INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team METISS

Modélisation et Expérimentation pour le Traitement des Informations et des Signaux Sonores

Rennes - Bretagne-Atlantique

Theme : Audio, Speech, and Language Processing

Activity Report

2009

Table of contents

METISS is a joint research group between CNRS, INRIA, Rennes 1 University and INSA.

1. Team

Research Scientist

Frédéric Bimbot [Team Leader, Research Director CNRS, HdR]
Rémi Gribonval [Research Director INRIA, HdR]
Guillaume Gravier [CR1 CNRS, HdR]
Emmanuel Vincent [CR2 INRIA]

Technical Staff

Simon Arberet [Contractual R&D Engineer - Until September 2009]
Nancy Bertin [Contractual R& D Engineer - Since September 2009]
Pierre Cauchy [Contractual Development Engineer - Until May 2009]
Florent Jaillet [Contractual R&D Engineer - Since June 2009]
Olivier Le Blouch [Contractual R&D Engineer]
Alexey Ozerov [Contractual R&D Engineer - Since September 2009]

PhD Student

Quang Khanh Ngoc Duong [INRIA Cordi Grant, 1st year]
Nobutaka Ito [Franco-Japanese Doctoral College, 1st Year]
Boris Mailhé [ENS Cachan (Bruz) - Defended December 2009]
Armando Muscariello [Regional Grant, 2nd year]
Gabriel Sargent [MENRT Grant - Since November 2009]
Prasad Sudhakaramurthy [INRIA Cordis Grant, 2nd year]

Post-Doctoral Fellow

Valentin Emiya [INRIA Grant]
Sangnam Nam [Contractual Postdoc]

Administrative Assistant

Stéphanie Lemaile

2. Overall Objectives

2.1. Presentation

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

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

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

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

2.2. Highlights

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

This year, our project-team organized and participated to the Signal Separation Evaluation Campaigns during the ICA conferences (International Conference on Independent Component Analysis and Signal Separation). The campaign has shown that the performance of the source separation algorithms developed by Metiss are at, or above, the level of the best state-of-the-art competitors [SiSEC09], for a large variety of performance criteria. For more details, refer to : [SiSEC09] http://sassec.gforge.inria.fr/SiSEC_underdetermined/test_eval.html, Algorithms 3,4 and 10.

The group was also involved in a number of evaluation campaigns in audio processing within the ESTER and QUAERO projects : our systems ranked at a state-of-the-art level in audio classification and in multipitch estimation.

3. Scientific Foundations

3.1. Introduction

Probabilistic approaches offer a general theoretical framework [73] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [69], 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 [75] have emerged as offering a powerful framework for the modeling of musical signals (for instance, [70], [77]).

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 [74]. 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 [68]. The potential applicative outcomes are numerous.

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

3.2. Probabilistic approach

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

3.2.1. Probabilistic formalism and modeling

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

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

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

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

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

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

3.2.2. Statistical estimation

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

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

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

3.2.3. Likelihood computation and state sequence decoding

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

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

3.2.4. Bayesian decision

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

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

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

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

3.2.5. Graphical models

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

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

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

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

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

The second problem, called the inference problem, consists in estimating the activity states of the model for a given signal in the maximum a posteriori sense. A number of techniques are available to achieve this goal (sampling methods, variational methods belief propagation, ...), whose challenge is to achieve a good compromise between tractability and accuracy [75].

3.3. Sparse representations

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

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

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

3.3.1. Redundant systems and adaptive representations

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

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

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

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

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

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

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

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

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

3.3.2. Sparsity criteria

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

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

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

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

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

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

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

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

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

3.3.3. Decomposition algorithms

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

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

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

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

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

3.3.4. Dictionary construction

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

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

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

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

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

3.3.5. Compressive sensing

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

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

4. Application Domains

4.1. Introduction

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

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

The main applicative fields targeted by these tasks are :

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

4.2. Spoken content processing

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

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

4.2.1. Robustness issues in speaker recognition

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

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

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

4.2.2. Speech recognition for multimedia analysis

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

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

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

4.3. Description and structuration of audio streams

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

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

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

4.3.1. Detecting and tracking sound classes and events

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

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

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

4.3.2. Describing multi-modal information

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

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

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

4.3.3. Recurrent audio pattern detection

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

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

4.4. Advanced processing for music information retrieval

4.4.1. Audio signal analysis and decomposition

The standards within the MPEG family, notably MPEG-4, introduce several sound description and transmission formats, with the notion of a "score", *i.e.* a high-level MIDI-like description, and an "orchestra", *i.e.* a set of "instruments" describing sonic textures. These formats promise to deliver very low bitrate coding, together with indexing and navigation facilities. However, it remains a challenge to design methods for transforming an arbitrary existing audio recording into a representation by such formats.

