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

Activity Report

2004

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 [CR CNRS - HDR]

Administrative assistant
Marie-Noëlle Georgeault [TR INRIA (with Dream and Symbiose teams)]

Research scientist (CNRS)
Guillaume Gravier [CR]

Research scientist (INRIA)
Rémi Gribonval [CR]

Project Technical Staff
Michaël Betser [Engineer, until April 2004]
Gilles Gonon [Engineer, since July 2004]
Sacha Krstulovic [Engineer, since April 2004]

Teaching Assistant
Ewa Kijak [until September 2004]

Ph.D. students
Mathieu Ben [MENRT grant, terminated November 2004]
Ewen Camberlein [External PhD Student with FTR&D-Rennes]
Mikaël Collet [FTR&D-Lannion funding, 1st year]
Robert Forthofer [CIFRE funding with TMM, 1st year]
Stéphane Huet [MENRT, since October 2004, also with TEXMEX]
Sylvain Lesage [MENRT Grant, 1st year]
Alexey Ozerov [FTR&D-Rennes funding, 1st year]
Amadou Sall [Regional Grant, 1st year]

2. Overall Objectives

The research objectives of the METISS research group are dedicated to the 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 (Domus Videum, Technolangues) European projects (INSPIRED) and industrial contracts with various companies (Thomson Multi-Media, France Télécom R&D, ...).

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 [56] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [41], 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 [58]. 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

The main application domains of the METISS project-team are in speaker authentification, audio indexing, and audio source separation.

4.2. Speaker characterisation

Keywords: *speaker recognition*, *user authentication*, *voice signature*.

Participants: Mathieu Ben, Frédéric Bimbot, Mikaël Collet, Gilles Gonon, Sacha Krstulovic.

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 [42], 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.

Another key issue in practice is the non-controlable deviation of the models from the exact probability density functions, which requires a step of normalisation before comparing the verification score to a decision threshold.

In the context of speaker verification, the METISS project puts particular effort on these robustness issues. Algorithmic approaches are also developed to contribute to the scalability, the complexity reduction and the distribution of processes so as to specifically address needs related to the implementation of this technology on personal devices.

Various other topics of speaker characterisation are linked to speaker recognition and verification, in particular speaker elicitation, i.e. how to select a representative subset of speakers from a larger population and speaker representation, namely how to represent a new speaker in reference to a given speaker population.

4.3. Detecting, tracking and searching 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, Mathieu Ben, Guillaume Gravier, Michaël Betser, Mikaël Collet.

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

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

Participants: Guillaume Gravier, Michaël Betser, Frédéric Bimbot, Robert Forthofer.

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 using heterogeneous information

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

Participants: Guillaume Gravier, Michaël Betser, Ewa Kijak.

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 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 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. The Bayes detection theory also 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 also offer a great potential which has been experimented in audiovisual speech recognition so far [43][44] [61].

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*, *speech modeling*, *speech recognition*.

Participants: Guillaume Gravier, Stéphane Huet.

Many audio documents contain speech from which useful information concerning the document content can be extracted. However, extracting information from speech requires specific processing such as speech recognition or word spotting. Though speech recognition is not the main activity of METISS, some research efforts are made in the areas of acoustic modeling of speech signals and automatic speech transcription, mainly in order to complement our know-how in terms of audio segmentation and indexing within a realistic setup.

In particular, speech recognition is complementary with 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. Speaker segmentation and tracking is often used to produce a *rich* transcription of an audio document, typically broadcast news, where the transcription contains speaker and topic indices in addition to the transcription. 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.

4.4. 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, 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, 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.

Additional tracks consist in investigating dictionaries of probabilistic functions.

5. Software

5.1. Speech signal processing toolkit

Keywords: *analysis*, *audio*, *processing*, *signal*, *speech*.

Participant: Guillaume Gravier.

The SPro toolkit provides standard front-end analysis algorithms for speech 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 agreement. It is used by several other french laboratories working in the field of speech processing.
Contact : guillaume.gravier@irisa.fr

5.2. Audio segmentation and classification toolkit

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

Participants: Guillaume Gravier, Michaël Betser, Mathieu Ben.

In the framework of our activities on audio indexing and speaker recognition, audioseg, a toolkit for the segmentation of audio streams is developed and maintained. This toolkit provides generic tools for the segmentation and indexing of audio streams under Unix, 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.

The audioseg toolkit has been used to develop a new speaker verification platform, validated with our participation to the NIST speaker recognition evaluation this year [36]. It was also extensively used for various work and developments, in particular for the detection of audio events in video sound tracks [57][40].
Contact : guillaume.gravier@irisa.fr

5.3. Speech recognition search engine, Sirocco

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

Participant: Guillaume Gravier.

In collaboration with the computer science dept. at ENST, METISS actively participates in the development of the freely available Sirocco large vocabulary speech recognition software [48] based on the algorithm described in [60]. The Sirocco project started as an INRIA Concerted Research Action now works on the basis of voluntary contributions.

In the METISS group, the Sirocco speech recognition software is used to validate algorithms within an entire indexing system. In particular, it has been used to study noise robustness of speech recognition using source separation techniques [38]. We are also currently using Sirocco as the heart of a broadcast news indexing system to illustrate the know-how of METISS in terms of segmentation into sound classes and into speakers. Our broadcast news indexing system will be evaluated in the framework of the national evaluation campaign ESTER (Broadcast News Rich Transcription System Evaluation).
Contact : guillaume.gravier@irisa.fr

5.4. A broadcast news indexing system

Keywords: *analysis*, *audio*, *processing*, *signal*, *speech*.

