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 COG

Activity Report

2008

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]
Guillaume Gravier [CR1 CNRS]
Rémi Gribonval [CR1 INRIA, HdR]
Emmanuel Vincent [CR2 INRIA]

Technical Staff

Pierre Cauchy [Contractual Development Engineer - Started November 2008]
Benjamin Roy [Contractual Development Engineer - Until August 2008]

PhD Student

Simon Arberet [CNRS & Region Grant, 3rd year]
Boris Mailhé [ENS Cachan (Bruz), 3rd year]
Armando Muscariello [Regional Grant, 1st year]
Prasad Sudhakaramurthy [INRIA Cordis Grant, 1st year]
Quang Khanh Ngoc Duong [INRIA Cordi Grant - Started October 2008]
Wen Xuan Teng [Telisma Funding, 4th year]
Klara Trakas [Orange FTR&D CIFRE Funding - Terminated September 2008]

Post-Doctoral Fellow

Valentin Emiya [INRIA Grant]

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 multi-media sector, the musical and audio-visual 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 research projects European projects and 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 participated to The Star Challenge, an international competition on information retrieval tasks in multimedia contents, as part of an international joint venture with the National Institute of Informatics (NII, Tokyo) and the TEXMEX project-team. Our joint team passed the qualifying round and took part of the grand final held in Singapore, where it ranked second out of the five finalist competitors.

TEXMEX and METISS were involved in audio retrieval tasks, such as the "Query by IPA" one which consists in retrieving shots containing a given string of phonemes in a multilingual collection of video shots.

3. Scientific Foundations

3.1. Introduction

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

Probabilistic approaches offer a general theoretical framework [60] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [56], 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 [62] have emerged as offering a powerful framework for the modeling of musical signals (for instance, [57], [65]).

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 [61]. 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 [55]. 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

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

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 [62].

3.3. Sparse representations

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

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.

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

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

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

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

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

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

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

3.3.2. Sparsity criteria

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

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

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

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

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

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

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

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

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

3.3.3. Decomposition algorithms

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

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

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

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

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

3.3.4. Dictionary construction

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

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

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

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

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

3.3.5. Compressive sensing

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

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

4. Application Domains

4.1. Introduction

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

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

The main applicative fields targeted by these tasks are :

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

4.2. Spoken content processing

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

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 multi-modal indexing purposes

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

Keywords: *audio descriptors*, *audio detection*, *audio segmentation*, *audio stream*, *audio tracking*, *audiovisual descriptors*, *audiovisual integration*, *information fusion*, *multimedia indexing*, *multimodality*.

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

Keywords: *audio object*, *multi-level representations*, *music description*, *music language modeling*.

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

Keywords: *compressive sensing*, *multi-channel audio*, *source characterization*, *source localization*, *source separation*.

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 [11].

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

Keywords: *analysis*, *audio*, *audio indexing*, *audio stream*, *detection*, *processing*, *segmentation*, *signal*, *speaker verification*, *speech*, *tracking*.

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

Keywords: *HMM*, *Viterbi*, *beam-search*, *broadcast news indexing*, *speech modeling*, *speech recognition*.

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 [58]. 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.

This year has been dedicated to a major refactoring of the ASR part ot the system. On the one hand, the architecture of the IRENE transcription system has been redesigned to be part of the multimedia indexing platform jointly developped by the TEXMEX, METISS and VISTA project-teams, thus enabling the large scale transcription of TV programs.

On the other hand, a new version of the IRENE transcription system was developped in collaboration with the TEXMEX project-team, using a larger amount of training data. In particular, we have upgraded the language model (LM) of the ASR system in order to enable vocabulary adaptation on top of LM adaptation. Significant improvements were benchmarked in the framework of the ESTER 2 evaluation campaign on the transcription radio shows in the French language.

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

5.3. MPTK: the Matching Pursuit Toolkit

Participants: Rémi Gribonval, Benjamin Roy.

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. Speaker modeling and characterisation

Keywords: *Gaussian Mixture Models (GMM)*, *affective computing*, *cognitive state*, *emotion*, *model interpolation*, *psychoacoustic*, *speaker adaptation*, *speaker characterisation*, *speaker selection*, *voice interaction*.

