Approximations parcimonieuses structurées pour le traitement de signaux audio

23 avril 2004

Résumé

L'objectif de ce projet est de lever les principaux verrous théoriques concernant l'identifiabilité des modèles parcimonieux structurés pour l'approximation de signaux avec des dictionnaires redondants. On étudiera en particulier l'identifiabilité en présence de bruit et on analysera les performances d'algorithmes d'approximation structurée en termes de codage et de séparation de sources sonores. Des journées de rencontres seront organisées à l'issue du projet afin de disséminer l'usage de ces méthodes pour le traitement du signal audio.

1 Equipes participantes :

Equipe-projet METISS, IRISA, Rennes

- Frédéric Bimbot (Responsable, CR CNRS), à 10 %;
- Rémi Gribonval (CR INRIA), à 30 %;
- Sylvain Lesage (doctorant MENRT), à 50 %;

Groupe traitement du signal, Laboratoire d'Analyse, Topologie et Probabilités
Université de Provence, Marseille

- Bruno Torrésani (Responsable, Professeur), à 15 %;
- Clothilde Melot (Maître de conférences), à 15 %;
- Florent Jaillet (doctorant CIFRE, société Genesis), à 20 %;
- Peter Balazs (PostDoc, réseau Européen HASSIP), à 20 %;
- Damián Marelli (PostDoc, réseau Européen HASSIP), à 20 %;

Laboratoire d'Acoustique Musicale,Université Pierre et Marie Curie, Paris

- Laurent Daudet (Maître de conférences, Université Paris VI), à 30 %
- Pierre Leveau (étudiant en DEA ATIAM, doctorant MENRT potentiel), à 50 %

2 Moyens demandés:

Le projet s'inscrit dans une perspective globale à moyen terme (trois à quatre ans). Les moyens demandés ici ne couvrent que la première année, qui doit servir d'amorçage pour d'autres formes de travail (ACI ...). Le budget prévisionnel suivant concerne donc le fonctionnement sur un an:

- réunions de travail (tous les deux mois environ) : 5000 Euros
- participation à des congrès nationaux/internationaux : 4500 Euros
- organisation de journées (invitation et prise en charge des frais d'une dizaine de conférenciers sur deux jours): 5000 Euros.

- échange de doctorants/post-docs, séjours : 5000 Euros
- consommables (livres, etc.) : 500 Euros

TOTAL: environ 20 000 Euros

3 Descriptif du projet

Les méthodes de traitement du signal basées sur des approximations parcimonieuses sont extrêmement prometteuses en termes d'applications aussi diverses que la compression, le débruitage, la séparation de sources, l'identification de systèmes et l'extraction d'information pour l'indexation.

Ces toutes dernières années, les équipes qui formulent aujourd'hui ce projet –et qui collaborent régulièrement via le GDR ISIS et le réseau européen de formation (RTN) HASSIP– ont obtenu des progrès substantiels non seulement en terme d'exploitation effective de ces méthodes (notamment sous forme d'un prototype de codeur audio ou pour la séparation de sources audio) mais aussi pour ce qui est de la compréhension des limites et des propriétés des outils mathématiques sur lesquels elle reposent. Le contexte international est assez dynamique puisque des équipes comme le Signal Processing Lab de Cambridge, le Département de Statistiques de Stanford ou le Vision and Image Sciences Laboratory du Technion Institute travaillent activement sur ces questions. Les résultats théoriques et pratiques obtenus notamment par les équipes candidates à ce projet ont permis d'identifier des verrous essentiellement liés au fait qu'une "bonne" approximation d'un signal soit envisagée uniquement sous l'angle de la parcimonie. De fait, en cherchant des approximations parcimonieuses *structurées* (en arbre, en chaîne, . . .), il y a aujourd'hui de bons indices qui laissent à penser que ces verrous pourront être levés.

