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

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
Marie-Noëlle Georgeault [until May 2006]
Stéphanie Lemaile [since June 2006]

Research scientist (CNRS)
Guillaume Gravier [CR1 CNRS]

Research scientist (INRIA)
Rémi Gribonval [CR1 INRIA]
Emmanuel Vincent [CR2 INRIA - Since November 2006]

Project Technical Staff
Mathieu Ben [Contractual Research Engineer - Since August 2006]
Gilles Gonon [Contractual Research Engineer]
Sacha Krstulovic [Contractual Research Engineer - until May 2006]
Benjamin Roy [Contractual Development Engineer - since September 2006]

Post-Doc
Daniel Moraru [until February 2006]

Ph.D. students
Mikaël Collet [FTR&D-Lannion Funding - Terminated September 2006]
Sylvain Lesage [MENRT Grant, 3rd year]
Alexey Ozerov [FTR&D-Rennes Funding - Terminated December 2006]
Amadou Sall [Regional Grant, 3rd year]
Stéphane Huet [MENRT Grant, 2nd year - also with TEXMEX]
Simon Arberet [CNRS & Region Grant, 1st year]
Boris Mailhé [ENS Cachan (Bruz), 1st year]
Ewen Camberlein [FTR&D-Rennes Funding, 3rd year]
Wen Xuan Teng [Telisma Funding, 2nd year]

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 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.

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 measuring our progress within the framework of evaluation campaigns, to disseminate software resources which we develop and to share our efforts with other partner laboratories.

On a regular basis, METISS is involved in bilateral or multilateral partnerships, within the framework of consortia (ELISA), networks (HASSIP), thematic groups (MathSTIC), national research projects (Technolangues) European projects (INSPIRED) and industrial contracts with various companies (Thomson Multi-Media, France Télécom R&D, Telisma, ...).

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 [50], 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).

In practice, however, the use of the theoretical tools 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 the adaptive representations of signals in redundant systems [71]. 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: *EM algorithm*, *Hidden Markov Model*, *Viterbi algorithm*, *acoustic parameterisation*, *beam search*, *classification*, *gaussian mixture model*, *gaussian model*, *hypotheses testing*, *maximum a posteriori*, *maximum likelihood*, *probability density function*.

For more than a decade, 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 $\hat{P}$ 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 processing (and audio signal) 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 : a mean value and a variance for a GM, p means, variances and weights for a GMM with p Gaussian, q states, q^2 transition probabilities and $p \times q$, means, variances and weights for an HMM with q states the PDF of which being GMMs with p Gaussians. Λ_X will denote the vector of parameters for class X, and in this case, the following notation will be used :

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

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 (number of Gaussian p, number of states q, etc.), 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 :

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

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 $\hat{X}$ and $\hat{\overline{X}}$ 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 :

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

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 :

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

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

3.3. Adaptive 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:

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

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 :

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

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

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

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

Keywords: *normalisation*, *representation and adaptation*, *scalability*, *speaker elicitation*, *speaker recognition*, *user authentication*, *voice signature*.

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 [51], 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 with two intentions : speaker indexation in the context of information retrieval and speaker selection in the context of 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.

4.2.1. Speaker model and test normalisation

Participants: Mathieu Ben, Frédéric Bimbot, Guillaume Gravier.

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.

4.2.2. Scalability and complexity reduction

Participants: Gilles Gonon, Frédéric Bimbot, Rémi Gribonval.

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.

4.2.3. Speaker representation, selection and adaptation

Participants: Mikaël Collet, Sacha Krstulovic, Wen Xuan Teng, Frédéric Bimbot.

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) and speaker adaptation for speech recognition.

4.3. Modeling, detecting and structuring information in audio streams

Automatic tools to locate events in audio documents, structure them and browse through them as in textual documents are key issues in order to fully exploit most of the available audio documents (radio and television programmes and broadcasts, conference recordings, etc). In this respect, defining and extracting meaningful characteristics from an audio stream aim at obtaining a more or less 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. In particular, speaker detection, tracking, clustering as well as speaker change detection are studied. We also maintain some background activities in speech recognition.

4.3.1. Speaker detection

Keywords: *audio stream*, *detection*, *segmentation*, *speaker recognition*, *tracking*.

Participants: Frédéric Bimbot, Guillaume Gravier, Mikaël Collet, Daniel Moraru.

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

Keywords: *audio indexing*, *audio stream*, *detection*, *segmentation*, *tracking*.

Participants: Guillaume Gravier, Daniel Moraru, Frédéric Bimbot, Gilles Gonon, Mathieu Ben.

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. Indexing multi-modal information

Keywords: *audio stream*, *audiovisual integration*, *information fusion*, *multimedia indexing*, *multimodality*.

Participant: Guillaume Gravier.

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. Speech modeling and recognition

Keywords: *beam-search*, *broadcast news indexing*, *rich transcription*, *speech modeling*, *speech recognition*, *spoken document*.

