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 3A

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]

Teaching Assistant
Ewa Kijak [since October 1st, 2003]

Ph.D. students
Mathieu Ben [MENRT grant, 2nd year]
Laurent Benaroya [MENRT grant, terminated June 2003]
Lorcan Mc Donagh [INRIA grant, 3rd year]
Sylvain Lesage [MENRT Grant, since Oct. 2003]
Alexey Ozerov [FTR&D-Rennes funding, since Nov. 2003]
Amadou Sall [Regional Grant, since Oct. 2003]
Mikael Collet [FTR&D-Lannion funding, since Nov. 2003]
Robert Forthofer [CIFRE funding with TMM, since Dec. 2003]

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 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), working groups (GDR ISIS), national research projects (Domus Videum, Technolangues) European projects (BANCA) and industrial contracts with various companies (Thomson Multi-Media, France Télécom R&D, ...).

3. Scientific Foundations

3.1. Introduction

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

Probabilistic approaches offer a general theoretical framework [53] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [37], 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 [55]. 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

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

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

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

The large family of audio signals includes a wide variety of temporal and frequential structures, objects of variable durations, ranging from almost stationary regimes (for instance, the note of a violin) to short transients (like in a percussion). The spectral structure can be mainly harmonic (vowels) or noise-like (fricative consonants). More generally, the diversity of timbers results in a large variety of fine structures for the signal and its spectrum, as well as for its temporal and frequential envelope.

In addition, a majority of audio signals are composite, i.e. they result from the mixture of several sources (voice and music, mixing of several tracks, useful signal and background noise). Audio signals may have undergone various types of distortion, recording conditions, media degradation, coding and transmission errors, etc.

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

3.3.1. Redundant systems and adaptive representations

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

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

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

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

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

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

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

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

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

3.3.2. Sparsity criteria

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

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

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

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

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

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

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

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

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

3.3.3. Decomposition algorithms

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

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

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

The “Matching Pursuit” approach consists in optimizing incrementally the decomposition of the signal, by searching at each stage the element of the dictionary which has the best correlation with the signal to be decomposed, and then by subtracting from the signal the contribution of this element. This procedure

is repeated on the residue thus obtained, until the number of (linearly independent) components is equal to the dimension of the signal. The coefficients α can then be reevaluated on the basis thus obtained. This greedy algorithm is sub-optimal but it has good properties for what concerns the decrease of the error and the flexibility of its implementation.

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

3.3.4. Dictionary construction

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

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

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

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

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

3.3.5. Signal separation

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

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

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

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

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

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

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

Key words: *speaker recognition*, *user authentication*, *voice signature*.

Participants: Mathieu Ben, Frédéric Bimbot.

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 [21], many factors (morphological, physiological, psychological, sociological, ...) have an influence on a person's voice. The activities of the METISS group in this domain are mainly focused on speaker verification, i.e the task of accepting or rejecting an identity claim made by the user of a service with access control.

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 deviation of the models from the exact probability density functions, which requires a step of score normalisation before comparing the likelihood ratio to a decision threshold.

The specific areas on which the METISS project puts particular effort are these two robustness issues.

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 characterisation and tracking in audio documents for a wide variety of features and documents.

4.3.1. Speaker detection

Participants: Frédéric Bimbot, Mathieu Ben, Guillaume Gravier.

Key words: *audio stream*, *detection*, *tracking*, *segmentation*, *speaker recognition*.

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 neutralize the variability factors inherent to this task.

4.3.2. Detecting and tracking sound classes

Participants: Guillaume Gravier, Michaël Betser, Frédéric Bimbot, Rémi Gribonval.

Key words: *audio stream*, *detection*, *tracking*, *segmentation*, *audio indexing*.

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

Participants: Guillaume Gravier, Michaël Betser, Ewa Kijak.

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

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 [38][41] [14].

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

Participant: Guillaume Gravier.

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

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

Key words: *source separation*, *audio events*, *indexing*, *multi-channel sound*, *granular models*.

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: Laurent Benaroya, Rémi Gribonval, 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: Lorcan Mc Donagh, 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 synthesis and sound compression. They attempt to provide such decompositions by representing audio signals as linear sums of elementary signals (or "atoms") from a "dictionary", which can be seen as the instruments. 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.