Audio object coding is an extension of the notion of parametric coding, where the signal is decomposed into meaningful sound objects such as notes, chords and instruments, described using high-level attributes. As well as offering the potential for very low bitrate compression, this coding paradigm leads to many other potential applications, including browsing by content, source separation and interactive signal manipulation.

4.4.2. 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.

4.4.3. Multi-level representations for music information retrieval

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

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

4.5. Audio scene analysis

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

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

4.5.1. Audio source separation

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

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

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

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

4.5.2. Compressive sensing of acoustic fields

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

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

5. Software

5.1. Audio signal processing, segmentation and classification toolkits

Participant: Guillaume Gravier.

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

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

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

5.2. Irene: a speech recognition and transcription platform

Participant: Guillaume Gravier.

In collaboration with the computer science dept. at ENST, METISS has actively participated in the past years in the development of the freely available Sirocco large vocabulary speech recognition software [71]. 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, Florent Jaillet, Boris Mailhé.

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

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

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

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

Contact : remi.gribonval@irisa.fr

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

6. New Results

6.1. Audio segmentation and classification

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

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

New developments were given to audio segmenting and clustering tasks, w.r.t to new merging and stopping criteria for the clustering process. The main goal was to enhance the robustness of the BIC approach and to get a more reliable stopping criterion, thus leading to a better segmentation of audio documents.

We also developed and evaluated speech and music detection and classification algorithms on the QUAERO audio corpus (80 hours, among which 40 hours were annotated at IRISA). Test were carried out with several configurations of the AudioSeg toolkit and in various combinations of blind source separation methods.

Our best system performed at a state-of-the-art level. Source separation approaches did occasionnaly improve the results but they did not show yet a systematic advantage.

6.2. Speech recognition for multimedia structuring and indexing

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

Participant: Guillaume Gravier.

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

Speech can be used to structure and index large collections of spoken documents (videos, audio streams...) based on semantics. This is typically achieved by first transforming speech into text using automatic speech recognition (ASR), before applying natural language processing (NLP) techniques on the transcriptions. Our research focuses on the integration of ASR and NLP techniques in the framework of large scale analysis of multimedia document collections.

In 2009, several aspects were considered, namely topic segmentation using semantic relations, unsupervised topic adaptation and semantic verification of TV programs.

We improved our transcript-based topic segmentation method based on an extension of the initial work of Utiyama and Isahara [76] that accounts for additional knowledge sources such as acoustic cues or semantic relations between words [12]. In particular, we further investigated the use of semantic relations, implementing a mathematically rigorous framework to account for such relations and comparing several methods for their automatic corpus-based acquisition. We demonstrated on a TV news corpus that directly using automatically generated semantic relations increases precision on topic boundaries to the expanse of a lower recall. This result points out the need for a careful selection of the relations to be considered.

Given thematically homogeneous segments, we pursued our work on unsupervised topic adaptation of the ASR system language model. Elaborating on our previous work based on the automatic acquisition of adaptation data from the Web, we investigated constraint selection in MDI adaptation. Experiments reported in [43] have shown that considering only a small number of terms in MDI contraints, *i.e.*, topic-specific words, is sufficient to perform an efficient adaptation. In addition to this result, it has also been shown that these terms can be automatically extracted from a small topic-specific corpus without any prior knowledge.

Finally, we extended our method for the semantic validation of automatic alignments of TV streams with an electronic program guide (EPG). The method is based on a comparison of the speech transcripts with the short program description provided by the EPG to validate the alignment. The comparison combines lexical and phonetic information retrieval techniques to define a distance between transcripts and descriptions. In 2009, we validated the method on a large dataset, introducing time-based constraints to limit computation [41].

6.3. Audio motif and structure discovery

6.3.1. Audio motif and structure discovery

Keywords: *Bayesian networks*, *data mining*, *pattern discovery*.

Participants: Frédéric Bimbot, Guillaume Gravier, Armando Muscariello.

6.3.1.1. Audio motif discovery

Audio motif discovery aims at finding repeating patterns from large audio streams in an unsupervised manner. Extending the segmentation framework defined in [72], we proposed a motif discovery method tolerant to variations in both the spectral and temporal domains. Our method relies on a dynamic time warping algorithm with relaxed boundary constraints to search for repetitions of a seed block of signal in the near future. Repeating motifs are found by extending the seed when a match is found in the near future are iterativeloy stored in a library of motifs for long-term matching. The algorithm has been used in a word-discovery task which demonstrates the effectiveness of the approach to retrieve repeating motifs (fillers, words, locutions) in radio broadcast news data.