3.1 Objectifs et contexte

Ce projet a pour but de fédérer les forces des trois équipes pour franchir une étape décisive dans l'usage des représentations parcimonieuses et structurées pour le traitement du signal sonore. Il s'agit de **briser effectivement un certain nombre des verrous** mentionnés ci-dessus et d'identifier des stratégies pour s'attaquer à ceux qui demeurent. L'objectif scientifique général est de comprendre les conditions d'identifiabilité des modèles parcimonieux structurés et de proposer et analyser des algorithmes de décomposition permettant l'identification de ces modèles. D'un point de vue plus technologique, le but est d'apprendre dans quelles conditions on peut exploiter ces méthodes pour d'une part le **codage**, d'autre part la **séparation aveugle** de signaux sonores. Mais il s'agit aussi de **favoriser l'usage et la dissémination** des méthodes d'approximation parcimonieuse structurées dans la communauté du traitement du signal sonore. C'est pourquoi le projet se conclura par des **journées** sur les "Modèles parcimonieux, modèles structurés et décompositions adaptatives pour le traitement du signal audio", si possible co-organisée avec le GDR ISIS, afin de **diffuser le plus largement auprès des industriels et de la communauté STIC** les connaissances sur les possibilités et les limites de ces outils.

Si l'approximation de signaux par des combinaisons de peu d'atomes choisis dans un dictionnaire redondant a prouvé son efficacité pratique (notamment pour le codage bas-débit d'images et la séparation de sources sonores), en revanche les résultats théoriques à son sujet sont récents et peu nombreux, et son utilisation pour le codage de signaux audio soulève encore des problèmes aussi bien théoriques que pratiques.

Pour une classe donnée de signaux telle que les signaux musicaux, les signaux de parole ou plus généralement les signaux audio, la plupart des techniques actuelles d'analyse ou de codage sont basées sur l'utilisation de transformées de type Fourier à court terme, dont l'échelle d'analyse (la taille de la fenêtre) est essentiellement fixée. Cependant, ces signaux sont constitués de plusieurs **couches** superposées dont les **échelles** sont assez différentes, comme les parties "stationnaires" (ou "tonales") et les transitoires, et un certain nombre de travaux montrent l'intérêt des représentations adaptatives à base de **dictionnaires redondants** (tels que l'union d'une base d'ondelette et d'une base de Fourier locale) pour traiter de tels signaux.

Parmi l'infinité d'approximations

$$x(t) \approx \sum_{k \in I} c_k g_k(t) \tag{1}$$

d'un signal $x(t)$ comme combinaison linéaire d'atomes d'un dictionnaire redondant $\mathcal{D} := \{g_k(t), k \in K\}$ (surdimensionné par rapport à une base), on aimerait trouver la plus **parcimonieuse** (*i.e* avec le plus petit nombre d'atomes card(I)) pour une distortion maximale admissible, mais sa recherche est en général un problème de **complexité combinatoire**. En pratique, on doit donc recourir à des algorithmes heuristiques tels que les poursuites (Matching Pursuit et Basis Pursuit), *a priori* sous-optimales. Récemment, des résultats théoriques sont venu étayer les arguments heuristiques pour l'emploi de ces algorithmes: si le signal admet une "bonne" approximation avec "suffisament peu" d'atomes, celle-ci est unique –le modèle parcimonieux est donc **identifiable**– et essentiellement retrouvée par les algorithmes en question. Cependant, les hypothèses qu'il faut imposer pour prouver l'identifiabilité sont assez restrictives, notamment en termes de parcimonie et de qualité supposée de la "bonne" approximation.

Pour un dictionnaire qui possède une certaine **structure**, la prise en compte de cette structure et de celle de l'approximant peut s'avérer payante pour prouver l'identifiabilité et caractériser le comportement des algorithmes. Ainsi, pour l'union de bases orthonormales incohérentes [5], nous avons prouvé via des principes d'incertitude généralisés que le modèle parcimonieux demeure identifiable sous des conditions beaucoup plus faibles que la contrainte générale de parcimonie. Par exemple, dans le cas d'une union de deux bases incohérentes telles que la base de Dirac $\{\delta_k\}$ et celle de Fourier discrète $\{e_k\}$, dont la cohérence M est minimale et vaut $1/\sqrt{N}$ où N est la dimension de l'espace signal, le modèle est identifiable dès que le signal peut s'écrire $x(t) = \sum_{k \in I_1} c_k \delta_k(t) + \sum_{k \in I_2} c_k e_k(t)$ où la taille $K_i = \text{card}(I_i)$ des "couches" de type Dirac et Fourier satifait $2MK_2/(1 + MK_2) < 1/(1 + MK_1)$.