Participants: Guillaume Gravier, Stéphane Huet, Daniel Moraru.

Many audio documents contain speech from which useful information concerning the document content and semantics can be extracted. However, extracting information from speech is a key application domain which requires specific processing such as speech recognition or word spotting. METISS maintains a know-how and develops some 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.

Moreover, speech modeling and recognition is complementary with other speech related activities in the group, such as audio segmentation, speaker recognition and transaction security. 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.

4.3.5. Music modeling

Keywords: *audio-object extraction*, *harmony*, *melody*, *music language modeling*.

Participants: Amadou Sall, Frédéric Bimbot.

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 two 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.

It is expected that progress in these two areas will yield, at the medium term, a music description and recognition scheme that allows to take into account both the acoustic variability and the syntagmatic constraints that exist in music pieces, borrowing ideas, models and algorithms from the field of speech recognition.

4.4. Advanced audio signal processing

Keywords: *audio events*, *indexing*, *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

Participants: Rémi Gribonval, Simon Arberet, Sylvain Lesage, Sacha Krstulovic, Boris Mailhé, Alexey Ozerov, Frédéric Bimbot.

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 and decomposition

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

The norms 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.

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*.

Participants: Guillaume Gravier, Daniel Moraru.

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: *Viterbi*, *beam-search*, *best path*, *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 [60]. 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, Sacha Krstulovic, 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.

METISS efforts this year have been targeted at adding new atom classes and improving the robustness and portability of the code. Newly added dictionaries include Chirps atoms, Anywave/Nyquist/Constant atoms, and MDCT/MDST/MCLT atoms, the latter being contributed by Emmanuel Ravelli from the Laboratoire d'Acoustique Musicale (University of Paris VII, Jussieu). A description of the various dictionaries and atoms implemented in MPTK can be found in its documentation at [65], [63], [64], [61]. A description of the algorithmic speed up can be found in [40].

An INRIA software development operation (Opération de Développement Logiciel, ODL) started in September 2006 with the aim of optimizing MPTK to ease its distribution by improving its portability to different platforms and simplifying its developpers' API. The main change at the moment is the replacement of automake/autoconf by Cmake for the build system, which will enable easier portability notably on Windows systems, and automated tests with DART. A multithread version is under development to exploit parallel (multicore) computer architectures.

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

Contact : remi.gribonval@irisa.fr

Relevant links : http://mptk.gforge.inria.fr.

5.4. BSS_EVAL: A toolbox for performance measurement in (blind) source separation

Participants: Rémi Gribonval, Emmanuel Vincent.

BSS_EVAL is a MATLAB toolbox to compute performance measures in (blind) source separation within an evaluation framework where the original sources are available as ground truth. BSS_EVAL has been developped in collaboration with C. Févotte and E. Vincent with the support of the French GdR-ISIS/CNRS Workgroup "Resources for Audio Source Separation".

The measures implemented in BSS_EVAL are based on a decomposition of each estimated source into four contributions corresponding to the target source, interferences of unwanted sources, remaining additive noise and artifacts such as "musical noise". They are valid for all usual types of signals, such as real-valued audio or biomedical signals or complex-valued subbands of these signals. A more detailed description of the BSS_EVAL methodology as well as a reference manual can be found in [28], [57] and [62].

Contact : remi.gribonval@irisa.fr

Relevant links : http://bass-db.gforge.inria.fr/bss_eval/.

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

Participants: Rémi Gribonval, Emmanuel Vincent.

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. An extended version of BSS_ORACLE has been developped this year in collaboration with E. Vincent and M.D. Plumbley. The toolbox provides oracle estimators for four classes of algorithms (time-invariant beamforming, 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). These estimators are defined in [77] and [48]. A reference manual can be found online at http://bass-db.gforge.inria.fr/bss_oracle/.

Other relevant references include [28], [57] and [62].

Contact : remi.gribonval@irisa.fr

Relevant links : http://bass-db.gforge.inria.fr/bss_oracle/, http://bass-db.gforge.inria.fr/bss_eval/.

6. New Results

6.1. Speaker characterisation

Keywords: *Anchor Models*, *Classification and Regression Trees (CART).* , *Gaussian Mixture Models (GMM)*, *mormalisation*, *speaker characterisation*, *speaker selection*, *speaker verification*.

6.1.1. Relative speaker information and related metrics

Participants: Mikaël Collet, Frédéric Bimbot.

The representation of speaker information relatively to a set of other speaker models (anchor models) yields a compact representation of the speaker information. This representation can be advantageous for speaker segmentation, indexing, tracking and adaptation.

In this framework, the speaker-related properties of a speech segment can be represented as a vector of likelihood ratio values (SCV) corresponding to the speech observations being scored by a pre-determined collection of reference (anchor) speaker models.

