IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Activity Report 2012

Project-Team METISS

Speech and sound data modeling and processing

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

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Audio, Speech, and Language Processing

Table of contents

Project-Team METISS

Keywords: Audio, Sparse Representations, Statistical Methods, Speech, Machine Learning

Creation of the Project-Team: November 01, 2001 .

1. Members

Research Scientists

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

Engineers

Grégoire Bachman [Contractual R&D Engineer - Until June 2012]
Laurence Catanese [Contractual R&D Engineer]
Jules Espiau de Lamaestre [Contractual R&D Engineer]
Guylaine Le Jan [Contractual R&D Engineer - Until May 2012]
Dimitris Moreau [Contractual R&D Engineer - Since May 2012]
Sangnam Nam [Contractual R&D Engineer - Until July 2012]
Jérémy Paret [Contractual R&D Engineer - Since September 2012]
Nathan Souviraà-Labastie [Contractual R&D Engineer]

PhD Students

Alexis Benichoux [MENRT Grant, 2nd Year]
Anthony Bourrier [Technicolor, 2nd Year]
Corentin Guichaoua [MENRT Grant - 1st Year]
Srdjan Kitic [ERC Please Grant - 1st Year]
Nobutaka Ito [Franco-Japanese Doctoral College, Defended October 2012]
Gabriel Sargent [MENRT Grant - 3rd Year]
Stefan Ziegler [CNRS & Regional Grant, 2nd Year]

Post-Doctoral Fellows

Kamil Adiloglu [Inria - Until April 2012]
Cagdas Bilen [Inria - Since September 2012]
Stanislaw Raczynski [Inria]
Laurent Simon [Inria]
Joachim Thiemann [CNRS]

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 of the Year

The 2nd Prize of the Rennes 1 Foundation was given to Ngoc Duong for his PhD co-supervised by Emmanuel Vincent and Rémi Gribonval.

For his contributions to the field, Emmanuel Vincent was awarded the 2012 SPIE ICA Unsupervised Learning Pioneer Award and gave a keynote at the SPIE DSS conference [49].

Emmanuel Vincent defended his Habilitation à Diriger des Recherches [31].

Reaching the end of its maximum lifespan, the Metiss project-team terminated at the end of the year 2012 and led to the creation of a new project-team Panama, headed by Rémi Gribonval.

3. Scientific Foundations

3.1. Introduction

probabilistic modeling, statistical estimation, bayesian decision theory gaussian mixture modeling, Hidden Markov Model, adaptive representation, redundant system, sparse decomposition, sparsity criterion, source separation

Probabilistic approaches offer a general theoretical framework [92] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [89], 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 [94] have emerged as offering a powerful framework for the modeling of musical signals (for instance, [90], [95]).

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 [93]. 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 [88]. 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

probability density function, gaussian model, gaussian mixture model, Hidden Markov Model, Bayesian network, maximum likelihood, maximum a posteriori, EM algorithm, inference, Viterbi algorithm, beam search, classification, hypotheses testing, acoustic parameterisation

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 [94].

3.3. Sparse representations

wavelet, dictionary, adaptive decomposition, optimisation, parcimony, non-linear approximation, pursuit, greedy algorithm, computational complexity, Gabor atom, data-driven learning, principal component analysis, independant component analysis

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.

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

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:

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

We will denote as $\alpha = \{\alpha_i\}_{1 \le i \le N}$, 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:

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

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

3.3.2. Sparsity criteria

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

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

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

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

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

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

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

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

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

3.3.3. Decomposition algorithms

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

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

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

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

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

3.3.4. Dictionary construction

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

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

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

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

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

3.3.5. Compressive sensing

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

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

4. Application Domains

4.1. Introduction

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

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

The main applicative fields targeted by these tasks are :

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

4.2. Spoken content processing

speaker recognition, user authentication, voice signature, speaker adaptation, spoken document, speech modeling, speech recognition, rich transcription, beam-search, broadcast news indexing, audio-based multimodal structuring

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

audio stream, audio detection, audio tracking, audio segmentation, audio descriptors, multimedia indexing, audiovisual integration, multimodality, information fusion, audio-visual descriptors,

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

audio object, music description, music language modeling, multi-level representations

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

source separation, multi-channel audio, source characterization, source localization, compressive sensing

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

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

4.5.1. Audio source separation

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

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

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

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

4.5.2. Compressive sensing of acoustic fields

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

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

5. Software

5.1. Audio signal processing, segmentation and classification toolkits

Participant: Guillaume Gravier.

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

speech, audio, signal, analysis, processing, audio stream, detection, tracking, segmentation, audio indexing, speaker verification

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

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

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

5.2. Irene: a speech recognition and transcription platform

Participant: Guillaume Gravier.

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

speech modeling, speech recognition, broadcast news indexing, beam-search, Viterbi, HMM

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 [91]. 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, Jules Espiau.

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

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

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

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

Contact : remi.gribonval@irisa.fr

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

5.4. FASST

Participants: Emmanuel Vincent [correspondant], Nancy Bertin, Frédéric Bimbot.

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

5.5. NACHOS

Participants: Nancy Bertin [correspondant], Rémi Gribonval.

The software and associated database were developed within the ANR ECHANGE project, with the participation of Gilles Chardon, Laurent Daudet, François Ollivier and Antoine Peillot.

NACHOS (Nearfield Acoustic HOlography with Sparse regularization) is a downloadable companion software for the journal paper [38], distributed to comply with the "reproducible research" principle. It performs the reconstruction of operational deflection shapes of a vibrating structure, from acoustic measurements of the generated sound field. The software consists in Matlab source code, and automatically downloads the needed database. It allows to reproduce all results and figures of the paper, and to experiment some additional settings. It is distributed under GPL 3.0 license.

6. New Results

6.1. Audio and speech content processing

Audio segmentation, speech recognition, motif discovery, audio mining

6.1.1. Audio motif discovery

Participants: Frédéric Bimbot, Laurence Catanese.

