IN PARTNERSHIP WITH:
CNRS
Université Rennes 1

Activity Report 2011

Project-Team METISS

Speech and audio data modeling and processing

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Audio, Speech, and Language Processing

Table of contents

Project-Team METISS

Keywords: Audio, Speech, Sparse Representations, Recognition, Statistical Methods, Signal Processing, Machine Learning, Perception

1. Members

Research Scientists

Frédéric Bimbot [Team Leader, Senior Researcher (DR2) CNRS, HdR]
Nancy Bertin [Junior Researcher (CR2) CNRS]
Rémi Gribonval [Senior Researcher (DR2) INRIA, HdR]
Emmanuel Vincent [Junior Researcher (CR1) INRIA]

Technical Staff

Grégoire Bachman [Contractual R&D Engineer - Since February 2011]
Yannick Benezeth [Contractual R&D Engineer - Until August 2011]
Charles Blandin [Contractual Development Engineer - Until February 2011]
Laurence Catanese [Contractual R&D Engineer - Since November 2011]
Valentin Emiya [Contractual R&D Engineer - Until August 2011]
Jules Espiau de Lamaestre [Contractual R&D Engineer]
Ronan Le Boulch [Contractual R&D Engineer]
Guylaine Le Jan [Contractual R&D Engineer - Since February 2011]
Armando Muscariello [Contractual R&D Engineer - Until September 2011 (formerly PhD Student)]
Sangnam Nam [Contractual R&D Engineer - Since September 2011 (formerly Post-Doc)]
Alexey Ozerov [Contractual R&D Engineer - Until October 2011]
Nathan Souviraà-Labastie [Contractual R&D Engineer - Since March 2011]

PhD Students

Alexis Benichoux [MENRT Grant, 2nd year]
Anthony Bourrier [Technicolor, 1st Year]
Quang Khanh Ngoc Duong [INRIA Cordi Grant, Defended October 2011]
Nobutaka Ito [Franco-Japanese Doctoral College, 3rd Year]
Gabriel Sargent [MENRT Grant - 2nd year]
Prasad Sudhakar [INRIA Cordis Grant - Defended March 2011]
Stefan Ziegler [CNRS & Regional Grant, 1st Year]

Post-Doctoral Fellows

Kamil Adiloglu [INRIA]
Stanislaw Raczynski [INRIA - Since December 2011]
Laurent Simon [INRIA - Since April 2011]
Nikolaos Stefanakis [INRIA]

Administrative Assistant

Stéphanie Lemaile

2. Overall Objectives

2.1. Presentation

The research interests of the METISS group are centered on audio, speech and music signal processing and cover a number of problems ranging from sensing, analysis and modelling sound signals to detection, classification and structuration of audio content.

Primary focus is put on information detection and tracking in audio streams, speech and speaker recognition, music analysis and modeling, source separation and "advanced" approaches for audio signal processing such as compressive sensing. All these objectives contribute to the more general area of audio scene analysis.

The main industrial sectors in relation with the topics of the METISS research group are the telecommunication sector, the Internet and multimedia sector, the musical and audiovisual production sector, and, marginally, the sector of education and entertainment.

On a regular basis, METISS is involved in bilateral or multilateral partnerships, within the framework of consortia, networks, thematic groups, national and European research projects, as well as industrial contracts with various local companies.

2.2. Highlights

Rémi Gribonval was awarded the 2011 Blaise Pascal Award of the GAMNI-SMAI by the French Academy of Sciences.

Rémi Gribonval obtained in 2011 a Starting Grant from the European Research Council.

Our group organized and participated in the PASCAL 'CHiME' Speech Separation and Recognition Challenge, aiming to evaluate speech separation, feature extraction and speech recognition algorithms in everyday listening conditions. The challenge attracted 13 groups worldwide, which is a major success compared to previous events in this field. For datasets and detailed results, please see http://www.dcs.shef.ac.uk/spandh/chime/challenge.html.

For his contributions to the field, Emmanuel Vincent will be awarded the 2012 SPIE ICA Unsupervised Learning Pioneer Award.

3. Scientific Foundations

3.1. Introduction

Probabilistic approaches offer a general theoretical framework [114] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [111], the probabilistic framework indeed provides a solid formalism which makes it possible to formulate various problems of segmentation, detection and classification. Coupled to statistical approaches, the probabilistic paradigm makes it possible to easily adapt relatively generic tools to various applicative contexts, thanks to estimation techniques for training from examples.

A particularily productive family of probabilistic models is the Hidden Markov Model, either in its general form or under some degenerated variants. The stochastic framework makes it possible to rely on well-known algorithms for the estimation of the model parameters (EM algorithms, ML criteria, MAP techniques, ...) and for the search of the best model in the sense of the exact or approximate maximum likelihood (Viterbi decoding or beam search, for example).

More recently, Bayesian networks [116] have emerged as offering a powerful framework for the modeling of musical signals (for instance, [112], [117]).

In practice, however, the use of probabilistic models must be accompanied by a number of adjustments to take into account problems occurring in real contexts of use, such as model inaccuracy, the insufficiency (or even the absence) of training data, their poor statistical coverage, etc...

Another focus of the activities of the METISS research group is dedicated to sparse representations of signals in redundant systems [115]. The use of criteria of sparsity or entropy (in place of the criterion of least squares) to force the unicity of the solution of a underdetermined system of equations makes it possible to seek an economical representation (exact or approximate) of a signal in a redundant system, which is better able to account for the diversity of structures within an audio signal.

The topic of sparse representations opens a vast field of scientific investigation : sparse decomposition, sparsity criteria, pursuit algorithms, construction of efficient redundant dictionaries, links with the non-linear approximation theory, probabilistic extensions, etc... and more recently, compressive sensing [110]. The potential applicative outcomes are numerous.

This section briefly exposes these various theoretical elements, which constitute the fundamentals of our activities.

3.2. Probabilistic approach

For several decades, the probabilistic approaches have been used successfully for various tasks in pattern recognition, and more particularly in speech recognition, whether it is for the recognition of isolated words, for the retranscription of continuous speech, for speaker recognition tasks or for language identification. Probabilistic models indeed make it possible to effectively account for various factors of variability occuring in the signal, while easily lending themselves to the definition of metrics between an observation and the model of a sound class (phoneme, word, speaker, etc...).

3.2.1. Probabilistic formalism and modeling

The probabilistic approach for the representation of an (audio) class X relies on the assumption that this class can be described by a probability density function (PDF) $P(.|X)$ which associates a probability $P(Y|X)$ to any observation Y.

In the field of speech processing, the class X can represent a phoneme, a sequence of phonemes, a word from a vocabulary, or a particular speaker, a type of speaker, a language, ... Class X can also correspond to other types of sound objects, for example a family of sounds (word, music, applause), a sound event (a particular noise, a jingle), a sound segment with stationary statistics (on both sides of a rupture), etc.

In the case of audio signals, the observations Y are of an acoustical nature, for example vectors resulting from the analysis of the short-term spectrum of the signal (filter-bank coefficients, cepstrum coefficients, time-frequency principal components, etc.) or any other representation accounting for the information that is required for an efficient separation of the various audio classes considered.

In practice, the PDF P is not accessible to measurement. It is therefore necessary to resort to an approximation $\widehat{P}$ of this function, which is usually refered to as the likelihood function. This function can be expressed in the form of a parametric model.

The models most used in the field of speech and audio processing are the Gaussian Model (GM), the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM). But recently, more general models have been considered and formalised as graphical models.

Choosing a particular family of models is based on a set of considerations ranging from the general structure of the data, some knowledge on the audio class making it possible to size the model, the speed of calculation of the likelihood function, the number of degrees of freedom of the model compared to the volume of training data available, etc.

3.2.2. Statistical estimation

The determination of the model parameters for a given class is generally based on a step of statistical estimation consisting in determining the optimal value for model parameters.

The Maximum Likelihood (ML) criterion is generally satisfactory when the number of parameters to be estimated is small w.r.t. the number of training observations. However, in many applicative contexts, other estimation criteria are necessary to guarantee more robustness of the learning process with small quantities of training data. Let us mention in particular the Maximum a Posteriori (MAP) criterion which relies on a prior probability of the model parameters expressing possible knowledge on the estimated parameter distribution for the class considered. Discriminative training is another alternative to these two criteria, definitely more complex to implement than the ML and MAP criteria.

In addition to the fact that the ML criterion is only one particular case of the MAP criterion, the MAP criterion happens to be experimentally better adapted to small volumes of training data and offers better generalization capabilities of the estimated models (this is measured for example by the improvement of the classification performance and recognition on new data). Moreover, the same scheme can be used in the framework of incremental adaptation, i.e. for the refinement of the parameters of a model using new data observed for instance, in the course of use of the recognition system.

3.2.3. Likelihood computation and state sequence decoding

During the recognition phase, it is necessary to evaluate the likelihood function of the observations for one or several models. When the complexity of the model is high, it is generally necessary to implement fast calculation algorithms to approximate the likelihood function.

In the case of HMM models, the evaluation of the likelihood requires a decoding step to find the most probable sequence of hidden states. This is done by implementing the Viterbi algorithm, a traditional tool in the field of speech recognition. However, when the acoustic models are combined with a syntagmatic model, it is necessary to call for sub-optimal strategies, such as beam search.

3.2.4. Bayesian decision

When the task to solve is the classification of an observation into one class among several closed-set possibilities, the decision usually relies on the maximum a posteriori rule.

In other contexts (for instance, in speaker verification, word-spotting or sound class detection), the problem of classification can be formulated as a binary hypotheses testing problem, consisting in deciding whether the tested observation is more likely to be pertaining to the class under test or not pertaining to it. In this case, the decision consists in acceptance or rejection, and the problem can be theoretically solved within the framework of Bayesian decision by calculating the ratio of the PDFs for the class and the non-class distributions, and comparing this ratio to a decision threshold.

In theory, the optimal threshold does not depend on the class distribution, but in practice the quantities provided by the probabilistic models are not the true PDFs, but only likelihood functions which approximate the true PDFs more or less accurately, depending on the quality of the model of the class.

The optimal threshold must be adjusted for each class by modeling the behaviour of the test on external (development) data.

3.2.5. Graphical models

In the past years, increasing interest has focused on graphical models for multi-source audio signals, such as polyphonic music signals. These models are particularly interesting, since they enable a formulation of music modelisation in a probabilistic framework.

It makes it possible to account for more or less elaborate relationship and dependencies between variables representing multiple levels of description of a music piece, together with the exploitation of various priors on the model parameters.

Following a well-established metaphore, one can say that the graphical model expresses the notion of modularity of a complex system, while probability theory provides the glue whereby the parts are combined. Such a data structure lends itself naturally to the design of efficient general-purpose algorithms.

The graphical model framework provides a way to view a number of existing models (including HMMs) as instances of a common formalism and all of them can be addressed via common machine learning tools.

A first issue when using graphical models is the one of the model design, i.e. the chosen variables for parameterizing the signal, their priors and their conditional dependency structure.

The second problem, called the inference problem, consists in estimating the activity states of the model for a given signal in the maximum a posteriori sense. A number of techniques are available to achieve this goal (sampling methods, variational methods belief propagation, ...), whose challenge is to achieve a good compromise between tractability and accuracy [116].

3.3. Sparse representations

Over the past decade, there has been an intense and interdisciplinary research activity in the investigation of sparsity and methods for sparse representations, involving researchers in signal processing, applied mathematics and theoretical computer science. This has led to the establishment of sparse representations as a key methodology for addressing engineering problems in all areas of signal and image processing, from the data acquisition to its processing, storage, transmission and interpretation, well beyond its original applications in enhancement and compression. Among the existing sparse approximation algorithms, L1-optimisation principles (Basis Pursuit, LASSO) and greedy algorithms (e.g., Matching Pursuit and its variants) have in particular been extensively studied and proved to have good decomposition performance, provided that the sparse signal model is satisfied with sufficient accuracy.

The large family of audio signals includes a wide variety of temporal and frequential structures, objects of variable durations, ranging from almost stationary regimes (for instance, the note of a violin) to short transients (like in a percussion). The spectral structure can be mainly harmonic (vowels) or noise-like (fricative consonants). More generally, the diversity of timbers results in a large variety of fine structures for the signal and its spectrum, as well as for its temporal and frequential envelope. In addition, a majority of audio signals are composite, i.e. they result from the mixture of several sources (voice and music, mixing of several tracks, useful signal and background noise). Audio signals may have undergone various types of distortion, recording conditions, media degradation, coding and transmission errors, etc.

Sparse representations provide a framework which has shown increasingly fruitful for capturing, analysing, decomposing and separating audio signals

3.3.1. Redundant systems and adaptive representations

Traditional methods for signal decomposition are generally based on the description of the signal in a given basis (i.e. a free, generative and constant representation system for the whole signal). On such a basis, the representation of the signal is unique (for example, a Fourier basis, Dirac basis, orthogonal wavelets, ...). On the contrary, an adaptive representation in a redundant system consists of finding an optimal decomposition of the signal (in the sense of a criterion to be defined) in a generating system (or dictionary) including a number of elements (much) higher than the dimension of the signal.