Participant: Guillaume Gravier.

The concept of rich transcription consists in generating an orthographic transcription of a show enriched with side information concerning the speakers, the presence of background music, the topics and any information related to the structure of the show.

METISS, in collaboration with the Computer Science Dept. at ENST Paris, developed a radio broadcast news rich transcription system that was evaluated during the dry-run phase of the national evaluation campaign ESTER (Broadcast News Rich Transcription System Evaluation). This system will also be evaluated in the official test campaign of ESTER, scheduled early 2005.

In the 2004 dry-run evaluation, METISS participated in the following tasks : transcription, audio event tracking and speaker tracking. The transcription was based on the Sirocco speech recognition software. Audio event tracking permitted the validation of results previously obtained in the framework of the Domus Videum project, concerning segmentation and simultaneous event detection. Finally, the speaker tracking system was adapted from the system used in the last NIST evaluation campaigns on speaker recognition.

In 2004, a huge effort has been put on the development of this system which is now fully operational. The system is used to validate the Sirocco speech recognition engine and our know-how in speaker characterisation. Some new approaches and results produced by the research group (as reported in the next section) were incorporated to the IRENE system.
Contact : guillaume.gravier@irisa.fr

5.5. Matching Pursuit and Short Time Fourier Transform packages for LastWave

Participants: Rémi Gribonval, Sacha Krstulovic.

METISS regularly contributes to the development of the *LastWave* signal-processing software, the kernel of which is developed by Emmanuel Bacry at the Center for Applied Mathematics of the Ecole Polytechnique. *LastWave* is published under a free software license model (GNU General Public License), runs on Windows, MacOS and Unix platforms and boasts a figure of nearly 300 registered users.

LastWave is an object-oriented signal processing software, which consists in several packages. METISS mainly contributes to the development, maintenance and publicity of the *Matching Pursuit* and *Short-Term Fourier Transform* packages. These modules have also been incorporated, independently of *LastWave*, into Fabien Brachere's *Guimauve* software, from the Midi-Pyrénées Astrophysics Lab/Observatory in Toulouse. METISS efforts this year have been targeted at extending the functionalities of the packages to deal with multichannel audio signals and source separation. A description of the various algorithms implemented in the packages can be found in [52][50][51][49].
Contact : remi.gribonval@irisa.fr
Relevant links :
http://www.irisa.fr/metiss/gribonval/LastWave/
http://www.cmap.polytechnique.fr/~bacry/LastWave/
http://webast.ast.obs-mip.fr/people/fbracher/.

6. New Results

6.1. Speaker and speech recognition

Keywords: *Bayesian adaptation*, *denoising*, *hierarchical clustering*, *source separation*, *speaker recognition*, *speech recognition*, *structural models*.

6.1.1. Comparison, normalisation and adaptation of speaker models

Participants: Mathieu Ben, Frédéric Bimbot, Mikaël Collet, Guillaume Gravier.

In speaker recognition, Bayesian adaptation of Gaussian Mixture Models (GMM) [62] with the Maximum A Posteriori (MAP) criterion have shown to be more efficient than the Maximum Likelihood (ML) estimation, because it limits over-adaptation on the training data by assuming a prior distribution for the model parameters. However, this technique is not sufficient in practice to compensate for the lack of training data, and the statistical behaviour of the score provided by the likelihood ratio test is not consistent with the Bayesian theory.

This problem is usually dealt with by score normalisation techniques, such as z-norm, t-norm, etc... [1]. In the framework of his PhD [9], Mathieu Ben has established formal relations between the statistics of likelihood ratio scores, the Kullback-Leibler distance between GMM models and the Euclidean distance between GMM parameters (under specific yet realistic hypotheses). These results have then been used to substitute to the concept of score normalisation, the approach of *model normalisation* which proves to be as efficient in terms of speaker recognition performance and much more advantageous in terms of speaker representation and score computation complexity. These results should also impact more recent work on anchor speaker models.

We have also studied a structural adaptation scheme which assumes a hierarchical structure of speech common to all speakers. We introduce multi-resolution GMMs in which the mean vectors are structured in a binary tree, with coarse-to-fine resolution when going down the tree. Bayesian adaptation [45] is then performed in a hierarchical way, propagating the estimated values of the coarsest GMM means down the tree via linear regression between contiguous depth. This allows some of the mean of the finest resolution speaker GMM which are not observed in the training set to be adapted according to their parent (or ancestor) node. As in the classical Bayesian adaptation approach, the parameters of the multi-resolution prior background GMMs are estimated using prior data. However, except offering a more general formalism as the conventional approach, the hierarchical scheme has not yielded yet a clear advantage in practice [22].

6.1.2. Denoising speech using single sensor source separation techniques

Participants: Guillaume Gravier, Rémi Gribonval, Alexey Ozerov.

Real-life speech material often contains speech with background noise. In particular for broadcast news, it is common to hear a jingle in the background when listening to the headline titles. Detecting the presence of background music and being able to remove it from the speech signal is of utmost importance in order to obtain a better automatic transcription.

Both detection and removal of background music can be stated in terms of source separation using a single sensor, where one source is the speech signal while the second one is the background music signal.