6.1.1. Rapid Speaker Adaptation by variable Reference Model Interpolation

Participants: Wen Xuan Teng, Guillaume Gravier, Frédéric Bimbot.

This work has taken place in the context of an industrial PhD with TELISMA.

Rapid adaptation of acoustic models for automatic speech recognition requires some form of a priori knowledge to guide the estimation of a new speaker model. Most techniques are based on linear combinations in a speaker subspace derived from fixed, a priori, speaker models (cf. the eigenvoice approach). In this context, the adaptation process may not provide robust solutions for a particular adaptation target, expecially when the number of reference models is small.

The approach investigated in [12] involves using variable subspaces at runtime for different adaptation targets. This yields a novel approach called variable RMI (Reference Model Interpolation) based on an a posteriori selection of reference models, with various possible selection criteria.

The proposed tehcnique has been applied and tested on phoneme decoding and LVCSR (Large Vocabulary Continuous Speech Recognition) tasks, and evaluated both in supervised and unsupervised adaptation modes. Experiments on three distinc databases (IDIOLOGOS, PAIDIALOGOS and ESTER) have shown the effectiveness of the variable RMI approach with utterance bu utterance on-line adaptation.

6.1.2. Voice modelling for emotion and cognitive state classification

Participants: Klara Trakas, Frédéric Bimbot.

This work has taken place in the context of an industrial PhD with Orange FTR&D Labs.

A growing interest has emerged in the field of speaker characterisation for approaches able to describe and classify voice expressions such as emotion, cognitive state and, more generally, any type of information conveyed by the voice of a speaker voice and indicative of his/her state of mind.

The first year of PhD of Klara Trakas has been focused on the analysis of a speech corpus composed of client's calls expressing their opinion on a hotline service. Human auditors were asked to give their perception of the emotional state of the speaker together with other impressions not related to emotions, so as to examine correlations between different classes of voice and speaker features.

This preliminary work was intended to lead to a robust system for emotion detection, investigating descriptors and models for representing speaker's characteristics at several linguistic and para-linguistic levels, together with training algorithms and decision strategies which enable the fusion multiple sources of information. However, the PhD was interrupted after one year (September 2008).

6.2. Speech recognition for multimedia structuring and indexing

Keywords: *multimedia indexing*, *natural language processing*, *semantic verification*, *speech recognition*, *topic detection*, *video structuring*.

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

Participants: Pierre Cauchy, Guillaume Gravier.

Work done in close collaboration with the TEXMEX *project-team of IRISA, in particular with Pascale Sébillot and Fabienne Moreau.*

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 [44].

6.2.1.1. Topic segmentation and adaptation

We improved our former extension of the text-based topic segmentation method of Utiyama and Isahara [64] to take into account additional knowledge such as semantic relations between words, discourse markers (like "`and now, thank you`"), and acoustic cues [38]. Results obtained on radio broadcast news make it possible to apply the method to large scale TV streams, eventually in conjunction with image-based features, as considered in C. Guinaudeau's Ph. D. thesis.

We also investigated efficient methods for the extraction of keywords to characterize thematic segments, in order to improve the language model of the ASR system using related texts retrieved on the Internet. Experiments reported in [27] have shown that topic adaptation is more effective when included in the early recognition stages. Thus, we focused on keyword extraction at the very beginning of the transcription using confusion networks rather than a single sentence.

6.2.1.2. Semantic verification of TV programmes

We investigated the use of automatic transcriptions of TV programs to validate labels automatically obtained from an electronic program guide (EPG). Given an online TV program guide, we can associate the phonetic or textual transcription of the soundtrack with descriptions extracted from the TV guide, using techniques inspired from the information retrieval field. Names obtained from the TV guide are then compared with the respective labels obtained from the EPG alignment. The phonetic and textual methods implemented allow to make a decision for 40 % of the segments and to decrease the labeling error rate by 3.5 %.

6.2.2. Audio information retrieval in multilingual audiovisual contents

Participant: Guillaume Gravier.