Let y be a monodimensional signal of length T and D a redundant dictionary composed of $N > T$ vectors g_i of dimension T.

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

If D is a generating system of R^T, there is an infinity of exact representations of y in the redundant system D, of the type:

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

We will denote as $\alpha = \{\alpha_i\}_{1 \leq i \leq N}$, the N coefficients of the decomposition.

The principles of the adaptive decomposition then consist in selecting, among all possible decompositions, the best one, i.e. the one which satisfies a given criterion (for example a sparsity criterion) for the signal under consideration, hence the concept of adaptive decomposition (or representation). In some cases, a maximum of T coefficients are non-zero in the optimal decomposition, and the subset of vectors of D thus selected are refered to as the basis adapted to y. This approach can be extended to approximate representations of the type:

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

with $M < T$, where ϕ is an injective function of $[1, M]$ in $[1, N]$ and where $e(t)$ corresponds to the error of approximation to M terms of $y(t)$. In this case, the optimality criterion for the decomposition also integrates the error of approximation.

3.3.2. Sparsity criteria

Obtaining a single solution for the equation above requires the introduction of a constraint on the coefficients α_i. This constraint is generally expressed in the following form :

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

Among the most commonly used functions, let us quote the various functions L_γ :

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

Let us recall that for $0 < \gamma < 1$, the function L_γ is a sum of concave functions of the coefficients α_i. Function L_0 corresponds to the number of non-zero coefficients in the decomposition.

The minimization of the quadratic norm L_2 of the coefficients α_i (which can be solved in an exact way by a linear equation) tends to spread the coefficients on the whole collection of vectors in the dictionary. On the other hand, the minimization of L_0 yields a maximally parsimonious adaptive representation, as the obtained solution comprises a minimum of non-zero terms. However the exact minimization of L_0 is an untractable NP-complete problem.

An intermediate approach consists in minimizing norm L_1, i.e. the sum of the absolute values of the coefficients of the decomposition. This can be achieved by techniques of linear programming and it can be shown that, under some (strong) assumptions the solution converges towards the same result as that corresponding to the minimization of L_0. In a majority of concrete cases, this solution has good properties of sparsity, without reaching however the level of performance of L_0.

Other criteria can be taken into account and, as long as the function F is a sum of concave functions of the coefficients α_i, the solution obtained has good properties of sparsity. In this respect, the entropy of the decomposition is a particularly interesting function, taking into account its links with the information theory.

Finally, let us note that the theory of non-linear approximation offers a framework in which links can be established between the sparsity of exact decompositions and the quality of approximate representations with M terms. This is still an open problem for unspecified redundant dictionaries.

3.3.3. Decomposition algorithms

Three families of approaches are conventionally used to obtain an (optimal or sub-optimal) decomposition of a signal in a redundant system.

The “Best Basis” approach consists in constructing the dictionary D as the union of B distinct bases and then to seek (exhaustively or not) among all these bases the one which yields the optimal decomposition (in the sense of the criterion selected). For dictionaries with tree structure (wavelet packets, local cosine), the complexity of the algorithm is quite lower than the number of bases B, but the result obtained is generally not the optimal result that would be obtained if the dictionary D was taken as a whole.

The “Basis Pursuit” approach minimizes the norm L_1 of the decomposition resorting to linear programming techniques. The approach is of larger complexity, but the solution obtained yields generally good properties of sparsity, without reaching however the optimal solution which would have been obtained by minimizing L_0.

The “Matching Pursuit” approach consists in optimizing incrementally the decomposition of the signal, by searching at each stage the element of the dictionary which has the best correlation with the signal to be decomposed, and then by subtracting from the signal the contribution of this element. This procedure is repeated on the residue thus obtained, until the number of (linearly independent) components is equal to the dimension of the signal. The coefficients α can then be reevaluated on the basis thus obtained. This greedy algorithm is sub-optimal but it has good properties for what concerns the decrease of the error and the flexibility of its implementation.

Intermediate approaches can also be considered, using hybrid algorithms which try to seek a compromise between computational complexity, quality of sparsity and simplicity of implementation.

3.3.4. Dictionary construction

The choice of the dictionary D has naturally a strong influence on the properties of the adaptive decomposition : if the dictionary contains only a few elements adapted to the structure of the signal, the results may not be very satisfactory nor exploitable.

The choice of the dictionary can rely on a priori considerations. For instance, some redundant systems may require less computation than others, to evaluate projections of the signal on the elements of the dictionary. For this reason, the Gabor atoms, wavelet packets and local cosines have interesting properties. Moreover, some general hint on the signal structure can contribute to the design of the dictionary elements : any knowledge on the distribution and the frequential variation of the energy of the signals, on the position and the typical duration of the sound objects, can help guiding the choice of the dictionary (harmonic molecules, chirplets, atoms with predetermined positions, ...).

Conversely, in other contexts, it can be desirable to build the dictionary with data-driven approaches, i.e. training examples of signals belonging to the same class (for example, the same speaker or the same musical instrument, ...). In this respect, Principal Component Analysis (PCA) offers interesting properties, but other approaches can be considered (in particular the direct optimization of the sparsity of the decomposition, or properties on the approximation error with M terms) depending on the targeted application.

In some cases, the training of the dictionary can require stochastic optimization, but one can also be interested in EM-like approaches when it is possible to formulate the redundant representation approach within a probabilistic framework.

Extension of the techniques of adaptive representation can also be envisaged by the generalization of the approach to probabilistic dictionaries, i.e. comprising vectors which are random variables rather than deterministic signals. Within this framework, the signal $y(t)$ is modeled as the linear combination of observations emitted by each element of the dictionary, which makes it possible to gather in the same model several variants of the same sound (for example various waveforms for a noise, if they are equivalent for the ear). Progress in this direction are conditioned to the definition of a realistic generative model for the elements of the dictionary and the development of effective techniques for estimating the model parameters.

3.3.5. Compressive sensing

The theoretical results around sparse representations have laid the foundations for a new research field called compressed sensing, emerging primarily in the USA. Compressed sensing investigates ways in which we can sample signals at roughly the lower information rate rather than the standard Shannon-Nyquist rate for sampled signals.

In a nutshell, the principle of Compressed Sensing is, at the acquisition step, to use as samples a number of random linear projections. Provided that the underlying phenomenon under study is sufficiently sparse, it is possible to recover it with good precision using only a few of the random samples. In a way, Compressed Sensing can be seen as a generalized sampling theory, where one is able to trade bandwidth (i.e. number of samples) with computational power. There are a number of cases where the latter is becoming much more accessible than the former; this may therefore result in a significant overall gain, in terms of cost, reliability, and/or precision.

4. Application Domains

4.1. Introduction

This section reviews a number of applicative tasks in which the METISS project-team is particularily active :

- spoken content processing
- description of audio streams
- audio scene analysis
- advanced processing for music information retrieval

The main applicative fields targeted by these tasks are :

- multimedia indexing
- audio and audio-visual content repurposing
- description and exploitation of musical databases
- ambient intelligence
- education and leisure

4.2. Spoken content processing

A number of audio signals contain speech, which conveys important information concerning the document origin, content and semantics. The field of speaker characterisation and verification covers a variety of tasks that consist in using a speech signal to determine some information concerning the identity of the speaker who uttered it.

In parallel, METISS maintains some know-how and develops new research in the area of acoustic modeling of speech signals and automatic speech transcription, mainly in the framework of the semantic analysis of audio and multimedia documents.

4.2.1. Robustness issues in speaker recognition

Speaker recognition and verification has made significant progress with the systematical use of probabilistic models, in particular Hidden Markov Models (for text-dependent applications) and Gaussian Mixture Models (for text-independent applications). As presented in the fundamentals of this report, the current state-of-the-art approaches rely on bayesian decision theory.

However, robustness issues are still pending : when speaker characteristics are learned on small quantities of data, the trained model has very poor performance, because it lacks generalisation capabilities. This problem can partly be overcome by adaptation techniques (following the MAP viewpoint), using either a speaker-independent model as general knowledge, or some structural information, for instance a dependency model between local distributions.

METISS also adresses a number of topics related to speaker characterisation, in particular speaker selection (i.e. how to select a representative subset of speakers from a larger population), speaker representation (namely how to represent a new speaker in reference to a given speaker population), speaker adaptation for speech recognition, and more recently, speaker's emotion detection.

4.2.2. Speech recognition for multimedia analysis

In multimodal documents, the audio track is generally a major source of information and, when it contains speech, it conveys a high level of semantic content. In this context, speech recognition functionalities are essential for the extraction of information relevant to the taks of content indexing.

As of today, there is no perfect technology able to provide an error-free speech retranscription and operating for any type of speech input. A current challenge is to be able to exploit the imperfect output of an Automatic Speech Recognition (ASR) system, using for instance Natural Language Processing (NLP) techniques, in order to extract structural (topic segmentation) and semantic (topic detection) information from the audio track.

Along the same line, another scientific challenge is to combine the ASR output with other sources of information coming from various modalities, in order to extract robust multi-modal indexes from a multimedia content (video, audio, textual metadata, etc...).

4.3. Description and structuration of audio streams

Automatic tools to locate events in audio documents, structure them and browse through them as in textual documents are key issues in order to fully exploit most of the available audio documents (radio and television programmes and broadcasts, conference recordings, etc).

In this respect, defining and extracting meaningful characteristics from an audio stream aim at obtaining a structured representation of the document, thus facilitating content-based access or search by similarity.

Activities in METISS focus on sound class and event characterisation and tracking in audio contents for a wide variety of features and documents.

4.3.1. Detecting and tracking sound classes and events

Locating various sounds or broad classes of sounds, such as silence, music or specific events like ball hits or a jingle, in an audio document is a key issue as far as automatic annotation of sound tracks is concerned. Indeed, specific audio events are crucial landmarks in a broadcast. Thus, locating automatically such events enables to answer a query by focusing on the portion of interest in the document or to structure a document for further processing. Typical sound tracks come from radio or TV broadcasts, or even movies.

In the continuity of research carried out at IRISA for many years (especially by Benveniste, Basseville, André-Obrecht, Delyon, Seck, ...) the statistical test approach can be applied to abrupt changes detection and sound class tracking, the latter provided a statistical model for each class to be detected or tracked was previously estimated. For example, detecting speech segments in the signal can be carried out by comparing the segment likelihoods using a speech and a “non-speech” statistical model respectively. The statistical models commonly used typically represent the distribution of the power spectral density, possibly including some temporal constraints if the audio events to look for show a specific time structure, as is the case with jingles or words. As an alternative to statistical tests, hidden Markov models can be used to simultaneously segment and classify an audio stream. In this case, each state (or group of states) of the automaton represent one of the audio event to be detected. As for the statistical test approach, the hidden Markov model approach requires that models, typically Gaussian mixture models, are estimated for each type of event to be tracked.

In the area of automatic detection and tracking of audio events, there are three main bottlenecks. The first one is the detection of simultaneous events, typically speech with music in a speech/music/noise segmentation problem since it is nearly impossible to estimate a model for each event combination. The second one is the not so uncommon problem of detecting very short events for which only a small amount of training data is available. In this case, the traditional 100 Hz frame analysis of the waveform and Gaussian mixture modeling suffer serious limitations. Finally, typical approaches require a preliminary step of manual annotation of a training corpus in order to estimate some model parameters. There is therefore a need for efficient machine learning and statistical parameter estimation techniques to avoid this tedious and costly annotation step.

4.3.2. Describing multi-modal information

Applied to the sound track of a video, detecting and tracking audio events can provide useful information about the video structure. Such information is by definition only partial and can seldom be exploited by itself for multimedia document structuring or abstracting. To achieve these goals, partial information from the various media must be combined. By nature, pieces of information extracted from different media or modalities are heterogeneous (text, topic, symbolic audio events, shot change, dominant color, etc.) thus making their integration difficult. Only recently approaches to combine audio and visual information in a generic framework for video structuring have appeared, most of them using very basic audio information.

Combining multimedia information can be performed at various level of abstraction. Currently, most approaches in video structuring rely on the combination of structuring events detected independently in each media. A popular way to combine information is the hierarchical approach which consists in using the results of the event detection of one media to provide cues for event detection in the other media. Application specific heuristics for decision fusions are also widely employed. The Bayes detection theory provides a powerful theoretical framework for a more integrated processing of heterogeneous information, in particular because this framework is already extensively exploited to detect structuring events in each media. Hidden Markov models with multiple observation streams have been used in various studies on video analysis over the last three years.

The main research topics in this field are the definition of structuring events that should be detected on the one hand and the definition of statistical models to combine or to jointly model low-level heterogeneous information on the other hand. In particular, defining statistical models on low-level features is a promising idea as it avoids defining and detecting structuring elements independently for each media and enables an early integration of all the possible sources of information in the structuring process.

4.3.3. Recurrent audio pattern detection

A new emerging topic is that of motif discovery in large volumes of audio data, i.e. discovering similar units in an audio stream in an unsupervised fashion. These motifs can constitue some form of audio "miniatures" which characterize some potentially salient part of the audio content : key-word, jingle, etc...