In a previous work [38], we demonstrated that, assuming statistics on the power spectral densities of the jingle and speech signals are known, the jingle can be efficiently removed from the speech material using adaptive Wiener filtering. On the other hand, classical methods such as spectral subtraction or time-frequency shrinkage gave poor results because of the non-stationarity of both the noise and speech signals. However, non-linearities introduced by source separation algorithm limits the benefit in terms of speech recognition.

Previous experiments were carried out on a limited corpus of 50 read sentences. In 2004, we validated these results on a larger corpus and we showed that a robust front-end using normalized cepstral coefficients can partially compensate for the non-linearities introduced in the denoising process [27]. However, performances after denoising are still far from that on the original clean signal and a more realistic setup where the spectral characteristics of the noise is not known a priori must be investigated.

6.2. Audio information extraction

Keywords: *HMM*, *audio information extraction*, *audiovisual integration*, *multimedia*, *statistical hypothesis tests*.

6.2.1. Detecting simultaneous events in audio tracks

Participants: Guillaume Gravier, Michaël Betser.

One common problem in sound event detection is the existence of simultaenous superposed events in complex auditive scenes.

To tackle this problem, we had already proposed to extend a baseline HMM-based system by adding states for all the possible combination of superposed events. As no sufficient data is available for a reliable estimation of the state conditional probability distributions for those states that correspond to multiple events, we proposed several methods to combine models of isolated events into a model for the superimposed events [40].

In 2004, we experimented a new approach that outperformed the previous HMM approach [23][24]. The new approach is based on a maximum a posteriori criterion to detect the events present in a portion of the document. The sound track is first segmented into homogeneous parts and detection is carried out in each segment and for each event of interest. The proposed MAP criterion is strongly related to statistical hypothesis tests but enables the use of a unique decision threshold for all the events considered. This approach was validated on tennis broadcast sound tracks to detect events such as speech, applause or ball hits, and on broadcast news material for speech and music detection.

Though efficient, this approach outlined the limitation of the classical segmentation algorithms, such as the Bayesian Information Criterion one, to detect changes in complex audio scenes (*e.g.* changes from speech to speech+music). An approach combining hypothesis testing and HMMs [23] was studied to solve this problem but achieved the same performance as the MAP criterion.

The results of event detection in tennis videos is exploited for video abstracting [26] in collaboration with VISTA and for video structuring in collaboration with Tex-Mex (see below).

6.2.2. Using audio cues for video structuring

Participants: Guillaume Gravier, Michaël Betser, Ewa Kijak, Robert Forthofer, Stéphane Huet.

The problem of detecting highlights in (sport) videos has so far been seen mainly from the image point of view with some authors using audio cues to select relevant portions of the video. Based on our work on the extraction of audio information (see above), we investigated how the latter can be combined with visual information in order to structure the tennis videos in terms of games, sets an points.

A previous work based on HMMs demonstrated the potential of the Markovian formalism to integrate multimodal (sound and image) information [57] [20] as well as prior structural knowledge. However, this work also demonstrated the limits of this formalism where a single observation is associated to one state. Due to such a constraint, the analysis of the different media must be synchronised to have sequences of descriptors sampled at exactly the same rate for each media stream. In the work of Ewa Kijak, this constraint leads to an analysis stongly drivn by a shot segmentation, even though this segmentation has no meaning from the sound track point of view!

To overcome this problem, we investigate on segment models whose principle is to associate a sequence of observations, aka segment, to a state of the Markov process. Such models were originally proposed for speech modeling. In this case, a state corresponds to a semantic event with its own duration, modeled at the state level, and to which a model is associated in order to compute the probability of a sequence of observations.

This framework was exploited for multimodal tennis video structuring with several, possibly asynchronous, sequences of observations per state. The state conditional probability of a sequence of visual descriptors is given by a HMM as in our previous work. However, the state conditional probability of a sequence of audio events is given by a bigram model thus enabling to take into account the dynamics of audio events. Preliminary results showed significant improvements over the previous HMM approach [1]

More general data structures and elaborate modeling strategies are currently being studied in the framework on two PhDs in their early stage.

6.3. Advanced audio signal processing

Keywords: *dictionary construction*, *granular models*, *source separation*, *sparse decomposition*.

6.3.1. Nonlinear approximation and sparse decompositions

Keywords: *Basis Pursuit*, *Matching Pursuit*, *linear programming*, *redundant dictionnaries*, *sparsity*.

Participant: Rémi Gribonval.

Research on nonlinear approximation of signals and images with redundant dictionaries has been carried out over the past few years in collaboration with Morten Nielsen, from the University of Aalborg in Denmark, and more recently with Pierre Vandergheynst, from the Swiss Federal Institute of Technology in Lausanne (EPFL).

