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

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

1 Composition de l'équipe

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]
Post-Doctorant
Guillaume Gravier [jusqu'au 31 mars 2001]
Ingénieur-Expert
Fabienne Porée [depuis le 1er septembre 2001]
Chercheurs doctorants
Mathieu Ben [allocataire MENRT (et moniteur) à partir du 1er octobre 2001]
Laurent Benaroya [allocataire MENRT, 3e année]
Raphaël Blouet [bourse INRIA, jusqu'au 30 novembre 2001]
Lorcan Mc Donagh [bourse INRIA, 2e 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 source). 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 de laboratoires (ELISA), d'actions de recherche (AUF-B1, SIROCCO), de projets nationaux (AGIR) ou européens (PICASSO, DiVAN, BANCA) et de contrats industriels (BULL/CP8).

3 Fondements scientifiques

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[Jel98] qui a été à l'origine de progrès considérables dans différents domaines de la reconnaissance des formes, et notamment en traitement de la parole[Boi00]. 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 Modèle 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[Mal99]. 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.

3.1 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

[Jel98] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Massachussetts, 1998.

[Boi00] R. Boite, *Traitement de la Parole*, Presses Polytechniques et Universitaires Romandes, 2000.

[Mal99] S. Mallat, *A Wavelet Tour of Signal Processing*, édition 2, Academic Press, San Diego, 1999.

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.).

3.1.1 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 de cette fonction $\hat{P}$, 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 ré-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.

3.1.2 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) :

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

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.

3.1.3 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, que le nombre de classes est élevé et que 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.

3.1.4 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,...

3.2 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.

3.2.1 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, des 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.

3.2.2 Critères de parcimonie

L'obtention d'une solution unique pour l'équation X 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.

3.2.3 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.

3.2.4 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.

3.2.5 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 source 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'applications

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 source. Certains aspects de la reconnaissance de la parole (modélisation et décodage) viennent accessoirement renforcer ces 3 thèmes principaux.

4.1 Vérification du locuteur

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

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

Résumé : *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.*

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 du locuteur). 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 la 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*, liées 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.).

4.2 Détection et suivi d'information dans les flux sonores

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

Participants : Frédéric Bimbot, Raphaël Blouet.

Résumé : *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.*

4.2.1 Détection de locuteur

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.

4.2.2 Suivi de classe de son

Les approches présentées au paragraphe précédent pour la détection de présence, la localisation et la segmentation de locuteur peuvent être réutilisées pour d'autres tâches similaires, notamment la détection de classes de sons. Ainsi, 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, certains jingles, etc. Ces différents éléments sont en effet des points de repère essentiels dans une émission ou une série d'émissions et leur localisation automatique, accompagnée d'une représentation visuelle appropriée, permet de focaliser immédiatement les recherches manuelles ou automatiques sur les plages sonores d'intérêt.

L'approche par rapport de vraisemblance 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'agit de classes générales, comme parole, musique, etc. ou d'événements particuliers, comme un jingle).

La détection de jingles fonctionne de façon très fiable, mais des problèmes tels que les jingles avec de la parole superposée (par exemple, le carillon d'Europe 1) peuvent encore poser quelques difficultés. Des performances acceptables sont obtenues en suivi de classe de son, avec une certaine variabilité selon la classe considérée et le type de données traitées. Les principaux problèmes à résoudre proviennent des cas où l'on est en présence de faibles volumes de données d'apprentissage et de scènes sonores complexes, par exemple les publicités lors desquelles différentes classes de signaux sont couramment mélangées.

4.3 Traitement avancé de signaux sonores

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

Résumé : *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 source et de la représentation de signaux sonores.

4.3.1 Séparation de sources

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

En toute généralité, le problème de la séparation de source 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 source, l'un des signaux étant le signal comprimé, l'autre le résidu de compression. Ainsi, le problème de la séparation de source 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 source 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 (stéréo) pour $n > 2$ sources.

4.3.2 Représentation de signaux sonores

Participants : Rémi Gribonval, Lorcan Mc Donagh.

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.