6.3.1.2. Discovering audiovisual structuring events in videos

Work carried out in collaboration with M. Ben and S. Campion from the TEXMEX *project-team.*

We have developed a cross-modal technique for the automatic discovery of audiovisual structuring events in TV programs, using only little prior knowledge for the definition of the targeted events. The algorithm is based on two separate hierarchical clustering processes, one for audio segments and one for video shots. The two resulting clustering trees are then correlated by measuring the mutual information between each pair of audio/video (A/V) clusters. The most correlated pair of cluster provides an initial segmentation into structuring events whose content is coherent both from the audio and visual viewpoint. Experiments on several kinds of TV programs have shown that the technique is able to extract the most relevant parts of the video, from a structuring point of view: anchorperson shots for TV news and report programs, audio/video jingles separating the reports for flash news programs.

6.4. Recent results on sparse representations

The team has had a substantial activity ranging from theoretical results to algorithmic design and software contributions in the field of sparse representations, which is at the core of the Equipe Associée SPARS (see Section 8.1.1) initiated in 2006 between METISS and the LTS2 lab at EPFL as well as the FET-Open European project (FP7) SMALL (Sparse Models, Algorithms and Learning for Large-Scale Data, to begin in 2009, see Section 7.2.1) and the ANR project ECHANGE (ECHantillonnage Acoustique Nouvelle GEnération, see, Section 6.5.1).

6.4.1. Algorithmic breakthrough in sparse approximation : LoCOMP

Participants: Boris Mailhé, Rémi Gribonval, Frédéric Bimbot.

Main collaborations: Pierre Vandergheynst (EPFL), Thomas Blumensath (Univ. Edinburgh), Emmanuel Ravelli, Laurent Daudet (LAM, Université Pierre et Marie Curie, Paris 6)

Our team had already made a substantial breakthrough in 2005 when first releasing the Matching Pursuit ToolKit (MPTK, see Section 5.3) which allowed for the first time the application of the Matching Pursuit algorithm to large scale data such as hours of CD-quality audio signals. In 2008, we designed a variant of Matching Pursuit called LoCOMP (ubiquitously for LOw Complexity Orthogonal Matching Pursuit or Local Orthogonal Matching Pursuit) speifically designed for shift-invariant dictionaries. LoCOMP has been shown to achieve an approximation quality very close to that of a full Orthonormal Matching Pursuit while retaining a much lower computational complexity of the order of that of Matching Pursuit. The complexity reduction is substantial, from one day of computation to 15 minutes for a typical audio signal [44], [61], and the algorithm is being integrated into MPTK, in collaboration with Dr Thomas Blumensath. Moreover, joint experiments have been performed together with Dr Emmanuel Ravelli and Pr Laurent Daudet to assess the impact of this new algorithm on the audio codec developed at LAM which is based on MPTK.

6.4.2. Theoretical results on dictionary learning

Participant: Rémi Gribonval.

Main collaboration: Karin Schnass (EPFL)

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 [10] 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. This second result was published at the conference EUSIPCO 2008 [9], and a journal paper has been submitted [66].

6.4.3. Theoretical results on identification of sparse representations

Participant: Rémi Gribonval.

Main collaboration: Mike Davies (Univ. Edinburgh), Simon Foucart (Univ. Paris VI)

We pursued our investigation of conditions on an overcomplete dictionary which guarantee that certain ideal sparse decompositions can be recovered by some specific optimization principles. Our results from the previous years [7], [2], [8] concentrated on positive results for greedy algorithms and convex optimization (ℓ^1-minimization).

In contrast, in 2008, in collaboration with Pr Michael Davies, we concentrated on ℓ^p-minimization, $0 < p \leq 1$, and our results highlighted the pessimistic nature of sparse recovery analysis when recovery is predicted based on the restricted isometry constants (RIC) of the associated matrix (published in [24], [36]). This year, we extended our analysis of the role of RIC to characterize the stability of ℓ^p minimization with respect to the approximate recovery of vectors which are not exactly sparse [35]. Moreover, in collaboration with Dr Simon Foucart, we iidentified and solve an overlooked problem about the characterization of underdetermined systems of linear equations for which sparse solutions have minimal l1-norm. This characterization is known as the null space property. When the system has real coefficients, sparse solutions can be considered either as real or complex vectors, leading to two seemingly distinct null space properties. We proved that the two properties actually coincide by establishing a link with a problem about convex polygons in the real plane. Incidentally, we also show the equivalence between stable null space properties which account for the stable reconstruction by l1-minimization of vectors that are not exactly sparse [67].

6.4.4. Shift-invariant dictionary learning algorithms and experiments with atrial signal extraction in ECG.

Participants: Boris Mailhé, Rémi Gribonval, Frédéric Bimbot.

Main collaborations: Pierre Vandergheynst and Matthieu Lemay (EPFL)

In addition to our pioneering theoretical work on dictionary identifiability, we amplified the effort begun in 2007 on the design of dictionary learning algorithms for structured shift-invariant dictionaries. This work, performed in the framework of the Ph.D. of Boris Mailhé, was published at the conference EUSIPCO 2008 [14]. The proposed approach was further developed to study the problem of ventricular cancellation and atrial modelling in the ECG of patients suffering from atrial fibrillation, in collaboration with Mathieu Lemay from EPFL [45].

