INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team metiss

Speech and sound data modeling and processing

Rennes - Bretagne-Atlantique

Theme : Audio, Speech, and Language Processing

Table of contents

1. Team

Research Scientists

Frédéric Bimbot [Team Leader, Senior Researcher (DR2) CNRS, HdR]
Rémi Gribonval [Senior Researcher (DR2) INRIA, HdR]
Nancy Bertin [Junior Researcher (CR2) CNRS - Since Oct. 2010 (formerly Post-Doctorant)]
Guillaume Gravier [Junior Researcher (CR1) CNRS, HdR]
Emmanuel Vincent [Junior Researcher (CR1) INRIA]

Technical Staff

Yannick Benezeth [Contractual R&D Engineer - Since November 2010]
Jules Espiau de Lamaestre [Contractual R&D Engineer - Since March 2010]
Olivier Le Blouch [Contractual R&D Engineer - Until April 2010]
Ronan Le Boulch [Contractual R&D Engineer - Since March 2010]
Alexey Ozerov [Contractual R&D Engineer]
Charles Blandin [Contractual Development Engineer - Since October 2010]

PhD Students

Alexis Benichoux [MENRT Grant, Since October 2010]
Quang Khanh Ngoc Duong [INRIA Cordi Grant, 2nd year]
Nobutaka Ito [Franco-Japanese Doctoral College, 2nd Year]
Armando Muscariello [Regional Grant, 3rd year]
Gabriel Sargent [MENRT Grant - 1st year]
Prasad Sudhakar [INRIA Cordis Grant, 3rd year]
Stefan Ziegler [CNRS & Regional Grant - Since October 2010]

Post-Doctoral Fellows

Kamil Adiloglu [Post-Doctorant - Since February 2010]
Valentin Emiya [Post-Doctorant]
Sangnam Nam [Post-Doctorant]
Nikolaos Stefanakis [Post-Doctorant - Since November 2010]
Boris Mailhé [ATER Univ. Rennes 1]

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

In addition to the dissemination of our work through publications in conferences and journals, our scientific activity is accompanied with the permanent concern of evaluation and assessment of our progress within the framework of evaluation campaigns.

This year, our project-team co-organized and participated to the 2010 Signal Separation Evaluation Campaigns, part of the LVA 2010 conference (International Conference on Latent Variable Analysis, formerly known as ICA). The algorithms submitted by our group ranked first in terms of overall perceptual quality (OPS) for the tasks "Under-determined speech and music mixtures" and "Professionally produced music recordings" and second for the task "Source separation in the presence of real-world background noise". For detailed results, please consult : http://sisec.wiki.irisa.fr.

In the framework of the Quaero project, Exalead, LIMSI and INRIA/IRISA (METISS and TEXMEX) collaborated to compete at the ACM Multimedia 2010 Grand Challenge with VoxaleadNews, where we have been awarded the Bronze medal (18 participants in total). VoxaleadNews is a video news web portal with search capabilities, aggregating news videos from several feeds and languages, providing search and browse capabilities based on speech transcription, natural language processing and information retrieval technologies. The spoken document topic segmentation technology is integrated in the demonstration portal, aiming at finding thematically coherent segments from LIMSI's automatic speech transcripts so as to provide structure information and facilitate information retrieval. In addition to topic segments, the IRISA News Topic Segmentation (irints) software enables a keyword based characterization of each segment, selecting those words which participates to the cohesion of the segment. For a live demo, see http://voxaleadnews.labs.exalead.com.

The group was also involved in a number of evaluation campaigns in music processing within the MIREX (Music Information Retrieval Evaluation eXchange) challenge, in particular in music structure segmentation, where our system ranked at a state-of-the-art level.

3. Scientific Foundations

3.1. Introduction

Probabilistic approaches offer a general theoretical framework [106] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [103], 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 [108] have emerged as offering a powerful framework for the modeling of musical signals (for instance, [104], [110]).

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 [107]. 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 [102]. 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 [108].

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 :

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

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

Participants: Guillaume Gravier, Olivier Le Blouch.

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.

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 [105]. 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, Boris Mailhé.

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. IRINTS : IRISA News Topic Segmenter

Participant: Guillaume Gravier.

IRISA News Topic Segmenter is a software dedicated to topic segmentation of texts and automatic transcripts, developed in collaboration with the Texmex team. In 2010, the software has been licensed to several industrial partners and was used in the awarded Voxalead's demonstration at ACM Multimedia.

6. New Results

6.1. Audio and speech content processing

6.1.1. Robust audio segmentation and classification

Participants: Olivier Le Blouch, Guillaume Gravier, Frédéric Bimbot.

This work has taken place in the context of the QUAERO Project.

Improvements were brought to audio segmenting and clustering tasks, w.r.t to robustness to variable quality conditions.

Tests carried out on audio data within the QUAERO evaluation campaign ranked our system first on medium quality audio data, while while keeping state-of-the art performance levels on higher quality audio material.

6.1.2. Speech based structuring and indexing of audio-visual documents

Participant: Guillaume Gravier.

