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

Table of contents

1. Team

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

Head of project-team
Frédéric Bimbot [CR1 CNRS, HdR]

Administrative assistant
Stéphanie Lemaile

Research scientist (CNRS)
Guillaume Gravier [CR1 CNRS]

Research scientist (INRIA)
Rémi Gribonval [CR1 INRIA, HdR]
Emmanuel Vincent [CR2 INRIA]

Project Technical Staff
Mathieu Ben [Contractual Research Engineer - Until August 2007]
Sylvain Busson [Contractual Research Engineer - Until June 2007]
Pierre Cauchy [Contractual Development Engineer - Since October 2007]
Gilles Gonon [Contractual Research Engineer - Until June 2007]
Benjamin Roy [Contractual Development Engineer]

Ph.D. students
Simon Arberet [CNRS & Region Grant, 2nd year]
Stéphane Huet [MENRT Grant, 3rd year - also with TEXMEX]
Boris Mailhé [ENS Cachan (Bruz), 2nd year]
Armando Muscariello [Regional Grant, 1st year - Started November 2007]
Sylvain Lesage [MENRT Grant, 3rd year - Terminated June 2007]
Amadou Sall [INRIA Grant, 4th year - Terminated March 2007]
Prasad Sudhakaramurthy [Cordis Grant, 1st year - Started December 2007]
Wen Xuan Teng [Telisma Funding, 3rd year]
Klara Trakas [Orange FTR&D Funding, 1st year - Started December 2007]

2. Overall Objectives

2.1. Overall Objectives

The research objectives of the METISS research group are dedicated to audio signal and speech processing and are organised along three main axes: speaker characterization, information detection and tracking in audio streams and "advanced" processing of audio signals (in particular, source separation). Some aspects of speech recognition (modeling and decoding) are also addressed so as to reinforce these three principal topics. 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 (with voice authentification), the Internet and multi-media sector (with audio indexing), the musical and audio-visual production sector (with audio signal processing), and, marginally, the sector of educational softwares, games and toys.

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. We also widely disseminate software resources corresponding to our latest developments.

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.

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 [66] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [62], 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 have emerged as offering a powerful framework for the modeling of musical signals.

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 [70]. 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.

This topic 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... 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 and 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).

In the rest of this text, we will denote as Λ the set of parameters which define the model under consideration. Λ_X will denote the vector of parameters for class X, and in this case, the following notation will be used :

$$\widehat{P}(Y|X) = P(Y|\Lambda_X)$$

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 X is generally based on a step of statistical estimation consisting in determining the optimal value for the vector of parameters Λ, i.e. the parameters that maximize a modeling criterion on a training set $\{Y\}_{tr}$ comprising observations corresponding to class X.

In some cases, the Maximum Likelihood (ML) criterion can be used :

$$\Lambda^*_{ML} = \arg\max_{\Lambda} P(\{Y\}_{tr}|\Lambda)$$

This approach 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 :

$$\Lambda^*_{MAP} = \arg\max_{\Lambda} P(\{Y\}_{tr}|\Lambda) \; . \; p(\Lambda)$$

which relies on a prior probability $p(\Lambda)$ of vector Λ, 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 (under the assumption of uniform prior probability for Λ), 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. In this case, the value of $p(\Lambda)$ is given by the model before adaptation and the MAP estimate uses the new data to update the model parameters.

Whatever criterion is considered (ML or MAP), the estimate of the parameters Λ is obtained with the EM algorithm (Expectation-Maximization), which provides a solution corresponding to a local maximum of the training criterion.

3.2.3. Likelihood computation and state sequence decoding

During the recognition phase, it is necessary to evaluate the likelihood function for the various class hypotheses X_k. When the complexity of the model is high - i.e when the number of classes is large and the observations to be recognized are multidimensional - it is generally necessary to implement fast calculation algorithms to approximate the likelihood function.

In addition, when the class model are HMMs, 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.

If, moreover, the observations consist of segments belonging to different classes, chained by probabilities of transition between successive classes and without a priori knowledge of the borders between segments (which is for instance the case in a continuous speech utterance), it is necessary to call for beam-search techniques to decode a (quasi-)optimal sequence of states at the level of the whole utterance.

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 :

$$\widehat{X}_k = \arg\max_{X_k} p(X_k) \; . \; \widehat{P}(Y|X_k)$$

where $\{X_k\}_{1 \leq k \leq K}$ denotes the set of possible classes.

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 X (denoted as hypothesis X) or not pertaining to it (i.e. pertaining to the "non-class", denoted as hypothesis $\overline{X}$). In this case, the decision consists in acceptance or rejection, respectively denoted $\widehat{X}$ and $\widehat{\overline{X}}$ in the rest of this document.

This latter problem can be theoretically solved within the framework of Bayesian decision by calculating the ratio S_X of the PDFs for the class and the non-class distributions, and comparing this ratio to a decision threshold :

$$S_X(Y) = \frac{P(Y|X)}{P(Y|\overline{X})} \begin{cases} \geq R & \text{hypothesis} \quad \widehat{X} \\ < R & \text{hypothesis} \quad \widehat{\overline{X}} \end{cases}$$

where the optimal threshold R does not depend on the distribution of class X, but only of the operating conditions of the system via the ratio of the prior probabilities of the two hypotheses and the ratio of the costs of false acceptance and false rejection.

In practice, however, the Bayesian theory cannot be applied straightforwardly, because 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 rule of optimal decision must then be rewritten :

$$\widehat{S}_X(Y) = \frac{\widehat{P}(Y|X)}{\widehat{P}(Y|\overline{X})} \begin{cases} \geq \Theta_X(R) & \text{hypothesis} \quad \widehat{X} \\ < \Theta_X(R) & \text{hypothesis} \quad \widehat{\overline{X}} \end{cases}$$

and the optimal threshold $\Theta_X(R)$ must be adjusted for class X, by modeling the behaviour of the ratio $\widehat{S}_X$ on external (development) data.

The issue of how to estimate the optimal threshold $\Theta_X(R)$ in the case of the likelihoo ratio test, can be formulated in an equivalent way as finding a normalisation of the likelihood ratio which brings back the optimal decision threshold to its theoretical value. Several transformations are now well known within the framework of speaker verification, in particular the Z-norm and the T-norm methods.

3.2.5. Bayesian networks

In the past years, increasing interest has focused on Bayesian models for multi-source signals, such as polyphonic music signals. These models are particularly interesting, since they enable a formulation of music information retrieval in a probabilistic modelling framework, together with the exploitation of various priors on the model parameters.

A first issue 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, whose challenge is to achieve a good compromise between tractability and accuracy.

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*.

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.

To account for these factors of diversity, our approach is to focus on techniques for decomposing signals on redundant systems (or dictionaries). The elementary atoms in the dictionary correspond to the various structures that are expected to be met in the signal.

3.3.1. Redundant systems and adaptive representations

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

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

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

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

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

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

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

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

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

3.3.2. Sparsity criteria

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

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

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

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

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

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

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

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

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

3.3.3. Decomposition algorithms

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

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

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

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

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

3.3.4. Dictionary construction

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

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

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

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

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

3.3.5. Signal separation

METISS is especially interested in source and signal separation in the underdetermined case, i.e. in the presence of a number of sources strictly higher than the number of sensors.

In the particular case of two sources and one sensor, the mixed (monodimensional) signal writes :

$$y = s_1 + s_2 + \epsilon$$

where s_1 and s_2 denote the sources and ϵ an additive noise.