6.5. Emerging activities on compressive sensing

6.5.1. Compressed sensing of Acoustic Wavefields (ECHANGE ANR project)

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

Main collaborations: Albert Cohen (Laboratoire Jacques-Louis Lions, Université Paris 6), Laurent Daudet, François Ollivier, Jacques Marchal (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. This research has began in the framework of the Ph.D. of Prasad Sudhakar (started in december 2007), and we have set up the ANR collaborative project ECHANGE (ECHantillonnage Acoustique Nouvelle GEnération) which is due to begin in January 2009. Rémi Gribonval is the coordinator of the project.

The main challenges are: a) to identify dictionaries of basic wavefield atoms making it possible to sparsely represent the wavefield in several acoustic scenarios of interest; b) to determine which types of (networks) of acoustic sensors maximise the identifiability of the sparse wavefield representation, depending on the acoustic scenario; c) to design scalable algorithms able to reconstruct the measured wavefields in a region of interest.

6.5.2. Compressed sensing of wideband signals

Participant: Rémi Gribonval.

Main collaborations: Laurent Jacques (EPFL & UCL Belgique), Pierre Vandergheynst (EPFL), Farid Nani Mohavedian

Compressed sensing is also the object of a collaboration with EPFL in the framework of the Equipe Associée SPARS 8.1.1. In the framework of the summer internship of Mr Farid Naini Mohavedian, we studied the application of compressed sensing to ultra wide-band signals. More precisely, we studied a model where the considered signals are sparse linear superpositions of shifts of a known, potentially wide-band, pulse. This signal model is key for applications such as Ultra Wide Band (UWB) communications or neural signal processing. We compared several acquisition strategies and showed that the approximations recovered via ℓ^1 minimization are greatly enhanced if one uses Spread Spectrum analog modulation prior to applying random Fourier measurements. We complemented our experiments with a discussion of possible hardware implementation of our technique, and checked that a simplified hardware implementation did not degrade the performance of the compressed sensing system. The results have been published at the conference ICASSP 2009 [48].

6.5.3. Wavelets on graphs

Participant: Rémi Gribonval.

Main collaboration: Pierre Vandergheynst, David Hammond (EPFL)

Within the framework of the Equipe Associée SPARS 8.1.1, we investigated the possibility of developping sparse representations of functions defined on graphs, by defining an extension to the traditional wavelet transform which is valid for data defined on a graph.

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

Computing the spectral decomposition can however be numerically expensive for large graphs. We have shown that, by approximating the spectrum of the wavelet generating operator with polynomial expansions, applying the forward wavelet transform and its transpose can be approximated through iterated applications of the graph Laplacian. Since in many cases the graph Laplacian is sparse, this results in a very fast algorithm. Our implementation also uses recurrence relations for computing polynomial expansions, which results in even faster algorithms. Finally, we have proved how numerical errors are precisely controlled by the properties of the desired spectral graph wavelets. Our algorithms have been implemented in a Matlab toolbox that will be released in parallel to the main theoretical article [28]. We also plan to include this toolbox in the SMALL project numerical platform.

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

6.6. Content description of music signals

6.6.1. Multi-pitch signal modeling

Participant: Emmanuel Vincent.

Main collaborations: N. Bertin and R. Badeau (Telecom ParisTech)

Music involves several levels of information, from the acoustic signal up to cognitive quantities such as composer style or key, through mid-level quantities such as a musical score or a sequence of chords. The dependencies between mid-level and lower- or higher-level information can be represented through acoustic models and language models, respectively. Our past work on nonnegative matrix factorization (NMF)-based acoustic models was finalized and led to several publications [32], [23], [33], [34]. These models represent an input short-term magnitude spectrum as a linear combination of magnitude spectra corresponding to different pitches, which are adapted to the input under harmonicity and temporal smoothness constraints. Besides, the convergence properties of NMF algorithms were analyzed [22].

6.6.2. Music language modeling

Participants: Emmanuel Vincent, Frédéric Bimbot.

Main collaboration: Ricardo Scholz (internship student),

We started working on the modeling of music as a language by studying N-gram models of chord sequences. We investigated various chord labelling schemes and various model smoothing techniques originally designed for spoken language processing. While state-of-the-art models consider N=2, we showed that more accurate models with N > 2 could be learned from a limited set of data [54].

Additional investigations (in the context of Christophe Hauser's internship) were carried out on how to integrate the language model with the acoustic level decoding, but did not reach sufficient maturity yet, to draw clear conclusions.

6.6.3. Music structuring

Participants: Frédéric Bimbot, Gabriel Sargent, Emmanuel Vincent.

In the context of the QUAERo Projec, we started investigating on various ways of describing the structure of musical content, with the double purpose of proposing a data model for annotation and a simple paradigm for automatic algorithms.