Work done in close collaboration with the TEXMEX *project-team.*

Speech can be used to structure and organize large collections of spoken documents (videos, audio streams...) based on semantics. This is typically achieved by first transforming speech into text using automatic speech recognition (ASR), before applying natural language processing (NLP) techniques on the transcripts. Our research focuses firstly on the adaptation of NLP methods designed for regular texts to account for the specificities of automatic transcripts. In particular, we investigate a deeper integration between ASR and NLP, i.e. between the transcription phase and the semantic analysis phase.

In 2010, we mostly focused on new ASR paradigms, unsupervised topic adaptation, and topic segmentation.

We started to investigate new paradigms for speech recognition exploiting broad-phonetic landmarks to guide the search for the best sentence hypothesis. This emerging work focused so far on model free segmentation techniques to detect spectrally stable regions likely to be landmarks.

We also worked on two aspects of unsupervised adaptation of the linguistic components of an ASR system, namely language model adaptation and adding words to the vocabulary [81]. Firstly, we pursued our work on MDI adaptation of the language model using terminologies, exploiting constraints based on simple or complex terms. Secondly, we proposed an original method to add out-of-vocabulary (OOV) words to the ASR system, combining syntactic and semantic aspects to define equivalences between the OOV word to add and in-vocabulary words.

Finally, we pursued our work on extending Utiyama and Isahara's topic segmentation method [109] for increased robustness to the peculiarities of ASR transcripts. to account for confidence measures, semantic relations and, in collaboration with Columbia University, acoustic cues [78]. We proposed new lexical cohesion measures including confidence measures, semantic relations and, in collaboration with Columbia University, prosodic features.

6.1.3. Audio motif and structure discovery

Participants: Frédéric Bimbot, Guillaume Gravier, Olivier Le Blouch, Armando Muscariello.

Audio motif discovery aims at finding repeating patterns from large audio streams in an unsupervised manner. In 2010, we extended previous work on the DTW-based discovery of word-like patterns [82]. In particular, we investigated distances between self-similarity matrices and demonstrated the robustness of this pattern matching technique to speaker variability. We also proposed speed-up techniques to apply the proposed algorithm to the discovery of patterns with limited variability (e.g., songs and ads) in large streams (several days of radio feeds).

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 submitted to ICASSP2011.

We also proposed an l1 criterion for learning for sparse signal representation in between the analysis and the synthesis model. Instead of directly searching for dictionary vectors that generate the data, our learning approach identifies vectors that are orthogonal to the subspaces in which the training data concentrate. We studied conditions on the synthesis coefficients of training data that guarantee that ideal normal vectors deduced from a synthesis dictionary are local optima of the criterion. We illustrated the behavior of the criterion on a 2D example, showing that the local minima correspond to ideal normal vectors when the number of training data is sufficient. We described an algorithm that can be used to optimize the criterion in higher dimension. These results have been published in ICASSP2010 [65] and have initiated a new line of research which is pursued within the SMALL project.

6.2.2. Theoretical results on sparse representations and dictionary learning

Participants: Rémi Gribonval, Sangnam Nam.

Main collaboration: Karin Schnass (EPFL), Mike Davies (University of Edinburgh), Volkan Cevher (EPFL), Simon Foucart (Université Paris 5, Laboratoire Jacques-Louis Lions)

Sparse representations and inverse problems. 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. Our results from the previous years [10], [2], [11] concentrated on positive results for greedy algorithms and convex optimization (ℓ^1-minimization). In contrast, in 2008-2009, in collaboration with Pr Michael Davies, we concentrated on ℓ^p-minimization, $0 < p \leq 1$, and our results highlighted the pessimistic nature of sparse recovery analysis when recovery is predicted based on the restricted isometry constants (RIC) of the associated matrix (published in [4], [5]), and we extended our analysis of the role of RIC to characterize the stability of ℓ^p minimization with respect to the approximate recovery of vectors which are not exactly sparse [3].

This year, in collaboration with Dr Simon Foucart, we identified and solved an overlooked problem about the characterization of underdetermined systems of linear equations for which sparse solutions have minimal l1-norm. This characterization is known as the null space property. When the system has real coefficients, sparse solutions can be considered either as real or complex vectors, leading to two seemingly distinct null space properties. We proved that the two properties actually coincide by establishing a link with a problem about convex polygons in the real plane. Incidentally, we also show the equivalence between stable null space properties which account for the stable reconstruction by l1-minimization of vectors that are not exactly sparse [34].

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], [23]. These results have been published in the journal paper [35].

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 submitted to IEEE Transactions on Signal Processing, 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 in preparation) 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 [15]. 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: Boris Mailhé, Rémi Gribonval, Frédéric Bimbot, Ronan Le Boulch.

Main collaborations: Pierre Vandergheynst (EPFL)

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 [19], [18]. The main effort this year has been to integrate this algorithm into MPTK to ensure its dissemination and exploitation, and a journal paper is under revision [99].

6.3. Emerging activities on compressive sensing and inverse problems

6.3.1. Nearfield acoustic holography (ECHANGE ANR project)

Participants: Rémi Gribonval, Prasad Sudhakar, Emmanuel Vincent, 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.

This year, 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 approach was submitted to ICASSP 2011 and a journal paper is in preparation.

6.3.2. 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.

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 the following paper: Amir Adler, Valentin Emiya, Maria G. Jafari, Michael Elad, Rémi Gribonval, Mark D. Plumbley, A Constrained Matching Pursuit Approach to Audio Declipping, submitted to ICASSP 2011.