Under a probabilistic framework, we can denote by θ_1, θ_2 and η the model parameters of the sources and of the noise. The problem of source separation then becomes :

$$(\widehat{s}_1, \widehat{s}_2) = \arg \max_{(s_1,s_2)} [P(s_1, s_2|y, \theta_1, \theta_2)]$$

By applying the Bayes rule and by assuming statistical independence between the two sources, the desired result can be obtained by solving :

$$(\widehat{s}_1, \widehat{s}_2) = \arg \max_{(s_1,s_2)} [P(y|s_1, s_2)P(s_1|\theta_1)P(s_2|\theta_2)]$$

The first of the three terms in the argmax can be obtained via the model noise :

$$P(y|s_1, s_2) \propto P(y - (s_1 + s_2)|\eta) = P(\epsilon|\eta)$$

The two other terms are obtained via likelihood functions corresponding to source models trained from examples, or designed from knowledge sources. For example, commonly used models are the Laplacian model, the Gaussian Mixture Model or the Hidden Markov Model.

These models can be linked to the distribution of the representation coefficients in a redundant system in which are pooled together several bases adapted to each of the sources present in the mixture.

4. Application Domains

4.1. Introduction

This section reviews a number of application domains in which the METISS project-team has been particularily active : speaker characterisation, audio description and indexing (including speech recognition) and advanced audio processing (in paticular, source separation).

4.2. Speaker characterisation and speech recognition

Keywords: *beam-search*, *broadcast news indexing*, *normalisation*, *representation and adaptation*, *rich transcription*, *scalability*, *speaker elicitation*, *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. Indeed, even though the voice characteristics of a person are not unique [63], many factors (morphological, physiological, psychological, sociological, ...) have an influence on a person's voice. One focus of the METISS group in this domain is speaker verification, i.e the task of accepting or rejecting an identity claim made by the user of a service with access control. We also dedicate some effort to the more general problem of speaker characterisation. 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.

4.2.2. Speaker model and test normalisation

A key issue, in many practical applications, is the non-controlable deviation of speaker models from the exact probability density functions. This requires a step of normalisation before comparing the verification score to a decision threshold. This issue has been a particular focus for our recent efforts in the domain of speaker verification and has led to the design and evaluation of various strategies of model and test normalisation.

4.2.3. Speaker representation, selection and adaptation

METISS also adresses a number of other 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.

4.2.4. Scalability and complexity reduction for speaker recognition

In order to address needs related to the implementation of speaker verification technology on personal devices, specific algorithmic approaches have to be developed to contribute to the scalability, the complexity reduction and the process distribution. In this context, speaker modelling approaches and classification procedures need to be designed, simulated and tested.

4.2.5. Speech modeling and recognition

Speech modeling and recognition is complementary with other speech related activities in the group, in particular, speaker recognition and audio description. In the first case, detecting speech segments in a continuous audio stream and segmenting the speech portions into pseudo-sentences is a preliminary step to automatic transcription. Detecting speaker changes and grouping together segments from the same speaker is also a crucial step for segmentation as for speaker adaptation, and can rely on acoustic as well as lexical

and linguistic features. Last, in speaker recognition for secured transactions over the telephone, recognizing the linguistic content of the message might be useful, for example to hypothesize an identity, to recognize a spoken password or to extract linguistic parameters that can benefit to the speaker models.

4.3. Description and structuration of audio and multimodal streams

Keywords: *audio descriptors*, *audio detection*, *audio segmentation*, *audio stream*, *audio tracking*, *audio-object extraction*, *audio-visual descriptors*, *audiovisual integration*, *harmony*, *information fusion*, *melody*, *multimedia indexing*, *multimodality*, *music language modeling*.

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 documents for a wide variety of features and documents.

4.3.1. Speaker detection

Speaker characteristics, such as the gender, the approximate age, the accent or the identity, are key indices for the indexing of spoken documents. So are information concerning the presence or not of a given speaker in a document, the speaker changes, the presence of speech from multiple speakers, etc.

More precisely, the above mentioned tasks can be divided into three main categories: detecting the presence of a speaker in a document (classification problem); tracking the portions of a document corresponding to a speaker (temporal segmentation problem); segmenting a document into speaker turns (change detection problem).

These three problems are clearly closely related to the field of speaker characterisation, sharing many theoretical and practical aspects with the latter. In particular, all these application areas rely on the use of statistical tests, whether it is using the model of a speaker known to the system (speaker presence detection, speaker tracking) or using a model estimated on the fly (speaker segmentation). However, the specificities of the speaker detection task require the implementation of adequate solutions to adapts to situations and factors inherent to this task.

4.3.2. 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.

4.3.3. Describing multi-modal information for indexing purposes

Applied to the sound track of a video, detecting and tracking audio events, as mentioned in the previous section, 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.

4.3.4. Music 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.

4.3.5. 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.

4.4. Source separation and advanced audio coding

Keywords: *audio events*, *audio objects*, *multi-channel sound*, *sound models*, *source separation*.

Speech signals are commonly found surrounded or superimposed with other types of audio signals in many application areas. The former are often mixed with musical signals or background noise. Moreover, audio signals frequently exhibit a composite nature, in the sense that they were originally obtained by combining several audio tracks with an audio mixing device. Audio signals are also prone to suffer from all kinds of degradations –ranging from non-ideal recording conditions to transmission errors– after having travelled through a complete signal processing chain.

Recent breakthrough developments in the field of voice technology (speech and speaker recognition) are a strong motivation for studying how to adapt and apply this technology to a broader class of signals such as musical signals.

The main themes discussed here are therefore those of source separation and audio signal representation.

4.4.1. 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.

4.4.2. Audio signal analysis, 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.

Atomic decomposition methods are yielding a rising interest in the field of sound representation, compression and synthesis. They attempt to provide such representation of audio signals as linear sums of elementary signals (or "atoms") from a "dictionary". In the classical model, "sonic grains" are deterministic functions (modulated sinusoïds, chirps, harmonic molecules, or even arbitrary waveforms stored in a wavetable, etc.). The reconstructed signal $y(t)$ is then the M-term adaptive approximation of the original signal from the dictionary D. Non-linear approximation theory and decomposition methods such as Matching Pursuit and derivatives respectively provide a mathematical framework and powerful tools to tackle this kind of problem.

4.4.3. Audio object coding

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.

5. Software

5.1. SPro and AudioSeg: 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

URL : http://gforge.inria.fr/projects/spro, http://gforge.inria.fr/projects/audioseg

5.2. Sirocco: a speech recognition search engine

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 actively participates in the development of the freely available Sirocco large vocabulary speech recognition software [65]. The Sirocco project started as an INRIA Concerted Research Action now works on the basis of voluntary contributions.

We use the Sirocco speech recognition software as the heart of the transcription modules whithin our spoken document analysis platform IRENE. In particular, it has been extensively used in our researches on ASR and NLP as well as for our work on phonetic landmarks in statistical speech recognition.
Contact : guillaume.gravier@irisa.fr
URL : http://gforge.inria.fr/projects/sirocco

5.3. MPTK: the Matching Pursuit Toolkit

Participants: Rémi Gribonval, Sylvain Lesage, 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 this year 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

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

5.4. BSS_ORACLE: A toolbox to compute oracle estimators for source separation

Participants: Emmanuel Vincent, Rémi Gribonval.

BSS_ORACLE is a MATLAB toolbox to compute the best performance achievable by a class of source separation algorithms in an evaluation framework where the true sources are known. Version 2.1 has been released this year. The toolbox provides oracle estimators defined in [38] and [52] for four classes of algorithms (time-invariant multichannel filtering, single-channel time-frequency masking, multichannel time-frequency masking and best basis masking), each with several variants (time-domain vs. frequency-domain, MDCT vs. STFT, etc).

Contact : emmanuel.vincent@irisa.fr

URL : http://bass-db.gforge.inria.fr/bss_oracle

6. New Results

6.1. Speaker characterisation

Keywords: *Gaussian Mixture Models (GMM)*, *affective computing.* , *speaker adaptation*, *speaker characterisation*, *speaker selection*.

6.1.1. Rapid Speaker Adaptation by Reference Model Interpolation

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

Acoustic model based adaptation techniques have become in recent years an important element in speech recognition systems to tune the system to the user's voice. Moreover, in somme applicative contexts, speaker's adaptation must take place on-line and rapidly.