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

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

Audio motif discovery is the task of finding out, without any prior knowledge, all pieces of signals that repeat, eventually allowing variability. The developed algorithms allows discovering and collecting occurrences of repeating patterns in the absence of prior acoustic and linguistic knowledge, or training material.

Former work extended the principles of seeded discovery to near duplicate detection and spoken document retrieval from examples [41].

In 2012, the work achieved consisted in consolidating previously obtained results with the motif discovery algorithm and making implementation choices regardless of the structure and the code, in order to minimize the computation time. This has lead to the creation of a software prototype called MODIS.

After the code has been thoroughly optimised, further optimizations to improve the system performances was to change the method used for the search of similarities between patterns. A new functionality has been added to get rid of unrelevant patterns like silence in speech. New versions of dynamic time warping have been implemented, as well as the possibility to downsample the input sequence during the process, which allows a huge gain of computation time.

The Inria/Metiss team has participated to the IRIT P5 evaluation for repetitive musical motifs discovery. The motif discovery software has been adapted to respect the input and output format defined for the task. The run has been made on a evaluation corpus comprised of French radio broadcast from YACAST.

This work has been carried out in the context of the Quaero Project.

6.1.2. Landmark-driven speech recognition

Participant: Stefan Ziegler.

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

Our previous studies indicate that acoustic-phonetic approaches to ASR, while they cannot achieve state-of-the-art ASR performance by themselves, can prevent HMM-based ASR from degrading, by integrating additional knowledge into the decoding.

In our previous framework we inserted knowledge into the decoding by detecting time frames (referred to as landmarks) which estimate the presence of the active broad phonetic class. This enables the use of a modified version of the viterbi decoding that favours states that are coherent with the detected phonetic knowledge[65].

In 2012 we focused on two major issues. First, we aimed at finding new ways to model and detect phonetic landmarks. Our second focus was on the extension of our landmark detector towards a full acoustic-phonetic framework, to model speech by a variety of articulatory features.

Our new approach for the classification and detection of speech units focuses on developping landmark-models that are different from existing frame-based approaches to landmark detection[64]. In our approach, we use segmentation to model any time-variable speech unit by a fixed-dimensional observation vector. After training any desired classifier, we can estimate the presence of a desired speech unit by searching for each time frame the corresponding segment, that provides the maximum classification score.

We used this segment-based landmark-detection inside a standalone acoustic-phonetic framework that models speech as a stream of articulatory features. In this framework we first search for relevant broad phonetic landmarks, before attaching each landmark with the full set of articulatory features.

Integrating these articulatory feature streams into a standard HMM-based speech recognizer by weighted linear combination improves speech recognition up to 1.5

Additionally, we explored the possibilities of using stressed syllables as an information to guide the viterbi decoding. This work was carried under the leaderhip of Bogdan Ludusan from the team TEXMEX at IRISA [56].

6.1.3. Speech-driven functionalities for interactive television

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

In the context of the collaborative ReV-TV project, the Metiss research group has contributed to technological solutions for the demonstration of new concepts of interactive television, integrating a variety of modalities (audio/voice, gesture, image, haptic feed-back).

The focus has been to provide algorithmic solutions to some advanced audio processing and speech recognition tasks, in particular : keywords recognition, lip synchronisation for an avatar, voice emotion recognition and interactive vocal control.

The main challenges adressed in the project have been to robustify state-of-the-art based technologies to the diversity of adverse conditions, to provide real-time response and to ensure the smooth integration of the various interactive technologies involved in the project.

The work of the project has resulted in a demonstration which was presented at the Forum Imagina 2012

6.2. Recent results on sparse representations

Sparse approximation, high dimension, scalable algorithms, dictionary design, graph wavelets

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 8.2.1.1), the ANR project ECHANGE (ECHantillonnage Acoustique Nouvelle GEnération, see section 8.1.1.2), and the ERC project PLEASE (projections, Learning and Sparsity for Efficient Data Processing, see section 8.2.1.2).

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

Participants: Rémi Gribonval, Sangnam Nam, Nancy Bertin, Srdjan Kitic.

Main collaboration: Mike Davies, Mehrdad Yaghoobi (Univ. Edinburgh), Michael Elad (The Technion).

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

These results have been published in conferences and in a journal paper [42]. Other algorithms based on iterative cosparse projections [83] as well as extensions of GAP to deal with noise and structure in the cosparse representation have been developed, with applications to toy MRI reconstruction problems and acoustic source localization and reconstruction from few measurements [58].

6.2.2. Theoretical results on sparse representations and dictionary learning

Participants: Rémi Gribonval, Sangnam Nam, Nancy Bertin.

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

Sparse recovery conditions for Orthogonal Least Squares : We pursued our investigation of conditions on an overcomplete dictionary which guarantee that certain ideal sparse decompositions can be recovered by some specific optimization principles / algorithms. We extended Tropp's analysis of Orthogonal Matching Pursuit (OMP) using the Exact Recovery Condition (ERC) to a first exact recovery analysis of Orthogonal Least Squares (OLS). We showed that when ERC is met, OLS is guaranteed to exactly recover the unknown support. Moreover, we provided a closer look at the analysis of both OMP and OLS when ERC is not fulfilled. We showed that there exist dictionaries for which some subsets are never recovered with OMP. This phenomenon, which also appears with ℓ_1 minimization, does not occur for OLS. Finally, numerical experiments based on our theoretical analysis showed that none of the considered algorithms is uniformly better than the other. This work has been submitted for publication in a journal [86]. More recently, we obtained simpler coherence-based conditions [85].

Performance guarantees for compressed sensing with spread spectrum techniques : We advocate a compressed sensing strategy that consists of multiplying the signal of interest by a wide bandwidth modulation before projection onto randomly selected vectors of an orthonormal basis. Firstly, in a digital setting with random modulation, considering a whole class of sensing bases including the Fourier basis, we prove that the technique is universal in the sense that the required number of measurements for accurate recovery is optimal and independent of the sparsity basis. This universality stems from a drastic decrease of coherence between the sparsity and the sensing bases, which for a Fourier sensing basis relates to a spread of the original signal spectrum by the modulation (hence the name "spread spectrum"). The approach is also efficient as sensing matrices with fast matrix multiplication algorithms can be used, in particular in the case of Fourier measurements. Secondly, these results are confirmed by a numerical analysis of the phase transition of the l1-minimization problem. Finally, we show that the spread spectrum technique remains effective in an analog setting with chirp modulation for application to realistic Fourier imaging. We illustrate these findings in the context of radio interferometry and magnetic resonance imaging. This work has been accepted for publication in a journal [45].