This problem naturally requires the definition of a robuste metric between audio segments, but a key issue relies in an efficient search strategy able to handle the combinatorial complexity stemming from the competition between all possible motif hypotheses. An additional issue is that of being able to model adequately the collection of instances corresponding to a same motif (in this respect, the HMM framework certainly offers a reasonable paradigm).

4.4. Advanced processing for music information retrieval

4.4.1. Music content modeling

Music pieces constitue a large part of the vast family of audio data for which the design of description and search techniques remain a challenge. But while there exist some well-established formats for synthetic music (such as MIDI), there is still no efficient approach that provide a compact, searchable representation of music recordings.

In this context, the METISS research group dedicates some investigative efforts in high level modeling of music content along several tracks. The first one is the acoustic modeling of music recordings by deformable probabilistic sound objects so as to represent variants of a same note as several realisation of a common underlying process. The second track is music language modeling, i.e. the symbolic modeling of combinations and sequences of notes by statistical models, such as n-grams.

4.4.2. Multi-level representations for music information retrieval

New search and retrieval technologies focused on music recordings are of great interest to amateur and professional applications in different kinds of audio data repositories, like on-line music stores or personal music collections.

The METISS research group is devoting increasing effort on the fine modeling of multi-instrument/multi-track music recordings. In this context we are developing new methods of automatic metadata generation from music recordings, based on Bayesian modeling of the signal for multilevel representations of its content. We also investigate uncertainty representation and multiple alternative hypotheses inference.

4.5. Audio scene analysis

Audio signals are commonly the result of the superimposition of various sources mixed together : speech and surrounding noise, multiple speakers, instruments playing simultaneously, etc...

Source separation aims at recovering (approximations of) the various sources participating to the audio mixture, using spatial and spectral criteria, which can be based either on a priori knowledge or on property learned from the mixture itself.

4.5.1. Audio source separation

The general problem of “source separation” consists in recovering a set of unknown sources from the observation of one or several of their mixtures, which may correspond to as many microphones. In the special case of *speaker separation*, the problem is to recover two speech signals contributed by two separate speakers that are recorded on the same media. The former issue can be extended to *channel separation*, which deals with the problem of isolating various simultaneous components in an audio recording (speech, music, singing voice, individual instruments, etc.). In the case of *noise removal*, one tries to isolate the “meaningful” signal, holding relevant information, from parasite noise.

It can even be appropriate to view audio compression as a special case of source separation, one source being the compressed signal, the other being the residue of the compression process. The former examples illustrate how the general source separation problem spans many different problems and implies many foreseeable applications.

While in some cases –such as multichannel audio recording and processing– the source separation problem arises with a number of mixtures which is at least the number of unknown sources, the research on audio source separation within the METISS project-team rather focusses on the so-called under-determined case. More precisely, we consider the cases of one sensor (mono recording) for two or more sources, or two sensors (stereo recording) for $n > 2$ sources.

We address the problem of source separation by combining spatial information and spectral properties of the sources. However, as we want to resort to as little prior information as possible we have designed self-learning schemes which adapt their behaviour to the properties of the mixture itself [1].

4.5.2. Compressive sensing of acoustic fields

Complex audio scene may also be dealt with at the acquisition stage, by using “intelligent” sampling schemes. This is the concept behind a new field of scientific investigation : compressive sensing of acoustic fields.

The challenge of this research is to design, implement and evaluate sensing architectures and signal processing algorithms which would enable to acquire a reasonably accurate map of an acoustic field, so as to be able to locate, characterize and manipulate the various sources in the audio scene.

5. Software

5.1. Audio signal processing, segmentation and classification toolkits

Participant: Guillaume Gravier.

Guillaume Gravier is now with the TEXMEX group but this software is being used by several members of the METISS group.

The SPro toolkit provides standard front-end analysis algorithms for speech signal processing. It is systematically used in the METISS group for activities in speech and speaker recognition as well as in audio indexing. The toolkit is developed for Unix environments and is distributed as a free software with a GPL license. It is used by several other French laboratories working in the field of speech processing.

In the framework of our activities on audio indexing and speaker recognition, AudioSeg, a toolkit for the segmentation of audio streams has been developed and is distributed for Unix platforms under the GPL agreement. This toolkit provides generic tools for the segmentation and indexing of audio streams, such as audio activity detection, abrupt change detection, segment clustering, Gaussian mixture modeling and joint segmentation and detection using hidden Markov models. The toolkit relies on the SPro software for feature extraction.

Contact : guillaume.gravier@irisa.fr
http://gforge.inria.fr/projects/spro, http://gforge.inria.fr/projects/audioseg

5.2. Irene: a speech recognition and transcription platform

Participant: Guillaume Gravier.

Guillaume Gravier is now with the TEXMEX group but this software is being used by several members of the METISS group.

In collaboration with the computer science dept. at ENST, METISS has actively participated in the past years in the development of the freely available Sirocco large vocabulary speech recognition software [113]. The Sirocco project started as an INRIA Concerted Research Action now works on the basis of voluntary contributions.

The Sirocco speech recognition software was then used as the heart of the transcription modules whithin a spoken document analysis platform called IRENE. In particular, it has been extensively used for research on ASR and NLP as well as for work on phonetic landmarks in statistical speech recognition.

In 2009, the integration of IRENE in the multimedia indexing platform of IRISA was completed, incorporating improvements benchmarked during the ESTER 2 evaluation campaign in december 2008. Additionnal improvements were alos carried out such as bandwidth segmentation and improved segment clustering for unsupervised acoustic model adaptation. The integration of IRENE in the multimedia indexing platform was mainly validated on large datasets extracted from TV streams.

Contact : guillaume.gravier@irisa.fr
http://gforge.inria.fr/projects/sirocco

5.3. MPTK: the Matching Pursuit Toolkit

Participants: Rémi Gribonval, Ronan Le Boulch.

The Matching Pursuit ToolKit (MPTK) is a fast and flexible implementation of the Matching Pursuit algorithm for sparse decomposition of monophonic as well as multichannel (audio) signals. MPTK is written in C++ and runs on Windows, MacOS and Unix platforms. It is distributed under a free software license model (GNU General Public License) and comprises a library, some standalone command line utilities and scripts to plot the results under Matlab.

MPTK has been entirely developed within the METISS group mainly to overcome limitations of existing Matching Pursuit implementations in terms of ease of maintainability, memory footage or computation speed. One of the aims is to be able to process in reasonable time large audio files to explore the new possibilities which Matching Pursuit can offer in speech signal processing. With the new implementation, it is now possible indeed to process a one hour audio signal in as little as twenty minutes.

Thanks to an INRIA software development operation (Opération de Développement Logiciel, ODL) started in September 2006, METISS efforts have been targeted at easing the distribution of MPTK by improving its portability to different platforms and simplifying its developpers' API. Besides pure software engineering improvements, this implied setting up a new website with an FAQ, developing new interfaces between MPTK and Matlab and Python, writing a portable Graphical User Interface to complement command line utilities, strengthening the robustness of the input/output using XML where possible, and most importantly setting up a whole new plugin API to decouple the core of the library from possible third party contributions.

Collaboration : Laboratoire d'Acoustique Musicale (University of Paris VII, Jussieu).

Contact : remi.gribonval@irisa.fr

http://mptk.gforge.inria.fr, http://mptk.irisa.fr

5.4. FASST

Participants: Emmanuel Vincent [correspondant], Alexey Ozerov, Frédéric Bimbot.

FASST is a Flexible Audio Source Separation Toolbox in Matlab, designed to speed up the conception and automate the implementation of new model-based audio source separation algorithms.

6. New Results

6.1. Audio and speech content processing

6.1.1. Audio motif discovery

Participants: Frédéric Bimbot, Laurence Catanese, Armando Muscariello.

This work was performed in close collaboration with Guillaume Gravier from the Texmex project-team.

As an alternative to supervised approaches for multimedia content analysis, where predefined concepts are searched for in the data, we investigate content discovery approaches where knowledge emerge from the data. Following this general philosophy, we pursued work on motif discovery in audio contents.

Audio motif discovery is the task of finding out, without any prior knowledge, all pieces of signals that repeat, eventually allowing variability. In 2011, we extended our recent work on seeded discovery to near duplicate detection and spoken document retrieval from examples. First, we proposed alogirhtmic speed ups for the discovery of near duplicate motifs (low variability) in large (several days long) audio streams, exploiting subsampling strategies [muscariello-cbmi-11]. Second, we investigated the use of previously proposed efficient pattern matching techniques to deal with motif variability in speech data [muscariello-icassp-11] in a different setting, that of spoken document retrieval from an audio example. We demonstrated the potential of model-free approaches for efficient spoken document retrieval on a variety of data sets, in particular in the framework of the Spoken Web Search task of the MediaEval 2011 international evaluation [muscariello-is-11, muscariello-mediaeval-11].

This work is carried out in the context of the Quaero Project.

6.1.2. Landmark-driven speech recognition

Participant: Stefan Ziegler.

This work is supervised by Guillaume Gravier and Bogdan Ludusan from the Texmex project-team.

Speech recognition is a key issue to access multimedia spoken contents. In this context, speech recognition faces several challenges among which robustness to acoustic and linguistic variability.

In 2011, we initiated research on landmark-driven speech recognition to increase robustness. The idea of this approach consists in accurately detecting in the signal landmarks corresponding to broad phonetic classes (vowels, nasals, etc.). These landmarks, which represent almost certain knowledge about the phonetic content of the signal, are then used to bias the search space in Viterbi decoding towards solutions consistent with the landmarks. We proposed a landmark detection system, which employs numerous attributes extracted from a segment based representation of speech. We use a decision tree for BPC classification, since this allows the evaluation of each BPC on its most informative attributes, selected from a large variety of attributes. Then, each segment is converted into a landmark and a probability estimate for each BPC is provided. Second, we extend a previously proposed landmark-driven decoding strategy by a more flexible implementation, which reinforces paths at the detected landmarks according to the obtained BPC probabilities. Results obtained on French broadcast news data show a relative improvement in word error rate of about 2 % with respect to the baseline.

6.2. Recent results on sparse representations

The team has had a substantial activity ranging from theoretical results to algorithmic design and software contributions in the field of sparse representations, which is at the core of the FET-Open European project (FP7) SMALL (Sparse Models, Algorithms and Learning for Large-Scale Data, see Section 7.2.1) and the ANR project ECHANGE (ECHantillonnage Acoustique Nouvelle GEnération, see, Section 6.3.1).

6.2.1. A new framework for sparse representations: analysis sparse models

Participants: Rémi Gribonval, Sangnam Nam.

Main collaboration: Mike Davies (Univ. Edinburgh), Michael Elad (The Technion), Hadi Zayyani (Sharif University)

In the past decade there has been a great interest in a synthesis-based model for signals, based on sparse and redundant representations. Such a model assumes that the signal of interest can be composed as a linear combination of *few* columns from a given matrix (the dictionary). An alternative *analysis-based* model can be envisioned, where an analysis operator multiplies the signal, leading to a *cosparse* outcome. Within the SMALL project, we initiated a research programme dedicated to this analysis model, in the context of a generic missing data problem (e.g., compressed sensing, inpainting, source separation, etc.). We obtained a uniqueness result for the solution of this problem, based on properties of the analysis operator and the measurement matrix. We also considered a number of pursuit algorithms for solving the missing data problem, including an L1-based and a new greedy method called GAP (Greedy Analysis Pursuit). Our simulations demonstrated the appeal of the analysis model, and the success of the pursuit techniques presented. These results have been published in international conferences [64] [63], and a journal paper is in preparation.

Our simulations demonstrated the appeal of the analysis model, and the success of the pursuit techniques presented. These results have been published in conferences [64], [91], [92] and a journal paper submitted to Applied and Computational Harmonic Analysis is under revision [103]. Other algorithms based on iterative cosparse projections [57] as well as extensions of GAP to deal with noise and structure in the cosparse representation have been developed, with applications to toy MRI reconstruction problems and acoustic source localization and reconstruction from few measurements (submitted to ICASSP 2012).

6.2.2. Theoretical results on sparse representations and dictionary learning

Participants: Rémi Gribonval, Sangnam Nam, Nancy Bertin.