We have designed a novel algorithm for fast speaker adaptation using small amounts of adaptation data. The approach is based on a set of representative speakers which can provide a priori knowledge to guide the estimation a new speaker's model in the speaker space.

The proposed scenario is based on an *a posteriori* selection of reference models as opposed to conventional techniques (such as eignevoices) which uses a fixed set of reference speakers. It calls for a user-dependent linear interpolation of the parameters of the reference speaker models

Comparisons of the proposed approach on the IDIOLOGOS and PAIDIALOGOS corpora have yields to slightly better performances tha eigenvoices on a phoneme recognition task, especially for atypical speakers such as children [50].

6.1.2. Voice characteristics modelling for emotion and cognitive state classification

Keywords: *cognitive state*, *emotion*, *psychoacoustic*, *voice interaction*.

Participants: Klara Trakas, Frédéric Bimbot.

This work is taking place in the context of an industrial PhD just starting with Orange FTR&D Labs.

Increased interest is noticeable 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.

Joint work between the Metiss Group is just starting to investigate descriptors and models for representing this type of 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.

6.2. Audio analysis and structuring for multimedia indexing and information extraction

Keywords: *audio and multimodal structuring broadcast news indexing*, *audio detection*, *audio segmentation*, *audiovisual integration*, *multimedia*, *rich transcription*, *speech recognition*, *statistical hypothesis testing*, *statistical hypothesis testing*.

6.2.1. Audio content processing and information extraction in sports events

Keywords: *audio detection*, *audio segmentation*, *audio-visual fusion*, *support vector machines*, *word spotting*.

Participants: Mathieu Ben, Gilles Gonon, Sylvain Busson, Guillaume Gravier, Frédéric Bimbot.

This work has been done in the context of the ITEA PELOPS project, in close cooperation with Thomson Multimedia).

Extracting relevant information in sports programmes (such as soccer matches) is a challenge which is closely linked to applicative considerations, such as automatic content summarization and fast post-production and repurposing. In this context, the activities of the Metiss group in the PELOPS Project were focused on 2 tasks :

- Generation of acoustic and semantic descriptors from audio soundtracks.
- Audio-visual information fusion and integration, for the classification of highlights in a sport event (collaboration with Thomson Multimedia who provides video descriptors).

A set of low levels audio descriptors have been setup, using statistical and pattern recognition techniques. BSS techniques are used as a preprocessing phase to separate the commentator track from the crowd and field ambiance. This preprocessing step improved the robustness of several audio descriptors, such as commentator pitch tracking, rate of commentator speech and cheering level.

The fusion and integration of audio and visual information addresses the problem of combining heterogenous descriptors with asynchronous streams. The events are modelled by means of contextual relations between time intervals, using different statistics on the descriptors (max, min, standard deviation). Support Vector Machine classifiers have been used to train the models and to score the test matches, as described in [17].

The resulting event classification is synchronized with the video shot segmentation and each shot is assigned a score for the considered events (goals, cards, goal attempts, other). The classification is evaluated using a precision-recall curve. In our experiments on a corpus of 12 soccer matches, 100of the shots with the highest estimated goal probability.

6.2.2. Integrating natural language processing and speech recognition

Keywords: *natural language processing*, *speech recognition*, *spoken document analysis*, *spoken document segmentation*.

Participants: Guillaume Gravier, Stéphane Huet.

This work has been done in close collaboration with the TEXMEX *project-team at IRISA and has led to a rising collaboration with the NLP group at the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE, Puebla, Mexico).*

Automatic speech recognition (ASR) systems aim at generating a textual transcription of a spoken document, usually for further analysis of the transcription with natural language processing (NLP) techniques. However, most current ASR systems solely rely on statistical methods and seldom use linguistic knowledge. In collaboration with the NLP group in the TEXMEX project-team of IRISA, we investigated several directions toward a better use of linguistic knwoledge such as morphology, syntax, semantics and pragmatics in ASR.

The works described here under were implemented in our Sirocco software and incorporated in our spoken document analysis platform IRENE. The proposed approaches were benchmarked on the ESTER French broadcast news corpus [8] which consitutes a reference in ASR for the French language.

6.2.2.1. Using morpho-syntactic knowledge in speech recognition

In 2006, we had demonstrated the interest of a score combining acoustic, language and morpho-syntactic information to rescore N-best sentence hypothesis lists. This year, we consolidated these results with various configuration of our ASR system and studied the impact of morpho-syntactic information for confidence measure computation. In particular, we demonstrated that confidence measures can be improved based on our combined score function [45].

6.2.2.2. Spoken document segmentation

Spoken document segmentation is a crucial step for the analysis of multimedia documents which requires the combination of linguistic and acoustic cues. To this end, we extended a statistical method based on lexical cohesion [73] for topic shift detection to take into account additional knowledge such as semantic relation between words, syntactic coherence and acoustic cues. Our technique enables us to improve segmentation, although a few parts —particularly those corresponding to the news headlines— have still to be refined.

6.2.2.3. Using information retrieval for language model adaptation

We proposed a method to adapt the language model of an ASR system for each segment resulting from the segmentation step described above. The method is completely unsupervised and uses neither *a priori* knowledge about topics nor a static collection of texts. The idea is to gather textual adaptation data for each segment, based on information retrieval (IR) methods to extract keywords which are used to retrieve documents from the Web. IR techniques, used both for keyword extraction and for document selection, have been adapted to tackle the specificities of automatic transcriptions (e.g. misrecognized words, named entities). Results indicate a large improvement of the language model, which finally yields a small improvement of the word error rate [20].

This preliminary work has demonstrated the potential of our approach to efficiently transcribe speech streams and suggests further work on language model and vocabulary adaptation based on IR methods to gather adaptation data from the Internet. The thesis of Gwénolé Lecorvé, which started in September 2007 in collaboration with the TEXMEX project-team, will be dedicated to language model and vocabulary adaptation for the robust transcription of multimedia streams.

6.2.3. Speech recognition based on phonetic landmarks

Keywords: *audio classification*, *phonetic classification*, *speech recognition*.

Participant: Guillaume Gravier.

HMM-based automatic speech recognition can hardly accomodate prior knowledge on the signal, apart from the definition of the topology of the phone-based elementary HMMs. In the previous years, we have shown that such knowledge can be efficiently used during decoding with the Viterbi algorithm as constraints on the best path.

Preliminary experiments have shown that accurately detecting broad phonetic landmarks, such as vowels or stops, can greatly benefit to ASR. Hence, we focused this year on the actual detection of such landmarks. Experiments on HMM-based landmark detection demonstrated that, if HMMs can be used to provide a segmentation into broad phonetic events, the classification rate is not high enough to benefit speech recognition. This is also due to the fact that the same paradigm (features and models) is used for both landmark detection and speech recognition. We therefore focused on the use of support vector machines to classify feature vectors into broad phonetic classes, achieving classification rates around 95 % for vowels, fricatives and nasals [42].

In the future, we plan to improve SVM-based landmark detection using different features and to demonstrate the actual feasibility of broad phonetic landmark-driven speech recognition.

6.2.4. Motif discovery in audio documents

Keywords: *data mining*, *dynamic time warping*, *motif discovery*, *pattern search*, *support vector machines*.

Participants: Guillaume Gravier, Armando Muscariello, Frédéric Bimbot.

Discovering repeating motifs —such as advertisments, jingles or even words— in audio streams or databases is a crucial task for the unsupervised strucuring of audio data collections and a necessary step toward the lightly supervised design of audio event recognition systems. Research in this field are oriented along two main axes, namely efficient search of a motif (querry) and efficient representation of a motif to deal with variability.

In 2007, our activity in the field of audio motif discovery mainly focuses on the study of sequence models for fast retrieval of audio sequences, in collaboration with the TEXMEX project-team at Irisa. Extending existing multidimensional indexing techniques is not possible as these were designed for description schemes in which the concept of sequence lacks. A solution is to summarize the sequence in a model before indexing and comparing models rather than sequences. To this end, we investigated the use of support vector machines as a prediction model and compared the SVM-based comparison of sequences with the more traditionnal feature-based dynamic time warping alignment method. Overall, we have shown that relying on models (instead of relying on descriptors) provides a better robustness to severe modifications of sequences, like temporal distortions for example [48], [49].