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: Emmanuel Vincent, Nancy Bertin, Kamil Adiloglu.

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 past work on nonnegative matrix factorization (NMF)-based acoustic models was published [44], [30] and the convergence properties of NMF algorithms were analyzed [51], [29]. These models represent 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 [76].

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 [45]. More generally, we started investigating the extraction of robust acoustic features from automatically separated sources by defining and exploiting confidence measures about the separation process.

6.4.2. Music language modeling

Participants: Emmanuel Vincent, Frédéric Bimbot.

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

We proposed a general dynamic Bayesian network structure integrating the most essential features of music into 5 layers: overall features, temporal organization features, symbolic features, performance features and acoustic features [74]. This work pinpointed the most challenging aspects of music modeling, including high dimensionality, large vocabulary, data-dependent vocabulary and long-term dependencies, and provided promising research directions to address them.

We started investigating in particular the modeling of polyphonic note sequences. We proposed a joint model of chord sequences and polyphonic note sequences and evaluated it both in terms of prediction capabilities (aka "perplexity") and polyphonic pitch transcription performance [71]. This model is the first to our knowledge to address joint modeling of "horizontal" (sequential) and "vertical" (simultaneous) dependencies between notes by interpolation of the corresponding conditional probabilities.

6.4.3. Music structuring

Participants: Frédéric Bimbot, Olivier Le Blouch, 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 proposed an approach called *decomposition into autonomous and comparable blocks*, based on principles inspired from structuralism and generativism. It specifies a methodology for producing music structure annotation by human listeners based on simple criteria and resorting solely to the listening experience of the annotator. We have shown on a development set that the proposed approach can provide a high level of concordance across annotators and we have produced a set of annotations on the RWC database, released to the MIR community [77], [52].

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. The algorithm consists in two parts : - segmentation of the audio file, based on localization of timbral breakdowns, short audio events (using the MFCC features) and repeated harmonic progressions (using chroma vectors). - labeling of the structural segments, grouped according to their timbral content by a hierarchical clustering with an adaptive number of clusters. Each group is then assigned to a different label [83].

Tested within the QUAERO project and during the MIREX 2010 campaign [72], the algorithm ranked very favourably for the inference of autonomous and comparable blocks on the RWC database.

6.5. Source separation

6.5.1. A general framework for audio source separation

Participants: Alexey Ozerov, Emmanuel Vincent, Ngoc Duong, Frédéric Bimbot, Rémi Gribonval.

Main collaboration: S. Arberet (Ecole Polytechnique Fédérale de Lausanne, CH)

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 several survey chapters [91], [88], [90], [43], [89], 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. Additional spectral cues must therefore be exploited.

To this aim, we proposed a general probabilistic framework for audio source separation, whereby each source is modeled as a zero-mean Gaussian random variable whose covariance matrix is factored into a scalar spectral variance and a spatial covariance matrix [92]. The spectral variance is itself factored into eight parameter subsets encoding the fine structure or the envelope of the excitation signal or the resonant filter over the frequency or the time axis. Source separation then consists of specifying suitable constraints over these parameter subsets, estimating the parameters in the Maximum A Posteriori (MAP) sense and deriving the source signals by Wiener filtering. This framework makes it possible to combine a range of existing spectral and spatial source models as well as to design novel advanced models, whose potential was evaluated in [42], [70], [50], [61], [69]. In addition, we showed the benefit of using the empirical mixture covariance and an auditory-motivated frequency scale as the input representation [56].

6.5.2. Improved spatial models for reverberant audio

Participants: Ngoc Duong, Nobutaka Ito, Emmanuel Vincent, Rémi Gribonval, Alexey Ozerov, Prasad Sudhakar, Alexis Benichoux.

Main collaborations: M. Kowalski (Laboratoire des Signaux et Systèmes, Supélec), N. Ono (University of Tokyo, JP), C. Blandin (MSc intern)

Besides the lack of spatial cues, another limitation of the state-of-the-art approach is that the acoustic path from the sources to the microphones is modeled as a complex-valued mixing matrix in each frequency bin. This so-called narrowband assumption does not hold for spatially diffuse or reverberated sources. In order to circumvent it, we proposed a family of wideband source separation methods relying on time-domain convolution filters, resulting in large performance improvements in reverberant environments when these filters are known [39].

Based on the theory of statistical room acoustics, we also proposed a family of probabilistic models of the acoustic path fitting into the above general source separation framework. The benefit of these models was demonstrated in large variety of situations, involving both Maximum Likelihood (ML) estimation [31], [55] and Maximum A Posteriori (MAP) estimation with suitable priors [57], [54]. Specific models were also derived for diffuse noise [63] in the line of [38] and exploited for improved source localization [64], [53].

To model not only the direct path (anechoic spatial model) but also the early reflections, we considered a spatial model associated to sparse acoustic filters. We proposed a framework for blind convolutive source localisation exploiting both the filter sparsity and the time-frequency disjointness of the sources [73],[25], [24]. We demonstrated that convex optimization can be used to blindly estimate the sparse filters associated to a source, provided that time-frequency regions where other sources are silent are known. Current work focuses on the harder challenge consisting in blindly estimating such time-frequency regions, as well as merging the sparse echoic model with the proposed probabilistic models of acoustic paths, in the framework of the thesis of Alexis Benichoux.