Initially, a multi-layered approach was considered, composed of seven parallel layers of information: key, color, tempo, melody, rhythm, harmony and lyrics, which together govern most of the structure of a piece of music. Some of these layers are characterized by the existence of statistical changes (key, color, tempo) whereas the others are mostly structured by the presence of recurrent patterns (melody, rhythm, harmony and lyrics).

Above this, it appears that, in many situations, the characterization of the structure of a music piece is governed by a high level structure linked to constant numbers of beats. This property serves as an anchor point for structure description and annotation and will be a central point of study in the PhD of Gabriel Sargent.

6.7. Source separation

6.7.1. Source separation via sparse and adaptive representations

Participants: Emmanuel Vincent, Remi Gribonval.

Main collaboration: Andrew Nesbit (Queen Mary, University of London), Matthieu Puigt (Laboratoire d'Astrophysique de Toulouse-Tarbes), Matthieu Kowalski (Laboratoire des Signaux et Systèmes, Supelec)

Source separation is the task of retrieving the source signals underlying a multichannel mixture signal, where each channel is the sum of filtered versions of the sources. The state-of-the-art approach, which we presented in a survey chapter [64], consists of representing the signals in a given time-frequency basis and estimating the source coefficients by sparse decomposition in that basis, based on narrowband approximation of the mixing process. This approach often provides limited performance due to poor approximation of the mixing process in reverberant environments and to the use of a time-frequency basis where the sources overlap. We proposed a family of wideband source separation methods that circumvent the narrowband assumption and result in large performance improvements in reverberant environments [30]. In parallel, we studied a range of adaptive lapped orthogonal time-frequency bases originally designed for audio coding and explained how to estimate the best basis in a source separation context [63], [50], [49]. Finally, we provided an experimental validation of the implicit source independence assumption underlying the above approaches [51].

6.7.2. New paradigms and new evaluation metrics for source separation

Participants: Emmanuel Vincent, Valentin Emiya, Ngoc Duong, Simon Arberet, Remi Gribonval, Nobutaka Ito.

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

In parallel with our work on sparse representations, we proposed a new generic probabilistic framework for audio source separation where each source is modeled as a zero-mean random variable whose parameters vary over the time-frequency plane [65], [58]. This framework makes it possible to combine a range of existing spectral and spatial source models as well as to design novel advanced models such as models of reverberated or spatially diffuse sources. The benefits of this framework were demonstrated both for the separation of instantaneous [57] and reverberant mixtures [25], [37], [38], [42].

In addition, ideas to model additional phase dependencies between neighboring time-frequency bins or to replace the usual ML learning framework by discriminative learning were investigated in [52], [53] and [39] respectively. Finally, the state-of-the-art audio source separation evaluation metrics previously developed by METISS were further improved using a computional auditory processing front-end and a neural network to fit subjective performance measurements [26]. Theoretical performance bounds were also proposed for source separation methods based on Gaussian Mixture Models (GMM) of the source spectra [60].

7. Contracts and Grants with Industry

7.1. National projects

7.1.1. ARC INRIA RAPSODIS: Syntactic and Semantic Information-Based Automated Speech Recognition

Participant: Guillaume Gravier.

Duration: 2 years, starting in February 2008. Partners: METISS, PAROLE, TALARIS *project-teams, CEA-LIST/LIC2M.*

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

In 2009, the objectives of the project were extended to include semantic knowledge acquisition and the use of such knowledge for spoken document processing in addition to speech transcription. In this extended framework, we have worked on corpus-based acquisition of semantic relations for topic segmentation of spoken documents. We compared various classical methods for relation acquisition and measured their impact on out topic segmentation system.

7.1.2. QUAERO CTC and Corpus Projects (OSEO)

Participants: Simon Arberet, Frédéric Bimbot, Guillaume Gravier, Olivier Le Blouch, Alexey Ozerov, Gabriel Sargent, Emmanuel Vincent.

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

Quaero is a European research and development program with the goal of developing multimedia and multilingual indexing and management tools for professional and general public applications (such as search engines). The project was approved by The European Commission on 11 March 2008.

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

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

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

7.1.3. ANR Attelage de Systèmes Hétérogènes

Participant: Guillaume Gravier.

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

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

7.1.4. ANR ECHANGE

Participants: Rémi Gribonval, Prasad Sudhakaramurthy, Emmanuel Vincent, Nancy Bertin, Florent Jaillet, Valentin Emiya.