A problem closely related to m-term approximation of a signal/function from an overcomplete dictionary is the computation of sparse representations of the signal in the dictionary. For the family of *localized frames* (which includes most Gabor and wavelet type systems) it is known [55] that the canonical frame expansion

[1] This work is a joint work with TEXMEX and is the prime focus of the PhD of Manolis Delakis, co-directed by Patrick Gros (TEXMEX), Pascale Sébillot (TEXMEX) and Guillaume Gravier (METISS).

provides a near-sparsest representation of any signal in the ℓ^τ sense, $1 \leq \tau \leq 2$. Last year, we have shown [53] that this property is also valid for $r < \tau < 1$ where r depends on the degree s of localization/decay of the frame, and combining it with our previous results [17] we showed that thresholding the canonical representation in a localized frame provided a predictable rate of m-term approximation. However, we disproved in [18] a conjecture of Gröchenig about the existence of a general *Bernstein inequality* for localized frames, by building a simple counter-example. Speaking in simpler words, we proved that for some localized frames, it is possible to find signals for which the ideal m-term approximation rate is infinitely better than what can be predicted from its sparsest representation (which turns out to be essentially its canonical frame expansion). This year, we proved that for *blockwise incoherent* dictionaries, a better behaviour can be expected, namely the rate of best m-term approximation never exceeds *twice* the rate predicted from its sparsest representation.

Many simple and yet interesting frames –such as the union of a wavelet basis and a Wilson basis– are not localized frames, and one cannot rely on the frame coefficients to obtain a near sparsest representation for various ℓ^τ measures. Last year, in [54][53], [30] we proposed several extensions of results by Donoho, Huo, Elad and Bruckstein on sparse representations of signals in a union of two orthonormal bases, by (1) relaxing the hypotheses on the structure of the dictionary and (2) replaced the $\ell 0$ and $\ell 1$ sparsity measures with a larger family of *admissible sparsity measures* (which includes all ℓ^τ norms, $0 \leq \tau \leq 1$), and we gave sufficient conditions for having a unique sparse representation of a signal from the dictionary w.r.t. such a sparsity measure. This year, we obtained results on sparse *approximations* (which include the case of sparse *representations*). We provided a simple test [33] that can be applied on the output of a sparse approximation algorithm to check whether it is nearly optimal, in the sense that no significantly different linear expansion from the dictionary can provide both a smaller approximation error and a better sparsity (in the sense of any *admissible* sparsity measure). As a by-product, we obtained results on the identifiability of sparse overcomplete models in the presence of noise, for the class of admissible sparse priors.

In a joint work with Pierre Vandergheynst from EPFL [34] we extended to the case of the Pure Matching Pursuit recent results by Gilbert *et al* [46][47] and Tropp [63] about exact recovery with Orthogonal Matching Pursuit. In particular, in incoherent dictionaries, our result extends a result by Villemoes [64] about Matching Pursuit in the Haar-Walsh wavepacket dictionary: if we start with a linear combination of sufficiently few atoms from an incoherent dictionary, Matching Pursuit will pick up at each step a “correct” atom and the residue will converge exponentially fast to zero. The rate of exponential convergence is controlled by the number of atoms in the initial expansion. We also obtained stability results of Matching Pursuit when the analyzed signal is well approximated by such a linear combination of few atoms.

6.3.2. Dictionary design for source separation

Keywords: *redundant dictionnaries*, *sparse coding*, *sparsity*.

Participants: Sylvain Lesage, Rémi Gribonval, Frédéric Bimbot.

Recent theoretical work has shown that Basis Pursuit or Matching Pursuit techniques can recover highly sparse representations of signals from *incoherent* redundant dictionaries, or structured (rather than sparse) representations from unions of orthonormal bases. To exploit these results we started last year a research project dedicated to the design of dictionaries structured as unions of orthonormal bases. We proposed a new method based on the SVD and thresholding to build dictionaries which are a union of orthonormal bases. The interest of such a structure is manifold. Indeed, it seems that many signals or images can be modeled as the superimposition of several layers with sparse decompositions in as many bases. Moreover, in such dictionaries, the efficient Block Coordinate Relaxation (BCR) algorithm can be used to compute sparse decompositions. We showed that it is possible to design an iterative learning algorithm that produces a dictionary with the required structure. Each step is based on the coefficients estimation, using a variant of BCR, followed by the update of one chosen basis, using Singular Value Decomposition. We assessed experimentally how well the learning algorithm recovers dictionaries that may or may not have the required structure, and to what extent the noise level is a disturbing factor.Besides its promising results, the method is flexible in that the sparsity measure which is optimized can easily be replaced with some other criterion.

6.3.3. 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 have started a study, in the context of a PhD, on 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 cooccurences of concurrent and consecutives notes, and we use these statistics to cluster music pieces into classes of models and to measure predictibility of notes within a class of models. Preliminary results have shown promising results that are currently being consolidated.

After simple n-gram models will have been investigated, we will evaluate more elaborate models such as Markov Fields. 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.

6.3.4. Underdetermined audio source separation

Keywords: *Gaussian Mixture Models*, *Hidden Markov Models*, *Kalman filtering*, *Wiener filter*, *clustering*, *degenerate blind source separation*, *denoising*, *masking*.

Participants: Alexey Ozerov, Frédéric Bimbot, Rémi Gribonval.

The problem of separating several audio sources mixed on one or more channels is now well understood and tackled in the determined cased, where the number of sources does not exceed the number of channels. Based on our work on statistical modeling and sparse decompositions of audio signals in redundant dictionaries (see above), we proposed in the past years techniques to deal with the degenerate case (monophonic and stereophonic), where it is not possible to merely estimate and apply a demixing matrix.