6.5.3. Perceptual evaluation metrics and artifact reduction techniques towards high-quality audio source separation

Participants: Valentin Emiya, Emmanuel Vincent.

Volker Hohmann (University of Oldenburg, DE), Jonathan Le Roux (NTT Communication Science Laboratories, JP)

Existing source separation techniques typically generate many "musical noise" artifacts due to discontinuities in the time-frequency representation of the estimated sources, which have prevented their use in real-world scenarios such as hearing aids or hi-fi audio so far. In order to overcome this limitation, we proposed two complementary time-frequency smoothing approaches operating over the source covariance matrices [75] and the Wiener filter [68], [80] respectively.

In addition, while the state-of-the-art quality metrics previously developed by METISS remain widely used in the community, the evaluation of such perceptually improved source separation algorithms calls for perceptually more relevant metrics. We proposed a dedicated subjective test protocol for the assessment of source separation quality and collected the scores of 20 subjects over 80 sounds. We then proposed a family of objective measures aiming to predict these subjective scores based on the decomposition of the estimation error into several distortion components and on the use of an auditory processing model. These measures increase correlation with subjective scores up to 0.5 compared to state-of-the-art source separation measures [33], [58].

7. Contracts and Grants with Industry

7.1. National projects

7.1.1. ARC INRIA RAPSODIS

Participant: Guillaume Gravier.

Ended January 2010. Partners: METISS, PAROLE, TALARIS *project-teams, CEA-LIST/LIC2M.*

This project, focused on "syntactic and semantic information-based automated Sspeech Reecognition was aimed at improving automatic speech recognition (ASR) by integrating linguistic information. Based on former work by S. Huet concerning the incorporation of morpho-syntactic knowledge in a post-processing stage of the transcription, we experimented, together with our partners, the deep insertion of automatically obtained semantic relations (especially paradigmatic ones) and syntactic knowledge within an ASR system.

During the project, the objectives were extended to include semantic knowledge acquisition and the use of such knowledge for spoken document processing in addition to speech transcription. In this context, we have worked on corpus-based acquisition of semantic relations for topic segmentation of spoken documents. We compared various classical methods for relation acquisition and measured their impact on out topic segmentation system.

7.1.2. QUAERO CTC and Corpus Projects (OSEO)

Participants: Kamil Adiloglu, Frédéric Bimbot, Guillaume Gravier, Olivier Le Blouch, 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). The project was approved by The European Commission on 11 March 2008.

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.

7.1.3. ANR Attelage de Systèmes Hétérogènes

Participant: Guillaume Gravier.

*Duration: 3 years, started in November 2009. Partners: IRISA/*METISS*, LIA, LIUM*

The project ASH (Automatic System Harnessing) aims at developing new collaborative paradigms for speech recognition. Many current ASR systems rely on an a posteriori combination of the output of several systems (e.g., confusion network combination). In the ASH project, we investigate new approaches in which three ASR systems work in parallel, exchanging information at every step of the recognition process rather than limiting ourselves to an a posteriori combination. What information is to be shared and how to share such information and make use of it are the key questions that the project is addressing. The collaborative paradigm is being extended to landmark-based speech recognition where detection of landmarks and speech transcription can be considered as two (or more) collaborative processes.

7.1.4. 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.

7.1.5. ANR Contint - ETAPE

Participant: Guillaume Gravier.

Duration: 2.5 years (2009-2012). Partners: LPP, LLF, AFCP, ELDA, DGA, LNE

As a continuation of the ESTER campaign, the ETAPE project aims at organizing evaluation campaigns in automatic speech transcription, while broadening the scope and the diversity of contents with respect to previous projects : spontaneous speech, simultaneous speakers, adverse conditions, heavy compression, etc...

The objective of the project is to provide a challenging evaluation framework and a set of spoken resources in French, both for the Engineering Sciences and the Humanities.

7.1.6. DGCIS REV-TV

Participants: Yannick Benezeth, Frédéric Bimbot, Guillaume Gravier.

Duration: 2.25 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.

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.

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. Other Grants and Activities

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 Rennes on August 17-18, 2010, and a total of 11 visits were made between the two teams in 2010. Several papers were published [64], [63], [71], [74] and submitted [45], [76], [54].

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/.

8.1.2. PHC Procope project with the University of Oldenburg

Participants: Emmanuel Vincent, Valentin Emiya.

Duration: 2 years, starting in January 2009.
Partner: Medical Physics section, the University of Oldenburg (DE)

We pursued the collaboration with Volker Hohmann's group at the University of Oldenburg in the framework of a PHC Procope project entitled "Statistical and perceptual modeling for versatile audio source separation". This collaboration has led to the definition of subjective and objective evaluation metrics for audio source separation, published in [33], [58].

9. Dissemination

9.1. Animation of the scientific community

Frédéric Bimbot and Emmanuel Vincent were part of the organizing committee of the 17th Journées d'Informatique Musicale (JIM), held in Rennes on May 18-20, 2010. This conference gathered about 50 participants from the French computer music community.