4.4 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 et la modélisation du langage car ces thématiques apportent des compléments indispensables pour améliorer l'impact des applications dans certains domaines, notamment en sécurisation de transactions et en indexation sonore.

5 Logiciels

5.1 Nouvelle version de la plate-forme ELISA

Participants : Raphaël Blouet, Frédéric Bimbot.

Une nouvelle version de la plate-forme ELISA, pour la participation à la campagne d'évaluation NIST 2001, a été mise en place, sur la base des modules développés à l'IRISA et dans les labos 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'ENST, l'IRISA, et le LIA ont été les principaux contributeurs à la version 2001 [22].

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.

5.2 Plate-forme SIROCCO

Participants : Guillaume Gravier, Frédéric Bimbot.

l'Action de Recherche Concertée SIROCCO de l'INRIA visant à la mise en place par plusieurs laboratoires français d'une plate-forme commune pour la reconnaissance de parole grand-vocabulaire permettant de canaliser les spécialités des uns et des autres au fur et à mesure des progrès accomplis.

Dans le cadre de cette ARC, METISS a fait porter ses efforts sur l'implémentation et l'amélioration de l'algorithme de décodage rapide (sous-optimal), connu sous le nom de *beam-search* (recherche en faisceaux). Cette technique permet de contrôler le compromis entre le temps de recherche et le risque de ne pas trouver la séquence d'états optimale. Une des contributions originales des travaux réalisés dans ce cadre est la gestion des variantes contextuelles de prononciation de façon intégrée avec la recherche en faisceaux [19].

La conclusion de l'ARC de l'INRIA SIROCCO a donné lieu à la distribution, au sein des partenaires de l'ARC, d'un ensemble de modules pour la reconnaissance de parole grand-vocabulaire. Le module de recherche en faisceau sera disponible publiquement sous license GNU.

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 par distance de Kullback

Participants : Mathieu Ben, Frédéric Bimbot, Raphaël Blouet.

Nos travaux récents ont débouché sur une nouvelle technique de normalisation des scores en vérification automatique du locuteur : la d-norm (pour « distance normalization »). L'avantage principal de cette normalisation est qu'elle ne nécessite aucune donnée de parole supplémentaire ni de population de locuteurs externes, contrairement aux techniques de l'état de l'art (z-norm, t-norm) [ACLT00]. La d-Norm est basée sur l'utilisation des distances de Kullback-Leibler entre modèles du locuteur et du non-locuteur. Ces distances sont estimées avec une méthode de type « Monte-Carlo »[RC99] et nous avons exploité la forte corrélation, observée expérimentalement, entre ces distances et les scores de vérification pour implémenter cette procédure de normalisation des scores.

Les performances de la d-norm ont été évaluées sur un sous-ensemble du corpus NIST'00 [Nat00] avec le système IRISA/ELISA [MCGB01]. Ces expériences ont montré que les résultats obtenus sont comparables aux performances d'une normalisation conventionnelle, la z-norm, qui elle, nécessite l'utilisation de données supplémentaires. La d-norm pourrait donc avantageusement remplacer la z-norm dans certains cas spécifiques pour lesquels l'utilisation d'un nouveau corpus de données n'est pas possible. De plus, l'utilisation de la d-norm permet d'alléger considérablement la procédure de normalisation, en la rendant plus rapide et beaucoup moins coûteuse que la t-norm, en puissance de calcul au moment du test.

Des travaux futurs devront consolider les résultats que nous avons obtenus, au travers de nouvelles expériences. Notamment, les performances de la d-norm devront être comparées à celles de la t-norm (celle-ci donnant en général des résultats comparables à la z-norm). On peut également envisager de combiner ces techniques entre elles et d'étudier l'impact sur les performances du système.

[ACLT00] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, « Score Normalization for Test-Independent Speaker Verification Systems », *Digital Signal Processing 10*, 1-3, 2000.

[RC99] C. Roberts, G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, 1999.

[Nat00] National Institute of Standards and Technology, « The 2000 NIST Speaker Recognition Evaluation », 2000, `http://www.nist.gov/speech/tests/spk/2000/`.

[MCGB01] 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.

6.1.2 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 (K) 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)[BFOS84]. 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[17, 18].