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

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

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

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

The first one uses a constrained optimization program based on L1 optimization. We derive a practical learning algorithm, based on projected subgradients, and demonstrate its ability to robustly recover a ground truth analysis operator, provided the training set is of sufficient size. A local optimality condition is derived, providing preliminary theoretical support for the well-posedness of the learning problem under appropriate conditions. Extensions to deal with noisy training samples are currently investigated, and a journal paper is under revision [87].

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

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

A first result, which we published last year, 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 $dP_X(x) \neq C \cdot \exp\left(-\phi(x)\right)dx$.

A second result, obtained in collaboration with Prof. Mike Davies and Prof. Volkan Cevher (a paper is under revision) characterizes the "compressibility" of various probability distributions with applications to underdetermined linear regression (ULR) problems and sparse modeling. We identified simple characteristics of probability distributions whose independent and identically distributed (iid) realizations are (resp. are not) compressible, i.e., that can be approximated as sparse. We prove that many priors which MAP Bayesian interpretation is sparsity inducing (such as the Laplacian distribution or Generalized Gaussian distributions with exponent p<=1), are in a way inconsistent and do not generate compressible realizations. To show this, we identify non-trivial undersampling regions in ULR settings where the simple least squares solution outperform oracle sparse estimation in data error with high probability when the data is generated from a sparsity inducing prior, such as the Laplacian distribution [39].

6.3. Emerging activities on compressive sensing, learning and inverse problems

Compressive sensing, acoustic wavefields, audio inpainting,

6.3.1. Nearfield acoustic holography (ECHANGE ANR project)

Participants: Rémi Gribonval, Nancy Bertin.

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

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

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

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

A paper about robustness issues and spatial multiplexing (an alternative to building antennas with random sensor position) was published in GRETSI last year and as a journal paper this year [38].

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

Participants: Rémi Gribonval, Nancy Bertin.

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

Underwater acoustic imaging is traditionally performed with beamforming: beams are formed at emission to insonify limited angular regions; beams are (synthetically) formed at reception to form the image. We proposed to exploit a natural sparsity prior to perform 3D underwater imaging using a newly built flexible-configuration sonar device. The computational challenges raised by the high-dimensionality of the problem were highlighted, and we described a strategy to overcome them. As a proof of concept, the proposed approach was used on real data acquired with the new sonar to obtain an image of an underwater target. We discussed the merits of the obtained image in comparison with standard beamforming, as well as the main challenges lying ahead, and the bottlenecks that will need to be solved before sparse methods can be fully exploited in the context of underwater compressed 3D sonar imaging. This work has been published in [61] and a journal paper is in preparation.

6.3.3. Audio inpainting (SMALL FET-Open project)

Participants: Rémi Gribonval, Nancy Bertin, Corentin Guichaoua.

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

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

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

The contributions consist of:

- defining audio inpainting as a general scheme where missing audio data must be estimated: it covers a number of existing audio processing tasks that have been addressed separately so far – click removal, declipping, packet loss concealment, unmasking in time-frequency;
- proposing algorithms based on sparse representations for audio inpainting (based on Matching Pursuit and on l_1 minimization);
- addressing the case of audio declipping (*i.e.* desaturation): thanks to the flexibility of our inpainting algorithms, they can be constrained so as to include the structure of signals due to clipping in the objective to optimize. The resulting performance are significantly improved. This work will appear as a journal paper [33].

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), dictionary learning for inpainting, and several applicative scenarios (declipping, time-frequency inpainting).

6.3.4. Blind Calibration of Compressive Sensing systems

Participants: Rémi Gribonval, Cagdas Bilen.

Main collaborations: Gilles Chardon, Laurent Daudet (Institut Langevin), Gilles Puy (EPFL)

We consider the problem of calibrating a compressed sensing measurement system under the assumption that the decalibration consists in unknown gains on each measure. We focus on blind calibration, using measures performed on a few unknown (but sparse) signals. A naive formulation of this blind calibration problem, using l1 minimization, is reminiscent of blind source separation and dictionary learning, which are known to be highly non-convex and riddled with local minima. In the considered context, we show that in fact this formulation can be exactly expressed as a convex optimization problem, and can be solved using off-the-shelf algorithms. Numerical simulations demonstrate the effectiveness of the approach even for highly uncalibrated measures, when a sufficient number of (unknown, but sparse) calibrating signals is provided. We observe that the success/failure of the approach seems to obey sharp phase transitions. This work has been published at ICASSP 2012 [54], and an extension dealing with the problem of phase-only decalibration, using techniques revolving around low-rank matrix recovery, has been submitted to ICASSP 2013. A journal version is in preparation.

6.3.5. Compressive Gaussian Mixture estimation

Participants: Rémi Gribonval, Anthony Bourrier.

Main collaborations: Gilles Blanchard (University of Potsdam), Patrick Perez (Technicolor R&D, FR)

When fitting a probability model to voluminous data, memory and computational time can become prohibitive. In this paper, we pro- pose a framework aimed at fitting a mixture of isotropic Gaussians to data vectors by computing a low-dimensional sketch of the data. The sketch represents empirical moments of the underlying probability distribution. Deriving a reconstruction algorithm by analogy with compressive sensing, we experimentally show that it is possible to precisely estimate the mixture parameters provided that the sketch is large enough. Our algorithm provides good reconstruction and scales to higher dimensions than previous probability mixture estimation algorithms, while consuming less memory in the case of numerous data. It also provides a privacy-preserving data analysis tool, since the sketch does not disclose information about individual datum it is based on. This work has been submitted for publication at ICASSP 2013.

