Ces travaux consistent à étudier les possibilités d'utilisation directe des cartes à puce pour stocker la référence caractéristique du locuteur et pour procéder à la vérification de son identité (collaboration avec CP8).

L'approche adoptée repose sur une technique originale qui comporte deux étapes. La première vise à obtenir une partition Q de l'espace des paramètres en régions indexées par un arbre de décision. La seconde consiste à affecter un score de décision à chacune des régions de Q.

Q est obtenue par utilisation de l'algorithme CART (Classification And Regression Tree)[23]. Cet algorithme permet d'obtenir incrémentalement un partage binaire récursif de l'espace des paramètres acoustiques (partitionnement). Durant la phase d'apprentissage, le score de décision est obtenu par estimation directe d'un rapport des densités de probabilités (estimées) entre l'hypothèse du locuteur et celle du non-locuteur dans chacune des partitions de l'espace des paramètres issues du partitionnement. Durant la phase de test, la structure en arbre permet d'accéder directement et efficacement à la partition, et donc au score de décision associé au vecteur acoustique[39][21].

La première mise en oeuvre de cette technique a été validée sur un sous-ensemble du corpus de l'évaluation NIST'01 [41] des systèmes de vérification du locuteur et est décrite dans [39]. Une deuxième série de travaux, actuellement en cours, consiste à consolider l'approche et à augmenter sa robustesse, d'une part en

améliorant l'estimation du score de décision en chacune des régions, d'autre part en renforçant la diversité et la représentativité des observations. L'utilisation du *boosting* [27] pour apprendre plusieurs arbres par locuteur, a permis d'améliorer notablement les performances obtenues.

Un brevet a été déposé couvrant l'ensemble des travaux décrits ci-dessus [3].

6.1.4. Incorporation de contraintes phonologiques dans la recherche en faisceaux

Participant : Guillaume Gravier.

Dans le cadre de l'ARC SIROCCO, l'équipe METISS de l'IRISA s'est focalisée sur le moteur de décodage, basé sur la technique dite de recherche en faisceaux [43] ou *beam-search*.

Nous nous sommes intéressés à l'introduction de contraintes phonologiques contextuelles dans l'algorithme classique de recherche. De telles contraintes permettent de rendre compte de l'influence que peut avoir la prononciation d'un mot sur la prononciation des mots voisins. Le phénomène de liaison en français, où la première lettre d'un mot conditionne la prononciation du dernier phonème du mot précédent, est un exemple typique de contrainte contextuelle.

De telles règles induisent des contraintes sur les séquences de prononciation qui viennent se superposer à celles induites pas le modèle de langage. Il convient cependant de gérer l'ensemble en demeurant dans un cadre théorique bien maîtrisé (notamment le cadre probabiliste).

Des contraintes correspondant à plusieurs phénomènes phonétiques (liaisons, élisions des e-muets, élisions des liquides finales, etc.) ont été étudiées. Les résultats obtenus mettent en évidence la cohérence de l'approche proposée mais soulignent également l'insuffisance des ressources utilisés (modèles de phonèmes simples, lexique de prononciation insuffisant, ...) et la limite des approximations faites dans le cadre de l'approche retenue [4].

6.1.5. Détection de classes sonores superposées

Participants : Michaël Betser, Guillaume Gravier, Rémi Gribonval.

Un des problèmes rencontrés dans le cadre de la détection d'événements sonores dans des documents audiovisuels est la détection de classes sonores simultanées dans des scènes auditives complexes.

Un système classique basé sur une modélisation par un MMC ergodique de la structure du document avec un état par classe de sons, ne permet pas, par nature, de détecter des événements simultanés. Par ailleurs, l'estimation d'un modèle pour chaque combinaison de classes sonores est exclu du fait du manque de données d'apprentissage. Dans un premier temps, nous avons donc proposé et comparé plusieurs méthodes de combinaison de modèles (concaténation et convolution) permettant d'obtenir des modèles de classe multiples à partir des modèles de chacune des classes isolées. Plusieurs variantes de l'approche par combinaison de modèles ont été comparées à une approche par décodage des N meilleurs chemins, ainsi qu'à une approche par tests binaires multiples. Cette dernière approche, basée sur les techniques de vérification du locuteur, consiste à tester la présence ou non d'une classe sonore indépendamment pour chaque classe. Les résultats obtenus mettent en évidence la validité des approches par combinaison de modèles et montrent l'inadéquation du décodage des N meilleurs chemins pour résoudre ce problème.