Last year we proposed [39][37] a series of methods to perform the separation of two sound sources from a single sensor. The methods were based on mixtures of Gaussian models to model the nonstationary data, and they involved a learning phase where the parameters of the models were estimated and a separation phase where a generalization of Wiener filtering was applied to estimate the sources. This year, in [27] we have applied these methods to the separation of music from speech in broadcast news for robust speech recognition and we have compared them to more classical denoising methods. Moreover, we have considered several new parametric models of nonstationary signal based on graphical models and mixtures of Gaussians, either in the spectral or in the log spectral domain. We are now beginning to understand experimentally the interplay between the choice of the modeling domain (spectral or log spectral), the estimation criteria used at the learning and separation phases (e.g., which (average) distortion is minimized) and the quality of the results in terms of a measured distortion.

6.3.5. Evaluation of audio source separation methods

Keywords: *Audio source separation*, *source to artefacts ratio*, *source to distortion ratio*, *source to interference ratio*, *source to noise ratio*.

Participant: Rémi Gribonval.

Because the success or failure of an algorithm for a practical task such as BSS cannot be assessed without agreed upon, pre-specified objective criteria, METISS took part in 2002-2003 to a GDR-ISIS (CNRS) workgroup [35] which goal was to “identify common denominators specific to the different problems related to audio source separation, in order to propose a toolbox of numerical criteria and test signals of calibrated difficulty suited for assessing the performance of existing and future algorithms”. The workgroup released an online prototype of a database of test signals together with an evaluation toolbox. This year, we have proposed a larger set of performance measures and an updated toolbox to deal with the fact that, depending on the exact application, different distortions can be allowed between an estimated source and the target true source. We considered four different sets of such allowed distortions, from time-invariant gains to time-varying filters. In each case we proposed to decompose the estimated source into a true source part plus error terms corresponding to interferences, additive noise and algorithmic artifacts. Then we derived a global performance measure using an energy ratio, plus a separate performance measure for each error term. These measures were computed and discussed on the results of several BASS problems with various difficulty levels. These proposals are the subject of a paper currently submitted to IEEE Trans. Speech and Audio Processing.

7. Contracts and Grants with Industry

7.1. Initiatives funded by the French Network RNRT

7.1.1. Projet Domus Videum (n° 2 02 C 0100 00 00 MPR 011)

Participants: Frédéric Bimbot, Guillaume Gravier, Michaël Betser.

The Domus Videum project is a national RNRT project which started in 2001 and which will terminate mid 2004.

Academic partners of the project are IRISA (VISTA, TEXMEX, TEMICS and METISS project-team) and Nantes University. Industrial partners are Thomson Multimedia, INA and SFRS.

The aim of the project is to design and implement techniques for the automatic summarization of audio-visual programmes (especially in the field of sports). Specific contributions of METISS are targeted towards the joint modeling of the audio and video information using Hidden Markov Model. METISS is also involved in evaluation activities.

7.1.2. Projets Technolangues (n° 2 03 C 0766 00 31 331 011, 2 03 C 0785 00 31 331 011

Participants: Sacha Krstulovic, Mathieu Ben, Frédéric Bimbot.

The Technolangue programme is dedicated to the developpent of software and data resources for research and development in speech and language research and engineering.

The NEOLOGOS project was dedicated to the selection of relevant linguistic material and a set of representative speakers for the definition and the recording of a multi-speaker speech database for speech recognition. The partners are : TELISMA, ELDA, DIALOCA, FTR&D-Lannion, LORIA and IRISA.

The AGILE-ALIZE project was dedicated to the design, development and test of a freeware speaker recognition platform based on the know-how of the ELISA Consortium. The partners are : ATLOG, Thalès, CLIPS, LIA, ENST, IRIT.

7.2. ACI actions

7.2.1. ACI Masse de Données : Demi-ton

Participants: Guillaume Gravier, Ewa Kijak, 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 groupp 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) and on interactions between speech recognition and natural language processing for the extraction of information on the structure of a spoken document (PhD Thesis of Stéphane Huet, jointly with Tex-Mex).

7.3. Initiatives funded by the European Commission

7.3.1. Projet FP6-IST-IP INSPIRED (n° 1 04 A 0115 00 47 622 005)

Participants: Gilles Gonon, Rémi Gribonval, Frédéric Bimbot.

The INSPIRED project is a European IP Project which started in January 2004.

The partners are Gemplus, Axalto, ATMEL, Gesiecke & Devriendt, Oberthur, Infineon, Univ. Catholique de Louvain, Univ. de Twente and INRIA.

The project aims at profiling, designing and prototyping new secure technologies and devices for user access control in fixed and mobile environments. The contribution of IRISA is focused on constrained architectures and algorithm for biometry.

8. Other Grants and Activities

8.1. National initiatives

8.1.1. MathSTIC national initiative on sparse and structured approximations in audio signal processing

Participants: Rémi Gribonval, Sylvain Lesage, Frédéric Bimbot.

The MathSTIC initiative (projet MathSTIC) "Sparse and structured approximations in audio signal processing" (Approximations parcimonieuses structurées pour le traitement de signaux audio) funded by CNRS is a collaboration between the METISS project-team at IRISA, the Signal Processing Group at LATP, Université de Provence, Marseille, and the Musical Acoustics Lab (LAM), Université Pierre et Marie Curie, Paris. The initiative started in June 2004 and will finish in December 2005. Its goal is to "solve the main theoretical issues about the identifiability of sparse structured models for the approximation of signals with overcomplete dictionaries". In December 2004, a three-day work group SPARS'04 will be held in CIRM, Marseille. It will gather members of the initiative as well as some external partners from the French GDR ISIS and European network HASSIP. In the course of 2005, student exchange between the groups is programmed, and to conclude the initiative, a larger scale international workshop (SPARS'05) will be organised in Rennes.