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

In the framework of our participation to The Star Challenge, we developed a system for phonetic-based information retrieval in multilingual collections of videos. Phonetic recognition is performed with French phoneme models and a classical boolean information retrieval model is used to index sequences of respectively 2, 3 and 4 phonemes [63]. Finally, the resulting rankings are merged using rank aggregation methods. Experiments with a database containing 4 languages demonstrated the effectiveness of the method accross languages using relatively simple models (context-independent 3 state HMM with 32 Gaussians / state). Complex models were inefficient unless from the target language. Finally, promising results were obtained using query expansion based on phonetic confusions.

6.3. Audio motif and structure discovery

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

6.3.1. Motif discovery and sequence description

Participants: Frédéric Bimbot, Guillaume Gravier, Armando Muscariello.

The work on sequence description corresponds to the Ph. D. Thesis of Romain Tavenard, in collaboration with Laurent Amsaleg from the TEXMEX *project-team.*

Audio motif discovery aims at finding repeating patterns from large audio streams in an unsupervised manner. Using the segmentation framework defined in [59], we proposed a motif discovery method tolerant to variations in both the spectral and temporal domains. Our algorithm relies on several extensions of the well-known dynamic time warping algorithm to a word-spotting framework where motif boundaries are unknown. Results on a word discovery task, to appear in [33], demonstrate the effectiveness of the method to retrieve repeating words or locutions in radio broadcast news data.

Audio discovery requires the fast comparison of sequences of audio descriptors in order to efficiently search for candidate motifs in a data stream. In the Ph. D. work of Romain Tavenard, we explore models of sequences of audio descriptors and the associated distances between models as an alternate approach to the traditional DTW-based solutions. In 2008, we dedicated ourselves to large scale experiments of the support vector regression (SVR) approach developped in 2007. Experimental results demonstrated the effectiveness of SVR to retrieve ads and jingles in a database containing 100 hours of radio broadcast news data. Moreover, SVR was found to be less sensitive to segmentation issues than DTW.

6.3.2. Structure learning in Bayesian networks

Participant: Guillaume Gravier.

Work carried out in the framework Siwar Baghdadi's Ph. D. Thesis with Thomson Multimedia and the TEXMEX *project-team.*

A key issue in statistical modeling is the design of that of model selection in order to select the best model according to the problem to solve. In the Ph. D. work of Siwar Baghdadi, we investigated several approaches to automate the model selection process in the framework of Bayesian networks applied to multimodal modeling of sport broadcasts. We emphasized the limits of the K2 structure learning algorithm (based on the Bayesian information criterion) for classification problems and investigated alternate algorithms based on discriminative criteria. Discriminative model selection algorithms were found more adequate for classification problems and eliminates the need for feature selection. Extensions of this work to structure learning in dynamic Bayesian networks were also investigated. Results on an action detection task in soccer videos demonstrate that the modeling step can be fully automated.

6.4. Recent results on sparse representations

Keywords: *dictionary design*, *high dimension*, *scalable algorithms*, *sparse approximation*.

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. This year, 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 (submitted to the conference ICASSP 2009 [31]) 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. A journal paper is in preparation.

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.

This year, 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 [24] 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 [23], and a journal paper is in preparation.

6.4.3. Theoretical results on identification of sparse representations

Participant: Rémi Gribonval.

Main collaboration: Mike Davies (Univ. Edinburgh)

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, published this year [16], [13], [17] concentrated on positive results for greedy algorithms and convex optimization (ℓ^1-minimization). In contrast this year, in collaboration with Pr Michael Davies, we concentrated on ℓ^p-minimization, $0 < p \leq 1$, and our results highlight the pessimistic nature of sparse recovery analysis when recovery is predicted based on the restricted isometry constants (RIC) of the associated matrix (submitted for publication [47]).

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 [32]. 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 (submitted to the conference ICASSP09 [30]).

6.5. Emerging activities on compressive sensing

Keywords: *acoustic wavefields*, *compressive sensing*, *graph wavelets*, *wideband signals*.