L'étude des mélanges de classe mérite d'être appronfondie car elle permet de traiter et de décrire des documents sonores complexes à partir d'une bibliothèque de modèles de taille restreinte. Ces résultats vont donner lieu à une publication prochaine.

6.1.6. Fusion d'information sonore et visuelle

Participants : Guillaume Gravier, Frédéric Bimbot.

Dans le contexte de l'indexation de documents audio-visuels, la collaboration entre les modalités audio et video peut se faire principalement à trois niveaux : au niveau des descripteurs (par exemple, descripteurs de mouvements pour la vidéo et coefficients cepstraux pour le signal), au niveau de la recherche de segments homogènes ou encore au moment de la classification de ces derniers.

Après une première étude montrant que, dans le type de document traité, aucune corrélation n'a pu être établie ni au niveau des descripteurs, ni au niveau des frontières de segments, nous avons proposé une approche

probabiliste permettant l'intégration des informations sonores lors de la classification des segments vidéo. Le modèle proposé se base sur l'utilisation des fréquences de co-occurrences des événements sonores et vidéo au sein d'un critère de maximum a posteriori ou encore intégré dans un modèle de Markov caché.

Ce travail, conduit en collaboration avec les équipes VISTA et TEX-MEX et Thomson Multimédia dans le cadre du projet RNRT Domus Videum, vise à proposer et tester des modèles de coopération entre modalités pour la génération automatique de résumé de documents sportifs audiovisuels.

6.2. Approximation de signaux sonores et séparation de sources

6.2.1. Matching Pursuit pour l'analyse de signaux sonores

Mots clés : *analyse de signaux musicaux*, *modèle de structures harmoniques*, *Matching Pursuit*.

Participants : Rémi Gribonval, Lorcan Mc Donagh.

Des travaux des années antérieures [29][30][31] ont montré l'intérêt et la flexibilité des méthodes de décomposition de signaux de type Matching Pursuit pour des applications sonores. En collaboration avec Emmanuel Bacry, du Centre de Mathématiques Appliquées de l'Ecole Polytechnique, nous avons mis au point un algorithme de « Matching Pursuit Harmonique » de faible complexité algorithmique. Nous avons pu montrer par des expériences que les décompositions obtenues permettent d'effectuer la détection de notes sur des enregistrements non polyphoniques mais pouvant être très réverbérés [28].

En s'inspirant de travaux similaires portant sur des décompositions en *chirplets*[42], nous avons mis en place une modélisation probabiliste des structures harmoniques des signaux sonores pour effectuer la décomposition selon le critère du maximum de vraisemblance plutôt que du maximum d'énergie. L'algorithme issu de ce modèle probabiliste a été incorporé dans la fonction de Matching Pursuit harmonique du module Matching Pursuit du logiciel LastWave.

6.2.2. Algorithmes itératifs pour les représentations parcimonieuses

Mots clés : *représentation parcimonieuse*, *méthode de Newton*, *Lagrangien*, *Lagrangien augmenté*, *méthode itérative*, *affine scaling*.

Participants : Laurent Benaroya, Frédéric Bimbot, Rémi Gribonval.

Nous avons cherché des méthodes itératives d'optimisation du problème pénalisé $\min_\omega \|Aw - b\|_2^2 + \gamma \sum_i f(\omega_i)$. Dans le cas où $f(x)$ est concave sur $[0, +\infty)$ et $(-\infty, 0]$, on obtient des solutions dites parcimonieuses.

L'idée première, dans la lignée de l'algorithme FOCUSS, a été de transformer ce problème en un problème de moindre carrés itérés pondérés. Cette méthode présente des limitations car il y a en général une matrice de grande taille à inverser à chaque étape. Cela nous a conduit a utiliser des algorithmes de type « scaled gradient » :

$$\omega^{(k+1)} = \omega^{(k)} - \nu_k \left[f^{(k)} \cdot A^T (A\omega^{(k)} - b) + e^{(k)} \right]$$

où $f_i^{(k)}$ est une fonction sigmoïdale de $\omega_i^{(k)}$ et $e^{(k)}$ est un terme de correction. Les propriétés théoriques de convergence et de vitesse de convergence de ce type d'algorithmes sont à l'étude actuellement.

Par ailleurs, nous nous intéressons au problème pénalisé sous contraintes, notamment de positivité des $\omega_i^{(k)}$ et/ou de nombre exact de composantes $\omega_i^{(k)}$ non nulles (problème de l'approximation non linéaire). Les solutions proposées consistent à modifier les poids $f_i^{(k)}$, en fonction de paramètres de Lagrange $\lambda_i^{(k)}$, ces paramètres étant eux-mêmes estimés de manière itérative (méthode de Harrow-Hurwicz, Lagrangien augmenté). Les premières expérimentations menées sur ces méthodes pénalisées sont prometteuses.