These encouraging results motivate further investigation on SVM-based models of audio sequences. In parallel, the thesis of Armando Muscariello, which started in October 2007, will focus on the practical application of sequence models for motif discovery in audio streams, aiming at the discovery of variable motifs.

6.2.5. Multimodal integration

Keywords: *dynamic Bayesian networks*, *multimodal integration*, *video structuring*.

Participant: Guillaume Gravier.

The work described in this section is carried out in the framework of the Ph. D. thesis of Siwar Baghdadi, in collaboration with the TEXMEX *project-team of IRISA and Thomson Multimedia Research.*

Bayesian networks provide an interesting framework for the joint modeling of multimodal information. Morevoer, unlike HMMs and segment models, it is possible to learn the structure of a Bayesian network, *i.e.* the relation between the variables describing the problem, from data.

We investigated the use of dynamic Bayesian networks and the potential of structure learning algorithms such as K2 [64] for multimodal integration in a commercial detection application. A video stream is considered as a succession of shots, where a shot is represented by a set of visual and audio features, which can be labeled either as commercial or not. We have shown that structure learning algorithms can efficiently learn the relations between the variables describing a shot. We investigated different approaches to model temporal relations between shots, in particular using an explicit duration model as in segment models.

Future work involves the extension of this approach to event detection in soccer games with a focus on structure learning, either static or temporal, in order to provide a framework for the lightly supervised development of new applications.

6.2.6. Polyphonic music transcription and coding

Keywords: *instrument identification*, *music*, *object coding*, *pitch transcription*.

Participant: Emmanuel Vincent.

Music signals can be described by a score consisting of several notes defined by their onset time, duration, pitch and instrument class. The task of estimating the notes underlying a given signal is termed polyphonic music transcription. It involves two substasks, namely pitch transcription and instrument identification. This task can also form the core of an "object-based" coder, encoding the signal in terms of resynthesis parameters for each note and allowing high-level manipulation of the signal.

Our previous work [39] focused on the modeling of music signals via Bayesian harmonic models. This year we proposed an improved inference method for such models allowing faster computation of the posterior probability of a set of notes on a given time frame.

We also investigated alternative methods addressing this task in the framework of sparse representations. The first method represents the signal in each time frame as a linear combination of harmonic atoms learnt on isolated notes from various instruments. The relevant atoms are selected by Matching Pursuit and additional structural constraints are used to extract sequences of atoms modeling individual notes. The second method represents the short-term magnitude spectrum as a linear combination of magnitude spectra corresponding to different pitches. These spectra are adapted from the signal alone ny minimizing the loudness of the residual under harmonicity constraints. This method provided similar pitch transcription accuracy as state-of-the-art methods, while allowing better generalization to unknown instruments.

Finally, we investigated the use of such note-based representations for bandwidth extension and "resolution-free" audio coding.

This work was conducted in collaboration with Mark D. Plumbley and Steve Welburn (Queen Mary, University of London), Pierre Leveau and Laurent Daudet (Université Paris 6) and Nancy Bertin and Roland Badeau (GET - Télécom Paris). Previous results have been published as journal articles [39], [31]. New results have been submitted to a journal [60] and published in the proceedings of a conference [56] and an evaluation campaign [51].

6.2.7. Statistical models of music

Keywords: *musical description*, *statistical models*.

Participants: Amadou Sall, Frédéric Bimbot.

Speech recognition is very advantageously guided by statistical language models : we hypothetise that music description, recognition and retranscription can strongly benefit from music models that express dependencies between notes within a music piece, due to melodic patterns and harmonic rules.

To this end, we have investigated the approximate modeling of syntactic and paradigmatic properties of music, through the use of n-grams models of notes, succession of notes and combinations of notes.

In practice, we consider a corpus of MIDI files on which we learn co-occurences of concurrent and consecutive notes, and we use these statistics to cluster music pieces into classes of models and to measure predictability of notes within a class of models.

The model is intended to be used in complement to source separation and acoustic decoding, to form a consistent framework embedding signal processing techniques, acoustic knowledge sources and music rules modeling. A publication is in preparation.

6.3. Source separation

6.3.1. Source separation using multichannel Matching Pursuit

Keywords: *Matching Pursuit*, *linear instantaneous*, *multichannel*, *sparse decomposition*, *underdetermined blind source separation*.

Participants: Sylvain Lesage, Sacha Krstulovic, Rémi Gribonval.

The source separation problem consists in retrieving unknown signals (the sources) form the only knowledge of one or more mixtures of these signals (the channels coming from each sensor). In the case we study, each channel is a linear combination of the sources, and there are more sources than channels, and at least two channels. Due to the underdeterminacy of the problem, knowing all the parameters of the mixing process is not sufficient to retrieve the sources. Focussing on the estimation of the sources –assuming the mixing process is known– we have studied methods to perform the separation based on sparse decomposition of the mixture with Matching Pursuit. Methods for the estimation of the mixing parameters are developped apart (see next section).

Last year we concentrated [68] on methods based on the difference in spatial direction between sources, assuming the source signals can be sparsely decomposed on a joint dictionary. This year, we explored the possibility of simultaneously exploiting spatial differences and "morphological" differences, by choosing a distinct dictionary to sparsely model each source signal in the spirit of [61]. For sources wich can be modeled sparsely in sufficiently distinct domains (e.g., drums and electric guitar), our experiments showed that this approach can drastically improve separation performance. While learning appropriate dictionaries for each source based on training data is straightforward, the problem of training adapted dictionaries based on the only knowledge of the mixture remains a challenge.

This work is has been presented in a workshop.

6.3.2. DEMIX anechoic: a robust algorithm to estimate the number of sources in a spatial anechoic mixture

Keywords: *clustering*, *linear instantaneous*, *multichannel*, *source localisation*, *underdetermined source separation*.

Participants: Simon Arberet, Rémi Gribonval, Frédéric Bimbot.

An important step for audio source separation consists in finding both the number of mixed sources and their directions in a multisensor mixture.

In complement to the separation methods based on Matching Pursuit, which we developed and evaluated assuming the mixing matrix is known, we proposed last year a robust technique to address this problem in the case of linear instantaneous mixtures [1], even with more sources than sensors. This year, we extended the approach to a more realistic setting of linear anechoic mixture (where the mixture involves not only intensity difference but also time delays between channels).

The method relies on the assumption that in the neighborhood of some time-frequency points, only one source contributes to the mixture. Such time-frequency points, located with a local confidence measure, provide estimates of the attenuation, as well as the phase difference at some frequency, of the corresponding source. Combining the phase differences at different frequencies, the time delay parameters are estimated, by a method similar to GCC-PHAT, on points having similar intensity differences. As a result, unlike DUET type methods, our method makes it possible to estimate time-delays higher than only one sample.

Experiments show that, in more than 65% of the cases, DEMIX Anechoic correctly estimates the number of directions until 6 sources. Moreover, it outperforms DUET in the accuracy of the estimation by a factor ten.

This work is currently submitted for publication.

6.3.3. Single channel source separation

Keywords: *Gaussian mixture model*, *Single channel source separation*, *Wiener filter*, *model adaptation*.

Participants: Alexey Ozerov, Rémi Gribonval, Frédéric Bimbot.

Probabilistic approaches can offer satisfactory solutions to source separation with a single channel, provided that the models of the sources match accurately the statistical properties of the mixed signals. However, it is not always possible in practice to construct and use such models.

To overcome this problem, we propose to resort to an adaptation scheme for adjusting the source models with respect to the actual properties of the signals observed in the mix. We develop a general formalism for source model adaptation. In a similar way as it is done for instance in speaker (or channel) adaptation for speech recognition, we introduce this formalism in terms of a Bayesian Maximum A Posteriori (MAP) adaptation criterion. We show then how to optimize this criterion using the EM (Expectation - Maximization) algorithm at different levels of generality.