Main collaboration: Karin Schnass (EPFL), Mike Davies (University of Edinburgh), Volkan Cevher (EPFL), Simon Foucart (Université Paris 5, Laboratoire Jacques-Louis Lions), Charles Soussen (Centre de recherche en automatique de Nancy (CRAN)) Jérôme Idier (Institut de Recherche en Communications et en Cybernétique de Nantes (IRCCyN)), Cédric Herzet (Equipe-projet FLUMINANCE (INRIA - CEMAGREF, Rennes)) Morten Nielsen (Department of Mathematical Sciences [Aalborg]), Gilles Puy, Pierre Vandergheynst, Yves Wiaux (EPFL) Mehrdad Yaghoobi, Rodolphe Jenatton, Francis Bach (Equipe-projet SIERRA (INRIA, Paris)) Boaz Ophir, Michael Elad (Technion) Mark D. Plumbley (Queen Mary, University of London)

Sparse recovery conditions for Orthogonal Least Squares : We pursued our investigation of conditions on an overcomplete dictionary which guarantee that certain ideal sparse decompositions can be recovered by some specific optimization principles / algorithms. This year, we extended Tropp's analysis of Orthogonal Matching Pursuit (OMP) using the Exact Recovery Condition (ERC) to a first exact recovery analysis of Orthogonal Least Squares (OLS). We showed that when ERC is met, OLS is guaranteed to exactly recover the unknown support. Moreover, we provided a closer look at the analysis of both OMP and OLS when ERC is not fulfilled. We showed that there exist dictionaries for which some subsets are never recovered with OMP. This phenomenon, which also appears with ℓ_1 minimization, does not occur for OLS. Finally, numerical experiments based on our theoretical analysis showed that none of the considered algorithms is uniformly better than the other. This work has been submitted for publication in a journal [108]

New links between the Restricted Isometry Property and nonlinear approximations : It is now well known that sparse or compressible vectors can be stably recovered from their low-dimensional projection, provided the projection matrix satisfies a Restricted Isometry Property (RIP). We establish new implications of the RIP with respect to nonlinear approximation in a Hilbert space with a redundant frame. The main ingredients of our approach are: a) Jackson and Bernstein inequalities, associated to the characterization of certain approximation spaces with interpolation spaces; b) a new proof that for overcomplete frames which satisfy a Bernstein inequality, these interpolation spaces are nothing but the collection of vectors admitting a representation in the dictionary with compressible coefficients; c) the proof that the RIP implies Bernstein inequalities. As a result, we obtain that in most overcomplete random Gaussian dictionaries with fixed aspect ratio, just as in any orthonormal basis, the error of best m-term approximation of a vector decays at a certain rate if, and only if, the vector admits a compressible expansion in the dictionary. Yet, for mildly overcomplete dictionaries with a one-dimensional kernel, we give examples where the Bernstein inequality holds, but the same inequality fails for even the smallest perturbation of the dictionary. This work has been submitted for publication in a journal [102].

Performance guarantees for compressed sensing with spread spectrum techniques : We advocate a compressed sensing strategy that consists of multiplying the signal of interest by a wide bandwidth modulation before projection onto randomly selected vectors of an orthonormal basis. Firstly, in a digital setting with random modulation, considering a whole class of sensing bases including the Fourier basis, we prove that the technique is universal in the sense that the required number of measurements for accurate recovery is optimal and independent of the sparsity basis. This universality stems from a drastic decrease of coherence between the sparsity and the sensing bases, which for a Fourier sensing basis relates to a spread of the original signal spectrum by the modulation (hence the name "spread spectrum"). The approach is also efficient as sensing matrices with fast matrix multiplication algorithms can be used, in particular in the case of Fourier measurements. Secondly, these results are confirmed by a numerical analysis of the phase transition of the l1-minimization problem. Finally, we show that the spread spectrum technique remains effective in an analog setting with chirp modulation for application to realistic Fourier imaging. We illustrate these findings in the context of radio interferometry and magnetic resonance imaging. This work has been presented at a conference [93] and accepted for publication in a journal [105].

Dictionary learning : An important practical problem in sparse modeling is to choose the adequate dictionary to model a class of signals or images of interest. While diverse heuristic techniques have been proposed in the litterature to learn a dictionary from a collection of training samples, there are little existing results which provide an adequate mathematical understanding of the behaviour of these techniques and their ability to recover an ideal dictionary from which the training samples may have been generated.

In 2008, we initiated a pioneering work on this topic, concentrating in particular on the fundamental theoretical question of the identifiability of the learned dictionary. Within the framework of the Ph.D. of Karin Schnass, we developed an analytic approach which was published at the conference ISCCSP 2008 [13] and allowed us to describe "geometric" conditions which guarantee that a (non overcomplete) dictionary is "locally identifiable" by ℓ^1 minimization.

In a second step, we focused on estimating the number of sparse training samples which is typically sufficient to guarantee the identifiability (by ℓ^1 minimization), and obtained the following result, which is somewhat surprising considering that previous studies seemed to require a combinatorial number of training samples to guarantee the identifiability: the local identifiability condition is typically satisfied as soon as the number of training samples is roughly proportional to the ambient signal dimension. The outline of the second result was published in conferences [12], [25]. These results have been published in the journal paper [15].

This year we have worked on extending the results to noisy training samples with outliers. A journal paper is in preparation, and the results will be presented at a workshop at NIPS 2011.

Analysis Operator Learning for Overcomplete Cosparse Representations : Besides standard dictionary learning, we also considered learning in the context of the cosparse model. We consider the problem of learning a low-dimensional signal model from a collection of training samples. The mainstream approach would be to learn an overcomplete dictionary to provide good approximations of the training samples using sparse synthesis coefficients. This famous sparse model has a less well known counterpart, in analysis form, called the cosparse analysis model. In this new model, signals are characterized by their parsimony in a transformed domain using an overcomplete analysis operator. We proposed two approaches to learn an analysis operator from a training corpus, both published in the conference EUSIPCO 2011 [79], [67].

The first one uses a constrained optimization program based on L1 optimization. We derive a practical learning algorithm, based on projected subgradients, and demonstrate its ability to robustly recover a ground truth analysis operator, provided the training set is of sufficient size. A local optimality condition is derived, providing preliminary theoretical support for the well-posedness of the learning problem under appropriate conditions. Extensions to deal with noisy training samples are currently investigated, and a journal paper is in preparation.

In the second approach, analysis "atoms" are learned sequentially by identifying directions that are orthogonal to a subset of the training data. We demonstrate the effectiveness of the algorithm in three experiments, treating synthetic data and real images, showing a successful and meaningful recovery of the analysis operator.

Connections between sparse approximation and Bayesian estimation: Penalized least squares regression is often used for signal denoising and inverse problems, and is commonly interpreted in a Bayesian framework as a Maximum A Posteriori (MAP) estimator, the penalty function being the negative logarithm of the prior. For example, the widely used quadratic program (with an ℓ^1 penalty) associated to the LASSO / Basis Pursuit Denoising is very often considered as MAP estimation under a Laplacian prior in the context of additive white Gaussian noise (AWGN) reduction.

A first result, which has been published in IEEE Transactions on Signal Processing [35], highlights the fact that, while this is *one* possible Bayesian interpretation, there can be other equally acceptable Bayesian interpretations. Therefore, solving a penalized least squares regression problem with penalty $\phi(x)$ need not be interpreted as assuming a prior $C \cdot \exp(-\phi(x))$ and using the MAP estimator. In particular, we showed that for *any* prior P_X, the minimum mean square error (MMSE) estimator is the solution of a penalized least square problem with some penalty $\phi(x)$, which can be interpreted as the MAP estimator with the prior $C \cdot \exp(-\phi(x))$. Vice-versa, for *certain* penalties $\phi(x)$, the solution of the penalized least squares problem is indeed the MMSE estimator, with a certain prior P_X. In general $dP_X(x) \neq C \cdot \exp(-\phi(x))dx$.

A second result, obtained in collaboration with Prof. Mike Davies and Prof. Volkan Cevher (a paper is under revision) characterizes the "compressibility" of various probability distributions with applications to underdetermined linear regression (ULR) problems and sparse modeling. We identified simple characteristics of probability distributions whose independent and identically distributed (iid) realizations are (resp. are not) compressible, i.e., that can be approximated as sparse. We prove that many priors which MAP Bayesian interpretation is sparsity inducing (such as the Laplacian distribution or Generalized Gaussian distributions with exponent p<=1), are in a way inconsistent and do not generate compressible realizations. To show this, we identify non-trivial undersampling regions in ULR settings where the simple least squares solution outperform oracle sparse estimation in data error with high probability when the data is generated from a sparsity inducing prior, such as the Laplacian distribution.

6.2.3. Wavelets on graphs

Participant: Rémi Gribonval.

Main collaboration: Pierre Vandergheynst, David Hammond (EPFL)

Within the framework of the SMALL project 7.2.1, we investigated the possibility of developping sparse representations of functions defined on graphs, by defining an extension to the traditional wavelet transform which is valid for data defined on a graph.

There are many problems where data is collected through a graph structure: scattered or non-uniform sampling, sensor networks, data on sampled manifolds or even social networks or databases. Motivated by the wealth of new potential applications of sparse representations to these problems, the partners set out a program to generalize wavelets on graphs. More precisely, we have introduced a new notion of wavelet transform for data defined on the vertices of an undirected graph. Our construction uses the spectral theory of the graph laplacian as a generalization of the classical Fourier transform. The basic ingredient of wavelets, multi-resolution, is defined in the spectral domain via operator-valued functions that can be naturally dilated. These in turn define wavelets by acting on impulses localized at any vertex. We have analyzed the localization of these wavelets in the vertex domain and showed that our multi-resolution produces functions that are indeed concentrated at will around a specified vertex. Our theory allowed us to construct an equivalent of the continuous wavelet transform but also discrete wavelet frames.

Computing the spectral decomposition can however be numerically expensive for large graphs. We have shown that, by approximating the spectrum of the wavelet generating operator with polynomial expansions, applying the forward wavelet transform and its transpose can be approximated through iterated applications of the graph Laplacian. Since in many cases the graph Laplacian is sparse, this results in a very fast algorithm. Our implementation also uses recurrence relations for computing polynomial expansions, which results in even faster algorithms. Finally, we have proved how numerical errors are precisely controlled by the properties of the desired spectral graph wavelets. Our algorithms have been implemented in a Matlab toolbox that has been released in parallel to the main theoretical article [16]. We also plan to include this toolbox in the SMALL project numerical platform.

We now foresee many applications. On one hand we will use non-local graph wavelets constructed from the set of patches in an image (or even an audio signal) to perform de-noising or in general restoration. An interesting aspect in this case, would be to understand how wavelets estimated from corrupted signals deviate from clean wavelets. In a totally different direction, we will also explore the applications of spectral graph wavelets constructed from brain connectivity graphs obtained from whole brain tractography. Our preliminary results show that graph wavelets yield a representation that is very well adapted to how the information flows in the brain along neuronal structures.

6.2.4. Algorithmic breakthrough in sparse approximation : LoCOMP

Participants: Rémi Gribonval, Frédéric Bimbot, Ronan Le Boulch.

Main collaborations: Pierre Vandergheynst (EPFL), Boris Mailhé (former team member, now with Queen Mary University, London)

Our team had already made a substantial breakthrough in 2005 when first releasing the Matching Pursuit ToolKit (MPTK, see Section 5.3) which allowed for the first time the application of the Matching Pursuit algorithm to large scale data such as hours of CD-quality audio signals. In 2008, we designed a variant of Matching Pursuit called LoCOMP (ubiquitously for LOw Complexity Orthogonal Matching Pursuit or Local Orthogonal Matching Pursuit) specifically designed for shift-invariant dictionaries. LoCOMP has been shown to achieve an approximation quality very close to that of a full Orthonormal Matching Pursuit while retaining a much lower computational complexity of the order of that of Matching Pursuit. The complexity reduction is substantial, from one day of computation to 15 minutes for a typical audio signal [20], [19]. The main effort this year has been to integrate this algorithm into MPTK to ensure its dissemination and exploitation, and a journal paper has been published [22].

6.3. Emerging activities on compressive sensing and inverse problems

6.3.1. Nearfield acoustic holography (ECHANGE ANR project)

Participants: Rémi Gribonval, Nancy Bertin.

Main collaborations: Albert Cohen (Laboratoire Jacques-Louis Lions, Université Paris 6), Laurent Daudet, Gilles Chardon, François Ollivier, Antoine Peillot (Institut Jean Le Rond d'Alembert, Université Paris 6)

Compressed sensing is a rapidly emerging field which proposes a new approach to sample data far below the Nyquist rate when the sampled data admits a sparse approximation in some appropriate dictionary. The approach is supported by many theoretical results on the identification of sparse representations in overcomplete dictionaries, but many challenges remain open to determine its range of effective applicability. METISS has chosen to focus more specifically on the exploration of Compressed Sensing of Acoustic Wavefields, and we have set up the ANR collaborative project ECHANGE (ECHantillonnage Acoustique Nouvelle GEnération) which began in January 2009. Rémi Gribonval is the coordinator of the project.

In 2010, the activity on ECHANGE has concentrated on Nearfield acoustic holography (NAH), a technique aiming at reconstructing the operational deflection shapes of a vibrating structure, from the near sound field it generates. In this application scenario, the objective is either to improve the quality of the reconstruction (for a given number of sensors), or reduce the number of sensors, or both, by exploiting a sparsity hypothesis which helps regularizing the inverse problem involved.

Contributions of the team in this task spans: notations and model definitions, experimental setting design and implementation, choice of an adapted dictionary in which the sparsity hypothesis holds, improved acquisition strategies through pseudo-random sensor arrays and/or spatial multiplexing of the inputs, experimental study of robustness issues, and theoretical study of potential success guarantees based on the restricted isometry property (which revealed being not verified in our case, despite improved experimental performance).

A paper about robustness issues and spatial multiplexing (an alternative to building antennas with random sensor position) was published in GRETSI [88]. A journal paper is under revision.

6.3.2. Sparse reconstruction for underwater acoustics (ECHANGE ANR project)

Participants: Rémi Gribonval, Valentin Emiya, Nikos Stefanakis, Nancy Bertin.