6.3.6. Nearest neighbor search for arbitrary kernels with explicit embeddings

Participants: Rémi Gribonval, Anthony Bourrier.

Main collaborations: Hervé Jégou (TEX-MEX team), Patrick Perez (Technicolor R&D, FR)

Many algorithms have been proposed to handle efficient search in large databases for simple metrics such as the Euclidean distance. However, few approaches apply to more sophisticated Positive Semi-Definite (PSD) kernels. In this document, we propose for such kernels to use the concept of explicit embedding and to cast the search problem into a Euclidean space. We first describe an exact nearest neighbor search technique which relies on bounds on the approximation of the kernel. We show that, in the case of SIFT descriptors, one can retrieve the nearest neighbor with probability 1 by computing only a fraction of the costly kernels between the query and the database vectors. We then propose to combine explicit embedding with a recent Euclidean approximate nearest neighbor search method and show that it leads to significant improvements with respect to the state-of-the-art methods which rely on an implicit embedding. The database vectors being indexed by short codes, the approach is shown to scale to a dataset comprising 200 million vectors on a commodity server. This work has been submitted for journal publication [74]

6.4. Music Content Processing and Music Information Retrieval

Acoustic modeling, non-negative matrix factorisation, music language modeling, music structure

6.4.1. Music language modeling

Participants: Frédéric Bimbot, Dimitris Moreau, Stanisław Raczyński, Emmanuel Vincent.

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

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.

We pursued our pioneering work on music language modeling, with a particular focus on the joint modeling of "horizontal" (sequential) and "vertical" (simultaneous) dependencies between notes by log-linear interpolation of the corresponding conditional distributions. We identified the normalization of the resulting distribution as a crucial problem for the performance of the model and proposed an exact solution to this problem [81]. We also applied the log-linear interpolation paradigm to the joint modeling of melody, key and chords, which evolve according to different timelines [80]. In order to synchronize these feature sequences, we explored the use of beat-long templates consisting of several notes as opposed to short time frames containing a fragment of a single note.

The limited availability of multi-feature symbolic music data is currently an issue which prevents the training of the developed models on sufficient amounts of data for the unsupervised probabilistic approach to significantly outperform more conventional approaches based on musicological expertise. We outlined a procedure for the semi-automated collection of large-scale multifeature music corpora by exploiting the wealth of music data available on the web (audio, MIDI, leadsheets, lyrics, etc) together with algorithms for the automatic detection and alignment of matching data. Following this work, we started collecting pointers to data and developing such algorithms.

6.4.2. Music structuring

Participants: Frédéric Bimbot, Gabriel Sargent, Emmanuel Vincent.

External collaboration: Emmanuel Deruty (as an independant consultant)

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

This year, our methodology for the *semiotic* annotation of music pieces has developed [72] and concretized into a set of principles, concepts and conventions for locating the boundaries and determining metaphoric labels of music segments [53] [71]. The method relies on a new concept for characterizing the inner organization of music segments called the System & Contrast (S&C) model [73]. At the time of writing this text, the annotation of over 400 music pieces is being finalized and will be released to the MIR scientific community.

In parallel to this work aiming at specifying the task of music structure description, we have designed, implemented and tested new algorithms for segmenting and labeling music into structural units. The segmentation process is formulated as a cost optimization procedure, accounting for two terms : the first one corresponds to the characterization of structural segments by means of the fusion of audio criteria, whereas the second term relies on a regularity constraint on the resulting segmentation. Structural labels are estimated as a probabilistic automaton selection process. A recent development of this work has included the S&C model in the algorithm.

Different systems based on these principles have been tested in the context of the Quaero Project and the MIREX international evaluation campaigns in 2010, 2011 and 2012 (see for instance [66], in 2012).

6.5. Source separation

Source separation, sparse representations, probabilistic model, source localization

6.5.1. A general framework for audio source separation

Participants: Frédéric Bimbot, Rémi Gribonval, Nobutaka Ito, Emmanuel Vincent.

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

Source separation is the task of retrieving the source signals underlying a multichannel mixture signal. The state-of-the-art approach 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. Recently, we proposed a general probabilistic framework for the joint exploitation of spatial and spectral cues [44], which generalizes a number of existing techniques including our former study on spectral GMMs [34]. This framework makes it possible to quickly design a new model adapted to the data at hand and estimate its parameters via the EM algorithm. As such, it is expected to become the basis for a number of works in the field, including our own.

Since the EM algorithm is sensitive to initialization, we devoted a major part of our work to reducing this sensitivity. One approach is to use some prior knowledge about the source spatial covariance matrices, either via probabilistic priors [75] or via deterministic subspace constraints [76]. The latter approach was the topic of the PhD thesis of Nobutaka Ito who defended this year [30]. A complementary approach is to initialize the parameters in a suitable way using source localization techniques specifically designed for environments involving multiple sources and possibly background noise [37].

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

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

Main collaboration: Simon Arberet (EPFL)

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

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

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

On a more theoretical side, we studied the frequency permutation ambiguity traditionnally incurred by blind convolutive source separation methods. We focussed on the filter permutation problem in the absence of scaling, investigating the possible use of the temporal sparsity of the filters as a property enabling permutation correction. The obtained theoretical and experimental results highlight the potential as well as the limits of sparsity as an hypothesis to obtain a well-posed permutation problem. This work has been published in a conference [52] and is accepted for publication as a journal paper, to appear in 2013.

6.5.3. Towards real-world separation and remixing applications

Participants: Nancy Bertin, Frédéric Bimbot, Jules Espiau de Lamaestre, Jérémy Paret, Laurent Simon, Nathan Souviraà-Labastie, Joachim Thiemann, Emmanuel Vincent.

Shoko Araki, Jonathan Le Roux (NTT Communication Science Laboratories, JP)

We participated in the organization of the 2011 Signal Separation Evaluation Campaign (SiSEC) [51], [59]. Following our founding role in the organization of this campaign, we wrote an invited paper summarizing the outcomes of the three first editions of this campaign from 2007 to 2010 [47]. While some challenges remain, this paper highlighted that progress has been made and that audio source separation is closer than ever to successful industrial applications. This is also exemplified by the ongoing i3DMusic project and the recently signed contracts with Canon Research Centre France and MAIA Studio.