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

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

7.2. European projects

7.2.1. FP7 FET-Open program SMALL

A joint research project called SMALL (Sparse Models, Algorithms and Learning for Large-scale data) has been setup with the groups of Pr Mark Plumbley (Centre for Digital Music, Queen Mary University of London, UK), Pr Mike Davies University of Edinburgh, UK), Pr Pierre Vandergheynst (EPFL, Switzeland) and Miki Elad (The Technion, Israel) in the framework of the European FP7 FET-Open call. SMALL was one of the eight selected projects among more than 111 submissions and is scheduled to begin in February 2009. The main objective of the project is to explore new generations of provably good methods to obtain inherently data-driven sparse models, able to cope with large-scale and complicated data much beyond state-of-the-art sparse signal modeling. The project will develop a radically new foundational theoretical framework for dictionary learning, and scalable algorithms for the training of structured dictionaries.

8. Other Grants and Activities

8.1. European initiatives

8.1.1. Associated Team SPARS with EPFL

Participants: Rémi Gribonval, Boris Mailhé, Simon Arberet, Frédéric Bimbot.

A strong partnership with the LTS2 lab lead by Pr. Pierre Vandergheynst at EPFL has been ongoing since 20606 and was formalized as the INRIA Equipe Associée SPARS in January 2007. The two groups share a common specialty on nonlinear and sparse approximation, with complementary expertise on audio (METISS) and image/video (LTS2). The Ph.D. of Boris Mailhé has been co-supervised by Frédéric Bimbot, Rémi Gribonval and Pierre Vandergheynst in this framework.

Since the official labelling of the Equipe Associée the academic exchanges between the groups have been further reinforced, with exchanges of Ph.D. students, crossed participations to Ph.D. jurys, and two two-month visits of Rémi Gribonval at EPFL as an "academic host" in the summers of 2006 and 2008. As a result of this collaboration there have been regular publications, including 3 publications in international peer-reviewed journals and 5 conference publications since the beginning of the Equipe Associée.

The Equipe Associée has also been the opportunity to jointly organize the SPARS'09 workshop (see Section 9.1 as well as to prepare the project SMALL (see Section 7.2.1).

9. Dissemination

9.1. Conference and workshop committees, invited conference

Rémi Gribonval was the co-organizer, together with Laure Blanc-Feraud (Projet ARIANA, I3S, Nice), of a the one day meeting on "sparsity". The meeting was held at Telecom ParisTech, Paris on April 17, 2008 and sponsored by the french GDR ISIS. It gathered twelve speakers and more than a hundred participants from all regions of France.

Rémi Gribonval was the General Chair of the workshop SPARS'09 on Signal Processing with Adaptive Sparse/Structured Representations, held in Saint-Malo, April 7-10 2009. This was the second edition of the workshop. The first edition was organized in Rennes in 2005 and gathered 65 international participants. This year we received more than 60 contributions and about 90 participants http://spars09.inria.fr.

Rémi Gribonval and Emmanuel Vincent are the General Chairs of the next edition of the international conference LVA/ICA on Latent Variable Analysis and Signal Separation, formerly known as the international conference on Independent Component Analysis and Signal Separation, to be held in Saint-Malo, September 27-30 2010. This is the 9th edition of the conference, and we are expecting around 150 participants http://lva2010.inria.fr.

Emmanuel Vincent gave a keynote talk on probabilistic paradigms for audio source separation at the 2009 ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP).

Frédéric Bimbot gave 3 tutorials in Speaker Recognition at WiSSAP-09 (Winter School on Speech and Audio Processing), 9612 JAnuary 2009, Kanpur, India.

9.2. Leadership within scientific community

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

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

Guillaume Gravier is a member of the Administration Board of the Association Francophone de la Communication Parlée (AFCP).

Guillaume Gravier was the organiser of the second ESTER evaluation campaign on the segmentation and transcription of audio contents.

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

Emmanuel Vincent was the chair of the first community-based Signal Separation Evaluation Campaign (SiSEC), co-organized with Shoko Araki (NTT, Japan) and Pau Bofill (University of Catalonia, Spain). The results of the campaign have been published in [56] and presented during a special session of the 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA 2009). Datasets, evaluation criteria and reference software are available at http://sisec.wiki.irisa.fr/.

Rémi Gribonval is a member of the International Steering Committee for the ICA conferences, and the Chair of the Steering Committee of the SPARS workshops.

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

9.3. Teaching

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

Guillaume Gravier has given 10 hours of lecture in Data Analysis and Statistical Modeling within the ADM module of the Master 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.

In 2009, Guillaume Gravier was a member of the Comité de Sélection (Selection Committee) in charge of examining applications for assistant professorship at INSA Rennes.

Rémi Gribonval gave a series of tutorial lectures on sparse decompositions and compressed sensing at the Peyresq09 summer school on Inverse Problems in Signal and Image Processing organized by the French association for signal and image processing, GRETSI.

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

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

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, Ph. D. Thesis.

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

[3] 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.

[4] 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", vol. 86, n^{o} 3, March 2006, p. 496–510.

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

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

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

[8] 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.", vol. 14, n$^{\text{o}}$ 5, December 2008, p. 655–687.