Dans l'optique du **codage**, des travaux menés au LATP [10, 11] ont également proposé d'imposer des contraintes de structure sur les approximants utilisés. L'heuristique est double: d'une part, le codage d'une représentation structurée est plus efficace (on dépense moins de bits pour décrire *quels* atomes employer); d'autre part, la représentation obtenue est plus robuste à l'ajout de bruit au signal. L'approximation sous la forme $x(t) = x_{\text{tonal}}(t) + x_{\text{transitoire}}(t) + x_{\text{residuel}}(t)$ a alors lieu en deux temps: la structure "tonale" est estimée et retranchée au signal, puis c'est le tour de la structure "transitoire". Les modèles de structures utilisés reposent sur des chaînes de Markov cachées (pour la persistance temporelle de lignes harmoniques dans la base de Fourier) et des arbres de Markov cachés (pour la structure localisée des transitoires dans la base d'ondelettes). Une prédiction des performances en termes de courbe débit-distortion est alors possible, mais l'une des grandes difficultés, non résolue à ce jour, est d'estimer la proportion relative tonale / transitoire pour allouer un budget de bits approprié à chacune des couches. Un "indice de transitoirité" récemment proposé [12] semble pouvoir servir à cet effet mais son comportement n'est pas encore bien contrôlé au niveau théorique. Par ailleurs

La recherche d'approximations à la fois parcimonieuses et structurées de signaux sonores est également très prometteuse pour la séparation "aveugle" de sources, où les méthodes fondées sur la parcimonie permettent de traiter le cas –naturel en audio– où l'on a plus de sources que de capteurs [2]. Il s'agit alors d'effectuer une approximation parcimonieuse *simultanée* des différents signaux (typiquement gauche/droite en stereo) constituant le mélange observé. L'expérience montre également que ces approximations/représentations peuvent fournir des outils pour "sculpter" un signal sonore [9, 14], c'est-à-dire le transformer de manière fine (changer l'intensité relative d'un instrument par rapport à un autre, transposer la hauteur d'un chanteur sans changer celle de l'orchestre) et en extraire des informations pour l'indexation audio.

Verrous. Des contraintes de structure permettant l'**identifiabilité en présence de bruit** restent à obtenir. On partira du cas d'une union de bases orthonormales sans bruit pour s'attaquer ensuite au problème avec le dictionnaire le plus prometteur en pratique, l'union d'un base d'ondelettes et d'une base de cosinus locaux. L'ajout de *chirplets* –qui sont peu corrélés avec les ondelettes et les cosinus locaux et bien adaptés à la représentation des signaux vocaux– est prometteur mais soulève son lot de difficultés.

Dans le cadre du codage, l'**analyse des algorithmes d'approximation structurée** intégrant des modèles de Markov est en cours. Il faut prédire leur comportement lorsque le signal analysé est "suffisament bien" approché par un approximant "suffisament structuré". Dans ce même cadre, une **analyse débit-distorsion** pour l'estimateur de maximum de vraisemblance reste à développer. Le problème du remplacement de l'algorithme d'estimation successive des composantes par un algorithme d'estimation conjointe de ces composantes reste entier.

En termes de **séparation de sources**, les performances des méthodes de séparation par la parcimonie sont aujourd'hui expliquées empiriquement, mais la **preuve de l'identifiabilité du modèle** sous des hypothèses de parcimonie+structure adéquates n'est plus loin. L'**analyse statistique** (en termes de bornes de Cramer-Rao et d'inégalités de type "oracle" notamment) sur l'erreur d'estimation dans le cas sous-déterminé et en présence de bruit est sensiblement plus difficile.

3.2 Situation dans le contexte national et international:

Ce projet se situe dans un contexte international très dynamique où l'activité s'intensifie autour de l'étude mathématique et statistique des représentations parcimonieuses de signaux avec des dictionnaires (D. Donoho à Stanford, V. Temlyakov et R. De Vore à l'Université de Caroline du Sud, M. Nielsen à l'Université d'Aalborg) et de ses applications en audio (S. Godsill et P. Wolfe du Signal Processing Group de Cambridge), en image et video (P. Vandergheynst du Laboratoire de Traitement des Signaux de l'EPFL) et en séparation de sources (M. Zibulevsky et Y. Zeevi du Technion Institute). Les équipes participantes à ce projet ont des liens étroits avec celles de Cambridge et de l'EPFL via le réseau européen de formation par la recherche HASSIP (Harmonic Analysis and Statistics in Signal and Image Processing). Ce réseau a notamment organisé en décembre 2003 à Marseille une mini-école "Approximations non-linéaires" à destination de doctorants et post-doctorants, et organisera également en septembre prochain à Cambridge sous le même format une rencontre dédiée aux méthodes de filtrage particulaire et au traitement du signal audio. Dans le cadre de ce réseau, une collaboration a par ailleurs été engagée avec H. Führ et F. Friedrichs du GSF à Munich sur l'estimation conjointe "tonal/transitoire" dans des modèles hybrides markoviens. Le projet s'inscrit également dans une dynamique nationale dont témoigne la journée "Décompositions adaptatives de signaux" organisée par le GDR ISIS début avril 2004. Le GDR ISIS a par ailleurs financé en 2003 un projet "Jeunes Chercheurs" piloté par Rémi Gribonval sur les "Ressources pour la séparation de signaux audiophoniques" dont les résultats (notamment en termes de données et de méthodes d'évaluation) seront exploités pour la partie de ce projet dédiée à la séparation de signaux sonores.