Frédéric Bimbot has been appointed by the ISCA Association, as General Chairman of the Interspeech 2013 Conference in Lyon (1200 participants expected).

Rémi Gribonval and Emmanuel Vincent were the General Chairs of the last edition of the international conference LVA/ICA on Latent Variable Analysis and Signal Separation, (formerly known as ICA), held in Saint-Malo, September 27-30 2010. This 9th edition of the conference, gathered 120 participants http://lva2010.inria.fr. The proceedings were published by Springer [93].

Emmanuel Vincent is part of the organizing committee and the scientific committee of the CHiME Workshop on Computational Hearing in Multisource Environments, to be held in Florence on September 1, 2011, as a satellite event of Interspeech 2011.

Emmanuel Vincent gave a tutorial on Music Source Separation and its Applications to Music Information Retrieval at ISMIR 2010 (11th International Society for Music Information Retrieval Conference), Utrecht, August 9-13, 2010.

Alexey Ozerov and Ngoc Duong were part of the organizing committee of the second community-based Signal Separation Evaluation Campaign (SiSEC 2010), whose first edition had been initiated by Metiss. The results of the campaign have been published in [48], [49] and presented during a panel session of the 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA 2010). Datasets, evaluation criteria and reference software are available at http://sisec.wiki.irisa.fr/.

Emmanuel Vincent is a co-organizer of the PASCAL 'CHiME' Speech Separation and Recognition Challenge, aiming to evaluate speech separation, feature extraction and speech recognition algorithms in everyday listening conditions. Datasets, evaluation criteria and reference software are available at http://www.dcs.shef.ac.uk/spandh/chime/challenge.html.

Guillaume Gravier is the Vice-President of the Association Francophone de la Communication Parlée (AFCP), acting as a liaison with the Intl. Speech Communication Association (ISCA).

Guillaume Gravier is a member of the scientific committee of Powedia, an IRISA start-up in the field of video diffusion on the Web.

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 was the co-organizer, together with Francis Bach (Projet WILLOW, INRIA-ENS, Paris), of a the one day meeting on "sparsity and machine learning". The meeting was held at Telecom ParisTech, Paris on November 10, 2010 and sponsored by the french GDR ISIS (CNRS). It gathered eight speakers and more than a hundred participants from all regions of France.

Guillaume Gravier is the Scientific Leader of the ANR project ETAPE.

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

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

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

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

9.2. Teaching

Frédéric Bimbot was the coordinator of the ARD module and has given 6 hours of lecture in speech and audio description within the FAV module of the Masters in Computer Science, Rennes I.

Guillaume Gravier is the coordinator and lecturer (20h) for the lecture Data Analysis and Statistical Modeling within the Master in Computer Science , Rennes 1.

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.

Guillaume Gravier was a member of the Comité de Sélection (Selection Committee) in charge of examining applications for assistant professorship at Laboratoire d'Informatique d'Avignon.

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 École Supérieure d'Applications des Transmissions (ESAT, Rennes).

Rémi Gribonval gave a series of tutorial lectures on sparse decompositions and compressed sensing at the Porquerolles10 spring school on Inverse Problems in Signal and Image Processing organized by the French association for signal and image processing, GRETSI.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] E. VINCENT, MARK 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

Articles in International Peer-Reviewed Journal

[28] S. ARBERET, R. GRIBONVAL, F. BIMBOT. *A robust method to count and locate audio sources in a multichannel underdetermined mixture*, in "IEEE Transactions on Signal Processing", jan 2010, vol. 58, n^{o} 1, p. 121–133, http://hal.inria.fr/inria-00305435/fr/.

[29] R. BADEAU, N. BERTIN, E. VINCENT. *Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization.*, in "IEEE Trans. on Neural Networks", 2010, vol. 21, n^{o} 11, p. 1869–1881.

[30] N. BERTIN, R. BADEAU, E. VINCENT. *Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, vol. 18, n^{o} 3, p. 538–549.

[31] N. DUONG, E. VINCENT, R. GRIBONVAL. *Under-determined reverberant audio source separation using a full-rank spatial covariance model*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, vol. 18, n^{o} 7, p. 1830–1840.

[32] V. EMIYA, R. BADEAU, B. DAVID. *Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle*, in "IEEE Transactions on Audio, Speech, and Language Processing", August 2010, vol. 18, n^{o} 6, p. 1643-1654 [*DOI :* 10.1109/TASL.2009.2038819], http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5356234, http://hal.inria.fr/inria-00510392/en.

[33] V. EMIYA, E. VINCENT, N. HARLANDER, V. HOHMANN. *Subjective and objective quality assessment of audio source separation*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, submitted.

[34] S. FOUCART, R. GRIBONVAL. *Real vs. Complex Null Space Properties for Sparse Vector Recovery*, in "Comptes Rendus de l'Academie des Sciences, Paris, Series I", aug 2010, vol. 348, n^{o} 15–16, p. 863–865.

[35] 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.

[36] D. K. HAMMOND, P. VANDERGHEYNST, R. GRIBONVAL. *Wavelets on graphs via spectral graph theory*, in "Applied and Computational Harmonic Analysis", apr 2010, vol. In Press, Corrected Proof [*DOI :* DOI: 10.1016/J.ACHA.2010.04.005], http://www.sciencedirect.com/science/article/B6WB3-4YYGH6T-1/2/a938fa59304fa0016915cf1a623448f7.