Granular techniques work by decomposing an audio signal into a great many "elementary" signals of short duration. Analysis methods drawing upon the concept of Gabor *atoms* rely on local-cosine type signals, with optional frequency-modulation. Granular synthesis techniques make it possible to compute highly complex sonic textures, with one notable drawback being the lack of user-control over the final result. We are working on an adaptive analysis method based on non-deterministic signals, called *prototypes* or models in this case, these signals being stochastic equivalents to the basis-vectors used in functional analysis. These prototypes are computed from the original signal, and they can subsequently be used to partially reconstruct (compress) the original signal, find the borders of the notes of a melody, re-synthesize the sound with control parameters easily tuned by the user.

5. Software

5.1. Audio segmentation and classification toolkit

Participants: Guillaume Gravier, Michaël Betser, Mathieu Ben.

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

In the framework of our activities on audio indexing and speaker recognition, audioseg, a toolkit for the segmentation of audio streams, was developed this year. 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 SPro, a free software developed by a member of our team, for feature extraction and should be itself distributed as a free software.

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 [16]. It was also extensively used for our work on the detection of audio events in video sound tracks [27][20].

Contact : guillaume.gravier@irisa.fr

5.2. Speech recognition search engine, Sirocco

Participant: Guillaume Gravier.

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

METISS actively participates in the development of the freely available Sirocco large vocabulary speech recognition software [46] based on the algorithm described in [56], in collaboration with the computer science department at ENST Paris. The Sirocco project started as an INRIA Concerted Research Action and now work with voluntary contributions.

We are using the Sirocco speech recognition software to validate our algorithms within an entire indexing system. In particular, it has been used to study noise robustness of speech recognition using source separation techniques [18]. We are also currently using Sirocco as the heart of a broadcast news indexing system to demonstrate the know-how of METISS in terms of segmentation into sound classes and into speakers.

Contact : guillaume.gravier@irisa.fr

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

Participants: Rémi Gribonval, Lorcan McDonagh.

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 [50][48][49][47].

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

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

6.1.1. Structural adaptation of speaker models

Participants: Mathieu Ben, Frédéric Bimbot, Guillaume Gravier.

In speaker recognition, Bayesian adaptation of GMMs [58] 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, when training data is very limited and sparse, this technique suffers from the fact that only components of the model which are observed in the training set are adapted, unseen components remaining unchanged [15][16].

We also study 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 [43] 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.

We use the same data to estimate the regression coefficients and the binary tree. The latter can be constructed by several hierarchical clustering methods. We used a hierarchical EM algorithm with a Gaussian splitting process. This structural Bayesian adaptation technique is currently under development and experimentation and has not yet shown improvement over classical Bayesian MAP adaptation.

6.1.2. Noise robust speech recognition using source separation techniques

Participants: Laurent Benaroya, Guillaume Gravier, Frédéric Bimbot, Rémi Gribonval, Alexey Ozerov.

Real-life speech material often contains speech with background noise. In particular for broadcast news, it is a common practice to have a jingle in the background when giving 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 good 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. This approach was validated on the BREF corpus [54] to which a jingle was artificially added at various signal-to-noise ratios. Assuming statistics on the power spectral densities of the jingle and speech signals are known, we were able to show [18] that the jingle can be efficiently removed from the speech material using adaptive Wiener filtering while 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, the non-linearities introduced by this type of algorithm limits the benefit in terms of speech recognition. The use of smoothed representations of the speech signal in the recognizer, such as RASTA filtering [52] or short-term gaussianization, can partially compensate for these non-linearities but more efficient spectral (or cepstral) smoothing techniques are still required.

6.2. Audio information extraction

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

6.2.1. Detecting simultaneous events in sport broadcast sound tracks

Participants: Guillaume Gravier, Michaël Betser, Rémi Gribonval.

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

To tackle this problem, we extended our baseline system based on ergodic hidden Markov modeling 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 multiple events states, we proposed several model combination techniques, namely convolution and concatenation, in order to derive a model of the superposed events from the models of the isolated events [20].