3.3 Motivations et historique de la collaboration visée:

Les équipes qui formulent aujourd'hui ce projet ont commencé à travailler sur les approximations adaptatives il y a une dizaine d'années, avec des liens forts dès l'origine (L. Daudet a effectué sa thèse sous la direction de B. Torrésani) ou qui se sont peu à peu noués via des jurys de thèse, visites ponctuelles, participations communes au réseau européen (RTN) HASSIP et au GDR ISIS. L'appel d'offres MathSTIC répond bien à notre désir de faire avancer une ambition et un projet commun, pour développer et disséminer la thématique des approximations parcimonieuses et structurées de signaux sonores. Nous trouvons particulièrement adapté de nouer cette collaboration sous la forme d'un projet resserré, porté par nos seules trois équipes, car c'est ainsi que se noue le plus efficacement un "noyau dur" où la synergie n'est pas un vain mot. Dans une deuxième phase où notre projet sera amené à prendre une envergure plus ambitieuse, avec d'autres collaborations, la présence de ce noyau dur qui aura l'habitude de travailler ensemble sera certainement un atout.

3.4 Description des champs visés, intérêt de la collaboration pour les champs visés:

Problèmes inverses mal posés, identification de systèmes. L'approximation d'un vecteur à partir d'un dictionnaire redondant est un exemple typique de problème linéaire inverse mal posé, dont la solution n'est pas unique. Les contraintes de parcimonie et/ou de structure peuvent garantir l'unicité de la solution (l'identifiabilité du modèle) et constituent une forme de régularisation. Si pour la parcimonie cette régularisation correspond à des approches classiques (dans une base de Fourier / ondelettes, la parcimonie est liée à la régularité au sens de Sobolev / Besov), la structure introduit de nouvelles formes non standard de régularisation.

Théorie de l'approximation. L'approximation sous contrainte est un champ de la théorie de l'approximation, où se pose notamment le problème des théorèmes directs (resp. inverses) d'approximation: quelles conditions de "régularité" –en un sens parfois abstrait– sont suffisantes (resp. nécessaires) pour qu'un signal puisse être approché sous contrainte avec un certain ordre de grandeur asymptotique de l'erreur d'approximation. De tels théorèmes font partie des résultats attendus de la collaboration, pour des classes à identifier de dictionnaires et de contraintes de structure.

Analyse-synthèse de signaux audio. L'analyse-synthèse de signaux audio consiste à changer la représentation d'un signal audio pour passer dans un domaine où l'extraction d'information (détection, reconnaissance d'instrument, indexation ...) et/ou la transformation (réduction de débit, changement de hauteur, de timbre, ...) sont plus aisément effectués. Les modèles parcimonieux structurés de signaux audio devraient fournir une nouvelle palette d'outils d'analyse-synthèse généralisant les modèles sinusoïdaux et le vocodeur de phase et permettant par exemple de changer la hauteur de la partie tonale sans changer le timbre des transitoires.

3.5 Résultats attendus:

Nous attendons plusieurs types de résultats très complémentaires de cette première phase de travail en commun, qui se place dans une perspective à plus long terme (trois ou quatre ans). L'ambition est qu'à terme, les outils d'approximation parcimonieuse structurée soient aussi naturels à utiliser en traitement du signal audio que peut l'être la transformée de Fourier à court terme aujourd'hui.