Main collaborations: Jacques Marchal, Pierre Cervenka (UPMC Univ Paris 06)

Underwater acoustic imaging is traditionally performed with beamforming: beams are formed at emission to insonify limited angular regions; beams are (synthetically) formed at reception to form the image. We proposed to exploit a natural sparsity prior to perform 3D underwater imaging using a newly built flexible-configuration sonar device. The computational challenges raised by the high-dimensionality of the problem were highlighted, and we described a strategy to overcome them. As a proof of concept, the proposed approach was used on real data acquired with the new sonar to obtain an image of an underwater target. We discussed the merits of the obtained image in comparison with standard beamforming, as well as the main challenges lying ahead, and the bottlenecks that will need to be solved before sparse methods can be fully exploited in the context of underwater compressed 3D sonar imaging. This work has been submitted to ICASSP 2012 and a journal paper is in preparation.

6.3.3. Audio inpainting (SMALL FET-Open project)

Participants: Rémi Gribonval, Valentin Emiya.

Main collaborations: Amir Adler, Michael Elad (Computer Science Department, The Technion, Israel); Maria G. Jafari, Mark D. Plumbley (Centre for Digital Music, Department of Electronic Engineering, Queen Mary University of London, U.K.).

Inpainting is a particular kind of inverse problems that has been extensively addressed in the recent years in the field of image processing. It consists in reconstructing a set of missing pixels in an image based on the observation of the remaining pixels. Sparse representations have proved to be particularly appropriate to address this problem. However, inpainting audio data has never been defined as such so far.

METISS has initiated a series of works about audio inpainting, from its definition to methods to address it. This research has begun in the framework of the EU Framework 7 FET-Open project FP7-ICT-225913-SMALL (Sparse Models, Algorithms and Learning for Large-Scale data) which began in January 2009. Rémi Gribonval is the coordinator of the project. The research on audio inpainting has been conducted by Valentin Emiya in 2010 and 2011.

The contributions consist of:

- defining audio inpainting as a general scheme where missing audio data must be estimated: it covers a number of existing audio processing tasks that have been addressed separately so far – click removal, declipping, packet loss concealment, unmasking in time-frequency;
- proposing algorithms based on sparse representations for audio inpainting (based on Matching Pursuit and on l_1 minimization);

- addressing the case of audio declipping (*i.e.* desaturation): thanks to the flexibility of our inpainting algorithms, they can be constrained so as to include the structure of signals due to clipping in the objective to optimize. The resulting performance are significantly improved. This work has been reported in [47] and it will appear as a journal paper [96].

Current and future works deal with developping advanced sparse decomposition for audio inpainting, including several forms of structured sparsity (*e.g.* temporal and multichannel joint-sparsity) and several applicative scenarios (declipping, time-frequency inpainting).

6.4. Music Content Processing and Music Information Retrieval

6.4.1. Acoustic music modeling

Participants: Nancy Bertin, Emmanuel Vincent.

Main collaborations: R. Badeau (Télécom ParisTech), J. Wu (University of Tokyo)

Music involves several levels of information, from the acoustic signal up to cognitive quantities such as composer style or key, through mid-level quantities such as a musical score or a sequence of chords. The dependencies between mid-level and lower- or higher-level information can be represented through acoustic models and language models, respectively.

Our acoustic models are based on nonnegative matrix factorization (NMF) and variants thereof. NMF models an input short-term magnitude spectrum as a linear combination of magnitude spectra, which are adapted to the input under suitable constraints such as harmonicity and temporal smoothness. While our previous work considered harmonic spectra only, we proposed the use of wideband spectra to represent attack transients and showed that this resulted in improved pitch transcription accuracy [77]. Our past work on the convergence properties of NMF was also disseminated [50].

We used the resulting model parameters to identify the musical instrument associated with each note, by means of a Support Vector Machine (SVM) classifier trained on solo data, and obtained improved instrument classification accuracy compared to state-of-the-art Mel-Frequency Cepstral Coefficient (MFCC) features [42], [78].

6.4.2. Music language modeling

Participants: Frédéric Bimbot, Emmanuel Vincent.

Main collaboration: S.A. Raczynski (University of Tokyo, JP)

We pursued our pioneering work on music language modeling, with a particular focus on the joint modeling of "horizontal" (sequential) and "vertical" (simultaneous) dependencies between notes by log-linear interpolation of the corresponding conditional distributions. We identified the normalization of the resulting distribution as a crucial problem for the performance of the model and proposed an exact solution to this problem.

We also applied the log-linear interpolation paradigm to the joint modeling of melody, key, chords and meter, which evolve according to different timelines. In order to synchronize these feature sequences, we explored the use of beat-long templates consisting of several notes as opposed to short time frames containing a fragment of a single note.

Both of these studies are ongoing.

6.4.3. Music structuring

Participants: Frédéric Bimbot, Gabriel Sargent, Emmanuel Vincent.

External collaboration: Emmanuel Deruty (as an independant consultant)

The structure of a music piece is a concept which is often referred to in various areas of music sciences and technologies, but for which there is no commonly agreed definition. This raises a methodological issue in MIR, when designing and evaluating automatic structure inference algorithms. It also strongly limits the possibility to produce consistent large-scale annotation datasets in a cooperative manner.

We have pursued our investigations on *autonomous and comparable blocks*, based on principles inspired from structuralism and generativism for producing music structure annotation. This work has allowed consolidating the methodology and producing additional annotations (over 400 pieces) [53].

We have also developed an algorithm aiming at the automatic inference of autonomous and comparable blocks using the timbral and harmonic content of music pieces, in combination with a regularity constraint [72].

Tested within the QUAERO project and during the MIREX 2011 campaign [94], the algorithm ranked among state-of-the-art methods.

6.5. Source separation

6.5.1. A general framework for audio source separation

Participants: Alexis Benichoux, Frédéric Bimbot, Charles Blandin, Ngoc Duong, Rémi Gribonval, Nobutaka Ito, Alexey Ozerov, Emmanuel Vincent.

Main collaborations: H. Tachibana (University of Tokyo, JP), N. Ono (National Institute of Informatics, JP)

Source separation is the task of retrieving the source signals underlying a multichannel mixture signal. The state-of-the-art approach, which we presented in a survey chapter [95], consists of representing the signals in the time-frequency domain and estimating the source coefficients by sparse decomposition in that basis. This approach relies on spatial cues, which are often not sufficient to discriminate the sources unambiguously. Last year, we proposed a general probabilistic framework for the joint exploitation of spatial and spectral cues [39] that was disseminated in several invited talks [43], [44]. This framework relies in particular on the thesis of Ngoc Duong, which was defended this year [30]. It makes it possible to quickly design a new model adapted to the data at hand and estimate its parameters via the EM algorithm. As such, it is expected to become the basis for a number of works in the field, including our own.

Since the EM algorithm is sensitive to initialization, we devoted a major part of our work to reducing this sensitivity. One approach is to set probabilistic priors over the model parameters, including spatial position priors [56] or temporal continuity priors [55]. A complementary approach is to initialize the parameters in a suitable way using source localization techniques specifically designed for environments involving multiple sources and possibly background noise [33], [54], [83]. In a longer-term perspective, we also investigated the design and exploitation of sparsity priors over time-domain acoustic transfer functions [52], [82].

6.5.2. Exploiting filter sparsity for source localization and/or separation

Participants: Alexis Benichoux, Prasad Sudhakar, Emmanuel Vincent, Rémi Gribonval, Frédéric Bimbot.

Main collaboration: Simon Arberet (EPFL)

Estimating the filters associated to room impulse responses between a source and a microphone is a recurrent problem with applications such as source separation, localization and remixing.

We considered the estimation of multiple room impulse responses from the simultaneous recording of several known sources. Existing techniques were restricted to the case where the number of sources is at most equal to the number of sensors. We relaxed this assumption in the case where the sources are known. To this aim, we proposed statistical models of the filters associated with convex log-likelihoods, and we proposed a convex optimization algorithm to solve the inverse problem with the resulting penalties. We provided a comparison between penalties via a set of experiments which shows that our method allows to speed up the recording process with a controlled quality tradeoff. This work was presented at two conferences [52], [82] and a journal paper including extensive experiments with real data is in preparation.

We also investigated the filter estimation problem in a blind setting, where the source signals are unknown. We proposed an approach for the estimation of sparse filters from a convolutive mixture of sources, exploiting the time-domain sparsity of the mixing filters and the sparsity of the sources in the time-frequency (TF) domain. The proposed approach is based on a wideband formulation of the cross-relation (CR) in the TF domain and on a framework including two steps: (a) a clustering step, to determine the TF points where the CR is valid; (b) a filter estimation step, to recover the set of filters associated with each source. We proposed for the first time a method to blindly perform the clustering step (a) and we showed that the proposed approach based on the wideband CR outperforms the narrowband approach and the GCC-PHAT approach by between 5 dB and 20 dB. This work has been published at ICASSP 2011 [49] and submitted for publication as a journal paper.

On a more theoretical side, we studied the frequency permutation ambiguity traditionnally incurred by blind convolutive source separation methods. We focussed on the filter permutation problem in the absence of scaling, investigating the possible use of the temporal sparsity of the filters as a property enabling permutation correction. The obtained theoretical and experimental results highlight the potential as well as the limits of sparsity as an hypothesis to obtain a well-posed permutation problem. This work has been submitted for publicatoin as a journal paper [99]

6.5.3. Towards real-world separation and remixing applications

Participants: Valentin Emiya, Alexey Ozerov, Laurent Simon, Emmanuel Vincent.

Shoko Araki (NTT Communication Science Laboratories, JP), Cédric Févotte (Télécom ParisTech, FR), Antoine Liutkus (Télécom ParisTech, FR), Volker Hohmann (University of Oldenburg, DE)

Following our founding role in the organization of a regular source separation evaluation campaign (SiSEC), we wrote an invited paper summarizing the outcomes of the three latest campaigns [41]. While some challenges remain, this paper highlighted that progress has been made and that audio source separation is closer than ever to successful industrial applications. This is also exemplified by the i3DMusic project and the contract recently signed with MAIA Studio.

In order to exploit our know-how for these real-world applications, we investigated issues such as how to implement our algorithms in real time and how best to exploit human input or metadata [68], [70]. In addition, while the state-of-the-art quality metrics previously developed by METISS remain widely used in the community, we proposed a new set of perceptually motivated metrics which greatly increase correlation with subjective assessments [34].

6.5.4. Source separation for multisource content indexing

Participants: Kamil Adiloglu, Alexey Ozerov, Emmanuel Vincent.

Main collaborations: J. Barker (University of Sheffield, UK), M. Lagrange (IRCAM, FR)

Another promising real-world application of source separation concerns information retrieval from multi-source data. Source separation may then be used as a pre-processing stage, such that the characteristics of each source can be separately estimated. The main difficulty is not to amplify errors from the source separation stage through subsequent feature extraction and classification stages. To this aim, we proposed a principled Bayesian approach to the estimation of the uncertainty about the separated source signals [45] and propagated this uncertainty to the features. We then exploited it in the training of the classifier itself, thereby greatly increasing classification accuracy [69].

While our work in this direction was initially motivated by music applications (e.g. artist recognition by separating the vocals from the accompaniment), we eventually applied it to noise-robust speech recognition, which is a better defined task [71]. In order to motivate further work byt the community, we created a new international evaluation campaign on that topic (CHiME) [86].

7. Contracts and Grants with Industry

7.1. National projects

7.1.1. QUAERO CTC and Corpus Projects (OSEO)

Participants: Kamil Adiloglu, Frédéric Bimbot, Laurence Catanese, Armando Muscariello, Alexey Ozerov, Gabriel Sargent, Emmanuel Vincent.

Main academic partners : IRCAM, IRIT, LIMSI, Telecom ParisTech

Quaero is a European research and development program with the goal of developing multimedia and multilingual indexing and management tools for professional and general public applications (such as search engines).

This program is supported by OSEO. The consortium is led by Thomson. Other companies involved in the consortium are: France Télécom, Exalead, Bertin Technologies, Jouve, Grass Valley GmbH, Vecsys, LTU Technologies, Siemens A.G. and Synapse Développement. Many public research institutes are also involved, including LIMSI-CNRS, INRIA, IRCAM, RWTH Aachen, University of Karlsruhe, IRIT, Clips/Imag, Telecom ParisTech, INRA, as well as other public organisations such as INA, BNF, LIPN and DGA.

METISS is involved in two technological domains : audio processing and music information retrieval (WP6). The research activities (CTC project) are focused on improving audio and music analysis, segmentation and description algorithms in terms of efficiency, robustness and scalability. Some effort is also dedicated on corpus design, collection and annotation (Corpus Project).

METISS also takes part to research and corpus activities in multimodal processing (WP10), in close collaboration with the TEXMEX project-team.

7.1.2. ANR ECHANGE

Participants: Rémi Gribonval, Prasad Sudhakar, Emmanuel Vincent, Nancy Bertin, Valentin Emiya, Nikolaos Stefanakis.