8.2. European initiatives

8.2.1. The ELISA Consortium

Participants: Mathieu Ben, Frédéric Bimbot, Guillaume Gravier.

The ELISA consortium was set up as a spontaneous non-funded initiative in 1997 by ENST, EPFL, IDIAP, IRISA and LIA.

Its objective is the development, maintenance and improvement of a speaker verification platform that is shared between the members of the Consortium and which is presented in the context of the NIST yearly evaluation in speaker recogntion and tracking.

In 2004, METISS has been participating for the 7th consecutive year to the NIST evaluation, with a system based on the ELISA platform, and obtained well-positioned performances. [59].

Since this year, a version of the ELISA platform is being consolidated in the context of the Technolangues AGILE project (ALIZE sub-package).

8.2.2. HASSIP Research Training Network

Participants: Rémi Gribonval, Sylvain Lesage.

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 will consist in new statistical models of audio signals for coding and source separation, as well as theoretical contributions on time-frequency/time-scale analysis and (highly) nonlinear approximation with redundant dictionaries.

9. Dissemination

9.1. Conference and workshop committees, invited conference

Frédéric Bimbot was a member of the Programme Committee for the Odyssey 2004 Workshop on Speaker Recognition.

Frédéric Bimbot was a member of the Reviewing Committee for the following conferences : ICSLP'04, JEP'04 and ICASSP'04.

Frédéric Bimbot has continued a cooperation with the University of Limerick (Rep. of Ireland), with Jacqueline Walker, started in the context of the Ulysses programme, on the topic of source separation.

Guillaume Gravier was a member of the Reviewing Committee for the JEP'04 conference.

Rémi Gribonval was an invited speaker at a workshop organized by the European network (RTN) HASSIP (Harmonic Analysis and Statistics in Signal and Image Processing) in Cambridge, 13-17 september 2004, and gave a lecture on Sparse Approximations.

Rémi Gribonval was an invited speaker at the "International Conference on Wavelet Theory and Applications: New Directions and Challenges" in Singapore, August 10-13 2004, and gave a lecture on Sparse Approximations.

Rémi Gribonval is the Local Chairman for the workshop SPARS'05 (Signal Processing with Adaptive Sparse Structured Representations) to be held in Rennes, November 16-18 2005. The workshop is organised in coordination with the MathSTIC initiative "Sparse and structured approximations in audio signal processing". Frédéric Bimbot is a member of the Local Organisation Committee.

9.2. Leadership within scientific community

Frédéric Bimbot is co-editor with Marcos Faundez and Renato De Mori, of a special issue of Speech Communication on non-linear speech processing.

Frédéric Bimbot and Guillaume Gravier are Board Members of the AFCP (Association Francophone de la Communication Parlée).

Frédéric Bimbot and Rémi Gribonval participate to the European Initiative COST-277 ("Nonlinear speech processing").

Guillaume Gravier is the coordinator, on behalf of AFCP, for the ESTER action on the evaluation of enriched transcription systems for broadcast news [29][28].

Rémi Gribonval is a member of the Editorial Board of the EURASIP (European Association for Signal, Speech and Image Processing) journal Signal Processing.

Rémi Gribonval is 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"

Rémi Gribonval participates to the MathSTIC initiative "Sparse and structured approximations for audio signal processing" funded by the French CNRS. The aim of the initiative is to "solve the main theoretical issues about the identifiability of sparse structured models for the approximation of signals with overcomplete dictionaries."

9.3. Teaching

Frédéric Bimbot has taught 18 hours in Speech Processing at ESIEA (Ecole Supérieure d'Informatique, d'Electronique et d'Automatique).

Frédéric Bimbot has also given two 2-hour lectures in Speech and Audio indexing within the TAIM Module of the Master in Computer Science, Rennes I.

Frédéric Bimbot gave a lecture on the topic of probabilistic models for audio signals at the "Ecole Chercheurs en Traitement du Signal" organised by IRISA.

Rémi Gribonval gave a lecture on time-frequency analysis at the "Ecole Chercheurs en Traitement du Signal" organised by IRISA.

10. Bibliography

Major publications by the team in recent years

[1] F. Bimbot. *Traitement Automatique du Langage Parlé*, collection Information - Commande - Communication (IC2), chap. Reconnaissance Automatique du Locuteur, Hermès, 2002, p. 79-114.

[2] F. Bimbot, R. Blouet, J.-F. Bonastre, et al.. *The ELISA systems for the NIST'99 evaluation in speaker detection and tracking*, in "Digital Signal Processing", vol. 10, n^o 1-3, janvier/avril/juillet 2000, p. 143-153.

[3] R. Blouet. *Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées*, Ph. D. Thesis, Université de Rennes 1, IRISA, Rennes, December 2002.

[4] G. Gravier, F. Yvon, B. Jacob, F. Bimbot. *Integrating contextual phonological rules in a large vocabulary decoder*, in "The Integration of Phonetic Knowledge in Speech Technology", W. van Dommelen, B. Barry (editors)., à paraître, Kluwer Academics, 2002.

[5] 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.