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

Participants: Rémi Gribonval, Prasad Sudhakar, Emmanuel Vincent.

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 submitted at the conference ICASSP 2009 [34].

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. The transform is based on spectral graph theory and allows the construction of families of multi-scale atoms which are well adapted to the specific connectivity of a graph.

We proved that it is possible to build a wavelet family which constitutes a frame, which is an important property to represent functions (i.e. signals). We also studied certain families of wavelets on graphs which generalize to the multiscale setting the notion of arbitrary powers of the Laplacian on the graph.

These wavelets could turn out to be very useful in application scenarios where signals are sampled non-uniformly (such as in sensor networks) or when graphs are inherently present in the model (e.g. in social networks). These results will be presented at the workshop "sparsity and large inverse problems" to be held in Cambridge in December 2008.

6.6. Content description of music signals

Keywords: *language model*, *music*, *n-gram*, *pitch transcription*.

6.6.1. Multi-pitch signal modeling

Participant: Emmanuel Vincent.

Main collaboration: P. Leveau, N. Bertin (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. Given some limitations of existing acoustic models, including our previous time-domain models [19], [20], we proposed a frequency-domain acoustic model that exploits the timbre of each instrument to increase the accuracy of the inferred musical score without relying on separate training data. This model represents 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 constraints [37].

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 [52].

6.7. Source separation

Keywords: *adaptive basis*, *probabilistic source model*, *source localization*, *source separation*, *sparse representation*.

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)

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 consists of representing the signals in a given time-frequency basis and estimating the source coefficients by sparse decomposition in that basis, under an exact mixture reconstruction constraint relying on a frequency-wise 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. Our previous work on adaptive stereo bases [18] showed promising results but suggested that the modeling of the mixing process and the choice of an adapted basis should be separately addressed so as to avoid overfitting issues. We investigated the replacement of the mixture reconstruction constraint by a quadratic penalty term computed from the true mixing process, resulting in improved separation performance in reverberant environments with large microphone spacing [26]. We also 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 [35], [50], [49]. Finally, we provided an experimental validation of the implicit source independence assumption underlying the above approaches [51].

6.7.2. A new probabilistic framework for source separation

Participants: Simon Arberet, Remi Gribonval, Emmanuel Vincent, Frédéric Bimbot.

Main collaboration: Alexey Ozerov (Telecom ParisTech)

In parallel with our work on sparse representations, we proposed a new framework for audio source separation where each source is modeled as a zero-mean Gaussian variable in the neighborhood of each time-frequency bin. This framework was first applied to the problem of source counting and localization and resulted in increased robustness by selection of the time-frequency bins with a single active source [46]. We subsequently investigated its use for the problem of source separation by defining two distinct models for the source variances: either a mild sparsity prior in each time-frequency bin [54] or a GMM prior introducing some dependencies between the variances in different frequency bins [48]. Both approaches were tested over instantaneous mixtures and provided respectively a significant improvement of the separation performance over all mixtures and an even larger improvement over music mixtures.

7. Contracts and Grants with Industry

7.1. National projects

7.1.1. ARC INRIA RAPSODIS

Participant: Guillaume Gravier.

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

The Concerted Research Action RAPSODIS (Syntactic and Semantic Information-Based Automated Speech Recognition) 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 2008, work has been mostly dedicated to the study of possible integration modes—reordering of n-best hypothesis lists is currently privileged—, to investigations about the impact of transcription errors on syntactic parsing—a correlation between the length of chunks in hypothesis and errors has been established—, to the acquisition of semantic relations from the Web, and to a major refactoring of our ASR system, using a larger amount of training data, in order to enable effective integration of linguistic information.

7.1.2. QUAERO CTC and Corpus Projects

Participants: Frédéric Bimbot, Guillaume Gravier, Emmanuel Vincent.

Main academic partners : IRCAM, IRIT, LIMSI, TelecomParisTech, Univ. Karlsruhe, CLIPS/Imag.

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.

7.2. 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, Switzerland) 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, Benjamin Roy, 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).

8.2. Visites, et invitations de chercheurs

8.2.1. Visit to the University of Nagoya and NII Tokyo

Participant: Guillaume Gravier.