La première mise en oeuvre de cette technique a été validée sur un sous-ensemble du corpus de l'évaluation NIST'01 [MP01] des systèmes de vérification du locuteur et est décrite dans [17]. Une deuxième vague de travaux, actuellement en cours, consistent 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* [FHT99] 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 [25].

6.1.3 Incorporation de contraintes phonologiques dans la recherche en faisceaux

Participants : Frédéric Bimbot, 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 [OH97] ou *beam-search*.

S'appuyant sur la version conventionnelle de l'algorithme, nous nous sommes intéressé à son extension dans le cas de l'intégration de contraintes phonologiques, c'est-à-dire de l'influence

[BFOS84] L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification And Regression Trees*, Wadsworth, 1984.

[MP01] A. Martin, M. Przybocki, « The NIST Year 2001 Speaker Recognition Plan Evaluation », 2001, `http://www.nist.gov/speech/tests/spk/2001/doc/index.htm`.

[FHT99] J. Friedman, T. Hastie, R. Tibshirani, « Additive Logistic Regression: a Statistical View of Boosting », *rapport de recherche*, Department of Statistics, Stanford University, 1999.

[OH97] S. Ortmanns, N. Hermann, « A word graph algorithm for large vocabulary continuous speech recognition », *Computer Speech and Language 11*, 1997, p. 43–72.

qu'un mot donné peut avoir sur la prononciation d'un autre. Un exemple (fréquent) en français est celui de la liaison entre mots d'un même groupe syntaxique.

Les règles de transcription contextuelle utilisées dans ce contexte 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).

Le module de beam-search développé dans SIROCCO incorpore cette fonctionnalité et une première validation expérimentale a été réalisée. Le résultats obtenus confirment la cohérence de l'approche et la validité de son implémentation [19]. Toutefois, ils mettent également en évidence l'insuffisance des ressources utilisées dans ces expériences (règles trop simples, modèles de phonèmes pas suffisamment précis, ...), induisant des performances qui méritent d'être améliorées.

6.2 Traitement du signal sonore

Participants : Laurent Benaroya, Frédéric Bimbot, Rémi Gribonval, Lorcan McDonagh.

6.2.1 Approximation non-linéaire

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

Participant : Rémi Gribonval.

Ce travail est conduit en coopération avec Morten Nielsen de l'Institut de Mathématiques Industrielles de l'Université de Caroline du Sud, dans le cadre du consortium de recherche Ideal Data Representation.

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[DL93] 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[DeV98]. 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[KT99] et sont appelées bases gloutonnes.

[DL93] R. A. DeVore, G. G. Lorentz, *Constructive approximation*, Springer-Verlag, Berlin, 1993.

[DeV98] R. A. DeVore, « Nonlinear approximation », *in : Acta numerica, 1998*, Cambridge Univ. Press, Cambridge, 1998, p. 51–150.

[KT99] S. V. Konyagin, V. N. Temlyakov, « A remark on greedy approximation in Banach spaces », *East J. Approx. 5*, 3, 1999, p. 365–379.

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

Nous avons obtenu un premier résultat qui précise 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. La caractérisation la plus précise est obtenue dans les bases gloutonnes [13].

Nous avons également obtenu tout récemment une des toutes premières caractérisation d'espaces d'approximation non-linéaire avec des dictionnaires redondants : nous avons montré que les espaces d'approximation à n termes dans $L_p(\mathbb{R})$, à partir de *framelets* splines[DHRS01], sont des espaces de Besov, comme dans le cas des ondelettes. Un élément essentiel de la preuve repose sur le fait que les *framelets* peuvent s'exprimer comme combinaison linéaire finie d'ondelettes splines bi-orthogonales[CW92]. Ces résultats sont valables même lorsque les conditions usuelles de moments nuls ne sont pas remplies, c'est-à-dire avec des *framelets* oscillant peu. Cela laisse entrevoir des applications prometteuses des *framelets* à la compression de signaux, en limitant le phénomène de Gibbs.

6.2.2 Séparation de sources dans des cas sous-déterminés

Mots clés : représentation parcimonieuse, ACP, ACI, estimation bayesienne.