6.2.3. Analyse granulaire

Mots clés : *analyse de signaux musicaux*, *algorithme EM*, *clustering*.

Participants : Lorcan Mc Donagh, Frédéric Bimbot, Rémi Gribonval.

Nous avons élaboré une technique d'analyse granulaire qui permet la décomposition d'un signal sur une famille adaptative de signaux stochastiques, dits *prototypes*, extraits du signal. Il s'agit d'une extension des atomes de Gabor au cas où les atomes sont des sous-signaux à support temporel restreint (appelés *grains*), extraits du signal analysé, et munis d'une structure probabiliste. Nous avons mis en place un cadre théorique ainsi qu'un ensemble d'algorithmes de clustering (notamment à base d'algorithme EM) que nous avons validé expérimentalement. La méthode a été appliquée sur des signaux réels pour des tâches comme la compression de données audio et la segmentation de notes.

Nous recherchons actuellement à améliorer la qualité du signal compressé, ce qui nous conduit à mettre en oeuvre une procédure d'optimisation globale du dictionnaire des *prototypes*. Nous travaillons également à intégrer l'algorithme dans la phase d'initialisation d'une méthode de séparation de sources monocapteur.

6.2.4. Approximation non-linéaire

Mots clés : *espace d'approximation*, *espace de Besov*, *espace de modulation*, *décomposition parcimonieuse*, *ondelette*, *analyse multirésolution*, *spline*, *framelets*.

Participant : Rémi Gribonval.

Ce travail est conduit en coopération avec Morten Nielsen de l'Université d'Aalborg au Danemark et porte sur des questions théoriques soulevées par l'utilisation de dictionnaires redondants pour l'approximation de signaux sonores ou plus généralement d'images.

Le problème traité est celui de la caractérisation des espaces d'approximation dans un espace de Banach X à partir d'éléments d'un système générateur $\mathcal{D}$ de vecteurs unitaires appelé dictionnaire. La théorie abstraite de l'approximation[26] ramène cette étude au calcul d'espaces d'interpolation entre X et certains sous-espaces, dès lors que l'on peut montrer les inégalités de Jackson et de Bernstein.

Lorsque $\mathcal{D}$ est une base d'ondelettes dans L_p, les espaces de meilleure approximation à n termes sont identifiés à des espaces de Besov que l'on peut caractériser par des conditions simples de décroissance des coefficients d'ondelettes[25]. Pratiquement, la meilleure approximation à n termes d'une fonction f dans une base d'ondelettes est essentiellement obtenue par l'algorithme glouton, c'est-à-dire en tronquant la décomposition $f = \sum_k c_k \psi_k$ de manière à ne garder que les n coefficients pour lesquels $\|c_k \psi_k\|_X$ est le plus grand. Les bases possédant cette propriété ont été identifiées[38] et sont appelées bases gloutonnes.

Nos travaux dans le domaine visent à généraliser les résultats obtenus avec les ondelettes au cadre plus large de dictionnaires éventuellement redondants.

Des résultats de l'année dernière [32] nous avaient permis de préciser le rapport entre l'approximation à n termes et l'approximation gloutonne à partir d'une base, ainsi que le rapport avec la décroissance des coefficients dans la base. Cette année, nous avons généralisé une partie des résultats valables dans une base au cas de dictionnaires : nous avont montré qu'il faut imposer une structure ℓ_1^p*-hilbertienne* au dictionnaire pour que l'existence d'une représentation parcimonieuse de f se traduise en vitesse d'approximation (c'est-à-dire pour avoir une inégalité de Jackson) [33].

Par ailleurs, dans le cas d'un dictionnaire de *framelets* splines [24] nous avons obtenu une caractérisation complète [34]. Celle-ci est la conséquence d'un résultat quelque peu surprenant au vu de la redondance du dictionnaire de framelets : il existe une décomposition *linéaire* (non adaptative) qui, seuillée, fournit des approximants à m termes convergeant à la vitesse optimale.