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

[7] M. Seck, R. Blouet, F. Bimbot. *The IRISA/ELISA speaker detection and tracking systems for the NIST'99 evaluation campaign*, in "Digital Signal Processing", vol. 10, n^o 13, janvier/avril/juillet 2000, p. 154-171.

[8] M. Seck. *Détection de ruptures et suivi de classe de sons pour l'indexation sonore*, Ph. D. Thesis, Université de Rennes 1, IRISA, Rennes, January 2001.

Doctoral dissertations and Habilitation theses

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

Articles in referred journals and book chapters

[10] L. BENAROYA, F. BIMBOT, R. GRIBONVAL. *Audio source separation with a single sensor*, in "IEEE Trans. On Speech and Audio Processing", to appear, 2005.

[11] 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° 4, April 2004, p. 430–451.

[12] 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.

[13] 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° 1–2, 2004.

[14] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Tight wavelet frames in Lebesgue and Sobolev spaces*, in "J. Function Spaces and Appl.", vol. 2, n° 3, 2004, p. 227–252.

[15] M. DUTAT, I. MAGRIN-CHAGNOLLEAU, F. BIMBOT. *Acoustic Modeling of Spoken Languages using Time-Frequency Principal Component Analysis and Hidden Markov Models : Application to Language Identification*, in "IEEE Trans. Signal and Audio Processing", to appear, 2005.

[16] G. GRAVIER, F. YVON, B. JACOB, F. BIMBOT. *Integrating contextual phonological rules in a large vocabulary decoder*, in "Integration of Phonetic Knowledge In Speech Technology", W. VAN DOMMELEN, B. BARRY (editors)., Kluwer Academic, 2004.

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

[18] R. GRIBONVAL, M. NIELSEN. *On a problem of Gröchenig about nonlinear approximation with localized frames*, in "J. of Fourier Anal. and Appl.", vol. 10, n° 4, 2004.

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

[20] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Audiovisual integration for tennis broadcast structuring*, in "Multimedia Tools and Application", 2004.

Publications in Conferences and Workshops

[21] M. BEN, M. BETSER, F. BIMBOT, G. GRAVIER. *Speaker Diarization using bottom-up clustering based on a Parameter-derived Distance between adapted GMMs*, in "Intl. Conf. on Speech and Language Processing", 2004.

[22] M. BEN, G. GRAVIER, F. BIMBOT. *Enhancing the robustness of Bayesian adaptation for text-independent speaker verification*, in "Odyssey'04 Speaker and Language Recognition Workshop", 2004.

[23] M. BETSER, G. GRAVIER. *Multiple events tracking in sound tracks*, in "Intl. Conf. on Multimedia and Exhibition", 2004.

[24] M. BETSER, G. GRAVIER. *Suivi d'événements sonores multiples dans les documents audiovisuels*, in "Compression et Représentation des Signaux Audiovisuels (CORESA)", 2004.

[25] J.-F. BONASTRE, F. BIMBOT, L.-J. BOË, J. CAMPBELL, D. REYNOLDS, I. MAGRIN-CHAGNOLLEAU. *Authentification des personnes par leur voix : un nécessaire devoir de précaution*, in "Journées d'Etude sur la Parole (JEP04), Fès, Maroc", 2004.

[26] F. COLDEFY, M. BETSER, G. GRAVIER, P. BOUTHÉMY. *Tennis video abstraction from audio and visual cues*, in "IEEE Intl. Workshop on Multimedia Signal Processing", 2004.

[27] G. GRAVIER, L. BENAROYA, A. OZEROV, R. GRIBONVAL, F. BIMBOT. *Séparation de sources à partir d'un seul capteur pour la reconnaissance robuste de la parole*, in "Journées d'Etude sur la Parole (JEP04), Fès, Maroc", 2004.

[28] G. GRAVIER, J. BONASTRE, S. GALLIANO, E. GEOFFROIS, K. M. TAIT, K. CHOUKRI. *The ESTER evaluation campaign of Rich Transcription of French Broadcast News*, in "Language Evaluation and Resources Conference", 2004.

[29] G. GRAVIER, J.-F. BONASTRE, S. GALLIANO, E. GEOFFROIS, K. M. TAIT, K. CHOUKRI. *ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques*, in "Journées d'Etude sur la Parole (JEP04)", 2004.

[30] R. GRIBONVAL, M. NIELSEN. *On the strong uniqueness of highly sparse expansions from redundant dictionaries*, in "Proc. Int Conf. Independent Component Analysis (ICA'04)", LNCS series, Springer-Verlag, September 2004.

[31] P. GROS, E. KIJAK, G. GRAVIER. *Automatic video structuring based on HMMs and audiovisual integration*, in "2nd International Symposium on Image/Video Communications over fixed and mobile networks", 2004.

[32] J. WALKER, F. BIMBOT, L. BENAROYA. *Experimental Evaluation of Audio Source Separation with One Sensor*, in "Mathematics in Signal Processing IV, Cirencester (UK)", December, 2004.

Internal Reports

[33] R. Gribonval, R. Figueras, P. Vandergheynst. *A simple test to check the optimality of a sparse signal approximation*, Technical report, n^o 1661, IRISA, November 2004.

[34] R. Gribonval, P. Vandergheynst. *On the exponential convergence of Matching Pursuits in quasi-incoherent dictionaries*, Technical report, n^o 1619, IRISA, April 2004.