Lately, a new approach was developed and outperformed the Viterbi based approach previously used. This method is based on statistical hypothesis tests to detect the presence or not of an event, similarly to what is done in speaker tracking and verification. The sound track is first segmented into homogeneous chunks and detection is carried out in each segment and for each event of interest. It was shown experimentally that using two models, *i.e.* the event and non-event models, can lead to better results than the maximum likelihood Viterbi-based approach. However, it was observed that the segmentation of complex audio scenes into homogeneous chunks is poor, thus causing errors in event detection. The latter is near perfect if a manual perfect segmentation is used. We are currently studying solutions to combine the advantages of model based segmentation (as in the HMM approach) and statistcal hypothesis tests.

6.2.2. Using audio cues for video structuring

Participants: Guillaume Gravier, Michaël Betser, Ewa Kijak.

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 detect highlights in a tennis video or to structure the video in terms of games, sets an points. Our approach is based on tracking broad sound classes in the video sound track, such as applause, ball hits (*i.e.* tennis noise), music or speech.

In the case of highlight detection, potentially relevant shots in the video are first selected based on movement information. These candidates shots are then filtered to select only the most interesting ones, based on the audio characterization of the shot. The filtering heuristics, of the type '*the shot contains ball hits and is followed by applauses*', were determined based on human knowledge.

In the case of video structuring of tennis video, audio cues are integrated as an additional observation stream for a parser based on hidden Markov models. The HMM parser models the syntax of a tennis game in terms of points and sets and classify shots accordingly. When using video attributes only, shots are represented by their similarity to a reference key frame and their duration. Audio information were incorporated into this framework by adding a third stream of observation which consists of a vector describing which sound classes are present in the shot. At training, the observation probabilities of each sound classes for each state are estimated. This strategy allowed for a significant increase of the shot classification rate provided a good detection of the audio events is available. However, this performance increase is less significant with automatic audio event detection and, in the future, audiovisual integration methods more robust to audio event detection errors must be devised [27][28].

This work was done in collaboration with Thomson Multimedia as well as with other teams at IRISA, namely VISTA and TEXMEX, in the framework of the Domus Videum project financed by the Réseau National de la Recherche en Télécommunication.

6.3. Advanced audio signal processing

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

6.3.1. Nonlinear approximation and sparse decompositions

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

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. The goal is to understand what classes of functions/signals can be approximated at a given rate by m-term expansions using various families of practical or theoretical approximation algorithms. Much is known when the dictionary is an orthonormal wavelet basis, and we focus on findind the right extension of the wavelet results to structured redundant dictionaries.

Last year, we completely characterized (in terms of Besov spaces) the best m-term approximation classes with spline-based redundant framelets systems in $L_p(\mathbb{R}^d)$ [12]. This year, we have shown that the characterization extends to more general framelet systems [40][39] in $L_p(\mathbb{R}^d)$, and that Sobolev spaces are also characterized of in terms of framelet coefficients [34]. The range of Besov smoothness for which the characterization holds with the frame expansion is limited by the number of vanishing moments of the functions in the dual frame. However, we proved in [33] that, for twice oversampled MRA-based framelets, the same results hold true with no restriction on the number of vanishing moments, but now the canonical frame expansion is replaced with another linear expansion. The trick is to prove a Jackson inequality by building a "nice" wavelet which has a highly sparse linear expansion in the twice oversampled framelet system.

A problem closely related to m-term approximation is the computation of sparse representations of a function in a redundant dictionary. For the family of *localized frames* (which includes most Gabor and wavelet type systems) it is known [51] that the canonical frame expansion provides a near-sparsest representation of any signal in the ℓ^τ sense, $1 \leq \tau \leq 2$. In [35] we have shown that this property is also valid for $r < \tau < 1$ where r depends on the degree s of localization/decay of the frame. However, we have disproved in [36] a conjecture of Gröchenig about the existence of a general Bernstein inequality for localized frames, by building a simple counter-example. 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. In [13][35] we extended a result by Donoho, Huo, Elad and Bruckstein on sparse representations of signals in a union of two orthonormal bases. In [13], we considered general (redundant) dictionaries in finite dimension, and derived sufficient conditions on a signal for having unique sparse representations (in the ℓ^0 and ℓ^1 sense) in such dictionaries. The special case where the dictionary is given by a union of several orthonormal bases was studied in more detail. In [35] we introduced a large class of admissible sparseness 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 sparseness measure, in finite or infinite dimension. Moreover, we gave sufficient conditions on the ℓ^0 sparseness of a signal such that the simple solution of a linear programming problem simultaneously solves all the non-convex (and generally hard combinatorial) problems of sparsest representation of the signal w.r.t. arbitrary admissible sparseness measures. In a joint work with Pierre Vandergheynst from EPFL (see in [24]) we extended to the case of the Pure Matching Pursuit recent results by Gilbert *et al* [44][45] and Tropp [59] about exact recovery with Orthogonal Matching Pursuit. In particular, in incoherent dictionaries, our result extends a result by Villemoes [60] about Matching Pursuit in the Haar-Walsh wavepacket dictionary: if we start with a linear combination of sufficiently few atoms of 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.