In previous work, several deterministic metrics (euclidean, angular and correlation) were investigated and evaluated for the comparison of speakers in the anchor space [76], [72], [52]. More recently, a probabilistic approach based on a speaker-dependent Gaussian modeling of the SCV was proposed [53] and yielded considerable improvement of the anchor speaker approach, making it competitive with respect to conventional GMMs [34], [32], [13].

This work was done in close collaboration with FTR&D-Lannion (Delphine Charlet).

6.1.2. Optimizing the speaker coverage of a speech database

Participants: Sacha Krstulovic, Mathieu Ben, Frédéric Bimbot.

The state of the art techniques in the various domains of Automatic Speech Processing (be it for Automatic Speaker Recognition, Automatic Speech Recognition or Text-To-Speech Synthesis) make extensive use of speech databases. Nevertheless, the problem of optimizing the contents of these databases to make them adequate to the development of a considered speech processing task has seldom been studied [73].

In this context, we have proposed a general database design method aiming at optimizing the contents of new speech databases by focusing the data collection on a selection of speakers chosen for its good coverage of the voice space. Such databases would be better adapted to the development of recent speech processing methods, such as those based on multi-models (e.g adaptation of speech recognition with specialized models, speaker recognition with anchor models, speech synthsis by unit selection, etc.). Such developments require indeed a much larger quantity of data per speaker than the traditional databases can offer [56]. Nevertheless, the increase in the collection cost for such newer and larger databases should be limited as much as possible, while preserving a good coverage of the speaker variability.

The corresponding work, led in the framework of the NEOLOGOS project[1], therefore re-thinks the design of speech databases in the following terms: it focuses on optimizing the contents of the speech databases in order to guarantee the diversity of the recorded voices, both at the segmental and supra-segmental levels, so that each of the recorded speakers can be precisely modeled and localized in an abstract space of speakers. In addition to this scientific objective, this method addresses the practical concern of reducing the collection costs for new speech databases.

[1] The following public, academic and industrial partners have participated in the NEOLOGOS project, funded by the French Ministry of Research in the framework of the TECHNOLANGUES program: ELDA, France Telecom R&D company/lab, IRISA-ENSSAT (Cordial), IRISA (Metiss), LORIA and TELISMA.

The resulting methodology proposes to operate a selection by optimizing a quality criterion defined in a variety of speaker similarity modeling frameworks. The selection can be operated and validated with respect to a unique similarity criterion, using classical clustering methods such as hierarchical or k-median clustering, or it can be operated and validated across several speaker similarity criteria, thanks to a newly developed clustering method that we called Focal Speakers selection [68]. In this framework, four different speaker similarity criteria have been tested, and three different speaker clustering algorithms have been compared. The outcome of this work has been used for the final specification of the list of speakers to be recorded in the NEOLOGOS database.

A manuscript detailing the methodology and the results of this speaker-driven database design was accepted for publication in an international journal [25].

6.1.3. Improved CART trees for fast speaker verification

Participants: Gilles Gonon, Rémi Gribonval, Frédéric Bimbot.

The main motivation for using decision trees in the context of speaker recognition comes from the fact that they can be directly applied in real-time implementations on a PC or a mobile device. Also, they are particularly suitable for embedded devices as they work without resorting to a log/exp calculus.

We address the problem of using decision trees in the rather general context of estimating the Log Likelihood Ratio (LLR) used in speaker verification, from two GMM models (speaker model and "background" model). Former attempts at using trees performed quite poorly compared to state of the art results with Gaussian Mixture Models (GMM). Two new solutions have been studied to improve the efficiency of the tree-based approach :

The first one is the introduction of a priori informations on the GMM used in state of the art techniques at the tree construction level. Taking into account the training method of the models with EM algorithms and maximum a posteriori techniques, it is possible to implicitly choose locally optimal hyperplane splits for some nodes of the trees. This is equivalent to building oblique trees using a specific set of oblique directions determined by the baseline GMM and thus limiting the complexity of the training phase.

The second one is the use of different complexity score functions within each leaf of the trees. These functions are computed after the creation of the trees, drawing data into the tree leaves and computing a regression function over the LLR scores. Mean score functions, linear score functions as well as quadratic score functions have been successfully tested resulting in more accurate trees.

These improvements applied to the classical classification and regression trees (CART) method in a speaker verification system allow to reduce more than 10 times the complexity of the LLR function computation. Considering a baseline state of the art system with an equal error rate (EER) of 11.6% on the NIST 2005 evaluation, a previous CART method provided typical EER ranging between 19% and 22% while the proposed improvements decrease the EER to 13.7 % [59].

This work was carried out in the framework of a feasibility study concerning security requirements for a "Trusted Personal Device" within the Inspired IST Project [58] and has been submitted for publication in an international journal.