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

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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", vol. 15, n$^{\text{o}}$ 5, juillet 2007, p. 1564–1578.

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

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

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

Year Publications

Doctoral Dissertations and Habilitation Theses

[19] G. GRAVIER. *Intégration de connaissances par modèles probabilistes pour l'analyse de documents multimédias*, Université de Rennes 1, 2009, Habilitation à Diriger des Recherches, Ph. D. Thesis.

[20] B. MAILHÉ. *Modèles et algorithmes pour la représentation parcimonieuse de signaux de grandes dimensions*, Université de Rennes I, december 2009, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

[21] S. Arberet, R. Gribonval, F. Bimbot. *A Robust Method to Count and Locate Audio Sources in a Multichannel Underdetermined Mixture*, in "IEEE Trans. on Signal Processing", vol. 58, 2010, 14 pages.

[22] R. Badeau, N. Bertin, E. Vincent. *On the stability of multiplicative update algorithms. Application to non-negative matrix factorization.*, in "IEEE Trans. on Neural Networks", 2010, To Appear.

[23] N. Bertin, R. Badeau, E. Vincent. *Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, To appear.

[24] M. E. Davies, R. Gribonval. *Restricted Isometry Constants where $\_ell^p$ sparse recovery can fail for $0 < p\_leq1$*, in "IEEE Trans. Inform. Theory", vol. 55, n^o 5, May 2009, p. 2203–2214 GB .

[25] N. Duong, E. Vincent, R. Gribonval. *Under-determined audio source separation with a new modeling of the convolutive mixing process*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, Submitted.

[26] V. Emiya, E. Vincent, N. Harlander, V. Hohmann. *Subjective assessment of audio source separation and objective measures using subband-based distortion extraction*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, Submitted DE .

[27] G. Gonon, F. Bimbot, R. Gribonval. *Probabilistic scoring using decision trees for fast and scalable speaker recognition*, in "Speech Communication", vol. 51, n^o 11, November 2009, p. 1065-1081.

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

[29] S. Huet, G. Gravier, P. Sébillot. *Morpho-syntactic post-processing with N-best lists for improved French automatic speech recognition*, in "Computer Speech and Language", vol. doi:10.1016/j.csl.2009.10.001, October 2009, 22 pages, doi:10.1016/j.csl.2009.10.001.

[30] M. Kowalski, E. Vincent, R. Gribonval. *Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, Submitted.

[31] R. Tavenard, L. Amsaleg, G. Gravier. *Model-based similarity estimation of multidimensional temporal sequences*, in "Annals of Telecommunications", vol. 64, n^o 5–6, 2009, p. 381-390.

[32] E. Vincent, N. Bertin, R. Badeau. *Adaptive harmonic spectral decomposition for multiple pitch estimation*, in "IEEE Trans. on Audio, Speech and Language Processing", 2010, To appear.

International Peer-Reviewed Conference/Proceedings

[33] N. Bertin, R. Badeau, E. Vincent. *Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription*, in "Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)", 2009.

[34] N. BERTIN, E. VINCENT, R. BADEAU. *Fast Bayesian constrained NMF for polyphonic pitch transcription*, in "Proc. 5th Music Information Retrieval Evaluation eXchange (MIREX)", International Society for Music Information Retrieval, 2009.

[35] 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 GB .

[36] 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 GB .

[37] N. DUONG, E. VINCENT, R. GRIBONVAL. *Spatial covariance models for under-determined reverberant audio source separation*, in "Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)", 2009.

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

[39] V. EMIYA, E. VINCENT, R. GRIBONVAL. *An investigation of discrete-state discriminant approaches to single-sensor source separation*, in "Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)", 2009.

[40] S. GALLIANO, G. GRAVIER, L. CHAUBARD. *The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts*, in "Conf. of the Intl. Speech Communication Association (Interspeech), Brighton, UK", September 2009, p. 2583–2586.

[41] C. GUINAUDEAU, G. GRAVIER, P. SÉBILLOT. *Can automatic speech transcripts be used for large scale TV stream description and structuring ?*, in "First International Workshop on Content-Based Audio/Video Analysis for Novel TV Services, San Diego, CA, USA", December 2009, In conjunction with the International IEEE Symposium on Multimedia.

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

[43] G. LECORVÉ, G. GRAVIER, P. SÉBILLOT. *Constraint selection for topic-based MDI adaptation of language models*, in "Proceedings of the International Conference on Speech and Language Technology (Interspeech'09), Brighton, UK", September 2009, p. 368–371.

[44] 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.

[45] B. MAILHÉ, M. LEMAY, R. GRIBONVAL, P. VANDERGHEYNST, J.-M. VESIN, F. BIMBOT. *Dictionary learning for the sparse modelling of atrial fibrillation in ECG signals*, in "Proc. IEEE ICASSP", April 2009.