6.3.2. Dictionary design for source separation

Key words: *redundant dictionnaries*, *sparsity*.

Participants: Sylvain Lesage, Laurent Benaroya, Frédéric Bimbot, Rémi Gribonval.

Recent theoretical work has shown that Basis Pursuit or Matching Pursuit techniques can recover highly sparse representations of signals from *incoherent* redundant dictionaries. To exploit these results we have started a research project dedicated to the design of incoherent dictionaries with the aim of performing source separation. First, we compared the sparsity of the decomposition in various orthonomal bases, both theorically

and experimentally. We observed that the cosine basis and the data-dependant Karhunen-Loeve basis provide the sparsest decomposition among the tested orthonormal bases. In the cosine basis, we noticed that the choice of the size of the signal frames which leads to the maximum gain in sparsity corresponds to the largest duration on which the signal can be considered stationary. Then we proposed five methods to "learn" a dictionary from training data so as to maximize the mean sparsity. We proposed a new method based on the SVD and thresholding to build dictionaries which are a union of orthonormal bases. 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. Granular models of audio signals

Key words: *musical signal analysis*, *granular synthesis*, *clustering*.

Participants: Lorcan Mc Donagh, Frédéric Bimbot, Rémi Gribonval.

The theoretical framework which is the foundation of our work on granular signal models is now established [30][19]. The model $s[n] = F(\gamma[k_n], \theta_n)$ is frame-based and of "hybrid" nature, in that it combines two different approaches. The non-parametric part consists of a dictionary element or *prototype* $\gamma[k_n]$ (plain waveforms, Fourier spectra, LPC excitation signals, etc.), where each prototype corresponds to one possible state of the model. The parametric part $F(\cdot, \theta)$ attempts to model how frame signals $s[n]$ deviate from a prototype. Clustering is then used to compute the dictionary and assign a state index k_n to each frame $s[n]$. The clustering algorithms attempts to maximize two antagonistic criteria, namely the quality and efficiency of the representation, respectively measured by the SNR and symbol-rate. Clustering algorithms have been specifically developped for our purpose. Comparative tests conducted on lossy compression of real-world signals showed that these performed better than a number of state-of-the-art algorithms. Various models such as LPC adaptive-codebook and various spectrum-based models have been investigated. Our current efforts concentrate on modelling local dependencies of amplitude and phase discrete spectra [57], both in the time/state and frequency-domain. In this "spectral" model, prototypes are spectrum-templates; we attempt to model amplitude and phase differences between the frame-spectra of a cluster and the corresponding prototype, only around energy-spectrum peaks. The aim is to build an efficient granular, object-based, time-frequency model of audio signals. Altough most of our work is directed towards compression, segmentation [42], content-based indexing, summary generation, musical resynthesis [61] etc. are some of many other possible applications.

6.3.4. Underdetermined audio source separation

Key words: *degenerate blind source separation*, *piecewise linear separation*, *sparse decomposition*, *nonlinear approximation*, *Best Basis*, *Matching Pursuit*, *denoising*, *Wiener filter*, *masking*, *clustering*, *Gaussian Mixture Models*, *Hidden Markov Models*.

Participants: Laurent Benaroya, 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 have proposed techniques to deal with the degenerate case (monophonic and stereophonic), where it is not possible to merely estimate and apply a demixing matrix.