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

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

6.2.1. Automatic speech recognition with broad phonetic landmarks

Keywords: *ASR*, *dynamic programming*, *landmarks*, *phonetic landmarks*, *speech recognition*.

Participants: Guillaume Gravier, Daniel Moraru.

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. We proposed a framework to incoporate such knowledge as constraints on the best path in a Viterbi based decoder. As HMM-based speech recognition can be viewed as the search for the best path in a treillis, knowledge of the broad phonetic content of the signal can be used to prune (or at least penalize) those paths which are inconsistent with the available prior knowledge. We refer to those places where prior information on the phonetic content of the signal is available as *l*andmarks. From a theoretical point of view, this can be seen as decoding with non stationary Markov models where the transition probabilities vary with time depending on the presence or not of landmarks. In practice, a confidence measure associated to automatically detected landmarks can be used to penalize transitions according to the confidence in the landmark, the lower the confidence, the lower the penalty.

We carried out a preliminary study to determine whether a local knowledge of the broad phonetic content of the signal can benefit the transcription system or not. The aim of the study is twofold: (i) validate the proposed approach to integrate landmarks in a statistical decoder and (ii) measure the benefit of broad phonetic landmarks. Broad phonetic landmarks of different types (vowels, fricatives, glides, etc.) were extracted from a reference phonetic alignment of the waveform and used as landmarks. Experimental results on broadcast news transcription show that each type of landmarks brings a small improvement to the system. Using all the landmarks simultaneously significantly yielded a considerable improvement of the transcription system along with a faster decoding, even when landmarks actually covers a small portion of the actual phone. This last result indicate that precisely detecting the landmark boundaries is not required. Finally, simulated miss detection errors showed that the performance gain scales linearly with the amount of detected landmarks [45].

These encouraging results triggered future work on the automatic detection of broad phonetic landmarks in order to validate these ideas in a realistic application framework. Application of this decision fusion scheme to audiovisual speech recognition is also foreseen, where the visual modality can be used to provide knowledge on the phonetic content for the audio modality.

6.2.2. Speech transcription with part-of-speech tagging

Keywords: *ASR*, *NLP*, *linguistics*, *natural language processing*, *part-of-speech*, *speech recognition*.

Participants: Stéphane Huet, Guillaume Gravier.

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. The thesis of Stéphane Huet, in collaboration with the Tex-Mex project, aims at using NLP techniques to improve the transcription of spoken documents.

In 2006, the work of Stéphane Huet has focused on incorporating linguistic knowledge for the rescoring of N-best sentence hypothesis lists. After a survey on the use of linguistic knowledge in ASR systems [47], we investigated N-best lists rescoring using part-of-speech (PoS) information. We first demonstrated that N-class based PoS taggers are robust to the specificities of spoken document transcriptions (lack of punctuation, no case in our ASR output, breath group instead of sentences, ASR errors). In particular, PoS taggers are robust to transcription errors, since they mostly rely on local decisions and many words are non ambiguous. We then showed that PoS information can be used to detect and correct transcription errors [38]. Finally, these two results enable the use of PoS taggers to rescore a list of sentence hypotheses based on a score combining acoustic, language and syntactic (PoS) information. The combined score was used in conjunction with several rescoring schemes, namely maximum a posteriori, minimum expected word error rate and consensus decoding, to rerank lists of 100 sentence hypotheses with a decrease of about 1 % of the word error rate in all cases [37]. Moreover, the resulting transcription exhibits, in most cases, a better grammatical structure, which is reflected by a decrease of the sentence error rate.

The corresponding algorithms were implemented in our Sirocco software and incorporated in our spoken document analysis platform IRENE.

Future work on this topic include exploiting the enhanced transcription along with PoS tags to segment the text into topically coherent stories caracterized by some automatically extracted keywords (which can in turn be used to adapt the vocabulary and the language model). The use of syntactic information for confidence measures is also a foreseen continuation of this work.

6.2.3. Multimodal segment models for video analysis

Keywords: *HMM*, *hidden Markov models*, *multimodal fusion*, *multimodality*, *segment models*, *video analysis*.

Participant: Guillaume Gravier.

This section and the next one describe a joint work with the Tex-Mex project, carried out in the framework of the Ph. D. thesis of E. Delakis [14] under the supervision of Guillaume Gravier and Patrick Gros (Tex-Mex).

In a previous work [24], we investigated the use of hidden Markov models (HMM) for the integration of audio and visual cues, applied to tennis video structuring. This work clearly demonstrated the potential of an HMM approach but also outlined the two main limitations of HMMs for such a task: the synchronization at the shot level of the descriptors of each media and the same underlying model for both modalities.