Les journées de rencontres prévues à l'issue du projet concrétiseront notre effort de dissémination des techniques et de structuration de la problématique. Elles serviront à amorcer une étape plus ambitieuse en termes de production technologique et scientifique et impliquant donc plus de partenaires et d'autres formes de financement. Technologiquement, il s'agira de construire un codeur audio "hybride" exploitant au maximum la parcimonie et la structure des signaux audio afin non seulement de les compresser mais aussi de les représenter sous une forme permettant leur manipulation simple par l'utilisateur final (remixage, écoute sélective d'un instrument) et leur indexation.

Sur l'année que devrait durer le projet, la production scientifique sera nécessairement d'ambition plus modeste: la collaboration interéquipe, à l'occasion de rencontres de travail courtes mais régulières où les doctorants seront très impliqués, permettra non seulement de débloquer les principaux verrous identifiés, mais aussi de partager les outils et de faire converger les approches pour faire émerger une vision commune et préciser les contours des projets pour l'avenir. Dans la mesure rendue possible par le financement du projet, l'accent sera mis sur les échanges de doctorants, qui est le garant d'une bonne dissémination technologique et scientifique. Les résultats scientifiques obtenus donneront lieu à des propositions de publications en commun (notamment dans les revues IEEE) et à des participations à des congrès. Pour répondre au besoin de dissémination, une proposition de publication à caractère plus didactique dans une revue moins spécialisée telle que le journal de de l'Audio Engineering Society sera également préparée.

Références

[1] R. Gribonval. Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps, IEEE Trans. Signal Proc. **49(5)**:994–1001 (2001).

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

[3] R. Gribonval et E. Bacry. Harmonic Decomposition of Audio Signals with Matching Pursuit, IEEE Trans. Signal Proc. **51(1)**:101–111 (2003).

[4] C. Jutten et R. Gribonval. L'analyse en composantes indépendantes: un outil puissant pour le traitement de l'information, Actes du colloque GRETSI03, Paris, (septembre 2003).

[5] R. Gribonval et M. Nielsen. Sparse decompositions in unions of bases, IEEE Trans. Inform. Theory, **49(12)**:3320–3325 (2003).

[6] R. Gribonval et M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure, *Appl. Comp. Harmonic Anal.* soumis (octobre 2003).

[7] R. Gribonval et M. Nielsen. On approximation with spline generated framelets, Constructive Approximation, **20(2)**:207–232 (2004).

[8] R. Gribonval et M. Nielsen. Nonlinear approximation with dictionaries. I. Direct estimates., J. Fourier Anal. and Appl., **10(1)**, *à paraître* (2004).

[9] F. Jaillet et B. Torrésani. Sculpture temps-fréquence des sons par modification de leur transformée en ondelettes, Actes du Congrès Français d'Acoustique (2002).

[10] L. Daudet et B. Torrésani. Hybrid Representations for Audiophonic Signal Encoding. *Signal Processing*, special issue on Image and Video Coding beyond Standards, **82** (2002), pp. 1595-1617.

[11] S. Molla et B. Torrésani. Hybrid audio scheme using hidden Markov models of waveforms. *Applied and Computational Harmonic Analysis*, soumis (Nov. 2003).

[12] S. Molla et B. Torrésani. Determining local transientness in Audio signals *IEEE Signal Processing Letters*, à paraître.

[13] S. Molla et B. Torrésani. Hidden Markov Trees of Wavelet Coefficients for Transient Detection in Audiophonic Signals, Actes de la conférence *Self-Similarity and Applications*, Annales de l'Université Blaise Pascal, A. Benassi Ed. (2002).

[14] F. Jaillet et B. Torrésani. Remarques sur l'adaptativité de représentations temps-fréquence. Actes du colloque GRETSI03, Paris (Septembre 2003).

[15] L. Daudet, S. Molla et B. Torrésani. Towards a hybrid audio coder. Actes de la conférence *Wavelets and Applications*, Chongqing, Chine (2004).

[16] B. Torrésani. Hybrid audio models and transform coding. Proceedings of the conference "Wavelets and coherent states", Louvain la Neuve, Belgium (2003).

[17] F. Jaillet et B. Torrésani. Adaptive time-frequency representations for sound analysis and processing. Proceedings of the joint symposium SFA/DAGA, Strasbourg (2004).

[18] M. Davies et L. Daudet. Fast sparse subband decomposition using FIRSP. soumis à EUropean SIgnal Processing COnference (EUSIPCO) (2004).