Enfin, nous avons également obtenu tout récemment une inégalité de Bernstein pour une large classe de dictionnaires structurés dans des espaces de Hilbert. Pour des dictionnaires *décomposables* et *séparés* [1], la vitesse d'approximation d'une fonction par des approximations à m termes se traduit « presque » par l'existence d'une représentation parcimonieuse. Nous nous penchons actuellement sur les dictionnaires constitués de l'union d'un base de cosinus locaux et d'une base d'ondelettes. Ces dictionnaires ont une

[1] l'exemple typique étant l'union de la base de Haar et du système de Walsh, ou bien le système trigonométrique et les ondelettes de Meyer périodiques, mais notre résultat couvre également des dictionnaires beaucoup plus redondants

structure très proche des dictionnaires décomposables séparés, et les espaces de parcimonie sont reliés à un mélange d'espaces de Besov et d'espaces de modulation.

6.2.5. Evaluation des algorithmes de séparation de sources sonores

Mots clés : *séparation de sources*, *rapport signal à bruit*, *interférences*, *diaphonie*, *artefacts*.

Participants : Rémi Gribonval, Laurent Benaroya, Frédéric Bimbot.

Ce travail est effectué dans le cadre d'une Action Jeunes Chercheurs du GDR ISIS sur les « Ressources pour la séparation de signaux audiophoniques », en collaboration avec l'équipe Analyse-Synthèse de l'IRCAM et l'équipe ADTS de l'IRCCyN. L'objectif de cette action est d'identifier des dénominateurs communs spécifiques aux problèmes soulevés par la séparation de sources audio, afin de proposer des critères numériques et une gamme de signaux tests pertinents et de difficulté calibrée pour l'évaluation des performances des algorithmes existants et futurs.

Les travaux de METISS dans ce domaine ont abouti cette année à la proposition de critères numériques permettant de mesure de distortion entre une sources sonore estimée et la source de référence. Le principal problème rencontré était lié à l'indétermination sur le gain liée à la nature du problème même de séparation de sources à l'aveugle. Nous avons proposé une approche qui nous a permis de distinguer, d'une part, la distortion due aux interférences des autres sources, d'autre part, celle due au bruit additif, et enfin la distortion due aux artefacts de l'algorithme. Une autre contribution de l'équipe a été la participation à la définition d'une typologie des tâches en séparation de sources sonores. Ces travaux vont faire l'objet de publications lors de conférences spécialisées.

6.2.6. Séparation de sources par modèles de signaux sonores monophoniques

Mots clés : *estimation bayésienne*, *modèle de Markov caché*, *mélange de Gaussiennes*.

Participants : Laurent Benaroya, Frédéric Bimbot, Rémi Gribonval.

Ce travail est axé avant tout sur la séparation de sources (audio) dans le cas où il y a moins de capteurs que de sources. La priorité est mise sur la séparation de sources à partir d'enregistrements monophoniques. Une première étude avait été menée sur l'utilisation de *décompositions parcimonieuses*[46] pour la séparation de deux sources avec un seul capteur et nous avions montré qu'une bonne séparation des sources était possible [20] sous des hypothèses de parcimonie. Les hypothèses de parcimonie nécessaires à cette approche étant trop fortes dans le cas de signaux audio, nous nous sommes orientés vers une piste connexe. Nous avons utilisé et généralisé le filtrage de Wiener pour la séparation de deux (ou plus) sources avec un seul capteur. Le filtrage de Wiener étant optimal pour des sources gaussiennes, nous avons généralisé cette approche à des multigaussiennes (mélanges de gaussiennes, modèles de Markov cachés), dans le cadre de l'estimation bayésienne (MAP, espérance conditionnelle), ainsi que des modèles de Markov cachés (ces travaux sont dans la lignée de ceux de S.T. Roweis [45]).

6.2.7. Séparation de sources par analyse de scènes multicanal

Mots clés : *représentation parcimonieuse*, *Matching Pursuit*, *clustering*, *stéréophonie*.

Participant : Rémi Gribonval.

Nous avons mis au point une méthode de séparation de sources à partir d'enregistrements stéréophoniques. La technique [18] est fondée sur une décomposition du signal stéréophonique en paires d'atomes de Gabor (stéréo) suivie d'un *clustering* des paramètres des atomes de la décomposition [35][37]. Elle a été validée sur des mélanges instantanés de signaux réels, en utilisant les critères définis dans le cadre de l'Action Jeunes Chercheurs du GDR ISIS sur les « Ressources pour la séparation de signaux audiophoniques » (voir 8.1). Sur les sources traitées, le niveau relatif des interférences résiduelles des autres sources est de -20 déciBels, et l'essentiel de la distortion est due aux « artefacts » de la séparation non-linéaire, qui se traduit par du « bruit musical » similaire à celui observé lors de la compression de signaux audio à bas débit.