Formulated in such a general way this adaptation formalism can be applied for different models (GMM, HMM, etc.) and using different types of priors (probabilistic laws, structural priors, etc.). Also, we extend this formalism by explaining how to integrate to the adaptation scheme any auxiliary information available in addition to the mix. This can be for example visual information, time segmentation of sound classes, some forms of incomplete separation, etc.

To show the use of model adaptation in practice, we apply this adaptation formalism to the problem of separating voice from music in popular songs. In 2005 we proposed some adaptation techniques based on some segmentation of the processed song into vocal and non-vocal parts. These techniques include learning of music model from the non-vocal parts and voice model filter adaptation from the vocal parts [72], [71].

We show that these adaptation techniques are just some particular forms of our general adaptation formalism. Furthermore, we introduce a new Power Spectral Density (PSD) gains adaptation technique, and we explain how to perform joint filter and PSD gains adaptation for voice model, which leads to better performance than filter adaptation alone. Finally, in addition to what was done in [72], [71], where a manual vocal / non-vocal segmentation was used, we have developed some automatic segmentation module.

Thus, we have developed a one microphone voice / music separation system based on adapted models. This system performs in a completely automatic manner, i.e. without any human intervention, and the computation load is quite reasonable (not more than 10 times real time). The obtained results show that for this task an adaptation scheme can significantly improve (at least by 5 dB) the separation performance in comparison with non-adapted models.

This work is accepted for publication [35] and is thoroughly detailed in Alexey Ozerov's Ph.D. manuscript [16]. It was done in close collaboration with FTR&D (Pierrick Philippe).

6.3.4. Source separation via sparse adaptive representations

Keywords: *adaptive basis*, *source separation*, *sparse representation*.

Participants: Rémi Gribonval, Emmanuel Vincent.

Source separation is the task of retrieving the source signals underlying a multichannel mixture signal, where each channel is the sum of scaled versions of the sources (instantaneous case) or filtered versions thereof (convolutive case). A popular approach is to assume that the sources admit a sparse representation in some (possibly overcomplete) basis. Separation can then be achieved by sparse decomposition of the mixture signal. Previous work in the group focussed on fixed time-frequency bases and source-adapted bases trained on isolated samples of each source.

This year we proposed two methods to adapt the bases directly from the mixture signal. The first method aims to find a time-frequency basis such that the source signals overlap as little as possible in this basis, so that separation can be performed by binary masking, i.e. associating each time-frequency bin with a single source. Such a basis is estimated by minimizing a quadratic overlap criterion, given the spatial directions of the sources. Experiments with Cosine Packet (CP) bases showed that this method outperformed binary masking on a fixed MDCT basis for the separation of stereo instantaneous mixtures of three sources.

The second method assumes that each time frame of the mixture signal can be represented as a sparse linear combination of multichannel atoms forming a complete basis, where each atom belongs to a single source. The best basis is found for all time frames by minimizing the lp norm of the combination weights. The spatial direction associated with each atom is then estimated using the GCC PHAT estimator and the set of atoms corresponding to each source is estimated by clustering of the directions. This method outperformed both convolutive ICA and DUET approaches on low-reverberation convolutive mixtures.

We also studied the minimization of the lp norm of the combination weights for complex-valued overcomplete bases. This optimization problem is difficult since it is nonconvex and theoretical results for real-valued data do not apply for complex-valued data. We characterized the local minima of the lp norm in a simple case and derived a fast algorithm for the estimation of the global minimum. This algorithm has been applied to the separation of stereo instantaneous and convolutive mixtures of three sources.

This work was conducted in collaboration with Maria G. Jafari and Mark D. Plumbley (Queen Mary, University of London) and Mike E. Davies (University of Edinburgh). The results have been published in the form of a journal article [29], a book chapter [24] and two conference papers [52], [55].

6.3.5. Evaluation of source separation algorithms

Keywords: *benchmark*, *blind source separation*, *evaluation*, *performance measure*.

Participants: Rémi Gribonval, Emmanuel Vincent.

Source separation of under-determined and/or convolutive mixtures is a difficult problem that has been tackled by many algorithms based on different source models. Their performance is usually limited by badly designed source models or local maxima of the function to be optimized. Moreover, it may be limited by algorithmic constraints, such as the length of the demixing filters or the number of frequency bins of the time-frequency masks. The best possible source signal that can be estimated under these constraints (in the ideal case where source models and optimization algorithms are perfect) is called an oracle estimator of the source. We have expressed and implemented oracle estimators for four classes of algorithms (time-invariant beamforming, single-channel time-frequency masking, multichannel time-frequency masking and best basis masking) and studied their performance on realistic speech and music mixtures. The results have led to interesting conclusions concerning the performance bounds of blind algorithms, the choice of the best class of algorithms and the assessment of the separation difficulty.

This work, which builds up on our previous contribution published in [74], was done in collaboration with Emmanuel Vincent and Mark D. Plumbley (Queen Mary, University of London). For more detail, please refer to [38] and [52].

6.4. Sparse decompositions: theory and algorithms

6.4.1. Learning of deformation-invariant atoms

Keywords: *Principal Component Analysis*, *Redundant dictionary learning*, *atom*, *shift invariance*, *sparsity*.

Participants: Sylvain Lesage, Boris Mailhé, Rémi Gribonval, Frédéric Bimbot.

Sparse approximation using redundant dictionaries is an efficient tool for many applications in the field of signal processing. The performances largely depend on the adaptation of the dictionary to the signal to decompose. As the statistical dependencies are most of the time not obvious in natural high-dimensional data, learning fundamental patterns is an alternative to analytical design of bases and has become a field of acute research. Most of the time, several different observed patterns can be viewed as different deformations of one generating function. For example, the underlying patterns of a class of signals can be found at any time, and in the design of a dictionary, this shift invariance property should be present. We developed a new algorithm for learning short generating functions, each of them building a set of atoms corresponding to all its translations. The resulting dictionary is highly redundant and shift invariant.

This algorithm learns the set of generating functions iteratively, from a set of learning signals. Each iteration is an alternate routine : we begin with a sparse decomposition of the learning signals on the dictionary generated by the learnt generating functions. We used Matching Pursuit for this step, mostly because of the availability of a fast implementation 5.3. Then, for each generating function, we get one signal patch for each occurrence of this function found by the decomposition and we update the function to obtain a least-square error approximation of the patches. Depending on whether you allow some decomposition coefficients to be updated or not during this step, the new function is given by the first principal component or the centroid of the corresponding patches. The first method gives a better approximation of the patches while the second one yields a lower algorithmic complexity. Then we iterate the same process.

On natural images, the learnt atoms are similar to what is generally found in the litterature. On other data, like ECG or EEG, typical waveforms are retrieved. We also show the results of a test on audio data, where the approximation using some learnt atoms is sparser than using local cosines.

This work, which extends our previous work with the MOTIF algorithm [67], was presented at a workshop It was done in collaboration with the group of Pierre Vandergheynst (EPFL, Lausanne). We are currently working on other deformation classes, such as phase shifts for audio signals, dilatation and rotation for images.

6.4.2. Learning multimodal dictionaries: applications to audiovisual data

Keywords: *Principal Component Analysis*, *Redundant dictionary learning*, *atom*, *audiovisual data*, *early fusion*, *multimodal data*, *shift invariance*, *sparsity*, *speaker localization*, *speaker tracking*.

Participants: Sylvain Lesage, Boris Mailhé, Rémi Gribonval.