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, situation notamment rencontrée sur la plupart des enregistrements monophoniques et stéréophoniques.

La priorité est mise sur la séparation de sources à partir d'un enregistrement monophonique. Une première étude a été menée sur l'utilisation de *décompositions parcimonieuses*[ZP01] pour la séparation de deux sources avec un seul capteur. Nous avons supposé que chacune des sources a une représentation fortement parcimonieuse dans une base à déterminer, c'est à dire que la source est bien approchée par un petit nombre de coefficients non nuls dans cette base. En supposant que ces deux bases bien « adaptées »aux sources (une pour chaque source) sont aussi bien « séparées », c'est-à-dire suffisament incohérentes, on a montré (dans le cadre des bases de Dirac/Fourier) que la séparation exacte des sources est possible [16].

Nous avons pris plus récemment une direction qui consiste à modéliser statistiquement chaque source (modèles AR, modèles de Markov cachés, ...) en vue d'une séparation par critère de Maximum a Posteriori (MAP).

Par ailleurs, nous avons mis au point une méthode de séparation de sources à partir d'enregistrements stéréo. La technique, fondée sur une décomposition du signal stéréo en paires

[DHRS01] I. Daubechies, B. Han, A. Ron, Z. Shen, « Framelets: MRA-based constructions of wavelet frames », *Preprint*, 2001.

[CW92] C. K. Chui, J.-Z. Wang, « On compactly supported spline wavelets and a duality principle », *Trans. Amer. Math. Soc. 330*, 2, 1992, p. 903–915.

[ZP01] M. Zibulevsky, B. Pearlmutter, « Blind Source Separation by Sparse Decomposition in a Signal Dictionary », *Neural Computations 13*, 4, 2001, p. 863–882, `http://citeseer.nj.nec.com/zibulevsky00blind.html`.

d'atomes de Gabor (stéréo) suivie d'un *clustering* des paramètres des atomes de la décomposition [Hul99,JRY00], a été validée sur des mélanges instantanés de signaux réels. 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.

6.2.3 Algorithmes gloutons pour l'analyse de signaux sonores

Mots clés : Matching Pursuit, analyse de signaux musicaux, analyse de signaux stéréo.

Participants : Rémi Gribonval, Lorcan Mc Donagh.

La partie théorique de ce travail est conduite en coopération avec Morten Nielsen de l'Institut de Mathématiques Industrielles de l'Université de Caroline du Sud, dans le cadre du consortium de recherche Ideal Data Representation. Des applications à l'analyse de signaux musicaux sont développées en collaboration avec Emmanuel Bacry, du Centre de Mathématiques Appliquées de l'Ecole Polytechnique, et devraient se prolonger dans le cadre de la thèse de Lorcan Mc Donagh.

Des travaux des années antérieures publiés cette année [14, 15, 20] ont montré l'intérêt et la flexibilité des méthodes de décomposition de signaux de type Matching Pursuit pour des applications sonores. En utilisant des résultats théoriques obtenus avec Morten Nielsen [12], nous avons prouvé la convergence du « Matching Pursuit Harmonique »dans des conditions de mise en œuvre souples permettant de réduire la 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.

Une prolongation de ces recherches est en cours dans le cadre de la thèse de Lorcan Mc Donagh. Il s'agit, en s'inspirant de travaux similaires portant sur des décompositions en *chirplets*[OFK00], de mettre 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.

7 Contrats industriels (nationaux, européens et internationaux)

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

[Hul99] M. V. Hulle, « Clustering approach to square and non-square blind source separation », *in : IEEE Workshop on Neural Networks for Signal Processing (NNSP99)*, p. 315–323, août 1999.

[JRY00] A. Jourjine, S. Rickard, O. Yilmaz, « Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures », *in : ICASSP00*, *5*, p. 2985–2988, Istanbul, Turkey, juin 2000.

[OFK00] J. O'Neill, P. Flandrin, W. Karl, « Sparse Representations with Chirplets via Maximum Likelihood Estimation », *IEEE Transactions on Signal Processing*, 2000, Soumis.

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.

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 [25, 17, 18].