Partners: Ichiro Ide, Associate professor at the University of Nagoya, and visiting associate professor at NII (National institute of informatics) Tokyo, Japan, and Shin'ichi Satoh, Professor at NII Tokyo, Japan.

In the context of our collaboration with the TEXMEX project-team and of the INRIA Associate team between TEXMEX, the University of Nagoya and NII Tokyo, Guillaume Gravier spent 2 weeks in Japan (both at Nagoya and Tokyo) in May 2008. During this period, he worked on various aspects of automatic structuring of TV streams: creation of a temporally-similar broadcast news corpus in Japanese and French to carry on similar investigations, relevance of automatic speech recognition in multimedia information retrieval, etc.

9. Dissemination

9.1. Conference and workshop committees, invited conference

Frédéric Bimbot was a member of the Programme Committee for the Odyssey 2008 Workshop on Speaker Recognition, in Stellenbosch, South Africa, January 21-25, 2008.

Frédéric Bimbot was a member of the Programme Committee (responsible for tutorials) for the Eusipco 2008 Conference, in Lausanne, Switzerland, August 25-29, 2008.

Frédéric Bimbot set up a proposal to ISCA for organizing the Interspeech 2011 Conference in Lyon.

Guillaume Gravier is part of the NOE MUSCLE.

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 is the general chair of the workshop SPARS'09 on Signal Processing with Adaptive Sparse/Structure Representations, to be held in Saint-Malo, April 6-9 2009. This is the second edition of the workshop. The first edition was organized in 2005 in Rennes and gathered 65 international participants. This year we received more than 60 contributions and are expecting about 100 participants.

9.2. Leadership within scientific community

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

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

Rémi Gribonval is a member of the ICA Steering Committee.

Emmanuel Vincent was the chair of the first community-based Signal Separation Evaluation Campaign (SiSEC 2008), co-organized with Shoko Araki (NTT, Japan) and Pau Bofill (University of Catalonia, Spain). The results of the campaign will be published in [53] 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/.

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.

Frédéric Bimbot visited three secondary schools in Britanny and gave presentations on speaker recognition to several classes, in the context of “A la découverte de la Recherche”.

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 was invited to give a plenary lecture at the international workshop "Workshop on sparsity and large inverse problems" at Robinson College, University of Cambridge, December 14-15, 2008.

Rémi Gribonval gave a tutorial lecture on sparse decompositions, compressed sensing and source separation at the one day meeting on sparsity of the french GDR ISIS at Telecom Paristech, Paris on April 17, 2008.

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

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

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

10. Bibliography

Major publications by the team in recent years

[1] 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.

[2] 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.

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

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

[5] R. GRIBONVAL, P. VANDERGHEYNST. *On the exponential convergence of Matching Pursuits in quasi-incoherent dictionaries*, in "IEEE Trans. Information Theory", vol. 52, n^{o} 1, January 2006, p. 255–261.

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

[7] 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^{o} 5, juillet 2007, p. 1564–1578.

[8] A. ROSENBERG, F. BIMBOT, S. PARTHASARATHY. *36*, in "Overview of Speaker Recognition", Y. H. J. BENESTY (editor), Springer, 2008, p. 725–741.

[9] E. VINCENT, R. GRIBONVAL, C. FÉVOTTE. *Performance measurement in Blind Audio Source Separation*, in "IEEE Trans. Speech, Audio and Language Processing", vol. 14, n^{o} 4, 2006, p. 1462–1469.

[10] 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^{o} 4, 2007, p. 1273–1282.

Year Publications

Doctoral Dissertations and Habilitation Theses

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

[12] W. TENG. *Adaptation rapide au locuteur par sous-espace variable de modèles de référence*, Ph. D. Thesis, Université de Rennes I, december 2008.

Articles in International Peer-Reviewed Journal

[13] 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.

[14] M. DELAKIS, G. GRAVIER, P. GROS. *Audiovisual Integration with Segment Models for Tennis Video Parsing*, in "Computer Vision and Image Understanding", vol. 111, n^{o} 2, August 2008, p. 142–154.