Real-world phenomena involve complex interactions between multiple signal modalities. As a consequence, humans are used to integrate at each instant perceptions from all their senses in order to enrich their understanding of the surrounding world. This paradigm can be also extremely useful in many signal processing and computer vision problems involving mutually related signals. The simultaneous processing of multi-modal data can in fact reveal information that is otherwise hidden when considering the signals independently. However, in natural multimodal signals, the statistical dependencies between modalities are in general not obvious. Learning fundamental multi-modal patterns could offer a deep insight into the structure of such signals. Typically, such recurrent patterns are shift invariant, thus the learning should try to find the best matching filters. In this paper we present an algorithm for iteratively learning multimodal generating functions that can be shifted at all positions in the signal. The learning is defined in such a way that it can be accomplished by iteratively solving a generalized eigenvector problem, which makes the algorithm fast, flexible and free of user-defined parameters. The proposed algorithm is applied to audiovisual sequences and we show that it is able to discover underlying structures in the data. In particular, it is possible to locate the mouse of a speaker based on the learnt multimodal dictionaries, even in adverse conditions where the audio is corrupted by noise and other speakers are visible (but not audible) who utter the same words as the target speaker. This work, which was done in collaboration with G. Monaci, P. Jost and P. Vandergheynst from EPFL was published in [15] and is currently submitted for possible journal publication.

6.4.3. Average case analysis of multichannel thresholding

Keywords: *average case*, *matching pursuit*, *multichannel signal analysis*, *recovery analysis*, *sensor networks*, *sparse decomposition*, *thresholding*, *worst case*.

Participants: Rémi Gribonval, Boris Mailhé.

Recent developments in sparse signal models mainly focus on analyzing sufficient conditions which which guarantee that various algorithms (matching pursuits, basis pursuit, ...) can “recover” a sparse signal representation. Typical conditions involve both basic properties of the representation itself (which should be sufficiently sparse or compressible) and of the dictionary used to represent the signal, which should satisfy some uniform uncertainty principle. Even though random dictionary models can be used to prove that strong uniform uncertainty principles are met by “most” dictionaries, it seems to remain combinatorial to check it for a specific dictionary, for which estimates based on the coherence provide very pessimistic recovery conditions.

In parallel to developments in sparse signal models, various application scenarios motivated renewed interest in processing not just a single signal, but many signals or channels at the same time. A striking example is sensor networks, where signals are monitored by low complexity devices whose observations are transfered to a central collector [69]. This central node thus faces the task of analyzing many, possibly high-dimensional, signals. Moreover, signals measured in sensor networks are typically not uncorrelated: there are global trends or components that appear in all signals, possibly in slightly altered forms.

We developped an analysis of the theoretical performance of two families of simultaneous sparse representation algorithms. First, we considered p-thresholding, a simple algorithm for recovering simultaneous sparse approximations of multichannel signals. Our analysis is based on studying the average behaviour in addition to the worst case one, and the spirit of our results is the following: given a not too coherent dictionary and signals with coefficients sufficiently large and balanced over the number of channels, p-thresholding can recover superpositions of up to $\mathcal{O}(d)$ atoms *with overwhelming probability* in dimension d. Our conditions on $\mathcal{D}$ are thus much less restrictive than in the worst case where only $\mathcal{O}(\sqrt{d}$ atoms can be recovered. Numerical simulations confirm our theoretical findings and show that p-thresholding is an interesting low complexity alternative to simultaneous greedy or convex relaxation algorithms for processing sparse multichannel signals with balanced coefficients.

This work was done in collaboration with Karin Schnass and Pierre Vandergheynst, EPFL, and Holger Rauhut, University of Vienna. A paper is in preparation and a conference paper was submitted for publication.

7. Contracts and Grants with Industry

7.1. ACI actions

7.1.1. ACI Masse de Données Demi-ton

Participants: Guillaume Gravier, Daniel Moraru, Stéphane Huet.

This project entitled "Multimodal description for automatic structuring of TV streams" started in Oct. 2004 and is funded by the ACI Masse de Données. The partners are the METISS and TEXMEX groups at IRISA and the DCA group at INA.

The aim of this project is to propose and evaluate algorithms to structure the video stream in order to automate this tedious part of the indexing process at INA. The main scientific objectives are the joint modeling of different medias (image, text, meta-data, sound, etc.) in a statistical framework and the use of prior information, mainly the program guide, in collaboration with a statistical model.

In the framework of this project, our team works on the use of segment models for video structuring as well as on the segmentation and transcription of the video stream soundtrack.

7.2. European Project supported by the French Authorities

7.2.1. Projet EUREKA/ITEA PELOPS

Participants: Mathieu Ben, Gilles Gonon, Sylvain Busson, Guillaume Gravier, Frédéric Bimbot.

The PELOPS project is a EUREKA-ITEA Project which started in 2005. IRISA joined the project in July 2006. The project terminated in June 2007.

The partners are Thomson Multimedia, Acotec, Barco, EVS, Leo Vision, MOG and Telefonica.

The project was targeted towards content creation and repurposing for live sports events.

The contribution of IRISA was focused on the conception of audio analysis tools and processes for content analysis, structuration and prioritisation, using statistical approaches for audio classification and source separation techniques.

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, Sylvain Lesage.

A bilateral collaboration with the Signal Processing group (LTS2) led by Pierre Vandergheynst at EPFL (Switzerland) was initiated a few years ago within the HASSIP European research training network. Since 2005, thanks to bilateral funding by the foreign affairs ministry, the collaboration has been reinforced, and has lead to several student exchanges and academic visits, including a two month visit of Rémi Gribonval at EPFL in the summer of 2006. Since the fall of 2005, a co-supervised Ph.D. thesis (Boris Mailhé) has started to reinforce even more the collaboration, and an INRIA Associated Team called SPARS officially started in January 2007 to strengthen and build upon this collaboration in the coming years. The collaboration resulted so far in joint theoretical contributions on sparse signal approximation, as well as on multimodal audiovisual signal analysis, using the complementary competences in audio (METISS) and image/video (LTS2) applications of sparse signal models.

8.2. Visites, et invitations de chercheurs

8.2.1. Exchanges within the Associated Team SPARS

In the framework of the INRIA Associated Team SPARS, several junior and senior researches from LTS2 (EPFL) visited the METISS group in 2007. A first visit by Pierre Vandergheynst and Karin Schnass was the occasion to complete a paper on the theoretical analysis of multichannel sparse approximation algorithms, which is currently submitted for publication. These results were presented at several international conferences this year. During a visit by Anna Llagostera and Gianluca Monaci, we experimented with multimodal signal models, using visual information from audiovisuel data to train audio source models for single channel source separation. Independently from the SPARS Associated Team, with the help of GDR ISIS - CNRS funding, we invited Matthieu Kowalski, a Ph.D. student with Bruno Torrésani at LATP, Université de Provence, for a one month visit where we studied iterative optimization algorithms for structured multichannel decompositions, experimenting their possible applications to convolutive source separation.

8.2.2. Visit to the LTL Lab in Mexico

Guillaume Gravier visited the Language Technology Lab (LTL) at the INAOE (Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico) for two month in July and August 2007 with the goal of a collaboration in the field of spoken document analysis. This stay has been the opportunity to investigate the application of the natural language processing techniques developed at LTL (text segmentation, clustering, summarization ...) on the output of the IRENE transcription system. Experiments have been carried out on text segmentation and on text pattern discovery. As a result of this exploratory visit (INRIA exploratory visit program 01AUTR5050-8), a more formal collaboration with the LTL is under study, targeting in particular a joint participation in the QAst (Question Answering on speech transcriptions) track of the CLEF evaluation campaign.

9. Dissemination

9.1. Conference and workshop committees, invited conference

Rémi Gribonval was an invited tutorial lecturer at the 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007), Londres, UK, september 2007.

Rémi Gribonval was a member of the Program Committee for the GRETSI french speaking Workshop on Signal and Image Processing which was held in Troyes, France in september 2007.

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

Frédéric Bimbot is a member of the Programme Committee for the Eusipco 2008 Conference, to be held in Lausanne, Switzerland, August 25-29, 2008.

Guillaume Gravier is part of the NOE MUSCLE.

Emmanuel Vincent was one of the panelists of the discussion session on evaluation campaigns held at the ICA Conference, London, september 2007.