Ces travaux devraient se poursuivre par l'évaluation des méthodes de clustering automatique appropriées et l'extension des résultats au mélange de sources avec délai et/ou réverbération. Un rapprochement avec la modélisation probabiliste conjointe des sources semble aussi souhaitable afin de diminuer les artefacts.

7. Contrats industriels

7.1. Conventions de Recherche

7.1.1. Contrat CP8 (n° 1 99 C 138 00 31321 01 2)

Participants : Raphaël Blouet, Frédéric Bimbot.

Des travaux sur la vérification du locuteur dans des contextes de faibles ressources de mémoire et de calcul ont fait l'objet d'une convention de recherche avec CP8 (ex-Bull) d'une durée de 36 mois, qui a pris fin en décembre 2001 mais dont les développements ont continué début 2002.

L'objectif visé est la vérification du locuteur rapide et distribuée. L'approche développée par METISS a consisté à utiliser des arbres de décision (CART) pour modéliser le score de vérification [3][39][21].

7.2. Actions financées par le RNRT

7.2.1. Projet Domus Videum (n° 2 02 C 0100 00 00 MPR 011)

Participants : Frédéric Bimbot, Guillaume Gravier, Michaël Betser.

Le Projet Domus Videum est un projet RNRT qui a débuté en 2001 et se terminera en 2003.

Les partenaires académiques du projet sont l'IRISA (VISTA, TEXMEX, TEMICS et METISS) et l'Université de Nantes. Les partenaires industriels sont Thomson Multimedia, l'INA, et SFRS.

Le but du projet est la conception et le développement de techniques de constitution automatique de résumés audio-visuels. Les contributions spécifiques de METISS portent sur la modélisation conjointe des flux audio et video par modèles de Markov cachés et l'évaluation des résultats.

7.3. Actions financées par la Commission Européenne

7.3.1. Projet BANCA (n° 1 01 C 0296 00 31331 00 5)

Participants : Fabienne Porée, Frédéric Bimbot.

Le projet BANCA (Biometric Access Control for Networked and e-Commerce Applications) est un projet européen issu du programme IST. Il a débuté en février 2000 et se terminera en février 2003. Les partenaires sont Ibermatica, EPFL, UniS, UCL, Thales, l'IDIAP, BBVA, Oberthur et UC3M.

Le projet vise à la conception d'un système sécurisé multi-modal pour des applications de télé-travail ou de service bancaire par Internet. METISS assume le rôle de Work-Package Manager des activités de recherche en vérification du locuteur.

8. Actions régionales, nationales et internationales

8.1. Actions nationales

8.1.1. Action Jeunes Chercheurs du GDR ISIS

Participants : Rémi Gribonval, Laurent Benaroya, Frédéric Bimbot.

L'Action Jeunes Chercheurs « Ressources pour la séparation de signaux audiophoniques »du GDR ISIS est une collaboration entre l'équipe METISS de l'IRISA, l'équipe Analyse-Synthèse de l'IRCAM et l'équipe ADTS de l'IRCCyN. Son objectif est d'identifier des dénominateurs communs spécifiques aux problèmes soulevés par la séparation de sources audio, afin de proposer des critères numériques et une gamme de signaux tests pertinents et de difficulté calibrée pour l'évaluation des performances des algorithmes existants et futurs.

Cette Action a débuté en mars 2002 et doit durer dix-huit mois.

8.2. Actions européennes

8.2.1. Consortium ELISA

Participants : Mathieu Ben, Raphaël Blouet, Frédéric Bimbot.

Le Consortium ELISA est un consortium d'initiative spontané, ne bénéficiant d'aucun financement spécifique. Il a été fondé en 1997 par l'ENST, l'EPFL, l'IDIAP, l'IRISA et le LIA.

Son objet est la mise en place et l'amélioration d'une plate-forme commune de vérification, détection et suivi du locuteur permettant aux membres du Consortium de participer tous les ans de façon coordonnée aux évaluations américaines NIST en reconnaissance du locuteur.

En 2002 l'IRISA, pour le consortium ELISA, a participé pour la cinquième année aux évaluations NIST. Le système IRISA/ELISA a terminé cette année en cinquième position en termes de taux d'erreur EER (*Equal Error Rate*) sur l'ensemble primaire d'évaluation, sur une vingtaine de participants aux évaluations NIST2002 [47].

8.2.2. Réseau HASSIP

Participant : Rémi Gribonval.