7.2 Actions financées par le RNRT

7.2.1 Projet AGIR (n°2 99 C 006 00 00 MPR 01)

Participant : Frédéric Bimbot.

Le Projet AGIR (Architecture Globale pour l'Indexation et la Recherche par le contenu de données multimédia) est un projet RNRT qui a débuté le 1er Janvier 1999 et s'est terminé le 28 Février 2001.

Les partenaires académiques du projet étaient le LIP6, l'INRIA Rhône-Alpes, l'INT, l'IRISA (VISTA et METISS) et l'IRIT. Les partenaires industriels étaient CS, l'INA, Arts Vidéo et Mémodata.

Le but du projet était la conception et le développement de techniques de recherche de documents multimédia, basé sur l'analyse automatique et sur un langage normalisé de description de contenus. Les contributions spécifiques de METISS ont porté sur la détection de classe de sons ainsi que la détection et le suivi de locuteur [23].

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. 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 ARC SIROCCO (n°39007)

Participants : Guillaume Gravier, Frédéric Bimbot.

L'Action de Recherche Coopérative de l'INRIA SIROCCO s'appuie sur deux projets de l'INRIA (METISS à l'IRISA et PAROLE au LORIA), ainsi que 3 équipes extérieures : l'ENST, l'IRIT et le LIA. Cette ARC a débuté en Janvier 2000 et s'est terminée en Mars 2001.

Son objectif était la mise en place d'une plate-forme commune de reconnaissance grand-vocabulaire. La contribution principale de METISS a été le développement d'un module de recherche en faisceaux (beam-search) ainsi que le conception et la réalisation d'un procédé de décodage intégrant les variations phonologiques contextuelles [19].

8.2 Actions européennes

8.2.1 Consortium ELISA

Participants : Raphaël Blouet, Mathieu Ben, 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 tout les ans de façon coordonnée aux évaluations américaines NIST en reconnaissance du locuteur.

L'année 2001 a vu la quatrième participation d'ELISA aux évaluations NIST. ELISA a terminé en troisième position à l'EER sur l'ensemble primaire d'évaluation [22].

9 Diffusion de résultats

9.1 Stages

Mathieu Ben, stage de DEA (STIR, Rennes 1), « Diagnostic et amélioration d'un système de vérification automatique du locuteur, par approches statistiques », du 1ermars 2001 au 15 juillet 2001.

Samuel Vermeulen, IFSIC (DIIC), « Représentation temps-fréquence à partir d'un fichier sonore au format MP3 », du 1erAvril au 31 Juillet 2001.

9.2 Participation à des colloques, séminaires, invitations

Frédéric Bimbot a co-organisé le Workshop ISCA-IEEE « 2001 : A Speaker Odyssey »à La Chanée, en Crète (Grèce), en juin 2001.

Rémi Gribonval a effectué un séjour à l'Institut de Mathématiques Industrielles (IMI) de l'Université de Caroline du Sud de mars à juin, dans le cadre du consortium de recherche Ideal Data Representation (IDR), pour une collaboration avec M. Nielsen, V. Temlyakov et R. DeVore sur le thème des approximations non-linéaires avec des systèmes redondants. Le séjour a donné lieu à la révision d'un article avec M. Nielsen sur les espaces d'approximation à m-termes et d'approximation gloutonne dans une base bien structurée, ainsi qu'à la préparation d'un article sur l'approximation avec des *framelets* construites sur une multirésolution spline.

9.3 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).

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

Rémi Gribonval et Frédéric Bimbot ont participé à la mise sur pied d'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.

9.4 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).

10 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] F. Bimbot, al., « The ELISA systems for the NIST'99 evaluation in speaker detection and tracking », *Digital Signal Processing 10*, 1-3, janvier/avril/juillet 2000, p. 143–153.

[2] F. Bimbot, *Traitement Automatique du Langage Parlé, collection Information - Commande - Communication (IC2)*, Hermès, 2001-2002, ch. Reconnaissance Automatique du Locuteur, à paraître.

[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*, INPI 0107913, juin 2001.

[4] R. Gribonval, M. Nielsen, « Approximate weak greedy algorithms », *Advances in Computational Mathematics 14*, 4, mai 2001, p. 361–378.