In [19][17] we proposed two new methods to perform the separation of two sound sources from a single sensor. The first method [19] generalizes the Wiener filtering with locally stationary, non gaussian, parametric source models. The method involves a learning phase for which we proposed three different algorithms. In the separation phase, we used a sparse non negative decomposition algorithm of our own. The second method [17] also generalizes the Wiener filtering but with Gaussian Mixture distributions and Hidden Markov Models. The method involves a training phase of the models parameters, which is done with the classical EM algorithm. We derived a new algorithm for the re-estimation of the sources with these mixture models, during the separation phase. In [18] we applied these methods to the separation of music from speech in broadcast news for robust speech recognition.

Following our work of last year [49] on Matching Pursuit based audio source separation in the stereophonic case, and building upon our new approaches for single channel separation (see above), we proposed a new framework [23], called piecewise linear separation, for blind source separation of possibly degenerate mixtures, including the extreme case of a single mixture of several sources. Its basic principle is to : 1/ decompose the observations into "components" using some sparse decomposition/nonlinear approximation technique; 2/ perform separation on each component using a "local" separation matrix. It covers many recently proposed techniques for degenerate BSS, as well as several new algorithms that we propose. We discussed two particular methods of multichannel decompositions based on the Best Basis and Matching Pursuit algorithms, as well as several methods to compute the local separation matrices (assuming the mixing matrix is known). Numerical experiments were used to compare the performance of various combinations of the decomposition and local separation methods. On the dataset used for the experiments, it seemed that Best Basis with either cosine packets of wavelet packets (Beylkin, Vaidyanathan, Battle3 or Battle5 filter) were the best choices in terms of overall performance because they introduce a relatively low level of artefacts in the estimation of the sources; Matching Pursuit introduces slightly more artefacts, but can improve the rejection of the unwanted sources.

6.3.5. Evaluation of blind audio source separation methods

Key words: *Blind audio source separation*, *source to distortion ratio*, *source to interference ratio*, *source to noise ratio*, *source to artefacts ratio*.

Participants: Laurent Benaroya, Frédéric Bimbot, 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 [62] 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.

In [32][31], we proposed a preliminary step towards the construction of a global evaluation framework for Blind Audio Source Separation (BASS) algorithms. BASS covers many potential applications that involve a more restricted number of tasks. An algorithm may perform well on some tasks and poorly on others. Various factors affect the difficulty of each task and the criteria that should be used to assess the performance of algorithms that try to address it. Thus a typology of BASS tasks greatly helps the building of an evaluation framework. We describe some typical BASS applications and propose some qualitative criteria to evaluate separation in each case. We then list some of the tasks to be accomplished and present a possible classification scheme.

In [22], we introduced several measures of distortion that take into account the gain indeterminacies of BSS algorithms. The total distortion includes interference from the other sources as well as noise and algorithmic artifacts, and we defined performance criteria that measure separately these contributions. The criteria are valid even in the case of correlated sources. When the sources are estimated from a degenerate set of mixtures by applying a demixing matrix, we proved that there are upper bounds on the achievable Source to Interference Ratio. We proposed these bounds as benchmarks to assess how well a (linear or nonlinear) BSS algorithm performs on a set of degenerate mixtures. We demonstrated on an example how to use these figures of merit to evaluate and compare the performance of BSS algorithms.

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.

7.2. Initiatives funded by the European Commission

7.2.1. Projet BANCA (n° 1 01 C 0296 00 31331 00 5)

Participant: Frédéric Bimbot.

The BANCA project (Biometric Access Control for Networked and e-Commerce Applications) is a European IST project which started in February 2000.

The partners of the project were : Ibermatica, EPFL, UniS, UCL, Thales, l'IDIAP, BBVA, Oberthur et UC3M.

The project aimed at the design of a secure multi-modal system (using face and voice recognition) for the tele-working and e-banking. Metiss was the Work-Package manager for the research activities in speaker verification.

The project terminated successfully in May 2003.

8. Other Grants and Activities

8.1. National initiatives

8.1.1. Junior researcher initiative "resources for audio source separation"

Participants: Rémi Gribonval, Laurent Benaroya, Frédéric Bimbot.