Motivated by this need for more efficient multimodal representations, the use of segmental features in the framework of Segment Models (SM) was previously proposed [55], [54], instead of the frame-based features of Hidden Markov Models. Considering each scene of the video as a segment, the synchronization points between different modalities are extended to the scene boundaries and a scene duration model is added. Conditionnally to a scene, the sequences of visual and auditory features are considered independent and different models can be used. This year, we studied various models for the auditory feature sequences, including a model based on cepstral coefficient to avoid the error-prone step which consists in tracking events like "ball hits" and "applause" in the soundtrack. Segment models yielded better performance than HMMs, mainly due to the scene duration model. Asynchronous audio-visual fusion at the scene level yielded no improvement compared to a synchronous fusion with SMs. This result is most probably due to the fact that strong correlations between visual and audio features at the scene level are disregarded in the asynchronous fusion scheme. Combining asynchronous and synchronous fusion resulted in a small performance gain [14].

Finally, we also explored the idea of using a hybrid SM-ANN approach using Recurrent Neural Networks as a segmental scorer. To this end, the newly introduced Long-Short Term Memeory (LSTM) topology was favorably compared to BPTT-trained RNNs and used in the hybrid model. The hybrid performed however visibly inferior to the standard SM but still with a promising performance. In fact, what makes the difference is that the HMM-based segmental scorers can use prior knowledge on the task directly into their topology, while the LSTM scorers were built from scratch [14].

These results illustrate the increased flexibility of SMs with respect to HMMs. However, the hypothesis of independence of the information streams is a clear limitation of the SM approach for multimodal integration. Exploiting he dynamic Bayesian framework to overcome this limitation and to relax the synchronization constraints are part of a new work following the Ph. D. of Emmanouil Delakis.

6.2.4. Score-oriented Viterbi search for sports audio and video analysis

Keywords: *HMM*, *hidden Markov models*, *multimodal fusion*, *multimodality*, *segment models*, *video analysis*.

Participant: Guillaume Gravier.

In sport videos, score annoucement are often displayed on screen, giving some valuable information on the high-level semantic of the video. However, current models can hardly accomodate such an information stream as they are highly synchronous and sparse (*i.e.* not always displayed). In tennis videos, using the presence of a displayed score as an extra feature, at the shot level in hidden Markov models, or at the scene level in segment models, result in a marginal performance improvement. The reason is that score announcements may appear a lot delayed or may not appear at all after a game event. The probability distributions of this feature become thus almost uniform, *i.e.*, they carry no useful information.

Instead, we studied a new decoding algorithm, the score-oriented Viterbi search, in order to fully exploit the semantic content of the score announcements. This algorithm aims at finding out the most likely path consistent with the score annoucements, at the expense of a computation overhead a little superior to the standard Viterbi decoding for both HMMs and SMs. The key idea is to perform a cascade of local optimization, penalizing local paths inconsistent with the number of points scored between two score announcements. Experimental results on our tennis video corpus demonstrated a significant performance improvement with both HMMs and SMs [35].

The scope of the proposed algorithm is not limited to tennis and extends to any constraints that can be formulated as "there are n events of a given kind between two instants a and b". In the particular case of tennis, the events are the number of points scored while the instants are the consecutive instants at which a score is displayed.

6.2.5. Statistical models of music

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

Participants: Amadou Sall, Frédéric Bimbot.

With analogy to speech recognition, which 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 are investigating 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. Preliminary results have shown promising results that are currently being consolidated. Bayesian networks will be investigated.

At the longer term, 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.

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 [69] 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 [49]. 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 [29], 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 [75], [74].

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 [75], [74], 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 [27] and is thoroughly detailed in Alexey Ozerov's Ph.D. manuscript [15]. It was done in close collaboration with FTR&D (Pierrick Philippe).

6.3.4. 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 [77], was done in collaboration with Emmanuel Vincent and Mark D. Plumbley (Queen Mary, University of London). It is currently published as a preprint [48] and submitted for journal publication.

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 [43] 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 [70]. 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 Tex-Mex 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 (joint supervision of the thesis of Manolis Delakis, Texmex) 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: Gilles Gonon, Mathieu Ben, Guillaume Gravier, Frédéric Bimbot.

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

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

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

The contribution of IRISA is 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. HASSIP Research Training Network

Participants: Rémi Gribonval, Sylvain Lesage, Boris Mailhé.

The HASSIP (Harmonic Analysis, Statistics in Signal and Image Processing) Research Training Network is a European network funded by the European Commission within the framework programme *Improving the Human Potential*. It started on October 1st 2002, with founding partners: Université de Provence/CNRS, University of Vienna, Cambridge University, Université Catholique de Louvain, EPFL, University of Bremen, University of Munich and Technion Institute.

One of the aims of the HASSIP network is to shorten the development cycle for new algorithms by bringing together those who are involved in this process: the mathematicians and physicists working on the foundations (with view towards applications), the partners doing applied research (mostly engineering departments), are more experienced when it comes to implementations. The main research goal is therefore to improve the link between the foundations and real word applications, by developing new nonstandard algorithms, by studying their behaviour on concrete tasks, and to look for innovative ways to circumvent shortcomings or satisfy additional request arising from the applications.