Duration: 3 years (started January 2009). Partners: A. Cohen, Laboratoire J. Louis Lions (Paris 6); F. Ollivier et J. Marchal, Laboratoire MPIA / Institut Jean Le Rond d'Alembert (Paris 6); L. Daudet, Laboratoire Ondes et Acoustique (Paris 6/7).

The objective of the ECHANGE project (ECHantillonage Acoustique Nouvelle GEnération) is to setup a theoretical and computational framework, based on the principles of compressed sensing, for the measurement and processing of complex acoustic fields through a limited number of acoustic sensors.

7.1.3. DGCIS REV-TV

Participants: Yannick Benezeth, Frédéric Bimbot, Guylaine Le Jan, Grégoire Bachman, Nathan Souviraà-Labastie.

Duration: 2.5 years (2010-2012). Partners: Technicolor (ex Thomson R&D), Artefacto, Bilboquet, Soniris, ISTIA, Télécom Bretagne, Cap Canal

The Rev-TV project aims at developping new concepts, algorithms and systems in the production of contents for interactive television based on mixed-reality.

In this context, the Metiss research group is focused on audio processing for the animation of an avatar (lip movements, facial expressions) and the control of interactive functionalities by voice and vocal noises.

7.2. European projects

7.2.1. FP7 FET-Open program SMALL

Participants: Rémi Gribonval, Ngoc Duong, Valentin Emiya, Jules Espiau de Lamaestre, Emmanuel Vincent, Nancy Bertin.

Duration: 2010-2012
Partners: Univ. Edimburg, Queen Mary Univ., EPFL, Technion Univ.

A joint research project called SMALL (Sparse Models, Algorithms and Learning for Large-scale data) has been setup with the groups of Pr Mark Plumbley (Centre for Digital Music, Queen Mary University of London, UK), Pr Mike Davies University of Edinburgh, UK), Pr Pierre Vandergheynst (EPFL, Switzeland) and Miki Elad (The Technion, Israel) in the framework of the European FP7 FET-Open call. SMALL was one of the eight selected projects among more than 111 submissions and began in February 2009.

The main objective of the project is to explore new generations of provably good methods to obtain inherently data-driven sparse models, able to cope with large-scale and complicated data much beyond state-of-the-art sparse signal modeling. The project will develop a radically new foundational theoretical framework for dictionary learning, and scalable algorithms for the training of structured dictionaries.

7.2.2. EUREKA Eurostars program i3DMusic

Participants: Emmanuel Vincent, Ngoc Duong, Rémi Gribonval, Laurent Simon.

Duration: 3 years, starting in October 2010.
Partners: Audionamix (FR), Sonic Emotion (CH), École Polytechnique Fédérale de Lausanne (CH)

A joint research project called i3DMusic (Real-time Interative 3D Rendering of Musical Recordings) has been setup with the SMEs Audionamix and Sonic Emotion and the academic partner EPFL. This project aims to provide a system enabling real-time interactive respatialization of mono or stereo music content. This will be achieved through the combination of source separation and 3D audio rendering techniques. Metiss is responsible for the source separation work package, more precisely for designing scalable online source separation algorithms and estimating advanced spatial parameters from the available mixture.

8. Partnerships and Cooperations

8.1. International initiatives

8.1.1. Associate Team VERSAMUS with the University of Tokyo

Participants: Emmanuel Vincent, Nobutaka Ito, Gabriel Sargent, Ngoc Duong, Frédéric Bimbot, Rémi Gribonval.

Duration: 3 years, starting in January 2010.
Partner: Lab#1, Department of Information Physics and Computing, the University of Tokyo (JP)

We initiated a partnership with Lab#1 of the Department of Information Physics and Computing of the University of Tokyo, led by Shigeki Sagayama and Nobutaka Ono. This collaboration was formalized as the INRIA Associate Team VERSAMUS in January 2010. The PhD of Nobutaka Ito is co-supervised by Nobutaka Ono, Emmanuel Vincent and Rémi Gribonval in this framework. A workshop was organized in Tokyo in June 2011, and a total of 5 visits were made between the two teams in 2011.

The aim of this collaboration is to investigate, design and validate integrated music representations combining many acoustic and symbolic feature levels. Tasks to be addressed include the design of a versatile Bayesian model structure, of a library of probabilistic feature models and of efficient algorithms for parameter inference and model selection. More details are available on http://versamus.inria.fr/.

9. Dissemination

9.1. Animation of the scientific community

Frédéric Bimbot is the Head of the "Digital Signals and Images, Robotics" in IRISA (UMR 6074).

Frédéric Bimbot has been appointed in the Comité National de la Recherche Scientifique - Section 07.

As the General Chairman of the Interspeech 2013 Conference in Lyon (1200 participants expected), Frédéric Bimbot chairs the Steering and Organisation Committees of the conference.

Frédéric Bimbot is the Scientific Leader of the Audio Processing Technology Domain in the QUAERO Project.

Rémi Gribonval was a plenary speaker at the international conference SPARS'11 on Signal Processing with Adaptive/Structured Representations

Rémi Gribonval is the organizer, together with Francis Bach, Mike Davies and Guillaume Obozinski, of a NIPS 2011 workshop on sparse representations, Granada, Spain, December 16, 2011.

Rémi Gribonval is in charge of the Action "Parcimonie" within the French GDR ISIS on Signal & Image Processing

Rémi Gribonval and Emmanuel Vincent are associate editors of the special issue on Latent Variable Analysis and Signal Separation of the Journal Signal Processing published by Elsevier.

Rémi Gribonval and Emmanuel Vincent are members of the International Steering Committee for the ICA conferences.

Emmanuel Vincent is an Associate Editor for the *IEEE Transactions on Audio, Speech, and Language Processing* (2011–2014).

Emmanuel Vincent is a Guest Editor of the special issue on Speech Separation and Recognition in Multisource Environments of the journal *Computer Speech and Language* published by Elsevier.

Emmanuel Vincent was elected a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing (2012–2014).

Emmanuel Vincent was a General Chair of the PASCAL 'CHiME' Speech Separation and Recognition Challenge and the 1st International Workshop on Computational Hearing in Multisource Environments, held in Florence on September 1, 2011, as a satellite event of Interspeech 2011.

Emmanuel Vincent is part of the organizing committee of the third community-based Signal Separation Evaluation Campaign (SiSEC 2011), whose first edition had been initiated by Metiss. The results of the campaign will be presented during a panel session at the 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA 2012). Datasets, evaluation criteria and reference software are available at http://sisec.wiki.irisa.fr/.

Emmanuel Vincent was nominated a titular member of the National Council of Universities (CNU section 61, 2012–2015).

Nancy Bertin and Frédéric Bimbot designed and animated a scientific stand (opened during 10 weeks) on speaker recognition at the "Palais de la Découverte", Paris, in the context of the programme "un chercheur, une manip".

9.2. Teaching

Rémi Gribonval gave a series of tutorial lectures on sparse decompositions and compressed sensing at the Machine Learning Summer School MLSS'11.

Rémi Gribonval gave lectures about sparse representations for inverse problems in signal and image processing for a total of 10 hours as part of the SISEA Masters in Signal & Image Processing, University of Rennes I.

Rémi Gribonval was the coordinator of the ARD module of the Masters in Computer Science, Rennes I.

Rémi Gribonval gave lectures about signal and image representations, time-frequency and time-scale analysis, filtering and deconvolution for a total of 8 hours as part of the ARD module of the Masters in Computer Science, Rennes I.

Emmanuel Vincent gave lectures about audio rendering, coding and source separation for a total of 6 hours as part of the CTR module of the Masters in Computer Science, Rennes I.

Emmanuel Vincent taught general tools for signal compression and speech compression for 10 hours within the DT SIC RTL course at the Ecole Supérieure d'Applications des Transmissions (ESAT, Rennes).

Emmanuel Vincent gave a tutorial on Music Source Separation at DAFx 2011 (14th Int. Conf. on Digital Audio Effects), Paris, September 19-23, 2011.

Nancy Bertin gave 6 hours of lecture in speech and audio description within the FAV module of the Masters in Computer Science, Rennes I.

10. Bibliography

Major publications by the team in recent years

[1] S. ARBERET. *Estimation robuste et apprentissage aveugle de modèles pour la séparation de sources sonores*, Université de Rennes I, december 2008.

[2] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Beyond coherence : recovering structured time-frequency representations*, in "Appl. Comput. Harmon. Anal.", 2008, vol. 24, n^{o} 1, p. 120–128.

[3] M. DAVIES, R. GRIBONVAL. *On Lp minimisation, instance optimality, and restricted isometry constants for sparse approximation*, in "Proc. SAMPTA'09 (Sampling Theory and Applications)", Marseille, France, may 2009.

[4] M. DAVIES, R. GRIBONVAL. *Restricted Isometry Constants where _ell^{p} sparse recovery can fail for* $0 < p \leq 1$, in "IEEE Trans. Inform. Theory", May 2009, vol. 55, n^{o} 5, p. 2203–2214.

[5] M. DAVIES, R. GRIBONVAL. *The Restricted Isometry Property and _ell^{p} sparse recovery failure*, in "Proc. SPARS'09 (Signal Processing with Adaptive Sparse Structured Representations)", Saint-Malo, France, April 2009.

[6] S. GALLIANO, E. GEOFFROIS, D. MOSTEFA, K. CHOUKRI, J.-F. BONASTRE, G. GRAVIER. *The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News*, in "European Conference on Speech Communication and Technology", 2005.

[7] R. GRIBONVAL, R. M. FIGUERAS I VENTURA, P. VANDERGHEYNST. *A simple test to check the optimality of sparse signal approximations*, in "EURASIP Signal Processing, special issue on Sparse Approximations in Signal and Image Processing", March 2006, vol. 86, n^{o} 3, p. 496–510.

[8] R. GRIBONVAL. *Sur quelques problèmes mathématiques de modélisation parcimonieuse*, Université de Rennes I, octobre 2007, Habilitation à Diriger des Recherches, spécialité "Mathématiques".

[9] R. GRIBONVAL, M. NIELSEN. *On approximation with spline generated framelets*, in "Constructive Approx.", January 2004, vol. 20, n^{o} 2, p. 207–232.

[10] R. GRIBONVAL, M. NIELSEN. *Beyond sparsity : recovering structured representations by ℓ^1-minimization and greedy algorithms*, in "Advances in Computational Mathematics", January 2008, vol. 28, n^{o} 1, p. 23–41.

[11] R. GRIBONVAL, H. RAUHUT, K. SCHNASS, P. VANDERGHEYNST. *Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms*, in "J. Fourier Anal. Appl.", December 2008, vol. 14, n^{o} 5, p. 655–687.

[12] R. GRIBONVAL, K. SCHNASS. *Dictionary identifiability from few training samples*, in "Proc. European Conf. on Signal Processing - EUSIPCO", August 2008.

[13] R. GRIBONVAL, K. SCHNASS. *Some recovery conditions for basis learning by l1-minimization*, in "3rd IEEE International Symposium on Communications, Control and Signal Processing - ISCCSP 2008", March 2008, p. 768–773.

[14] R. GRIBONVAL, P. VANDERGHEYNST. *On the exponential convergence of Matching Pursuits in quasi-incoherent dictionaries*, in "IEEE Trans. Information Theory", January 2006, vol. 52, n^{o} 1, p. 255–261, http://dx.doi.org/10.1109/TIT.2005.860474.

[15] R. GRIBONVAL, K. SCHNASS. *Dictionary Identifiability - Sparse Matrix-Factorisation via ℓ_1 minimisation*, in "IEEE Trans. Information Theory", jul 2010, vol. 56, n^{o} 7, p. 3523–3539.

[16] D. K. HAMMOND, P. VANDERGHEYNST, R. GRIBONVAL. *Wavelets on Graphs via Spectral Graph Theory*, in "Applied and Computational Harmonic Analysis", 2010, submitted.

[17] S. HUET, G. GRAVIER, P. SÉBILLOT. *Un modèle multi-sources pour la segmentation en sujets de journaux radiophoniques*, in "Proc. Traitement Automatique des Langues Naturelles", 2008, p. 49–58.

[18] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Audiovisual integration for tennis broadcast structuring*, in "Multimedia Tools and Application", 2006, vol. 30, n^{o} 3, p. 289–312.

[19] B. MAILHÉ, R. GRIBONVAL, F. BIMBOT, P. VANDERGHEYNST. *LocOMP: algorithme localement orthogonal pour l'approximation parcimonieuse rapide de signaux longs sur des dictionnaires locaux*, in "Proc. GRETSI", Septembre 2009.

[20] B. MAILHÉ, R. GRIBONVAL, P. VANDERGHEYNST, F. BIMBOT. *A low–complexity Orthogonal Matching Pursuit for Sparse Signal Approximation with Shift–Invariant Dictionaries*, in "Proc. IEEE ICASSP", April 2009.

[21] B. MAILHÉ, S. LESAGE, R. GRIBONVAL, P. VANDERGHEYNST, F. BIMBOT. *Shift–invariant dictionary learning for sparse representations : extending K–SVD*, in "Proc. European Conf. on Signal Processing - EUSIPCO", August 2008.