In order to exploit our know-how for these real-world applications, we investigated issues such as how to implement our algorithms in real time [60], how to reduce artifacts [40] and how best to exploit extra information or human input. In addition, while the state-of-the-art quality metrics previously developed by METISS remain widely used in the community, we proposed some improvements to the perceptually motivated metrics introduced last year [62].

6.5.4. Source separation for multisource content indexing

Participants: Kamil Adiloğlu, Emmanuel Vincent.

Main collaborations: Jon Barker (University of Sheffield, UK), Mathieu Lagrange (IRCAM, FR), Alexey Ozerov (Technicolor R&D, FR)

Another promising real-world application of source separation concerns information retrieval from multi-source data. Source separation may then be used as a pre-processing stage, such that the characteristics of each source can be separately estimated. The main difficulty is not to amplify errors from the source separation stage through subsequent feature extraction and classification stages. To this aim, we proposed a principled Bayesian approach to the estimation of the uncertainty about the separated source signals [50], [69], [68] and propagated this uncertainty to the features. We then exploited it in the training of the classifier itself, thereby greatly increasing classification accuracy [43].

This work was applied both to singer identification in polyphonic music [55] and to speech and speaker recognition in real-world nonstationary noise environments. In order to motivate further work by the community, we created a new international evaluation campaign on that topic (CHiME) in 2011 and analyzed the outcomes of the first edition [36].

Some work was also devoted to the modeling of similarity between sound events [32].

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

7.1.1. Contract with Canon Research Centre France SAS

Participants: Emmanuel Vincent, Joachim Thiemann, Nancy Bertin, Frédéric Bimbot.

Duration: 1.5 years (2012–2013). Partner: Canon Research Centre France SAS

This contract aims at transfering some of the research done within Metiss to products developed by Canon Inc.

7.1.2. Contract with Studio Maïa

Participants: Nancy Bertin, Frédéric Bimbot, Jules Espiau, Jérémy Paret, Emmanuel Vincent.

Duration: 3 years (2012–2014). Partners: Studio Maïa (Musiciens Artistes Interprètes Associés), Imaging Factory

This contract aims at transfering some of the research done within Metiss towards new services provided by Maïa Studio.

More specifically, the main objective is to adapt source separations algorithms and some other advanced signal processing techniques elaborated by Metiss in a user-informed context.

The objective is twofold:

- partial automation of some tasks which the user previously had to accomplish manually
- improved quality of separation and processing by exploiting user inputs and controls

The resulting semi-automated separation and processing will feed an integrated software used for the professional remastering of audiovisual pieces.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. National Projects

8.1.1.1. QUAERO CTC and Corpus Projects (OSEO)

Participants: Kamil Adiloglu, Frédéric Bimbot, Laurence Catanese, Gabriel Sargent, Emmanuel Vincent.

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

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

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

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

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

8.1.1.2. ANR ECHANGE

Participants: Rémi Gribonval, Emmanuel Vincent, Nancy Bertin.

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) was 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.

8.1.1.3. DGCIS REV-TV

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

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

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

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

8.1.2. Action de Développement Technologique

8.1.2.1. FASST

Participants: Nancy Bertin, Emmanuel Vincent, Frédéric Bimbot.

Duration: 2 years (2012–2014). Partners: Inria Teams Parole (Nancy) and Texmex (Rennes)

This Inria ADT aims to develop a new version of our FASST audio source separation toolbox in order to facilitate its large-scale dissemination in the source separation community and in the various application communities. A specific effort will be made towards the speech processing community by developing an interface with existing speech recognition software.

8.2. European Initiatives

8.2.1. FP7 Projects

8.2.1.1. SMALL

Participants: Rémi Gribonval, Jules Espiau de Lamaestre, Sangnam Nam, Emmanuel Vincent, Nancy Bertin.

Title: Sparse Models, Algorithms and Learning for Large-scale data

Type: COOPERATION (ICT)

Defi: FET Open

Instrument: Specific Targeted Research Project (STREP)

Duration: February 2009 - January 2012

Coordinator: Inria (France)

Others partners: Univ. Edimburg (UK), Queen Mary Univ. (UK), EPFL (CH), Technion Univ. (ISR)

See also: http://small-project.eu/

Abstract: The project has developed new foundational theoretical framework for dictionary learning, and scalable algorithms for the training of structured dictionaries.

8.2.1.2. PLEASE

Title: Projections, Learning and Sparsity for Efficient data processing.

Type: IDEAS ()

Instrument: ERC Starting Grant (Starting)

Duration: January 2012 - December 2016

Coordinator: Inria (France)

Principal investigator: Rémi Gribonval

Abstract: The Please ERC is focused on the extension of the sparse representation paradigm towards that of “sparse modeling”, with the challenge of establishing, strengthening and clarifying connections between sparse representations and machine learning

8.2.2. Collaborations in other European Programs

Program: Eureka - Eurostars

Project acronym: i3DMusic

Project title: Real-time Interative 3D Rendering of Musical Recordings

Duration: October 2010 - September 2013

Other partners: Audionamix (FR), Sonic Emotion (CH), École Polytechnique Fédérale de Lausanne (CH)

Abstract:The i3DMusic project (Real-time Interative 3D Rendering of Musical Recordings) has been setup with the SMEs Audionamix and Sonic Emotion and the academic partner EPFL 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.