[37] S. HUET, G. GRAVIER, P. SÉBILLOT. *Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition*, in "Computer Speech and Language", 2010, n^{o} 24, p. 663–684.

[38] N. ITO, N. ONO, S. SAGAYAMA. *Diffuse noise suppression using crystal-shaped microphone arrays*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, submitted.

[39] M. KOWALSKI, E. VINCENT, R. GRIBONVAL. *Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, vol. 18, n^{o} 7, p. 1818–1829.

[40] A. LLAGOSTERA CASANOVAS, G. MONACI, P. VANDERGHEYNST, R. GRIBONVAL. *Blind Audiovisual Source Separation Based on Sparse Representations*, in "IEEE Transactions on Multimedia", aug 2010, vol. 12, n^{o} 5, p. 358–371.

[41] A. OZEROV, W. KLEIJN. *Asymptotically optimal model estimation for quantization*, in "IEEE Trans. on Communications", 2010, to appear.

[42] A. OZEROV, E. VINCENT, F. BIMBOT. *A general flexible framework for the handling of prior information in audio source separation*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, submitted.

[43] M. D. PLUMBLEY, T. BLUMENSATH, L. DAUDET, R. GRIBONVAL, M. E. DAVIES. *Sparse Representations in Audio and Music: from Coding to Source Separation*, in "Proceedings of the IEEE.", June 2010, vol. 98, n^{o} 6, p. 995–1005.

[44] E. VINCENT, N. BERTIN, R. BADEAU. *Adaptive harmonic spectral decomposition for multiple pitch estimation*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, vol. 18, n^{o} 3, p. 528–537.

[45] 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", 2010, submitted.

International Peer-Reviewed Conference/Proceedings

[46] K. ADILOGLU, C. DRIOLI, P. POLOTTI, D. ROCCHESSO, S. DELLE MONACHE. *Physics-Based Spike-Guided Tools for Sound Design*, in "Conference on Digital Audio Effects", Autriche Graz, Institute of Electronic Music and Acoustics, September 2010, http://hal.inria.fr/inria-00545453/en.

[47] K. ADILOGLU, C. DRIOLI, P. POLOTTI, D. ROCCHESSO, S. D. MONACHE. *Physics-Based Spike-Guided Tools for Sound Design*, in "Conference on Digital Audio Effects", 2010, p. 153–160.

[48] S. ARAKI, A. OZEROV, V. GOWREESUNKER, H. SAWADA, F. THEIS, G. NOLTE, D. LUTTER, N. DUONG. *The 2010 Signal Separation Evaluation Campaign (SiSEC 2010): Audio Source Separation*, in "Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", 2010, p. 114–122.

[49] S. ARAKI, F. THEIS, G. NOLTE, D. LUTTER, A. OZEROV, V. GOWREESUNKER, H. SAWADA, N. DUONG. *The 2010 Signal Separation Evaluation Campaign (SiSEC 2010): Biomedical Source Separation*, in "Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", 2010, p. 123–130.

[50] S. ARBERET, A. OZEROV, N. DUONG, E. VINCENT, R. GRIBONVAL, F. BIMBOT, P. VANDERGHEYNST. *Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation*, in "Proc. 2010 IEEE Int. Conf. on Information Science, Signal Processing and their Applications (ISSPA)", 2010, p. 1–4.

[51] R. BADEAU, N. BERTIN, E. VINCENT. *Stability analysis of multiplicative update algorithms for non-negative matrix factorization*, in "Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2011, submitted.

[52] F. BIMBOT, O. LE BLOUCH, G. SARGENT, E. VINCENT. *Decomposition into autonomous and comparable blocks: A structural description of music pieces*, in "Proc. 2010 Int. Society for Music Information Retrieval Conf. (ISMIR)", 2010, p. 189–194.

[53] C. BLANDIN, E. VINCENT, A. OZEROV. *Multi-source TDOA estimation using SNR-based angular spectra*, in "Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2011, submitted.

[54] 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 "Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2011, submitted.

[55] N. DUONG, E. VINCENT, R. GRIBONVAL. *Under-determined convolutive blind source separation using spatial covariance models*, in "Proc. 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2010, p. 9–12.

[56] N. DUONG, E. VINCENT, R. GRIBONVAL. *Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation*, in "Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", 2010, p. 73–80.

[57] N. DUONG, E. VINCENT, R. GRIBONVAL. *An acoustically-motivated spatial prior for under-determined reverberant source separation*, in "Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2011, submitted.

[58] V. EMIYA, E. VINCENT, N. HARLANDER, V. HOHMANN. *Multi-criteria subjective and objective evaluation of audio source separation*, in "Proc. AES 38th Conf. on Sound Quality Evaluation", 2010, p. 251–259.

[59] J. FAYOLLE, F. MOREAU, C. RAYMOND, G. GRAVIER. *Reshaping automatic speech transcripts for robust high-level spoken document analysis*, in "Proc. Workshop on Analytics for Noisy Unstructured Text Data", 2010.

[60] J. FAYOLLE, F. MOREAU, C. RAYMOND, G. GRAVIER, P. GROS. *CRF-based Combination of Contextual Features to Improve A Posteriori Word-level Confidences Measures*, in "Proc. Annual Intl. Speech Communication Association Conference (Interspeech)", 2010.