The main contributions of the METISS project-team at IRISA consisted in new statistical models of audio signals for coding and source separation, theoretical contributions on time-frequency/time-scale analysis and (highly) nonlinear approximation with redundant dictionaries, as well as the Matching Pursuit ToolKit 5.3.

The HASSIP network final meeting took place in september 2006.

8.1.2. PAI Germaine de Stael with EPFL

Participants: Rémi Gribonval, Sylvain Lesage, Boris Mailhé.

A bilateral collaboration with the Signal Processing group (LTS2) led by Pierre Vandergheynst at EPFL (Switzerland) was initiated within the HASSIP European 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. The collaboration resulted so far in joint theoretical contributions on sparse signal approximation, as well as on multimodal audiovisual signal analysis. Since the fall of 2005, a co-supervised Ph.D. thesis (Boris Mailhé) has started to reinforce even more the collaboration. A proposal to build an INRIA Associated Team has been submitted to strengthen and build upon this collaboration in the coming years, one of the most interesting aspects being the complementary competences in audio (METISS) and image/video (LTS2) applications of sparse signal models.

8.2. Visites, et invitations de chercheurs

Participant: Rémi Gribonval.

Within the joint framework of the HASSIP network and the PAI Germaine de Staël, Rémi Gribonval visited the Signal Processing group (LTS2) lead by Pierre Vandergheynst in the summer of 2006 (August and September). As a result of this visit, new theoretical contributions on simultaneous sparse signal approximation were obtained (they are currently submitted for publication) as well as a project for a special issue of Signal Processing Magazine on the topic of “Sparse Representations in Signal and Image Processing”.

9. Dissemination

9.1. Conference and workshop committees, invited conference

Rémi Gribonval was an invited speaker at the European Symposium on Artificial Neural Networks (ESANN’06) in Brugges, Belgium, where he gave a tutorial lecture on Sparse Source Separation.

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

Frédéric Bimbot (chair) and Guillaume Gravier organized the Journées d'Étude sur la Parole, held in Dinard, June 12-16, 2006. They were also part of the scientific committe for this conference.

Frédéric Bimbot was a member of the Programme Committee for the Odyssey 2006 Workshop on Speaker Recognition, in Puerto-Rico, June 28-30, 2006.

Guillaume Gravier gave an invited talk on speech recognition at the Colloque sur les Technologies Vocales, help at ENSSAT, Nov. 2006.

9.2. Leadership within scientific community

Frédéric Bimbot is an associate editor for IEEE Signal Processing Letters.

Rémi Gribonval was a Guest Editor (together with Morten Nielsen of the Dept of Math. Sciences at the University of Aalborg) of a special issue of the EURASIP journal Signal Processing dedicated to "Sparse Approximations in Signal and Image Processing", published in July 2006.

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

Guillaume Gravier and Frédéric Bimbot are members of the administration board of the Association Francophone de la Communication Parlée (AFCP).

9.3. Teaching

Guillaume Gravier gave 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 has also given 4 hours of lecture in Speech and Audio indexing within the TAIM Module of the Master in Computer Science, Rennes I.

Mathieu Ben, Frédéric Bimbot, Guillaume Gravier, Rémi Gribonval and Gilles Gonon prepared demonstrations and actively participated to the Fête de La Science in Rennes (13-15 Oct. 2006)

10. Bibliography

Major publications by the team in recent years

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] 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

Doctoral dissertations and Habilitation theses

[13] M. COLLET. *Mesures de similarité robustes dans un espace de locuteurs de référence. Application pour l'indexation de documents audio*, Thèse de doctorat, Université de Rennes 1, IRISA, Rennes (France), September 2006.

[14] M. DELAKIS. *Multimodal Tennis Video Structure Analysis with Segment Models*, Ph. D. Thesis, University of Rennes 1, France, 2006.

[15] A. OZEROV. *Adaptation de modèles statistiques pour la séparation de sources mono-capteur. Application à la séparation voix/musique dans les chansons*, Thèse de doctorat, Université de Rennes 1, IRISA, Rennes (France), December 2006.

Articles in refereed journals and book chapters

[16] L. BENAROYA, F. BIMBOT, G. GRAVIER, R. GRIBONVAL. *Experiments in audio source separation with one sensor for robust speech recognition*, in "Speech Communication", vol. 48, n^{o} 7, 2006, p. 848–854.

[17] 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.

[18] F. BIMBOT, M. FAUNDEZ-ZANUY, R. D. MORI. *Editorial of the Special Issue on Non-Linear and Non-Conventional Speech Processing (NOLISP'03)*, in "Speech Communication", vol. 48, n^{o} 7, July 2006, 759.

[19] 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.