[46] A. MUSCARIELLO, G. GRAVIER, F. BIMBOT. *Audio keyword extraction by unsupervised word discovery*, in "Conf. of the Intl. Speech Communication Association (Interspeech), Brighton, UK", September 2009, p. 2843–2846.

[47] A. MUSCARIELLO, G. GRAVIER, F. BIMBOT. *Variability tolerant motif discovery*, in "Intl. Multimedia Model Conference", B. HUET, E. AL. (editors), 2009.

[48] F. M. NAINI, R. GRIBONVAL, L. JACQUES, P. VANDERGHEYNST. *Compressive sampling of pulse trains: Spread the spectrum !*, in "Proc. ICASSP", 2009.

[49] A. NESBIT, E. VINCENT, M. PLUMBLEY. *Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation*, in "Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2009 UK .

[50] A. NESBIT, E. VINCENT, M. PLUMBLEY. *Extension of sparse, adaptive signal decompositions to semi-blind audio source separation*, in "Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)", 2009, p. 605-612 UK .

[51] M. PUIGT, E. VINCENT, Y. DEVILLE. *Validity of the independence assumption for the separation of instantaneous and convolutive mixtures of speech and music sources*, in "Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)", 2009, p. 613-620.

[52] J. L. ROUX, H. KAMEOKA, E. VINCENT, N. ONO, K. KASHINO, S. SAGAYAMA. *Complex NMF under spectrogram consistency constraints*, in "Proc. of the Acoustical Society of Japan (ASJ) Autumn Meeting", 2009 JP .

[53] J. L. ROUX, H. KAMEOKA, E. VINCENT, N. ONO, K. KASHINO, S. SAGAYAMA. *Complex NMF with spectrogram consistency constraints*, in "Proc. 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2010, Submitted JP .

[54] R. SCHOLZ, E. VINCENT, F. BIMBOT. *Robust modeling of musical chord sequences using probabilistic N-grams*, in "Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2009.

[55] W. X. TENG, G. GRAVIER, F. BIMBOT, F. SOUFFLET. *Speaker Adaptation by Variable Reference Model Subspace and Application to large Vocabulary Speech Recognition*, in "IEEE Intl. Conf. on Acoustics, Speech and Signal Processing", April 2009, p. 4381–4384.

[56] E. VINCENT, S. ARAKI, P. BOFILL. *The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation*, in "Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)", 2009, p. 734-741 JP ES .

[57] E. VINCENT, S. ARBERET, R. GRIBONVAL. *Underdetermined instantaneous audio source separation via local Gaussian modeling*, in "Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)", 2009, p. 775-782.

[58] E. VINCENT. *Audio source separation using hierarchical phase-invariant models*, in "Proc. 2009 ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)", 2009.

National Peer-Reviewed Conference/Proceedings

[59] S. BAGHDADI, G. GRAVIER, C.-H. DEMARTY, P. GROS. *Apprentissage de structure dans les réseaux bayésiens pour la détection d'événements vidéo*, in "Traitement et Analyse de l'Information : Méthodes et Applications", 2009.

[60] V. EMIYA, E. VINCENT, R. GRIBONVAL. *Estimateurs oracles pour la séparation de sources monocapteur par approches spectrales à états discrets*, in "Proc. 22e colloque GRETSI sur le Traitement du Signal et des Images", 2009.

[61] 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.

Scientific Books (or Scientific Book chapters)

[62] F. BIMBOT. *9*, in "Automatic Speaker Recognition", Iste / John Wiley, 2009, p. 321–353.

[63] A. NESBIT, M. JAFARI, E. VINCENT, M. PLUMBLEY. *Audio source*, in "Audio source separation using sparse representations", IGI Global, 2010, Accepted subject to minor revisions UK .

[64] E. VINCENT, Y. DEVILLE. *Audio applications*, in "Handbook of Blind Source Separation, Independent Component Analysis and Applications", Academic Press, 2009.

[65] E. VINCENT, M. JAFARI, S. ABDALLAH, M. PLUMBLEY, M. DAVIES. *unknown*, in "Probabilistic modeling paradigms for audio source separation", IGI Global, 2010, Accepted subject to minor revisions UK .

Research Reports

[66] R. GRIBONVAL, K. SCHNASS. *Dictionary Identification - Sparse Matrix-Factorisation via _ell_1-Minimisation*, n^o 0904.4774, arXiv, April 2009, Technical reportCH.

Other Publications

[67] S. FOUCART, R. GRIBONVAL. *Real vs. Complex Null Space Properties for Sparse Vector Recovery*, oct 2009, submitted to Comptes Rendus de l'Académie des Sciences.

References in notes

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

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

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

[71] 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.

[72] C. HERLEY. *ARGOS: Automatically Extracting repeating objects from multimedia streams*, in "IEEE Trans. on Multimedia", vol. 8, n^{o} 1, February 2006, p. 115–129.

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

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

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

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

[77] 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.