8.3. International Initiatives

8.3.1. Inria Associate Teams

8.3.1.1. VERSAMUS

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

Title: Integrated probabilistic music representations for versatile music content processing

Inria principal investigator: Emmanuel Vincent

International Partner (Institution - Laboratory - Researcher):

Tokyo University (Japan) - Department of Physics and Computing

Duration: 2010 - 2012

See also: http://versamus.inria.fr/

Music plays a major role in everyday use of digital media contents. Companies and users are waiting for smart content creation and distribution functionalities, such as music classification, search by similarity, summarization, chord transcription, remixing and automatic accompaniment. So far, research efforts have focused on the development of specific algorithms and corpora for each functionality based on low-level sound features characterizing sound as a whole. Yet, music generally results from the superposition of heterogeneous sound components (e.g. voices, pitched musical instruments, drums, sound samples) carrying interdependent features at several levels (e.g. music genre, singer identity, melody, lyrics, voice signal). Integrated music representations combining all feature levels would make it possible to address all of the above functionalities with increased accuracy as well as to visualize and interact with the content in a musically relevant manner. The aim of this project was to investigate, design and validate such representations in the framework of Bayesian data analysis, which provides a rigorous way of combining separate feature models in a modular fashion. Tasks addressed in the project have included the design of a versatile model structure, of a library of feature models and of efficient algorithms for parameter inference and model selection.

9. Dissemination

9.1. Animation of the scientific community

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

Frédéric Bimbot was a member of the Comité National de la Recherche Scientifique (Section 07 and Inter-Disciplinary Commission 42)

Frédéric Bimbot is the General Chairman of the Interspeech 2013 Conference in Lyon (1200 participants expected).

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

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

Rémi Gribonval and Emmanuel Vincent were Guest Editors of the special issue on Latent Variable Analysis and Signal Separation of the journal *Signal Processing* published by Elsevier [46].

Rémi Gribonval and Emmanuel Vincent are members of the International Steering Committee for the LVA conference series.

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

Emmanuel Vincent was a Guest Editor of the special issue on Speech Separation and Recognition in Multisource Environments of the journal *Computer Speech and Language* published by Elsevier [35].

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

Emmanuel Vincent is the General Chair of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME), to be held in Vancouver as a satellite even of ICASSP 2013 on June 1, 2013, and an organizer of the 2nd CHiME Speech Separation and Recognition Challenge.

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

Nancy Bertin was elected a member of the IEEE Technical Committee on Audio and Acoustic Signal Processing (2013–2015).

9.2. Teaching

Rémi Gribonval gave a series of tutorial lectures on sparse decompositions and compressed sensing at two-day workshop on "Sparse Representations and Machine Learning" (Rennes, October 2012).

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

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

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

Emmanuel Vincent gave a tutorial on “Uncertainty handling for environment-robust speech recognition” with Ramon F. Astudillo and Li Deng at InterSpeech 2012 (13th Annual Conference of the International Speech Communication Association), Portland, OR, September 9–13, 2012.

Nancy Bertin gave lectures about audio rendering, coding, indexing, classification, speech processing and source separation, for a total of 12 hours as part of the CTR and FAV modules of the Masters in Computer Science, Université Rennes I.

10. Bibliography

Major publications by the team in recent years

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

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

[3] M. 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] R. Gribonval, K. Schnass. *Dictionary Identifiability - Sparse Matrix-Factorisation via ℓ_1 minimisation*, in "IEEE Trans. Information Theory", jul 2010, vol. 56, n^o 7, p. 3523–3539.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[30] N. ITO. *Robust microphone array signal processing against diffuse noise*, University of Tokyo, January 2012, http://hal.inria.fr/tel-00691931.

[31] E. VINCENT. *Contributions à la séparation de sources et à la description des contenus audio*, Université Rennes 1, November 2012, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00758517.

Articles in International Peer-Reviewed Journals

[32] K. ADILOGLU, A. ROBERT, W. ELIO, P. HENDRIK, O. KLAUS. *A Graphical Representation and Dissimilarity Measure for Basic Everyday Sound Events*, in "IEEE Transactions on Audio, Speech and Language Processing", January 2012, vol. 20, n^{o} 5, p. 1542-1552, http://hal.inria.fr/hal-00684620.

[33] A. ADLER, V. EMIYA, M. JAFARI, M. ELAD, R. GRIBONVAL, M. D. PLUMBLEY. *Audio Inpainting*, in "IEEE Transactions on Audio, Speech and Language Processing", March 2012, vol. 20, n^{o} 3, p. 922 - 932 [*DOI :* 10.1109/TASL.2011.2168211], http://hal.inria.fr/inria-00577079.

[34] S. ARBERET, A. OZEROV, F. BIMBOT, R. GRIBONVAL. *A tractable framework for estimating and combining spectral source models for audio source separation*, in "Signal Processing", August 2012, vol. 92, n^{o} 8, p. 1886-1901, http://hal.inria.fr/hal-00694071.

[35] J. BARKER, E. VINCENT. *Special Issue on Speech Separation and Recognition in Multisource Environments*, in "Computer Speech and Language", October 2012, http://hal.inria.fr/hal-00743532.

[36] J. BARKER, E. VINCENT, N. MA, H. CHRISTENSEN, P. GREEN. *The PASCAL CHiME Speech Separation and Recognition Challenge*, in "Computer Speech and Language", October 2012, http://hal.inria.fr/hal-00743529.

[37] C. BLANDIN, A. OZEROV, E. VINCENT. *Multi-source TDOA estimation in reverberant audio using angular spectra and clustering*, in "Signal Processing", March 2012, vol. 92, p. 1950-1960, Revised version including minor corrections in equations (17), (18) and Figure 1 compared to the version published by Elsevier [*DOI :* 10.1016/J.SIGPRO.2011.10.032], http://hal.inria.fr/inria-00630994.

[38] G. CHARDON, L. DAUDET, A. PEILLOT, F. OLLIVIER, N. BERTIN, R. GRIBONVAL. *Nearfield Acoustic Holography using sparsity and compressive sampling principles*, in "Journal of the Acoustical Society of America", 2012, Code & data for reproducing the main figures of this paper are available at http://echange.inria.fr/nah, http://hal.inria.fr/hal-00720129.

[39] R. GRIBONVAL, V. CEVHER, M. DAVIES. *Compressible Distributions for High-dimensional Statistics*, in "IEEE Transactions on Information Theory", 2012, Was previously entitled "Compressible priors for high-dimensional statistics" [*DOI :* 10.1109/TIT.2012.2197174], http://hal.inria.fr/inria-00563207.