[61] C. FÉVOTTE, A. OZEROV. *Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues*, in "Proc. 7th Int. Symp. on Computer Music Modeling and Retrieval (CMMR)", 2010.

[62] C. GUINAUDEAU, G. GRAVIER, P. SÉBILLOT. *Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations*, in "Proc. Annual Intl. Speech Communication Association Conference (Interspeech)", 2010.

[63] N. ITO, N. ONO, E. VINCENT, S. SAGAYAMA. *Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra*, in "Proc. 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2010, p. 2818–2821.

[64] N. ITO, E. VINCENT, N. ONO, R. GRIBONVAL, S. SAGAYAMA. *Crystal-MUSIC: accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays*, in "Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", 2010, p. 81–88.

[65] F. JAILLET, R. GRIBONVAL, M. D. PLUMBLEY, H. ZAYYANI. *An L1 criterion for dictionary learning by subspace identification*, in "Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)", Dallas, 2010, p. 5482–5485, http://dx.doi.org/10.1109/ICASSP.2010.5495206.

[66] H. JÉGOU, M. DOUZE, G. GRAVIER, C. SCHMID, P. GROS. *INRIA LEAR-TEXMEX: Video copy detection task*, in "Proc. of the TRECVid 2010 Workshop", 2010.

[67] M. LAGRANGE, R. BADEAU, B. DAVID, N. BERTIN, J. ECHEVESTE, O. DERRIEN, S. MARCHAND, L. DAUDET. *The DESAM toolbox: spectral analysis of musical audio*, in "International Conference on Digital Audio Effects (DAFx)", Autriche Graz, 2010, p. 254-261, http://hal.inria.fr/hal-00523319/en.

[68] J. LE ROUX, E. VINCENT, Y. MIZUNO, H. KAMEOKA, N. ONO, S. SAGAYAMA. *Consistent Wiener filtering: generalized time-frequency masking respecting spectrogram consistency*, in "Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", 2010, p. 89–96.

[69] A. OZEROV, C. FÉVOTTE, R. BLOUET, J.-L. DURRIEU. *Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation*, in "Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2011, submitted.

[70] A. OZEROV, E. VINCENT, F. BIMBOT. *A general modular framework for audio source separation*, in "Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", 2010, p. 33–40.

[71] S. RACZYNSKI, E. VINCENT, F. BIMBOT, S. SAGAYAMA. *Multiple pitch transcription using DBN-based musicological models*, in "Proc. 2010 Int. Society for Music Information Retrieval Conf. (ISMIR)", 2010, p. 363–368.

[72] G. SARGENT, F. BIMBOT, E. VINCENT. *A structural segmentation of songs using generalized likelihood ratio under regularity assumptions*, in "Proc. 2010 Music Information Retrieval Evaluation eXchange (MIREX)", 2010.

[73] P. SUDHAKAR, S. ARBERET, R. GRIBONVAL. *Double Sparsity: Towards Blind Estimation of Multiple Channels*, in "Latent Variable Analysis and Signal Separation", Saint-Malo, France, V. VIGNERON, V. ZARZOSO, E. MOREAU, R. GRIBONVAL, E. VINCENT (editors), Lecture Notes in Computer Science, Springer, sep 2010, vol. 6365, p. 571–578.

[74] E. VINCENT, S. RACZYNSKI, N. ONO, S. SAGAYAMA. *A roadmap towards versatile MIR*, in "Proc. 2010 Int. Society for Music Information Retrieval Conf. (ISMIR)", 2010, p. 662–664.

[75] E. VINCENT. *An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation*, in "Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", 2010, p. 157–164.

[76] J. WU, E. VINCENT, S. RACZYNSKI, T. NISHIMOTO, N. ONO, S. SAGAYAMA. *Multipitch estimation by joint modeling of harmonic and transient sounds*, in "Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2011, submitted.

National Peer-Reviewed Conference/Proceedings

[77] F. BIMBOT, O. LE BLOUCH, G. SARGENT, E. VINCENT. *Décomposition en blocs autonomes comparables - Une proposition de description et d'annotation de structure pour le traitement automatique des morceaux de musique*, in "Actes des Journées d'Informatique Musicale (JIM)", 2010, p. 187–196.

[78] C. GUINAUDEAU, G. GRAVIER, P. SÉBILLOT. *Indices utiles à la cohésion lexicale pour la segmentation thématique de documents oraux*, in "Journées d'Etude sur la Parole", 2010.

[79] C. GUINAUDEAU, G. GRAVIER, P. SÉBILLOT. *Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels*, in "Traitement Automatique du Language Naturel", 2010.

[80] J. LE ROUX, E. VINCENT, Y. MIZUNO, H. KAMEOKA, N. ONO, S. SAGAYAMA. *Consistent Wiener filtering: designing generalized time-frequency masks respecting spectrogram consistency*, in "Proc. ASJ Spring Meeting", 2010.

[81] G. LECORVÉ, G. GRAVIER, P. SÉBILLOT. *L'adaptation thématique d'un modèle de langue fait-elle apparaître des mots thématiques ?*, in "Journées d'Etude sur la Parole", 2010.