[20] R. GRIBONVAL, M. NIELSEN. *Beyond sparsity : recovering structured representations by L1-minimization and greedy algorithms*, in "Advances in Computational Mathematics", to appear, 2006.

[21] R. GRIBONVAL, M. NIELSEN. *Nonlinear approximation with dictionaries. II. Inverse estimates*, in "Constructive Approximation", vol. 24, n^{o} 2, September 2006, p. 157–173.

[22] R. GRIBONVAL, M. NIELSEN. *Sparse Approximations in Signal and Image Processing* - EDITORIAL, in "EURASIP Signal Processing, special issue on Sparse Approximations in Signal and Image Processing", vol. 86, n^{o} 3, March 2006, p. 415–416.

[23] 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.

[24] 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.

[25] S. KRSTULOVIC, F. BIMBOT, O. BOËFFARD, D. CHARLET, D. FOHR, O. MELLA. *Optimizing the coverage of a speech database through a selection of representative speaker recordings*, in "Speech Communication", vol. 48, n^{o} 10, October 2006, p. 1319-1348.

[26] 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.

[27] A. OZEROV, P. PHILIPPE, R. GRIBONVAL, F. BIMBOT. *Adaptation des modèles pour la séparation de voix chantée à partir d'un seul microphone*, in "Traitement du Signal", to appear, 2006.

[28] 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.

Publications in Conferences and Workshops

[29] S. ARBERET, R. GRIBONVAL, F. BIMBOT. *A Robust Method to Count and Locate Audio Sources in a Stereophonic Linear Instantaneous Mixture*, in "Proc. ICA'06", Springer-Verlag LNCS series, March 2006, p. 536–543.

[30] 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 Series, vol. 3889, Springer, March 2006, p. 536–543.

[31] E. CAMBERLEIN, P. PHILIPPE, F. BIMBOT. *Adaptive Filter Banks Using Fixed Size MDCT and Subband Merging for Audio Coding - Comparison with the MPEG AAC Filter Banks*, in "121st AES Convention, San Francisco, USA", 2006.

[32] M. COLLET, D. CHARLET, F. BIMBOT. *A Weighted Measure of Similarity for Speaker Tracking*, in "Proc. IEEE Odyssey Workshop 2006, Puerto-Rico, USA", June 2006.

[33] M. COLLET, D. CHARLET, F. BIMBOT. *Représentation du locuteur par modèles d'ancrage pour l'indexation de documents audio*, in "Journées d'Étude sur la Parole, Dinard, France", 2006.

[34] M. COLLET, D. CHARLET, F. BIMBOT. *Speaker Tracking by anchor models using speaker segment cluster information*, in "Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06), Toulouse, France", vol. 1, May 2006, p. I-1009 – I-1012.

[35] M. DELAKIS, G. GRAVIER, P. GROS. *Score oriented Viterbi search in sport video structuring using HMM and segment models*, in "IEEE Conf. on Multimedia Signal Processing", 2006.

[36] S. GALLIANO, E. GEOFFROIS, J.-F. BONASTRE, G. GRAVIER, D. MOSTEFA, K. CHOUKRI. *Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News*, in "Language Resources and Evaluation Conference", 2006.

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

[38] S. HUET, G. GRAVIER, P. SÉBILLOT. *Peut-on utiliser les étiquetteurs morpho-syntaxique pour la transcription automatique?*, in "Journées d'Étude sur la Parole, Dinard, France", 2006.

[39] P. JOST, P. VANDERGHEYNST, S. LESAGE, R. GRIBONVAL. *MoTIF : an Efficient Algorithm for Learning Translation Invariant Dictionaries*, in "Int. Conf. Acoust. Speech Signal Process. (ICASSP'06), Toulouse, France", May 2006.

[40] S. KRSTULOVIC, R. GRIBONVAL. *MPTK: Matching Pursuit made Tractable*, in "Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06), Toulouse, France", vol. 3, May 2006, p. III-496 – III-499.

[41] S. LESAGE, S. KRSTULOVIC, R. GRIBONVAL. *Under-determined source separation: comparison of two approaches based on sparse decompositions*, 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 Series, vol. 3889, Springer, March 2006, p. 633–640.

[42] Y. MAMI, F. BIMBOT. *Etude comparative de modélisation de langage par bigrams et par multigrams pour la reconnaissance de la parole*, in "Journées d'Étude sur la Parole, Dinard, France", 2006.

[43] G. MONACI, P. JOST, P. VANDERGHEYNST, B. MAILHE, 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.

[44] G. MONACI, P. JOST, P. VANDERGHEYNST, B. MAILHE, 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.

[45] D. MORARU, G. GRAVIER. *Ancres macrophonétiques pour la transcription automatique*, in "Journées d'Étude sur la Parole, Dinard, France", 2006.

[46] X. NATUREL, G. GRAVIER, P. GROS. *Fast structuring of large television streams using program guides*, in "Intl. Workshop on Adaptive Multimedia Retrieval", 2006.