[22] B. MAILHÉ, R. GRIBONVAL, P. VANDERGHEYNST, F. BIMBOT. *Fast orthogonal sparse approximation algorithms over local dictionaries*, INRIA, feb 2010, http://hal.archives-ouvertes.fr/hal-00460558/PDF/LocOMP.pdf.

[23] A. OZEROV, P. PHILIPPE, F. BIMBOT, R. GRIBONVAL. *Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs*, in "IEEE Trans. Audio, Speech and Language Processing", juillet 2007, vol. 15, n^{o} 5, p. 1564–1578.

[24] A. ROSENBERG, F. BIMBOT, S. PARTHASARATHY. *36*, in "Overview of Speaker Recognition", Springer, 2008, p. 725–741.

[25] K. SCHNASS, R. GRIBONVAL. *Basis Identification from Random Sparse Samples*, in "Proc. SPARS'09 (Signal Processing with Adaptive Sparse Structured Representations)", Saint-Malo, France, April 2009.

[26] P. SUDHAKAR, R. GRIBONVAL. *A sparsity-based method to solve the permutation indeterminacy in frequency domain convolutive blind source separation*, in "ICA 2009, 8th International Conference on Independent Component Analysis and Signal Separation", Paraty, Brazil, March 2009.

[27] P. SUDHAKAR, R. GRIBONVAL. *Sparse filter models for solving permutation indeterminacy in convolutive blind source separation*, in "Proc. SPARS'09 (Signal Processing with Adaptive Sparse Structured Representations)", Saint-Malo, France, April 2009.

[28] E. VINCENT, R. GRIBONVAL, C. FÉVOTTE. *Performance measurement in Blind Audio Source Separation*, in "IEEE Trans. Speech, Audio and Language Processing", 2006, vol. 14, n^o 4, p. 1462–1469, http://dx.doi.org/10.1109/TSA.2005.858005.

[29] E. VINCENT, M. D. PLUMBLEY. *Low bitrate object coding of musical audio using bayesian harmonic models*, in "IEEE Trans. on Audio, Speech and Language Processing", 2007, vol. 15, n^o 4, p. 1273–1282.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[30] N. DUONG. *Modélisation gaussienne de rang plein des mélanges audio convolutifs appliquée à la séparation de sources*, Université Rennes 1, November 2011, http://hal.inria.fr/tel-00646419/en.

[31] A. MUSCARIELLO. *Découverte de motifs variables dans les grandes volumes de données audio.*, Université Rennes 1, January 2011, http://hal.inria.fr/tel-00642956/en.

[32] P. SUDHAKARA. *Modèles Parcimonieux et Optimisation Convexe pour la Séparation Aveugle de Sources Convolutives*, Université Rennes 1, February 2011, http://hal.inria.fr/tel-00586610/en.

Articles in International Peer-Reviewed Journal

[33] C. BLANDIN, A. OZEROV, E. VINCENT. *Multi-source TDOA estimation in reverberant audio using angular spectra and clustering*, in "Signal Processing", October 2011, http://hal.inria.fr/inria-00630994/en.

[34] V. EMIYA, E. VINCENT, N. HARLANDER, V. HOHMANN. *Subjective and objective quality assessment of audio source separation*, in "IEEE Transactions on Audio, Speech, and Language Processing", September 2011, vol. 19, n^o 7, p. 2046-2057 [*DOI :* 10.1109/TASL.2011.2109381], http://hal.inria.fr/inria-00567152/en.

[35] R. GRIBONVAL. *Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?*, in "IEEE Transactions on Signal Processing", May 2011, vol. 59, n^o 5, p. 2405-2410, To appear in IEEE Transactions on Signal Processing [*DOI :* 10.1109/TSP.2011.2107908], http://hal.inria.fr/inria-00486840/en.

[36] D. K. HAMMOND, P. VANDERGHEYNST, R. GRIBONVAL. *Wavelets on graphs via spectral graph theory*, in "Applied and Computational Harmonic Analysis", March 2011, vol. 30, n^o 2, p. 129–150 [*DOI :* 10.1016/J.ACHA.2010.04.005], http://hal.inria.fr/inria-00541855/en.

[37] B. MAILHÉ, R. GRIBONVAL, P. VANDERGHEYNST, F. BIMBOT. *Fast orthogonal sparse approximation algorithms over local dictionaries*, in "Signal Processing", January 2011 [*DOI :* 10.1016/J.SIGPRO.2011.01.004], http://hal.inria.fr/hal-00460558/en.

[38] A. Ozerov, W. B. Kleijn. *Asymptotically optimal model estimation for quantization*, in "IEEE Transactions on Communications", 2011, vol. 59, n^{o} 4, p. 1031-1042, http://hal.inria.fr/inria-00553532/en.

[39] A. Ozerov, E. Vincent, F. Bimbot. *A General Flexible Framework for the Handling of Prior Information in Audio Source Separation*, in "IEEE Transactions on Audio, Speech and Language Processing", September 2011, p. 1 - 16, 16, http://hal.inria.fr/hal-00626962/en.

[40] V. Vigneron, V. Zarzoso, R. Gribonval, E. Vincent. *Preface to the special issue on latent variable analysis and signal separation*, in "Signal Processing", January 2012, http://hal.inria.fr/hal-00658459/en.

[41] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, N. Duong. *The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges*, in "Signal Processing", October 2011, http://hal.inria.fr/inria-00630985/en.

[42] J. Wu, E. Vincent, S. Raczynski, T. Nishimoto, N. Ono, S. Sagayama. *Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds*, in "IEEE Journal of Selected Topics in Signal Processing", May 2011, vol. 5, n^{o} 6, p. 1124-1132, http://hal.inria.fr/inria-00594965/en.

Invited Conferences

[43] E. Vincent. *A general flexible probabilistic framework for audio source separation*, in "Workshop "Computational Audition"", Delmenhorst, Germany, October 2011, http://hal.inria.fr/hal-00646367/en.

[44] E. Vincent. *Music source separation*, in "14th International Conference on Digital Audio Effects (DAFx-11)", Paris, France, September 2011, http://hal.inria.fr/inria-00614275/en.

International Conferences with Proceedings

[45] K. Adiloglu, E. Vincent. *An Uncertainty Estimation Approach for the Extraction of Source Features in Multisource Recordings*, in "European Signal Processing Conference (Eusipco 11)", Barcelona, Spain, Centre Tecnològic de Telecomunicacions de Catalunya, Universitat Politècnica de Catalunya, August 2011, http://hal.inria.fr/inria-00597615/en.

[46] K. Adiloglu, E. Vincent. *A General Variational Bayesian Framework for Robust Feature Extraction in Multisource Recordings*, in "IEEE International Conference on Acoustics, Speech and Signal Processing", Kyoto, Japan, March 2012, http://hal.inria.fr/hal-00656613/en.

[47] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, M. D. Plumbley. *A Constrained Matching Pursuit Approach to Audio Declipping*, in "Acoustics, Speech and Signal Processing, IEEE International Conference on (ICASSP 2011)", Prague, Czech Republic, IEEE, May 2011 [*DOI :* 10.1109/ICASSP.2011.5946407], http://hal.inria.fr/inria-00557021/en.

[48] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, A. Benichoux. *The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Audio source separation -*, in "10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", Tel Aviv, Israel, March 2012, http://hal.inria.fr/hal-00655394/en.

[49] S. Arberet, P. Sudhakar, R. Gribonval. *A Wideband Doubly-Sparse Approach for MITO Sparse Filter Estimation*, in "Acoustics, Speech and Signal Processing, IEEE International Conference on (ICASSP 2011)",

Prague, Czech Republic, IEEE, May 2011, To appear [*DOI :* 10.1109/ICASSP.2011.5947085], http://hal.inria.fr/inria-00567210/en.

[50] R. BADEAU, N. BERTIN, E. VINCENT. *Stability analysis of multiplicative update algorithms for non-negative matrix factorization*, in "International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Prague, Czech Republic, IEEE, May 2011, 4, http://hal.inria.fr/hal-00557789/en.

[51] A. BENICHOUX, P. SUDHAKAR, F. BIMBOT, R. GRIBONVAL. *Some uniqueness results in sparse convolutive source separation*, in "International Conference on Latent Variable Analysis and Source Separation", Tel Aviv, Israel, Springer, March 2012, http://hal.inria.fr/hal-00659913/en.

[52] A. BENICHOUX, E. VINCENT, R. GRIBONVAL. *A compressed sensing approach to the simultaneous recording of multiple room impulse responses*, in "WASPAA", United States, October 2011, 1, http://hal.inria.fr/hal-00612911/en.

[53] F. BIMBOT, E. DERUTY, G. SARGENT, E. VINCENT. *Methodology and Resources for The Structural Segmentation of Music Pieces into Autonomous and Comparable Blocks*, in "International Society for Music Information Retrieval Conference (ISMIR)", Miami, United States, October 2011.

[54] C. BLANDIN, E. VINCENT, A. OZEROV. *Multi-Source TDOA estimation using SNR-based angular spectra*, in "2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", Prague, Czech Republic, May 2011, p. 2616 - 2619, Revised version including bugfixes in SNR-APR and MUSIC and in Figure 2 compared to the original version published by the IEEE., http://hal.inria.fr/inria-00566706/en.

[55] N. DUONG, H. TACHIBANA, E. VINCENT, N. ONO, R. GRIBONVAL, S. SAGAYAMA. *Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity*, in "Acoustics, Speech and Signal Processing, IEEE Conference on (ICASSP'11)", Prague, Czech Republic, January 2011 [*DOI :* 10.1109/ICASSP.2011.5946376], http://hal.inria.fr/inria-00557145/en.

[56] N. DUONG, E. VINCENT, R. GRIBONVAL. *An acoustically-motivated spatial prior for under-determined reverberant source separation*, in "Acoustics, Speech and Signal Processing, IEEE Conference on (ICASSP'11)", Prague, Czech Republic, February 2011 [*DOI :* 10.1109/ICASSP.2011.5946315], http://hal.inria.fr/inria-00566868/en.

[57] R. GIRYES, S. NAM, R. GRIBONVAL, M. E. DAVIES. *Iterative Cosparse Projection Algorithms for the Recovery of Cosparse Vectors*, in "The 19th European Signal Processing Conference (EUSIPCO-2011)", Barcelona, Spain, 2011, http://hal.inria.fr/inria-00611592/en.

[58] R. GRIBONVAL, G. CHARDON, L. DAUDET. *BLIND CALIBRATION FOR COMPRESSED SENSING BY CONVEX OPTIMIZATION*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing", Kyoto, Japan, 2012, http://hal.inria.fr/hal-00658579/en.

[59] G. LE-JAN, Y. BENEZETH, G. GRAVIER, F. BIMBOT. *A study on auditory feature spaces for speech-driven lip animation*, in "Interspeech", Florence, Italy, August 2011, http://hal.inria.fr/inria-00598314/en.

[60] A. MUSCARIELLO, G. GRAVIER, F. BIMBOT. *An efficient method for the unsupervised discovery of signalling motifs in large audio streams*, in "International Workshop on Content-Based Multimedia Indexing", Madrid, Spain, June 2011, http://hal.inria.fr/inria-00572817/en.

[61] A. MUSCARIELLO, G. GRAVIER, F. BIMBOT. *Towards robust word discovery by self similarity matrix comparison*, in "IEEE International Conference on Acoustics, Speech and Signal Processing", Prague, Czech Republic, May 2011, http://hal.inria.fr/inria-00563418/en.

[62] A. MUSCARIELLO, G. GRAVIER, F. BIMBOT. *Zero-resource audio-only spoken term detection based on a combination of template matching techniques*, in "INTERSPEECH 2011: 12th Annual Conference of the International Speech Communication Association", Florence, Italy, 2011, spoken term detection, template matching, unsupervised learning, posterior features, http://hal.inria.fr/inria-00597907/en.

[63] S. NAM, M. DAVIES, M. ELAD, R. GRIBONVAL. *Recovery of Cosparse Signals with Greedy Analysis Pursuit in the Presence of Noise*, 2011.

[64] S. NAM, M. DAVIES, M. ELAD, R. GRIBONVAL. *Cosparse Analysis Modeling - Uniqueness and Algorithms*, in "Acoustics, Speech and Signal Processing, IEEE International Conference on (ICASSP 2011)", Prague, Czech Republic, IEEE, 2011 [*DOI :* 10.1109/ICASSP.2011.5947680], http://hal.inria.fr/inria-00557933/en.

[65] S. NAM, R. GRIBONVAL. *Physics-driven structured cosparse modeling for source localization*, in "Acoustics, Speech and Signal Processing, IEEE International Conference on (ICASSP 2012)", Kyoto, Japan, IEEE, 2012, http://hal.inria.fr/hal-00659405/en.

[66] G. NOLTE, D. LUTTER, A. ZIEHE, F. NESTA, E. VINCENT, Z. KOLDOVSKY, A. BENICHOUX, S. ARAKI. *The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Biomedical data analysis -*, in "10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", Tel Aviv, Israel, March 2012, http://hal.inria.fr/hal-00655801/en.