[15] G. GONON, F. BIMBOT, R. GRIBONVAL. *Probabilistic scoring using decision trees for fast and scalable speaker recognition*, in "Speech Communication", to appear, 2009.

[16] 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.

[17] R. GRIBONVAL, H. RAUHUT, K. SCHNASS, P. VANDERGHEYNST. *Atoms of all channels, unite ! average case analysis of multi-channel sparse recovery using greedy algorithms*, 2008.

[18] M. JAFARI, E. VINCENT, S. ABDALLAH, M. PLUMBLEY, M. DAVIES. *An adaptive stereo basis method for convolutive blind audio source separation*, in "Neurocomputing", vol. 71, n^{o} 10–12, 2008, p. 2087–2097.

[19] P. LEVEAU, E. VINCENT, G. RICHARD, L. DAUDET. *Instrument-specific harmonic atoms for mid-level music representation*, in "IEEE Transactions on Audio, Speech and Language Processing", vol. 16, n^{o} 1, 2008, p. 116–128.

[20] E. VINCENT, M. PLUMBLEY. *Efficient Bayesian inference for harmonic models via adaptive posterior factorization*, in "Neurocomputing", vol. 72, 2008, p. 79–87.

Articles in National Peer-Reviewed Journal

[21] A. BÜRKI, C. GENDROT, G. GRAVIER, G. LINARÈS, C. FOUGERON. *Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa*, in "Traitement Automatique des Langues", vol. 49, n^{o} 3, 2008.

International Peer-Reviewed Conference/Proceedings

[22] A. L. CASANOVAS, G. MONACI, P. VANDERGHEYNST, R. GRIBONVAL. *Blind Audiovisual Separation based on Redundant Representations*, in "Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP", 2008.

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

[24] 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.

[25] S. HUET, G. GRAVIER, P. SÉBILLOT. *Morphosyntactic Resources for Automatic Speech Recognition*, in "Intl. Conf. on Language, Resources and Evaluation", 2008.

[26] M. KOWALSKI, E. VINCENT, R. GRIBONVAL. *Under-determined source separation via mixed-norm regularized minimization*, in "Proc. European Signal Processing Conf. - EUSIPCO", 2008.

[27] G. LECORVÉ, G. GRAVIER, P. SÉBILLOT. *An unsupervied Web-based topic language model adaptation method*, in "IEEE Intl. Conf. on Acoustics, Speech and Signal Processing", 2008.

[28] G. LECORVÉ, G. GRAVIER, P. SÉBILLOT. *Using Internet as a corpus ...*, in "Intl. Conf. on Language, Resources and Evaluation", 2008.

[29] B. LECOUTEUX, G. LINARÈS, Y. ESTÈVE, G. GRAVIER. *Generalized driven decoding for speech recognition system combination*, in "IEEE Intl. Conf. on Acoustics, Speech and Signal Processing", 2008.

[30] B. MAILHÉ, R. GRIBONVAL, F. BIMBOT, M. LEMAY, P. VANDERGHEYNST, J.-M. VESIN. *Dictionary learning for the sparse modelling of atrial fibrillation in ECG signals*, in "ICASSP'09", 2009.

[31] B. MAILHÉ, R. GRIBONVAL, F. BIMBOT, P. VANDERGHEYNST. *A low complexity orthogonal matching pursuit for sparse signal approximation with shift-invariant dictionaries*, in "ICASSP'09", 2009.

[32] 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", 2008.

[33] A. MUSCARIELLO, G. GRAVIER, F. BIMBOT. *Variability tolerant motif discovery*, in "Intl. Multimedia Model Conference", 2009.

[34] F. NAINI, R. GRIBONVAL, L. JACQUES, P. VANDERGHEYNST. *Compressive sampling of pulse trains : Spread the spectrum !*, in "ICASSP'09", 2009.

[35] A. NESBIT, M. PLUMBLEY, E. VINCENT. *Oracle evaluation of flexible adaptive transforms for underdetermined audio source separation*, in "Proc. UK ICA Research Network International Workshop, University of Liverpool", 2008.