Le réseau HASSIP (Harmonic Analysis, Statistics in Signal and Image Processing) est un réseau (research training network) financé par la Commission Européenne dans le cadre du programme *Improving the Human Potential*. Lancé le 1eroctobre 2002, ce réseau a pour partenaires fondateurs l'Université de Provence/CNRS, l'Université de Vienne, l'Université de Cambridge, l'Université Catholique de Louvain, l'EPFL, l'Université de Brême, l'Université de Munich et le Technion Institute.

Son objet est la mise en place et le développement d'activités de recherche, de formation et de collaborations dans le domaine de l'analyse mathématique et statistique appliquée au traitement des signaux et des images. Le but est de réduire la durée du cycle de développement de nouveaux algorithmes en rassemblant d'une part les mathématiciens et physiciens qui travaillent sur les fondements théoriques (à vocation appliquée), d'autre part les partenaires plus expérimentés en ce qui concerne l'implémentation et l'évaluation.

Les principales contributions de l'IRISA devraient se situer au niveau de la modélisation des signaux sonores pour la compression et la séparation de sources, ainsi que sur les fondements théoriques de l'analyse temps-fréquence/temps-échelle et des approximations non-linéaires.

8.3. Participation à des colloques, séminaires, invitations

Rémi Gribonval a effectué un séjour d'une semaine au Département de Mathématiques de l'Université d'Aalborg, pour une collaboration avec M. Nielsen sur le thème des approximations non-linéaires avec des systèmes redondants. Le séjour a été précédé d'une intervention invitée au workshop « Wavelets, their generalizations and their applications » les 15-16 août à Aalborg, et a donné lieu à la préparation d'un article avec M. Nielsen sur les inégalités de Bernstein pour la caractérisation des espaces d'approximation à m-termes avec des dictionnaires décomposables et séparés.

8.4. Participation à des réunions, constructions de groupes de travail

Frédéric Bimbot est membre du Bureau de l'ISCA (International Speech Communication Association).

Frédéric Bimbot est Vice-Président de l'AFCP (Association Francophone pour la Communication Parlée).

Guillaume Gravier est membre du comité d'animation de l'AFCP.

Rémi Gribonval et Frédéric Bimbot participent à l'Action Européenne COST-277 (« Nonlinear speech processing »).

Rémi Gribonval pilote une Action Jeune Chercheur dans le cadre du GDR ISIS. L'Action vise à rassembler des « Ressources pour la séparation de sources audiophoniques ». Les partenaires sont l'équipe ADTS de l'IRCCyN, Nantes, d'une part, et l'équipe Analyse-Synthèse de l'IRCAM, Paris, d'autre part.

8.5. Enseignement

Frédéric Bimbot a enseigné 30 heures de Traitement de la la Parole à l'EISTI (Ecole Internationale Supérieure du Traitement de l'Information, Cergy-Pontoise) et 18 heures à l'ESIEA (Ecole Supérieure d'Informatique, d'Electronique et d'Automatique).

Mathieu Ben, en tant que moniteur de l'enseignement supérieur, a enseigné 26 heures en traitement du signal appliqué à l'image et 10 heures en asservissements échantillonnés à l'IFSIC (Institut de Formation Supérieure en Informatique et Communication), 36 heures d'électronique en DEUG SPM2 de l'université de Rennes I et 12 heures en asservissements échantillonnés en maîtrise EEA de l'université de Rennes I.

10. Bibliographie

Bibliographie de référence

[1] F. BIMBOT. *Traitement Automatique du Langage Parlé.* série collection Information - Commande - Communication (IC2), Hermès, 2002, chapitre Reconnaissance Automatique du Locuteur, pages 79-114.

[2] F. BIMBOT, R. BLOUET, J.-F. BONASTRE, ET AL.. *The ELISA systems for the NIST'99 evaluation in speaker detection and tracking.* in « Digital Signal Processing », numéro 1-3, volume 10, janvier/avril/juillet, 2000, pages 143-153.

[3] R. BLOUET, F. BIMBOT, C. GOIRE. *Procédé de vérification automatique de signaux de données biométriques, notamment générées par un locuteur, et architecture pour la mise en oeuvre.* in « Brevet d'Invention déposé par BULL-CP8, CNRS et INRIA », numéro INPI 0107913, juin, 2001, déposé en juin 2001, étendu en juin 2002.

[4] G. GRAVIER, F. YVON, B. JACOB, F. BIMBOT. *Integrating contextual phonological rules in a large vocabulary decoder.* éditeurs W. VAN DOMMELEN, B. BARRY., in « The Integration of Phonetic Knowledge in Speech Technology », Kluwer Academics, 2002, à paraître.