Internal Reports

[47] S. HUET, G. GRAVIER, P. SÉBILLOT. *Utilisation de la linguistique en reconnaissance de la parole : un état de l'art*, Technical report, n^{o} PI 1804, Irisa, May 2006, http://www.irisa.fr/doccenter/publis/PI/2006/irisapublication.2006-05-30.9598024893.

[48] E. VINCENT, R. GRIBONVAL, M. D. PLUMBLEY. *Oracle Estimators for the Benchmarking of Source Separation Algorithms*, 28 july 2006, Technical report, n^{o} C4DM-TR-06-03, Centre for Digital Music, Queen Mary, University of London, July 2006.

References in notes

[49] J. BOBIN, Y. MOUDDEN, J.-L. STARCK, M. ELAD. *Morphological Diversity and Source Separation*, in "to appear in the IEEE Signal Processing Letters", 2006.

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

[51] 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.

[52] M. COLLET, D. CHARLET, F. BIMBOT. *A Correlation metric for speaker tracking using anchor models*, in "Proc. IEEE-ICASSP (International Conference on Acoustics, Speech and Signal Processing)", vol. I, 2005, p. 713–716.

[53] M. COLLET, Y. MAMI, D. CHARLET, F. BIMBOT. *Probabilistic Anchor Models Approach for Speaker Verification*, in "Proc. Interspeech (Eurospeech, Lisbonne)", September 2005, p. 2005–2008.

[54] M. DELAKIS, G. GRAVIER, P. GROS. *Audiovisual fusion with segment models for video structure analysis*, in "2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies", 2005.

[55] M. DELAKIS, G. GRAVIER, P. GROS. *Multimodal segmental-based modeling of tennis video broadcasts*, in "Intl. Conf. on Multimedia and Exhibition", 2005.

[56] ELDA. *ELDA - Evaluations and Language resources Distribution Agency*, see http://www.elda.org/ for the specifications of the currently available SpeechDat databases, 2005, http://www.elda.org/.

[57] C. FÉVOTTE, R. GRIBONVAL, E. VINCENT. *BSS_EVAL Toolbox User Guide – Revision 2.0*, Technical report, n^{o} 1706, IRISA, Rennes (France), April 2005, http://www.irisa.fr/bibli/publi/pi/2005/1706/1706.html.

[58] G. GONON, F. BIMBOT, ET AL.. *Security requirements for TPD (Deliverable) – Chapter 8 : Enhanced User Authentication / Biometry for TPD*, Technical report, n^{o} D8, Inspired Consortium, IST-2003-507894, June 2005.

[59] G. GONON, R. GRIBONVAL, F. BIMBOT. *Decision Trees with Improved Efficiency for Fast Speaker Verification*, in "Proc. Interspeech'05 (Eurospeech, Lisbonne)", September 2005, p. 3077–3080.

[60] 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.

[61] R. GRIBONVAL, E. BACRY. *Harmonic Decomposition of Audio Signals with Matching Pursuit*, in "IEEE Trans. Signal Proc.", vol. 51, n^{o} 1, jan 2003, p. 101–111.

[62] R. GRIBONVAL, L. BENAROYA, E. VINCENT, C. FÉVOTTE. *Proposals for Performance Measurement in Source Separation*, in "Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation (ICA2003), Nara, Japan", April 2003, p. 763–768.

[63] R. GRIBONVAL. *Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps*, in "IEEE Trans. Signal Proc.", vol. 49, n^{o} 5, May 2001, p. 994-1001.

[64] R. GRIBONVAL. *Sparse decomposition of stereo signals with Matching Pursuit and application to blind separation of more than two sources from a stereo mixture*, in "Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'02), Orlando, Florida", May 2002.

[65] R. GRIBONVAL. *Approximations non-linéaires pour l'analyse de signaux sonores*, Ph. D. Thesis, Université Paris IX Dauphine, September 1999.

[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. KRSTULOVIC, F. BIMBOT, D. CHARLET, O. BOËFFARD. *Focal speakers : a speaker selection method able to deal with heterogeneous similarity criteria*, in "Proc. Interspeech'05 (Eurospeech, Lisbonne)", September 2005, p. 3057–3060.

[69] 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.

[70] 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.

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

[72] Y. MAMI, D. CHARLET. *Speaker identification by location in an optimal space of anchor models*, in "ICSLP", vol. 2, 2002, p. 1333-1336.

[73] NAGORSKI, BOVES, STEENEKEN. *Optimal Selection of Speech Data for Automatic Speech Recognition Systems*, in "ICSLP", 2002, p. 2473–2476.

[74] 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.

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

[76] D. STURIM, D. REYNOLDS, E. SINGER, J. CAMPBELL. *Speaker indexing in large audio databases using anchor models*, in "IEEE-ICASSP", 2001, p. 429–432.

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