[82] A. MUSCARIELLO, G. GRAVIER, F. BIMBOT. *Découverte non supervisée de mot(if)s dans le signal de parole*, in "Journées d'Etude sur la Parole", 2010.

[83] G. SARGENT, F. BIMBOT, E. VINCENT. *Un système de détection de rupture de timbre pour la description de la structure des morceaux de musique*, in "Actes des Journées d'Informatique Musicale (JIM)", 2010, p. 177–186.

Workshops without Proceedings

[84] A. ADLER, V. EMIYA, M. JAFARI, M. ELAD, R. GRIBONVAL, MARK D. PLUMBLEY. *Audio inpainting: problem statement, relation with sparse representations and some experiments*, in "9th Int. Conf. on Latent Variable Analysis and Signal Separation", France Saint-Malo, September 2010, http://hal.inria.fr/inria-00545480/en.

[85] V. EMIYA, E. VINCENT, N. HARLANDER, V. HOHMANN. *The PEASS Toolkit - Perceptual Evaluation methods for Audio Source Separation*, in "9th Int. Conf. on Latent Variable Analysis and Signal Separation", France Saint-Malo, September 2010, http://hal.inria.fr/inria-00545477/en.

[86] J. LAW-TO, G. GREFENSTETE, J.-L. GAUVAIN, G. GRAVIER, L. LAMEL, J. DESPRES. *VoxaleadNews: Robust Automatic Segmentation of Video Content into Browsable and Searchable Subjects*, in "ACM Multimedia", 2010.

[87] E. VINCENT, N. ONO. *Music source separation and its applications to MIR*, in "2010 Int. Society for Music Information Retrieval Conf. (ISMIR)", Pays-Bas Utrecht, 2010, http://hal.inria.fr/inria-00545508/en.

Scientific Books (or Scientific Book chapters)

[88] G. EVANGELISTA, S. MARCHAND, M. D. PLUMBLEY, E. VINCENT. *Sound source separation*, in "DAFX - Digital Audio Effects, 2nd Edition", U. ZOLZER (editor), Wiley, 2010, to appear.

[89] R. GRIBONVAL, M. ZIBULEVSKY. *Sparse Component Analysis*, in "Handbook of Blind Source Separation, Independent Component Analysis and Applications", P. COMON, C. JUTTEN (editors), Academic Press, 2010, chap. 10, p. 367–420.

[90] A. NESBIT, M. JAFARI, E. VINCENT, M. D. PLUMBLEY. *Audio source separation using sparse representations*, in "Machine Audition: Principles, Algorithms and Systems", W. WANG (editor), IGI Global, 2010, p. 246–265.

[91] E. VINCENT, Y. DEVILLE. *Audio applications*, in "Handbook of Blind Source Separation, Independent Component Analysis and Applications", P. COMON, C. JUTTEN (editors), Academic Press, 2010, p. 779–819.

[92] E. VINCENT, M. JAFARI, S. ABDALLAH, M. D. PLUMBLEY, M. E. DAVIES. *7*, in "Probabilistic modeling paradigms for audio source separation", W. WANG (editor), IGI Global, 2010, p. 162–185.

Books or Proceedings Editing

[93] V. VIGNERON, V. ZARZOSO, E. MOREAU, R. GRIBONVAL, E. VINCENT (editors). *LNCS 6365 - Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation*, Springer, 2010.

Research Reports

[94] F. BIMBOT, O. LE BLOUCH, G. SARGENT, E. VINCENT. *Decomposition into autonomous and comparable blocks : a structural description of music pieces*, INRIA, 2010, n° PI 1948, http://hal.inria.fr/inria-00473479/en.

[95] N. DUONG, E. VINCENT, R. GRIBONVAL. *Under-determined reverberant audio source separation using a full-rank spatial covariance model*, INRIA, 2010, http://hal.inria.fr/inria-00435807/en.

[96] V. EMIYA, N. BERTIN, B. DAVID, R. BADEAU. *MAPS - A piano database for multipitch estimation and automatic transcription of music*, INRIA, July 2010, http://hal.inria.fr/inria-00544155/en.

[97] V. EMIYA, E. VINCENT, N. HARLANDER, V. HOHMANN. *Subjective and objective quality assessment of audio source separation*, INRIA, May 2010, n° RR-7297, http://hal.inria.fr/inria-00485729/en.

[98] R. GRIBONVAL. *Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?*, INRIA, May 2010, n° RR-7484, http://hal.inria.fr/inria-00486840/en.

[99] 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.

[100] A. OZEROV, E. VINCENT, F. BIMBOT. *A General Flexible Framework for the Handling of Prior Information in Audio Source Separation*, INRIA, 2010, n° RR-7453, http://hal.inria.fr/inria-00536917/en.

[101] E. VINCENT. *An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation*, INRIA, April 2010, n° RR-7261, http://hal.inria.fr/inria-00474383/en.

References in notes

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

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

[104] 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.

[105] 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.

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

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

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

[109] M. UTIYAMA, H. ISAHARA. *A Statistical Model for Domain-Independent Text Segmentation*, in "Proceedings of the 39th Annual Meeting of Association for Computational Linguistics, ACL'01", Toulouse, France, July 2001, p. 491-498.

[110] 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.