[5] R. GRIBONVAL. *Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps.* in « IEEE Trans. Signal Proc. », numéro 5, volume 49, mai, 2001, pages 994-1001.

[6] R. GRIBONVAL. *Approximations non-linéaires pour l'analyse de signaux sonores.* thèse de doctorat, Université Paris IX Dauphine, septembre, 1999.

[7] M. SECK, R. BLOUET, F. BIMBOT. *The IRISA/ELISA speaker detection and tracking systems for the NIST'99 evaluation campaign.* in « Digital Signal Processing », numéro 13, volume 10, janvier/avril/juillet, 2000, pages 154-171.

[8] M. SECK. *Détection de ruptures et suivi de classe de sons pour l'indexation sonore.* thèse de doctorat, Université de Rennes 1, IRISA, Rennes, janvier, 2001.

Thèses et habilitations à diriger des recherche

[9] F. BIMBOT. *Contributions en modélisation et reconnaissance de locuteurs et de sources sonores.* Habilitation à Diriger des Recherches, Université de Rennes I, 2002, Soutenance prévue le 16 décembre 2002.

[10] R. BLOUET. *Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées.* thèse de doctorat, Université de Rennes 1, IRISA, Rennes, décembre, 2002.

Articles et chapitres de livre

[11] B. LIU, S. LING, R. GRIBONVAL. *Bearing failure detection using matching pursuit.* in « NDT & E International », numéro 4, volume 35, 2002, pages 255-262.

[12] I. MAGRIN-CHAGNOLLEAU, G. DUROU, F. BIMBOT. *Application of Time-Frequency Principal Component Analysis to Text-Independent Speaker Identification.* in « IEEE Trans. on Speech and Audio Processing », numéro 5, volume 10, jul, 2002, pages 1-10.

Communications à des congrès, colloques, etc.

[13] M. BEN, R. BLOUET, F. BIMBOT. *A Monte-Carlo Method for Score Normalization in Automatic Speaker Verification.* in « Proceedings of ICASSP'2002 », pages I-689 à I-692, Orlando, mai, 2002.

[14] G. GRAVIER, S. AXELROD, G. POTAMIANOS, C. NETI. *Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR.* in « Proc. IEEE Int. Conf on Acoust. Speech and Signal Proc. », mai, 2002.

[15] G. GRAVIER, G. POTAMIANOS, C. NETI. *Asynchronous multi-stream audio-visual speech recognition : from digits to large vocabulary.* in « Proc. Journées d'Étude sur la Parole », juin, 2002.

[16] G. GRAVIER, G. POTAMIANOS, C. NETI. *Asynchrony modeling for audio-visual speech recognition.* in « Human Language Technology Conference », mars, 2002.

[17] G. GRAVIER, F. YVON, B. JACOB, F. BIMBOT. *Sirocco, un système ouvert de reconnaissance de la parole.* in « Proc. Journées d'Étude sur la Parole », pages 273-276, Nancy, juin, 2002.

[18] R. GRIBONVAL. *Sparse decomposition of stereo signals with Matching Pursuit and application to blind separation of more than two sources from a stereo mixture.* in « Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'02) », Orlando, Florida, mai, 2002.

Bibliographie générale

[19] R. AUCKENTHALER, M. CAREY, H. LLOYD-THOMAS. *Score Normalization for Test-Independent Speaker Verification Systems.* in « Digital Signal Processing », numéro 1-3, volume 10, 2000.

[20] L. BENAROYA, R. GRIBONVAL, F. BIMBOT. *Représentations parcimonieuses pour la séparation de sources avec un seul capteur.* in « GRETSI 2001 », Toulouse, 2001.

[21] R. BLOUET, F. BIMBOT. *Tree-based score computation for speaker verification.* in « 7th European Conference on Speech Communication and Technology », pages 67 à 72, septembre, 2001.

[22] R. BOITE. *Traitement de la Parole.* Presses Polytechniques et Universitaires Romandes, 2000.

[23] L. BREIMAN, J. FRIEDMAN, R. OLSHEN, C. STONE. *Classification And Regression Trees.* Wadsworth, 1984.

[24] I. DAUBECHIES, B. HAN, A. RON, Z. SHEN. *Framelets : MRA-based constructions of wavelet frames.* in « Preprint », 2001.

[25] R. A. DEVORE. *Nonlinear approximation.* in « Acta numerica, 1998 », Cambridge Univ. Press, Cambridge, 1998, pages 51-150.