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

[6] R. Gribonval, « Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps », *IEEE Trans. Signal Proc. 49*, 5, mai 2001, p. 994–1001.

[7] R. Gribonval, *Temps-fréquence : concepts et outils, sous la coordination de F. Hlawatsch, F. Auger et J.-P. Ovarlez, collection Information-Commande-Communication (IC2)*, Hermès, 2001-2002, ch. Analyse temps-fréquence linéaire I : Représentations type Fourier, à paraitre.

[8] M. Seck, R. Blouet, F. Bimbot, « The IRISA/ELISA speaker detection and tracking systems for the NIST'99 evaluation campaign », *Digital Signal Processing 10*, 13, janvier/avril/juillet 2000, p. 154–171.

Thèses et habilitations à diriger des recherches

[9] 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.

Articles et chapitres de livre

[10] F. Bimbot, M. El-Bèze, S. Igounet, M. Jardino, K. Smaïli, I. Zitouni, « An alternative scheme for perplexity estimation and its assessment for the evaluation of language models », *Computer Speech and Language 15*, 1, January 2001, p. 1–13.

[11] F. Bimbot, *Traitement Automatique du Langage Parlé, Information - Commande - Communication*, Hermès, 2002, ch. Reconnaissance Automatique du Locuteur, à paraitre.

[12] R. Gribonval, M. Nielsen, « Approximate weak greedy algorithms », *Advances in Computational Mathematics 14*, 4, mai 2001, p. 361–378.

[13] R. Gribonval, M. Nielsen, « Some remarks on nonlinear approximation with Schauder bases », *East Journal on Approximations 7*, 2, 2001, p. 1–19.

[14] R. Gribonval, « A counter-example to the general convergence of partially greedy algorithms », *J. Approx. Theory 111*, 2001, p. 128–138.

[15] R. Gribonval, « Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps », *IEEE Trans. Signal Proc. 49*, 5, mai 2001, p. 994–1001.

Communications à des congrès, colloques, etc.

[16] L. Benaroya, R. Gribonval, F. Bimbot, « Représentations parcimonieuses pour la séparation de sources avec un seul capteur », *in : GRETSI 2001*, Toulouse, 2001.

[17] R. Blouet, F. Bimbot, « A tree-based approach for score computation in speaker verification », *in : Workshop Speaker Odyssey*, La Chanée (Crète), Grèce.

[18] R. Blouet, F. Bimbot, « Tree-based score computation for speaker verification », *in : 7th European Conference on Speech Communication and Technology*, p. 67 à 72, septembre 2001.

[19] G. Gravier, F. Yvon, B. Jacob, F. Bimbot, « Integrating contextual phonological rules in a large vocabulary decoder », *in : Eurospeech'2001*, p. 2293–2296, Aalborg, Danemark, septembre 2001.

[20] R. Gribonval, « Partially greedy algorithms », *in : Trends in Approximation Theory*, K. Kopotun, T. Lyche, M. Neamtu (éditeurs), Vanderbilt University Press, p. 143–148, Nashville, septembre 2001.

[21] S. Krstulovic, F. Bimbot, « Signal modeling with non-uniform topology lattice filters », *in : ICASSP'2001*, Salt Lake City, mai 2001.

[22] I. Magrin-Chagnolleau, G. Gravier, R. Blouet, « Overview of the 2000-2001 Elisa consortium research activities », *in : Workshop Speaker Odyssey*, p. 67 à 72, La Chanée (Crète), Grèce.

[23] M. Seck, I. Magrin-Chagnolleau, F. Bimbot, « Experiments on speech tracking in audio documents using Gaussian Mixture Modeling », *in : Proceedings of ICASSP'01, 1*, Salt Lake City, mai 2001.

Rapports de recherche et publications internes

[24] F. Bimbot, E. Bailly-Baillière-Gutiérrez, S. Bengio, J. Mariéthoz, B. Ruiz, « D72.2-SV - General assessment report on speaker verification (SV) », *rapport de recherche*, IRISA - IDIAP - UC3M, 2001.

Divers

[25] 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*, INPI 0107913, juin 2001.