Bibliography in notes

[35] Action Jeunes Chercheurs du GDR ISIS (CNRS). *Ressources pour la séparation de signaux audiophoniques*, 2002-2003, http://www.ircam.fr/anasyn/ISIS/.

[36] M. Ben, G. Gravier, A. Ozerov, F. Bimbot. *IRISA 2003 speaker recognition system - 1sp speaker detection, limited data*, in "Proc. NIST Workshop on Speaker Verification", 2003.

[37] L. Benaroya, F. Bimbot. *Wiener based source separation with HMM/GMM using a single sensor*, in "Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation (ICA2003), Nara, Japan", April 2003, p. 957–961.

[38] L. Benaroya, F. Bimbot, G. Gravier, R. Gribonval. *Audio source separation with one sensor for robust speech recognition*, in "ISCA Tutorial and Research Workshop on Non-Linear Speech Processing", 2003.

[39] L. Benaroya, L. McDonagh, F. Bimbot, R. Gribonval. *Non negative sparse representation for Wiener based source separation with a single sensor*, in "Proc. IEEE Intl. Conf. Acoust. Speech Signal Process (ICASSP'03), Hong-Kong", April 2003, p. 613–616.

[40] M. Betser, G. Gravier, R. Gribonval. *Extraction of information from video sound tracks - Can we detect simultaneous events?*, in "Conference on Content-Based Multimedia Indexing", 2003, p. 71–78.

[41] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, H. Leich. *Traitement de la Parole*, Presses Polytechniques et Universitaires Romandes, 2000.

[42] J.-F. Bonastre, F. Bimbot, L.-J. Boë, J. C. bell, D. Reynolds, I. Magrin-Chagnolleau. *Person Authentication by Voice : A Need For Caution*, in "Proc. Eurospeech'03, Genève", 2003.

[43] H. Bourlard, S. Dupont, C. Ris. *Multi-stream speech recognition*, Research Report, n^o RR 96-07, IDIAP, Dec. 1996.

[44] S. Dupont, J. Luettin. *Audio-Visual Speech Modeling for Continuous Speech Recognition*, in "IEEE Trans. on Multimedia", vol. 2, n^o 3, September 2000, p. 141–151.

[45] J.-L. Gauvain, C.-H. Lee. *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*, in "IEEE Trans. on Speech and Audio Processing", vol. 2, n^o 2, April 1994.

[46] A. GILBERT, S. MUTHUKRISHNAN, M. STRAUSS. *Approximation of Functions over Redundant Dictionaries Using coherence*, in "The 14th ACM-SIAM Symposium on Discrete Algorithms (SODA'03)", January 2003.

[47] A. GILBERT, S. MUTHUKRISHNAN, M. STRAUSS, J. TROPP. *Improved sparse approximation over quasi-incoherent dictionaries*, in "Int. Conf. on Image Proc. (ICIP'03), Barcelona, Spain", sep 2003.

[48] 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.

[49] R. GRIBONVAL, E. BACRY. *Harmonic Decomposition of Audio Signals with Matching Pursuit*, in "IEEE Trans. Signal Proc.", vol. 51, n^o 1, jan 2003.

[50] 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.

[51] 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.

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

[53] R. GRIBONVAL, M. NIELSEN. *Highly sparse representations from dictionaries are unique and independent of the sparseness measure*, submitted to Appl. Comp. Harmonic Anal., Technical report, n^o R-2003-16, Dept of Math. Sciences, Aalborg University, October 2003.

[54] R. GRIBONVAL, M. NIELSEN. *Sparse decompositions in unions of bases*, in "IEEE Trans. Inform. Theory", vol. 49, n^o 12, December 2003, p. 3320–3325.

[55] K. GRÖCHENIG. *Localization of frames, Banach frames, and the invertibility of the frame operator*, in "J. Fourier Anal. Appl.", to appear, 2003.

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

[57] E. KIJAK, G. GRAVIER, P. GROS, L. OISEL, F. BIMBOT. *HMM based structuring of tennis videos using visual and audio cues*, in "Proc. Intl. Conf. on Multimedia and Exhibition", 2003.

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

[59] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *The 2003 NIST Speaker Recognition Evaluation*, 2003, http://www.nist.gov/speech/tests/spk/2003/.

[60] S. ORTMANNS, H. NEY. *A word graph algorithm for large vocabulary continuous speech recognition*, in "Computer Speech and Language", vol. 11, 1997, p. 43-72.

[61] G. POTAMIANOS, C. NETI, G. GRAVIER, A. GARG, A. W. SENIOR. *Recent advances in the automatic recognition of audio-visual speech*, in "IEEE Proceedings", vol. 91, n^o 9, September 2003, p. 1306–1326.

[62] A. REYNOLDS, T. QUATIERI, R. DUNN. *Speaker Verification Using Adapted Gaussian Mixture Models*, in "Digital Signal Processing Vol 10,num 1-3", 2000.

[63] J. TROPP. *Greed is good : Algorithmic results for sparse approximation*, Technical report, Texas Institute for Computational Engineering and Sciences, 2003.

[64] L. VILLEMOES. *Nonlinear Approximation with Walsh Atoms*, in "Proceedings of “Surface Fitting and Multiresolution Methods”, Chamonix 1996", A. LE M'EHAUT'E, C. RABUT, L. SCHUMAKER (editors)., Vanderbilt University Press, 1997, p. 329–336.