[26] R. A. DEVORE, G. G. LORENTZ. *Constructive approximation.* Springer-Verlag, Berlin, 1993.

[27] J. FRIEDMAN, T. HASTIE, R. TIBSHIRANI. *Additive Logistic Regression : a Statistical View of Boosting.* rapport technique, Department of Statistics, Stanford University, 1999.

[28] R. GRIBONVAL, E. BACRY. *Harmonic Decomposition of Audio Signals with Matching Pursuit.* in « IEEE Trans. Signal Proc. », numéro 1, volume 51, jan, 2003, à paraître.

[29] R. GRIBONVAL. *A counter-example to the general convergence of partially greedy algorithms.* in « J. Approx. Theory », volume 111, 2001, pages 128-138, http://www.sciencedirect.com/science/journals, doi :10.1006/jath.2001.3556.

[30] R. GRIBONVAL. *Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps.* in « IEEE Trans. Signal Proc. », numéro 5, volume 49, mai, 2001, pages 994-1001.

[31] R. GRIBONVAL. *Partially greedy algorithms.* in « Trends in Approximation Theory », Vanderbilt University Press, éditeurs K. KOPOTUN, T. LYCHE, M. NEAMTU., pages 143-148, Nashville, septembre, 2001.

[32] R. GRIBONVAL, M. NIELSEN. *Some remarks on nonlinear approximation with Schauder bases.* in « East Journal on Approximations », numéro 2, volume 7, 2001, pages 1-19.

[33] R. GRIBONVAL, M. NIELSEN. *Nonlinear approximation with dictionaries. I. Direct estimates..* in « J. Fourier Anal. and Appl. », 2003, à paraître.

[34] R. GRIBONVAL, M. NIELSEN. *On approximation with spline generated framelets.* in « Constr. Approx. », 2003, à paraître.

[35] M. V. HULLE. *Clustering approach to square and non-square blind source separation.* in « IEEE Workshop on Neural Networks for Signal Processing (NNSP99) », pages 315-323, août, 1999.

[36] F. JELINEK. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, Massachussetts, 1998.

[37] A. JOURJINE, S. RICKARD, O. YILMAZ. *Blind Separation of Disjoint Orthogonal Signals : Demixing N Sources from 2 Mixtures.* in « ICASSP00 », volume 5, pages 2985-2988, Istanbul, Turkey, juin, 2000.

[38] S. V. KONYAGIN, V. N. TEMLYAKOV. *A remark on greedy approximation in Banach spaces.* in « East J. Approx. », numéro 3, volume 5, 1999, pages 365-379.

[39] I. MAGRIN-CHAGNOLLEAU, G. GRAVIER, R. BLOUET. *Overview of the 2000-2001 ELISA Consortium Research Activities.* in « Proceedings of 2001 : A Speaker Odyssey-The Speaker Recognition Workshop », 2001.

[40] S. MALLAT. *A Wavelet Tour of Signal Processing.* édition 2, Academic Press, San Diego, 1999.

[41] A. MARTIN, M. PRZYBOCKI. *The NIST Year 2001 Speaker Recognition Plan Evaluation.* 2001, http://www.nist.gov/speech/tests/spk/2001/doc/index.htm.

[42] J. O'NEILL, P. FLANDRIN, W. KARL. *Sparse Representations with Chirplets via Maximum Likelihood Estimation.* in « IEEE Transactions on Signal Processing », 2000, Soumis.

[43] S. ORTMANNS, N. HERMANN. *A word graph algorithm for large vocabulary continuous speech recognition.* in « Computer Speech and Language », volume 11, 1997, pages 43-72.

[44] A. REYNOLDS, T. QUATIERI, R. DUNN. *Speaker Verification Using Adapted Gaussian Mixture Models.* in « Digital Signal Processing Vol 10,num 1-3 », 2000.

[45] S. T. ROWEIS. *One Microphone Source Separation.* in « NIPS 2000 : Proc. Neural Information Processing Systems 2000 », pages 793-799, 2000.

[46] M. ZIBULEVSKY, B. PEARLMUTTER. *Blind Source Separation by Sparse Decomposition in a Signal Dictionary.* in « Neural Computations », numéro 4, volume 13, 2001, pages 863-882, http://citeseer.nj.nec.com/article/zibulevsky00blind.html.

[47] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *The 2002 NIST Speaker Recognition Evaluation.* 2002, http://www.nist.gov/speech/tests/spk/2002/.