[36] P. SUDHAKAR, R. GRIBONVAL. *A sparsity-based method to solve the permutation indeterminacy in frequency domain convolutive blind source separation*, in "ICA'09", submitted, 2009.

[37] E. VINCENT, N. BERTIN, R. BADEAU. *Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription*, in "Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", 2008, p. 109–112.

National Peer-Reviewed Conference/Proceedings

[38] 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.

[39] G. LECORVÉ, G. GRAVIER, P. SÉBILLOT. *Vers une adaptation thématique non supervisée de modèles de langage : utilisation d'Internet comme un corpus ouvert*, in "Journées d'Études sur la Parole", 2008.

[40] B. LECOUTEUX, G. LINARÈS, Y. ESTÈVE, G. GRAVIER. *Combinaison de systèmes par décodage guidé*, in "Journées d'Études sur la Parole", 2008.

Scientific Books (or Scientific Book chapters)

[41] F. BIMBOT. *9*, in "Automatic Speaker Recognition", J. MARIANI (editor), 35 pages – to appear, Hermès, 2009.

[42] M. DELAKIS, G. GRAVIER, P. GROS. *Stochastic models for multimodal video analysis*, in "Multimodal Processing and Interaction: Audio, Video, Text", P. MARAGOS, A. POTAMIANOS, P. GROS (editors), Springer Verlag, 2008.

[43] G. GRAVIER, J.-F. BONASTRE, S. GALLIANO, E. GEOFFROIS, D. MOSTEFA, K. CHOUKRI. *Évaluation des systèmes de transcription enrichie d'émissions radiophoniques*, in "L'évaluation des technologies de traitement de la langue", S. CHAUDIRON, K. CHOUKRI (editors), Cognition et traitement de l'information, chap. 7, Hermès Science, 2008, p. 165–182.

[44] S. HUET, G. GRAVIER, P. SÉBILLOT. *Toward the integration of NLP and ASR techniques: POS tagging and transcription*, in "Multimodal Processing and Interaction: Audio, Video, Text", P. MARAGOS, A. POTAMIANOS, P. GROS (editors), Springer Verlag, 2008.

[45] A. ROSENBERG, F. BIMBOT, S. PARTHASARATHY. *36*, in "Overview of Speaker Recognition", Y. H. J. BENESTY (editor), Springer, 2008, p. 725–741.

Research Reports

[46] S. ARBERET, R. GRIBONVAL, F. BIMBOT. *A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture*, Technical report, n^{o} 6593, INRIA Research Report, 2008.

[47] M. DAVIES, R. GRIBONVAL. *Restricted isometry constants where ℓ^p sparse recovery can fail for* $0 < p \leq 1$, Technical report, n^{o} 1899, IRISA-INRIA Technical Report, July 2008.

Other Publications

[48] S. ARBERET, A. OZEROV, R. GRIBONVAL, F. BIMBOT. *Blind spectral-GMM estimation for underdetermined instantaneous audio source separation*, submitted, 2009.

[49] A. NESBIT, E. VINCENT, M. PLUMBLEY. *Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation*, submitted, 2009.

[50] A. NESBIT, E. VINCENT, M. PLUMBLEY. *Extension of sparse, adaptive signal decompositions to semi-blind audio source separation*, submitted, 2009.

[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*, submitted, 2009.

[52] R. SCHOLZ, E. VINCENT, F. BIMBOT. *Robust modeling of musical chord sequences using probabilistic N-grams*, in "ICASSP'09", 2009.

[53] E. VINCENT, S. ARAKI, P. BOFILL. *The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation*, submitted, 2009.

[54] E. VINCENT, S. ARBERET, R. GRIBONVAL. *Underdetermined instantaneous audio source separation via local Gaussian modeling*, submitted, 2009.

References in notes

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

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

[57] 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.

[58] 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.

[59] 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.

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

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

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

[63] C. NG, R. WILKINSON, J. ZOBEL. *Experiments in spoken document retrieval using phoneme n-grams*, in "Speech Communication, Vol", vol. 32, 2000.

[64] 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.

[65] 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.