[67] B. OPHIR, M. ELAD, N. BERTIN, M. D. PLUMBLEY. *Sequential Minimal Eigenvalues - An Approach to Analysis Dictionary Learning*, in "The 19th European Signal Processing Conference (EUSIPCO-2011)", Barcelona, Spain, 2011, http://hal.inria.fr/inria-00577231/en.

[68] A. OZEROV, C. FÉVOTTE, R. BLOUET, J.-L. DURRIEU. *Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)", Prague, Czech Republic, May 2011, http://hal.inria.fr/inria-00564851/en.

[69] A. OZEROV, M. LAGRANGE, E. VINCENT. *GMM-based classification from noisy features*, in "International Workshop on Machine Listening in Multisource Environments (CHiME 2011)", Florence, Italy, September 2011, http://hal.inria.fr/inria-00598742/en.

[70] A. OZEROV, A. LIUTKUS, R. BADEAU, G. RICHARD. *Informed source separation: source coding meets source separation*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)", Mohonk, NY, United States, October 2011, http://hal.inria.fr/inria-00610526/en.

[71] A. OZEROV, E. VINCENT. *Using the FASST source separation toolbox for noise robust speech recognition*, in "International Workshop on Machine Listening in Multisource Environments (CHiME 2011)", Florence, Italy, September 2011, http://hal.inria.fr/inria-00598734/en.

[72] G. SARGENT, F. BIMBOT, E. VINCENT. *A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs*, in "International Society for Music Information Retrieval Conference (ISMIR)", Miami, United States, October 2011, http://hal.inria.fr/inria-00616274/en.

[73] L. S. R. SIMON, R. MASON. *Spaciousness Rating of 8-channel Stereophony-based Microphone Arrays*, in "Audio Engineering Society 130th convention", London, United Kingdom, March 2011, http://hal.inria.fr/inria-00590518/en.

[74] L. S. R. SIMON, E. VINCENT. *A general framework for online audio source separation*, in "International conference on Latent Variable Analysis and Signal Separation", Tel-Aviv, Israel, March 2012, http://hal.inria.fr/hal-00655398/en.

[75] N. STEFANAKIS, J. MARCHAL, V. EMIYA, N. BERTIN, R. GRIBONVAL, P. CERVENKA. *Sparse underwater acoustic imaging: a case study*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing", Kyoto, Japan, March 2012, http://hal.inria.fr/hal-00661526/en.

[76] E. VINCENT. *Improved perceptual metrics for the evaluation of audio source separation*, in "10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", Tel Aviv, Israel, March 2012, http://hal.inria.fr/hal-00653196/en.

[77] J. WU, E. VINCENT, S. RACZYNSKI, T. NISHIMOTO, N. ONO, S. SAGAYAMA. *Multipitch estimation by joint modeling of harmonic and transient sounds*, in "2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", Prague, Czech Republic, May 2011, p. 25 - 28, http://hal.inria.fr/inria-00567175/en.

[78] J. WU, E. VINCENT, S. RACZYNSKI, T. NISHIMOTO, N. ONO, S. SAGAYAMA. *Musical instrument identification based on new boosting algorithm with probabilistic decisions*, in "Int. Symp. on Computer Music Modeling and Retrieval (CMMR)", Bhubaneswar, India, March 2011, http://hal.inria.fr/inria-00562115/en.

[79] M. YAGHOOBI, S. NAM, R. GRIBONVAL, M. E. DAVIES. *Analysis Operator Learning for Overcomplete Cosparse Representations*, in "European Signal Processing Conference (EUSIPCO'11)", Barcelona, Spain, 2011, http://hal.inria.fr/inria-00583133/en.

[80] M. YAGHOOBI, S. NAM, R. GRIBONVAL, M. DAVIES. *NOISE AWARE ANALYSIS OPERATOR LEARNING FOR APPROXIMATELY COSPARSE SIGNALS*, in "ICASSP - IEEE International Conference on Acoustics, Speech, and Signal Processing - 2012", Kyoto, Japan, IEEE, 2012, http://hal.inria.fr/hal-00661549/en.

National Conferences with Proceeding

[81] Y. BENEZETH, G. GRAVIER, F. BIMBOT. *Étude comparative des différentes unités acoustiques pour la synchronisation labiale*, in "GRETSI, Groupe d'Etudes du Traitement du Signal et des Images", Bordeaux, France, 2011, http://hal.inria.fr/inria-00593756/en.

[82] A. BENICHOUX, E. VINCENT, R. GRIBONVAL. *Optimisation convexe pour l'estimation simultanée de réponses acoustiques*, in "23e Colloque du GRETSI", Bordeaux, France, May 2011, http://hal.inria.fr/inria-00594252/en.

[83] N. ITO, E. VINCENT, N. ONO, R. GRIBONVAL, S. SAGAYAMA. *Diffuse noise robust multiple source localization based on matrix completion via trace norm minimization*, in "ASJ Spring Meeting", Tokyo, Japan, March 2011, http://hal.inria.fr/inria-00596138/en.

Conferences without Proceedings

[84] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, M. D. Plumbley. *A Reproducible Research Framework for Audio Inpainting*, in "Workshop on Signal Processing with Adaptive Sparse Structured Representations", Edinburgh, United Kingdom, June 2011, http://hal.inria.fr/inria-00587688/en.

[85] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, M. D. Plumbley. *Audio inpainting: problem statement, relation with sparse representations and some experiments*, in "SMALL Workshop on Sparse Dictionary Learning", London, United Kingdom, January 2011, http://hal.inria.fr/inria-00560110/en.

[86] J. Barker, E. Vincent, N. Ma, H. Christensen, P. Green. *Overview of the PASCAL CHiME Speech Separation and Recognition Challenge*, in "1st International Workshop on Machine Listening in Multisource Environments (CHiME 2011)", Firenze, Italy, September 2011, http://hal.inria.fr/hal-00646370/en.

[87] A. Benichoux, P. Sudhakar, R. Gribonval. *Well-posedness of the frequency permutation problem in sparse filter estimation with lp minimization*, in "Signal Processing with Adaptive Sparse Structured Representations", Edinburgh, United Kingdom, June 2011, http://hal.inria.fr/inria-00587789/en.

[88] G. Chardon, N. Bertin, L. Daudet. *Multiplexage spatial aléatoire pour l'échantillonnage compressif - application à l'holographie acoustique*, in "XXIIIe Colloque GRETSI", Bordeaux, France, September 2011, http://hal.inria.fr/hal-00647780/en.

[89] J. Espiau De Lamaëstre. *Project SMALL: Overview*, in "SMALL Workshop on Sparse Dictionary Learning", London, United Kingdom, QMUL, 2011, http://hal.inria.fr/inria-00574129/en.

[90] R. Gribonval. *An Overview of Analysis vs Synthesis in Low-Dimensional Signal Models*, in "SPARS'11", Edinburgh, United Kingdom, June 2011, http://hal.inria.fr/inria-00624355/en.

[91] S. Nam, M. E. Davies, M. Elad, R. Gribonval. *Cosparse Analysis Modeling*, in "The 9th International Conference on Sampling Theory and Applications", Singapore, Singapore, May 2011, http://hal.inria.fr/inria-00591779/en.

[92] S. Nam, M. E. Davies, M. Elad, R. Gribonval. *Cosparse Analysis Modeling*, in "Workshop on Signal Processing with Adaptive Sparse Structured Representations", Edinburgh, United Kingdom, June 2011, http://hal.inria.fr/inria-00587943/en.

[93] G. Puy, P. Vandergheynst, R. Gribonval, Y. Wiaux. *Spread Spectrum for Universal Compressive Sampling*, in "SPARS'11", Edinburgh, United Kingdom, 2011, http://hal.inria.fr/inria-00582817/en.

[94] G. Sargent, S. Raczynski, F. Bimbot, E. Vincent, S. Sagayama. *A music structure inference algorithm based on symbolic data analysis*, in "MIREX - ISMIR 2011", Miami, United States, October 2011, http://hal.inria.fr/hal-00618141/en.

Scientific Books (or Scientific Book chapters)

[95] G. Evangelista, S. Marchand, M. D. Plumbley, E. Vincent. *Sound source separation*, in "DAFX - Digital Audio Effects, 2nd Edition", U. Zölzer (editor), Wiley, 2011, http://hal.inria.fr/inria-00544043/en.

Research Reports

[96] A. ADLER, V. EMIYA, M. JAFARI, M. ELAD, R. GRIBONVAL, M. D. PLUMBLEY. *Audio Inpainting*, INRIA, March 2011, n^{o} RR-7571, http://hal.inria.fr/inria-00577079/en.

[97] S. ARBERET, A. OZEROV, F. BIMBOT, R. GRIBONVAL. *A Tractable Framework for Estimating and Combining Spectral Source Models for Audio Source Separation*, INRIA, March 2011, n^{o} RR-7556, http://hal.inria.fr/inria-00572249/en.

[98] Y. BENEZETH, G. BACHMAN, G. LE-JAN, N. SOUVIRAÀ-LABASTIE, F. BIMBOT. *BL-Database: A French audiovisual database for speech driven lip animation systems*, INRIA, August 2011, n^{o} RR-7711, http://hal.inria.fr/inria-00614761/en.

[99] A. BENICHOUX, P. SUDHAKAR, F. BIMBOT, R. GRIBONVAL. *Well-posedness of the permutation problem in sparse filter estimation with lp minimization*, INRIA, November 2011, n^{o} RR-7782, http://hal.inria.fr/hal-00640198/en.

[100] F. BIMBOT, E. DERUTY, G. SARGENT, E. VINCENT. *Methodology and resources for the structural segmentation of music pieces into autonomous and comparable blocks*, May 2011, http://hal.inria.fr/inria-00596431/en.

[101] C. BLANDIN, A. OZEROV, E. VINCENT. *Multi-source TDOA estimation in reverberant audio using angular spectra and clustering*, INRIA, April 2011, n^{o} RR-7566, http://hal.inria.fr/inria-00576297/en.

[102] R. GRIBONVAL, M. NIELSEN. *The restricted isometry property meets nonlinear approximation with redundant frames*, INRIA, February 2011, n^{o} RR-7548, This work has been submitted for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible., http://hal.inria.fr/inria-00567801/en.

[103] S. NAM, M. E. DAVIES, M. ELAD, R. GRIBONVAL. *The Cosparse Analysis Model and Algorithms*, 2011, Submitted, http://hal.inria.fr/inria-00602205/en.

[104] A. OZEROV, M. LAGRANGE, E. VINCENT. *Uncertainty-based learning of Gaussian mixture models from noisy data*, INRIA, January 2012, n^{o} RR-7862, http://hal.inria.fr/hal-00660689/en.

[105] G. PUY, P. VANDERGHEYNST, R. GRIBONVAL, Y. WIAUX. *Universal and efficient compressed sensing by spread spectrum and application to realistic Fourier imaging techniques*, 2011, http://hal.inria.fr/inria-00582432/en.

[106] E. VINCENT, S. ARAKI, F. J. THEIS, G. NOLTE, P. BOFILL, H. SAWADA, A. OZEROV, B. V. GOWREESUNKER, D. LUTTER, N. DUONG. *The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges*, INRIA, March 2011, n^{o} RR-7581, La version 2 de ce rapport de recherche correspond à une rétractation, pour raisons scientifiques. Le fichier pdf reste disponible dans la version 1., http://hal.inria.fr/inria-00579398/en.

Other Publications

[107] R. GRIBONVAL, V. CEVHER, M. DAVIES. *Compressible Distributions for High-dimensional Statistics*, 2011, Submitted. Was previously entitled "Compressible priors for high-dimensional statistics", http://hal.inria.fr/inria-00563207/en.

[108] C. SOUSSEN, R. GRIBONVAL, J. IDIER, C. HERZET. *Sparse recovery conditions for Orthogonal Least Squares*, 2011, http://hal.inria.fr/hal-00637003/en.

[109] N. SOUVIRAÀ-LABASTIE. *Prédiction du mouvement des lèvres à partir d'un signal de parole pour l'animation d'un avatar*, INSA, September 2011, http://hal.inria.fr/inria-00628856/en.

References in notes

[110] R. BARANIUK. *Compressive sensing*, in "IEEE Signal Processing Magazine", July 2007, vol. 24, n^{o} 4, p. 118–121.

[111] R. BOITE, H. BOURLARD, T. DUTOIT, J. HANCQ, H. LEICH. *Traitement de la Parole*, Presses Polytechniques et Universitaires Romandes, 2000.

[112] M. DAVY, S. J. GODSILL, J. IDIER. *Bayesian Analysis of Polyphonic Western Tonal Music*, in "Journal of the Acoustical Society of America", 2006, vol. 119, n^{o} 4, p. 2498–2517.

[113] G. GRAVIER, F. YVON, B. JACOB, F. BIMBOT. *Sirocco, un système ouvert de reconnaissance de la parole*, in "Journées d'étude sur la parole", Nancy, June 2002, p. 273-276.

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

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

[116] K. MURPHY. *An introduction to graphical models*, 2001, http://www.cs.ubc.ca/~murphyk/Papers/intro_gm.pdf.

[117] N. WHITELEY, A. T. CEMGIL, S. J. GODSILL. *Sequential Inference of Rhythmic Structure in Musical Audio*, in "Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2007, p. 1321–1324.