9.2. Leadership within scientific community

Rémi Gribonval participates to the CNRS expert committee “methods in signal and image processing”.

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.

Emmanuel Vincent was the chair of first Stereo Audio Source Separation Evaluation Campaign (SASSEC).

9.3. Teaching

Rémi Gribonval has given 8 hours of lecture on signal and image representation within the ARD module of the Masters in Computer Science, Université de Rennes 1.

Guillaume Gravier has given two 2-hour conferences on Voice Technologies at the École Supérieure d’Applications des Transmissions (ESAT, Rennes) and the Institut de Formation Supérieure en Informatique et Communication (IFSIC, Univ. Rennes 1).

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.

Guillaume Gravier has given 2 lectures (4 h) at the Ermites 2007 summer school (Ecole Recherche Multimodale d’informations) on automatic speech recognition and on multimodal information fusion.

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).

The project-team prepared demonstrations for the 40th anniversary of INRIA, (Lille, 10-11 December 2007) under the technical coordination of Gilles Gonon.

10. Bibliography

Major publications by the team in recent years

[1] S. ARBERET, R. GRIBONVAL, F. BIMBOT. *A Robust Method to Count and Locate Audio Sources in a Stereophonic Linear Instantaneous Mixture*, in "Proc. of the Int'l. Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2006), Charleston, South Carolina, USA", J. ROSCA, D. ERDOGMUS, J. PRÍNCIPE, S. HAYKIN (editors), LNCS, vol. 3889, Springer, March 2006, p. 536–543.

[2] M. BEN. *Approches robustes pour la vérification automatique du locuteur par normalisation et adaptation hiérarchique*, Thèse de doctorat, Université de Rennes 1, IRISA, Rennes (France), November 2004.

[3] L. BENAROYA, F. BIMBOT, R. GRIBONVAL. *Audio Source Separation With a Single Sensor*, in "IEEE Trans. Audio, Speech and Language Processing", vol. 14, n^{o} 1, January 2006, p. 191–199.

[4] F. BIMBOT, J.-F. BONASTRE, C. FREDOUILLE, G. GRAVIER, I. MAGRIN-CHAGNOLLEAU, S. MEIGNIER, T. MERLIN, J. ORTEGA-GARCIA, D. A. REYNOLDS. *A tutorial on text-independent speaker verification*, in "EURASIP Journal on Applied Signal Processing", vol. 2004, n^{o} 4, April 2004, p. 430–451.

[5] F. BIMBOT, G. GRAVIER. *Evaluation des systèmes de reconnaissance de la parole*, in "Evaluation des systèmes de traitement de l'information", Traité des Sciences et Techniques de l'Information, chap. 8, Hermes Science Publications, 2004, p. 189–213.

[6] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Bi-framelet systems with few vanishing moments characterize Besov spaces*, in "Appl. Comp. Harmonic Anal. (special issue on frames in harmonic analysis)", vol. 17, n^{o} 1–2, 2004.

[7] 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, p. 1149–1152.

[8] 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.

[9] 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.

[10] R. GRIBONVAL, M. NIELSEN. *Nonlinear approximation with dictionaries. I. Direct estimates*, in "J. of Fourier Anal. and Appl.", vol. 10, n^{o} 1, 2004.

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

[12] 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.

[13] S. HUET, G. GRAVIER, P. SÉBILLOT. *Are morpho-syntaxic taggers suitable to improve automatic transcription*, in "Intl. Workshop on Text, Speech and Dialogue", 2006.

[14] 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.

[15] G. MONACI, P. JOST, P. VANDERGHEYNST, B. MAILHÉ, S. LESAGE, R. GRIBONVAL. *Learning Multi-Modal Dictionaries: Application to Audiovisual Data*, in "Proc. of International Workshop on Multimedia Content Representation, Classification and Security (MCRCS'06)", LNCS, vol. 4105, Springer-Verlag, September 2006, p. 538–545.

[16] A. OZEROV. *Adaptation de modèles statistiques pour la séparation de sources mono-capteur. Application à la séparation voix / musique dans les chansons.*, Ph. D. Thesis, Université de Rennes I, December 2006.

[17] C. G. M. SNOEK, M. WORRING. *Time Interval Maximum Entropy based Event Indexing in Soccer Video*, in "Proc. IEEE International Conference on Multimedia & Expo", July 2003, p. 481–484.

[18] 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.

Year Publications

Books and Monographs

[19] M. DELAKIS, G. GRAVIER, P. GROS. *Stochastic models for multimodal video analysis*, to be published, Springer Verlag, 2007.

[20] S. HUET, G. GRAVIER, P. SÉBILLOT. *Toward the integration of NLP and ASR techniques: POS tagging and transcription*, to be published, Springer Verlag, 2007.

Doctoral dissertations and Habilitation theses

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

[22] S. LESAGE. *Apprentissage de dictionnaires structurés pour la modélisation parcimonieuse des signaux multicanaux*, Ph. D. Thesis, Université de Rennes I, avril 2007.

Articles in refereed journals and book chapters

[23] F. BIMBOT. *Description des documents sonores*, in "L'indexation multimédia : description et recherche automatique", P. GROS (editor), Traité IC2, chap. 5, Hermès, 2007, p. 137–161.

[24] M. DAVIES, M. JAFARI, S. ABDALLAH, E. VINCENT, M. PLUMBLEY. *3*, in "Blind source separation using space-time independent component analysis", S. MAKINO, T.-W. LEE, H. SAWADA (editors), Springer, 2007.

[25] M. DELAKIS, G. GRAVIER, P. GROS. *Audiovisual Integration with Segment Models for Tennis Video Parsing*, in "Computer Vision and Image Understanding", accepted for publication, 2007.

[26] G. GONON, F. BIMBOT. *De la reconnaissance automatique du locuteur à la signature vocale*, in "Interstices", 2007, http://interstices.info.

[27] G. GRAVIER, J.-F. BONASTRE, S. GALLIANO, E. GEOFFROIS, D. MOSTEFA, K. CHOUKRI. *Évaluation des systèmes de transcription enrichie d'émissions radiophoniques*, in "Les campagnes d'évaluation EVALDA", S. CHAUDIRON (editor), (à paraître), Hermès Science, 2007.

[28] G. GRAVIER. *Description multimodale multimedia*, in "L'indexation multimédia : description et recherche automatique", P. GROS (editor), Traité IC2, chap. 7, Hermès, 2007, p. 191–214.

[29] M. JAFARI, E. VINCENT, S. ABDALLAH, M. PLUMBLEY, M. DAVIES. *An adaptive stereo basis method for convolutive blind audio source separation*, in "Neurocomputing", to appear, 2007.

[30] S. KRSTULOVIC, F. BIMBOT, O. BOËFFARD, D. CHARLET, D. FOHR, O. MELLA. *Selecting Representative Speakers for a Speech Database on the Basis of Heterogeneous Similarity Criteria*, C. MÜLLER (editor), Springer, Berlin / Heidelberg, 2007, p. 276–292.

[31] P. LEVEAU, E. VINCENT, G. RICHARD, L. DAUDET. *Instrument-specific harmonic atoms for mid-level music representation*, in "IEEE Trans. on Audio, Speech and Language Processing", to appear, 2007.

[32] G. MONACI, P. JOST, P. VANDERGHEYNST, B. MAILHÉ, S. LESAGE, R. GRIBONVAL. *Learning Multi-Modal Dictionaries*, in "IEEE Trans. Image Processing", vol. 16, n^{o} 9, septembre 2007, p. 2272-2283.

[33] 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.

[34] A. OZEROV, P. PHILIPPE, R. GRIBONVAL, F. BIMBOT. *Choix et adaptation de modèles statistiques pour la séparation de voic chantée à partir d'un seul microphone*, in "Revue Française de Traitement du Signal", vol. 24, n^{o} 3, 2007.