[40] J. LE ROUX, E. VINCENT. *Consistent Wiener filtering for audio source separation*, in "IEEE Signal Processing Letters", October 2012, http://hal.inria.fr/hal-00742687.

[41] A. MUSCARIELLO, F. BIMBOT, G. GRAVIER. *Unsupervised Motif Acquisition in Speech via Seeded Discovery and Template Matching Combination*, in "IEEE Transactions on Audio, Speech and Language Processing", September 2012, vol. 20, n^{o} 7, p. 2031 - 2044 [*DOI :* 10.1109/TASL.2012.2194283], http://hal.inria.fr/hal-00740978.

[42] S. NAM, M. E. DAVIES, M. ELAD, R. GRIBONVAL. *The Cosparse Analysis Model and Algorithms*, in "Applied and Computational Harmonic Analysis", 2012, Preprint available on arXiv since 24 Jun 2011 [*DOI :* 10.1016/J.ACHA.2012.03.006], http://hal.inria.fr/inria-00602205.

[43] A. OZEROV, M. LAGRANGE, E. VINCENT. *Uncertainty-based learning of acoustic models from noisy data*, in "Computer Speech and Language", July 2012, http://hal.inria.fr/hal-00717992.

[44] A. OZEROV, E. VINCENT, F. BIMBOT. *A General Flexible Framework for the Handling of Prior Information in Audio Source Separation*, in "IEEE Transactions on Audio, Speech and Language Processing", May 2012, vol. 20, n^o 4, p. 1118 - 1133, 16, http://hal.inria.fr/hal-00626962.

[45] G. PUY, P. VANDERGHEYNST, R. GRIBONVAL, Y. WIAUX. *Universal and efficient compressed sensing by spread spectrum and application to realistic Fourier imaging techniques*, in "EURASIP Journal on Advances in Signal Processing", 2012 [*DOI :* 10.1186/1687-6180-2012-6], http://hal.inria.fr/inria-00582432.

[46] V. VIGNERON, V. ZARZOSO, R. GRIBONVAL, E. VINCENT. *Latent variable analysis and signal separation*, in "Signal Processing", January 2012, vol. 92, p. 1765-1766 [*DOI :* 10.1016/J.SIGPRO.2012.01.001], http://hal.inria.fr/hal-00658459.

[47] E. VINCENT, S. ARAKI, F. J. THEIS, G. NOLTE, P. BOFILL, H. SAWADA, A. OZEROV, B. V. GOWREESUNKER, D. LUTTER, N. DUONG. *The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges*, in "Signal Processing", March 2012, vol. 92, p. 1928-1936 [*DOI :* 10.1016/J.SIGPRO.2011.10.007], http://hal.inria.fr/inria-00630985.

Invited Conferences

[48] R. F. ASTUDILLO, E. VINCENT, L. DENG. *Uncertainty handling for environment-robust speech recognition*, in "Interspeech", Portland, États-Unis, September 2012, http://hal.inria.fr/hal-00664091.

[49] E. VINCENT. *Advances in audio source separation and multisource audio content retrieval*, in "SPIE Defense, Security, and Sensing", Baltimore, États-Unis, April 2012, http://hal.inria.fr/hal-00664090.

International Conferences with Proceedings

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

[51] S. ARAKI, F. NESTA, E. VINCENT, Z. KOLDOVSKY, G. NOLTE, A. ZIEHE, A. BENICHOUX. *The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Audio source separation -*, in "10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", Tel Aviv, Israël, March 2012, p. 414-422, http://hal.inria.fr/hal-00655394.

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

[53] F. BIMBOT, E. DERUTY, G. SARGENT, E. VINCENT. *Semiotic structure labeling of music pieces: Concepts, methods and annotation conventions*, in "13th International Society for Music Information Retrieval Conference (ISMIR)", Porto, Portugal, October 2012, http://hal.inria.fr/hal-00758648.

[54] R. GRIBONVAL, G. CHARDON, L. DAUDET. *Blind Calibration for Compressed Sensing by Convex Optimization*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing", Kyoto, Japon, 2012, http://hal.inria.fr/hal-00658579.

[55] M. LAGRANGE, A. OZEROV, E. VINCENT. *Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning*, in "13th International Society for Music Information Retrieval Conference (ISMIR)", Porto, Portugal, October 2012, http://hal.inria.fr/hal-00709826.

[56] B. LUDUSAN, S. ZIEGLER, G. GRAVIER. *Integrating Stress Information in Large Vocabulary Continuous Speech Recognition*, in "Interspeech", Portland, U.S.A., 2012, http://hal.inria.fr/hal-00758622.

[57] F. METZE, N. RAJPUT, X. ANGUERA, M. DAVEL, G. GRAVIER, C. VAN HEERDEN, G. MANTENA, A. MUSCARIELLO, K. PRADHALLAD, I. SZÖKE, J. TEJEDOR. *The Spoken Web Search task at MediaEval 2011*, in "IEEE International Conference on Acoustics, Speech and Signal Processing", France, 2012, 4, http://hal.inria.fr/hal-00671011.

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

[59] G. NOLTE, D. LUTTER, A. ZIEHE, F. NESTA, E. VINCENT, Z. KOLDOVSKY, A. BENICHOUX, S. ARAKI. *The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Biomedical data analysis -*, in "10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", Tel Aviv, Israël, March 2012, p. 423-429, http://hal.inria.fr/hal-00655801.

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

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

[62] E. VINCENT. *Improved perceptual metrics for the evaluation of audio source separation*, in "10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)", Tel Aviv, Israël, March 2012, p. 430-437, http://hal.inria.fr/hal-00653196.

[63] M. YAGHOOBI, S. NAM, R. GRIBONVAL, M. DAVIES. *Noise Aware Analysis Operator Learning for Approximately Cosparse Signals*, in "ICASSP - IEEE International Conference on Acoustics, Speech, and Signal Processing - 2012", Kyoto, Japon, IEEE, 2012, http://hal.inria.fr/hal-00661549.