[35] A. OZEROV, P. PHILIPPE, R. GRIBONVAL, F. BIMBOT. *Choix et adaptation de modèles statistiques pour la séparation de voic chantée à partir d'un seul microphone*, in "Revue Française de Traitement du Signal", vol. 24, n^{o} 3, 2007.

[36] A. ROSENBERG, F. BIMBOT, S. PARTHASARATHY. *36*, in "Overview of Speaker Recognition", J. BENESTY, M. M. SONDHI, Y. HUANG (editors), Springer, 2007, p. 725–741.

[37] E. VINCENT, R. GRIBONVAL, M. D. PLUMBLEY. *Oracle Estimators for the Benchmarking of Source Separation Algorithms*, in "Signal Processing", vol. 87, n^{o} 8, August 2007, p. 1933–1950.

[38] E. VINCENT, R. GRIBONVAL, M. PLUMBLEY. *Oracle estimators for the benchmarking of source separation algorithms*, in "Signal Processing", vol. 87, n^{o} 8, 2007, p. 1933–1950.

[39] 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.

Publications in Conferences and Workshops

[40] S. ARBERET, R. GRIBONVAL, F. BIMBOT. *A Robust Method to Count and Locate Audio Sources in a Stereophonic Linear Anechoic Mixture*, in "Proc. IEEE Intl. Conf. Acoust. Speech Signal Process (ICASSP'07)", avril 2007.

[41] D. CHARLET, M. COLLET, F. BIMBOT. *VZ-Norm : an extension of Z-norm to the Multivariate Case for Anchor Model based Speaker Verification*, in "European Conf. on Speech Communication and Technology – Interspeech", 2007.

[42] G. GRAVIER, D. MORARU. *Towards phonetically-driven hidden Markov models: Can we incorporate phonetic landmarks in HMM-based ASR?*, in "Proc. ISCA Tutorial and Research Workshop on Non Linear Speech Processing", M. CHETOUANI, ET AL (editors), Lecture Notes in Artificial Intelligence, vol. 4885, Springer Verlag, 2007, p. 161–168.

[43] R. GRIBONVAL, B. MAILHÉ, H. RAUHUT, K. SCHNASS, P. VANDERGHEYNST. *Average Case Analysis of Multichannel Thresholding*, in "Proc. IEEE Intl. Conf. Acoust. Speech Signal Process (ICASSP'07)", avril 2007.

[44] R. GRIBONVAL, B. MAILHÉ, H. RAUHUT, K. SCHNASS, P. VANDERGHEYNST. *Multichannel thresholding with sensing dictionaries*, in "Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP'07)", 2007.

[45] S. HUET, G. GRAVIER, P. SÉBILLOT. *Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation*, in "European Conf. on Speech Communication and Technology – Interspeech", 2007.

[46] D. MORARU, G. GRAVIER. *Landmark Based Large Vocabulary Continuous Speech Recognition*, in "Proc. Conf. on Speech Technology and Human-Computer Dialogue", 2007.

[47] K. SCHNASS, P. VANDERGHEYNST, R. GRIBONVAL, H. RAUHUT. *Average case analysis of multichannel sparse approximations using p-thresholding*, in "SPIE Optics and Photonics, Wavelet XII, San Diego", 2007.

[48] R. TAVENARD, L. AMSALEG, G. GRAVIER. *Estimation de similarité entre séquences de descripteurs à l'aide de machines à vecteurs supports*, in "Proc. Conf. Base de Données Avancées", 2007.

[49] R. TAVENARD, L. AMSALEG, G. GRAVIER. *Machines à vecteurs supports pour la comparaison de séquences de descripteurs*, in "Proc. Journées d'étude et d'échange COmpression et REprésentation des Signaux Audiovisuels", 2007.

[50] W. TENG, G. GRAVIER, F. BIMBOT, F. SOUFFLET. *Rapid Speaker Adaptation by Reference Model Interpolation*, in "European Conf. on Speech Communication and Technology – Interspeech", 2007.

[51] E. VINCENT, N. BERTIN, R. BADEAU. *Two non-negative matrix factorization methods for polyphonic pitch transcription*, in "Proc. Music Information Retrieval Evaluation eXchange (MIREX)", 2007.

[52] E. VINCENT, R. GRIBONVAL. *Blind criterion and oracle bound for instantaneous audio source separation using adaptive time-frequency representations*, in "Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)", 2007.

[53] E. VINCENT, R. GRIBONVAL. *Blind criterion and oracle bound for instantaneous audio source separation using adaptive time-frequency representations*, in "Proc. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)", IEEE, oct 2007.

[54] E. VINCENT, H. SAWADA, P. BOFILL, S. MAKINO, J. ROSCA. *First stereo audio source separation evaluation campaign: data, algorithms and results*, in "Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)", 2007, p. 552–559.

[55] E. VINCENT. *Complex nonconvex lp norm minimization for underdetermined source separation*, in "Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)", 2007, p. 430–437.

[56] S. WELBURN, M. PLUMBLEY, E. VINCENT. *Object-coding for resolution-free musical audio*, in "Proc. AES Int. Conf. on new directions in high resolution audio", 2007.

Internal Reports

[57] V. ALLIE-CROCITTI, M. BEN, F. BIMBOT, S. BUSSON, J. CHICO, D. DEGRAEVE, W. D. NEVE, G. GONON, G. NUYTTENS, P. SCHMOUKER, C. SERRE, D. V. DEURSEN. *WP3 Status Report on Indexation and Content Processing Activities*, Technical report, 2007.

[58] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Beyond coherence : recovering structured time-frequency representations*, Technical report, n^{o} 1833, IRISA, feb 2007.

[59] R. GRIBONVAL, H. RAUHUT, K. SCHNASS, P. VANDERGHEYNST. *Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms*, Preprint, n^{o} PI 1848, IRISA, mai 2007.

[60] E. VINCENT, M. PLUMBLEY. *Efficient Bayesian inference for harmonic models via adaptive posterior factorization*, Technical report, n^{o} PI 1841, 2007.

References in notes

[61] J. BOBIN, Y. MOUDDEN, J.-L. STARCK, M. ELAD. *Morphological Diversity and Source Separation*, in "IEEE Signal Processing Letters", n^{o} 7, 2006, p. 409–412.

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

[63] J.-F. BONASTRE, F. BIMBOT, L.-J. BOË, J. CAMPBELL, D. REYNOLDS, I. MAGRIN-CHAGNOLLEAU. *Person Authentication by Voice : A Need For Caution*, in "Proc. Eurospeech'03, Genève", 2003.

[64] G. F. COOPER, E. HERSKOVITS. *A Bayesian method for the induction of probabilistic networks from data*, in "machine Learning", 1992.

[65] 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.

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

[67] P. JOST, P. VANDERGHEYNST, S. LESAGE, R. GRIBONVAL. *Learning redundant dictionaries with translation invariance property : the MoTIF algorithm*, in "SPARS, Rennes", 2005.

[68] S. LESAGE, S. KRSTULOVIC, R. GRIBONVAL. *Séparation de sources dans le cas sous-déterminé : comparaison de deux approches basées sur des décompositions parcimonieuses*, in "Proc. GRETSI", 2005.

[69] Z. LUO, M. GASPAR, J. LIU, A. SWAMI. *Distributed signal processing in sensor networks*, in "IEEE Signal processing magazine", vol. 23, n^{o} 4, July 2006, p. 14-15.

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

[71] A. OZEROV, R. GRIBONVAL, P. PHILIPPE, F. BIMBOT. *Séparation voix / musique à partir d'enregistrements mono : quelques remarques sur le choix et l'adaptation des modèles*, in "Proc. GRETSI", 2005.

[72] A. OZEROV, P. PHILIPPE, R. GRIBONVAL, F. BIMBOT. *One microphone singing voice separation using source-adapted model*, in "Proc. WASPAA", 2005.

[73] 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.

[74] E. VINCENT, R. GRIBONVAL. *Construction d'estimateurs oracles pour la séparation de sources*, in "Proc. GRETSI", 2005.