[64] S. ZIEGLER, B. LUDUSAN, G. GRAVIER. *Towards a new speech event detection approach for landmark-based speech recognition*, in "IEEE Workshop on Spoken Language Technology", États-Unis, 2012, http://hal.inria.fr/hal-00758424.

[65] S. ZIEGLER, B. LUDUSAN, G. GRAVIER. *Using Broad Phonetic Classes to Guide Search in Automatic Speech Recognition*, in "Interspeech", Portland, U.S.A., 2012, http://hal.inria.fr/hal-00758427.

Conferences without Proceedings

[66] G. Sargent, F. Bimbot, E. Vincent. *A Music Structure Inference Algorithm Based on Morphological Analysis*, in "The Music Information Retrieval Evaluation eXchange (MIREX), ISMIR 2012", Porto, Portugal, October 2012, http://hal.inria.fr/hal-00727791.

Scientific Books (or Scientific Book chapters)

[67] B. Durand, F. Bimbot, I. Bloch, A. Charara, et al.. *Sciences et Technologies de l'Information (Informatique, Automatique, Signal et Communication)*, in "Rapport de Conjoncture 2010 du Comité National du CNRS", CNRS, 2012, n^{o} 7, p. 121–137.

Research Reports

[68] K. Adiloglu, E. Vincent. *Supporting Technical Report for the Article "Variational Bayesian Inference for Source Separation and Robust Feature Extraction"*, Inria, April 2012, n^{o} RT-0423, 25, http://hal.inria.fr/hal-00687162.

[69] K. Adiloglu, E. Vincent. *Variational Bayesian Inference for Source Separation and Robust Feature Extraction*, Inria, August 2012, n^{o} RT-0428, http://hal.inria.fr/hal-00726146.

[70] A. Benichoux, L. S. R. Simon, E. Vincent, R. Gribonval. *Convex regularizations for the simultaneous recording of room impulse responses*, Inria, November 2012, n^{o} RR-8130, http://hal.inria.fr/hal-00749585.

[71] F. Bimbot, E. Deruty, G. Sargent, E. Vincent. *Complementary Report to the Article "Semiotic structure labeling of music pieces : concepts, methods and annotation conventions" (Proceedings ISMIR 2012)*, IRISA, June 2012, n^{o} PI 1996, http://hal.inria.fr/hal-00713196.

[72] F. Bimbot, E. Deruty, G. Sargent, E. Vincent. *Methodology and conventions for the latent semiotic annotation of music structure*, IRISA, February 2012, n^{o} PI-1993, http://hal.inria.fr/hal-00676509.

[73] F. Bimbot, E. Deruty, G. Sargent, E. Vincent. *System & Contrast : a Polymorphous Model of the Inner Organization of Structural Segments within Music Pieces*, IRISA, December 2012, n^{o} PI 1999.

[74] A. Bourrier, F. Perronnin, R. Gribonval, P. Pérez, H. Jégou. *Nearest neighbor search for arbitrary kernels with explicit embeddings*, Inria, August 2012, n^{o} RR-8040, http://hal.inria.fr/hal-00722635.

[75] N. Q. K. Duong, E. Vincent, R. Gribonval. *Spatial location priors for Gaussian model-based reverberant audio source separation*, Inria, September 2012, n^{o} RR-8057, http://hal.inria.fr/hal-00727781.

[76] N. Ito, E. Vincent, N. Ono, S. Sagayama. *Robust estimation of directions-of-arrival in diffuse noise based on matrix-space sparsity*, Inria, October 2012, n^{o} RR-8120, http://hal.inria.fr/hal-00746271.

[77] R. Jenatton, R. Gribonval, F. Bach. *Local stability and robustness of sparse dictionary learning in the presence of noise*, Polytechnique, October 2012, 41, http://hal.inria.fr/hal-00737152.

[78] J. Le Roux, E. Vincent. *Consistent Wiener filtering for audio source separation*, Inria, August 2012, n^{o} RR-8049, http://hal.inria.fr/hal-00725350.

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

[80] S. RACZYNSKI, S. FUKAYAMA, E. VINCENT. *Melody harmonisation with interpolated probabilistic models*, Inria, October 2012, n^{o} RR-8110, http://hal.inria.fr/hal-00742957.

[81] S. RACZYNSKI, E. VINCENT, S. SAGAYAMA. *Dynamic Bayesian networks for symbolic polyphonic pitch modeling*, Inria, September 2012, n^{o} RT-0430, http://hal.inria.fr/hal-00728771.

Other Publications

[82] V. EMIYA, N. STEFANAKIS, J. MARCHAL, N. BERTIN, R. GRIBONVAL, P. CERVENKA. *Underwater acoustic imaging: sparse models and implementation issues*, May 2012, Projet ANR : ANR-09-EMER-001, http://hal.inria.fr/hal-00677287.

[83] R. GIRYES, S. NAM, M. ELAD, R. GRIBONVAL, M. E. DAVIES. *Greedy-Like Algorithms for the Cosparse Analysis Model*, http://hal.inria.fr/hal-00716593.

[84] C. GUICHAOUA. *Dictionary Learning for Audio Inpainting*, Computer Science, Rennes 1, June 2012, http://dumas.ccsd.cnrs.fr/dumas-00725263.

[85] C. HERZET, C. SOUSSEN, J. IDIER, R. GRIBONVAL. *Coherence-based Partial Exact Recovery Condition for OMP/OLS*, http://hal.inria.fr/hal-00759433.

[86] C. SOUSSEN, R. GRIBONVAL, J. IDIER, C. HERZET. *Joint k-step analysis of Orthogonal Matching Pursuit and Orthogonal Least Squares*, 2012, http://hal.inria.fr/hal-00637003.

[87] M. YAGHOOBI, S. NAM, R. GRIBONVAL, M. E. DAVIES. *Constrained Overcomplete Analysis Operator Learning for Cosparse Signal Modelling*, 2012, 28 pages, 11 figures, http://hal.inria.fr/hal-00699556.

References in notes

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

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

[90] 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.

[91] 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.

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

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

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

[95] 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.