The *Junior researcher initiative* (Action Jeunes Chercheurs) "Resources for audio source separation" (Ressources pour la séparation de signaux audiophoniques) of the french GDR ISIS (CNRS) is a collaboration between the METISS project-team at IRISA, the Analysis-Synthesis group at IRCAM, Paris and the ADTS group at IRCCyN, Nantes. The initiative started in march 2002 for a duration of 18 month. Its 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. Several joint publications were published [32][22][31] in international conferences (ICA'03, GRETSI'03) and a paper is in preparation for submission in a journal.

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 2003, METISS has been participating for the 6th consecutive year to the NIST evaluation, with a system based on the ELISA platform. The system ranked 5th out of 15 international participants [63].

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 the Local Chairman of the NOLISP'03 workshop on Non-Linear Speech Processing, organised by IRISA in Le Croisic, 20-23 May 2003. Guillaume Gravier and Rémi Gribonval participated actively to the Organisation Committee.

Frédéric Bimbot was an invited speaker to the 3rd AVBPA Conference on Audio-Visual Biometric Person Authentication in Guildford (UK), 9-11 June 2003 and gave a one hour presentation on Speaker Recognition.

Guillaume Gravier was a member of the CBMI'03 workshop on Multimedia Indexation organised in Rennes, 22-24 September 2003

Frédéric Bimbot was a member of the Eurospeech'03, 3rd AVBPA, NOLISP'03 and CBMI'03 Reviewing Committees.

Rémi Gribonval made a short (one week) visit to Pierre Vandergheynst at the Laboratoire des Signaux et Système, EPFL, for a collaboration on Matching Pursuit techniques for audio and video coding. A joint paper is in preparation as a result of the starting collaboration.

Rémi Gribonval visited the Department of Mathematical Sciences at the University of Aalborg for a three month collaboration with Morten Nielsen on the theme of nonlinear approximations with redundant systems. During the visit, Rémi Gribonval participated as an invited speaker to the workshop "Wavelets and their generalizations" on August 15-16 at Aalborg. At the end of the visit, three journal papers had been prepared and submitted for publication with Morten Nielsen and Lasse Borup [33][35][36].

Frédéric Bimbot has set up a cooperation with the University of Limerick (Rep. of Ireland), C/° Jacqueline Walker, in the context of the Ulysses programme, on the topic of source separation (as a first step for music retranscription).

Rémi Gribonval organized a special session on "Applications of Independent Components Analysis and Blind Source Separation" at the conference GRETSI'03, Paris.

Rémi Gribonval was an invited speaker to give a lecture series on nonlinear approximation at a seminar organized in CIRM (Marseille) by the European Network HASSIP in December 2003.

9.2. Leadership within scientific community

Frédéric Bimbot was a member of ISCA Board (International Speech Communication Association), in charge of membership services, until the end of his term (September 2003).

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

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

Rémi Gribonval is the leader of a Junior Researcher Initiative (Action Jeune Chercheur) within the French GDR ISIS (CNRS). The aim of the initiative is to gather “Resources for audio source separation”. Besides the METISS project-team, the partners are the ADTS group at IRCCyN, Nantes and the Analysis-Synthesis group at IRCAM, Paris.

Guillaume Gravier coordonne l’action ESTER d’évaluation des systèmes de transcription enrichie des émissions radiophoniques, pour l’AFCP.

9.3. Teaching

Frédéric Bimbot has taught 40 hours in Speech Processing in EISTI (Ecole Internationale Supérieure du Traitement de l’Information, Cergy-Pontoise) and 36 hours in ESIEA (Ecole Supérieure d’Informatique, d’Electronique et d’Automatique - Paris and Laval).

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

Mathieu Ben, as a teaching assistant, has been teaching 26 hours of signal processing for image applications in IFSIC (Institut de Formation Supérieur en Informatique et Communication), 36 hours of electronics for DEUG SPM2, University Rennes I and 22 hours in sampled slave processes in Maîtrise EEA, University Rennes I.

10. Bibliography

Major publications by the team in recent years

[1] F. BIMBOT. *Traitement Automatique du Langage Parlé.* series collection Information - Commande - Communication (IC2), Hermès, 2002, chapter Reconnaissance Automatique du Locuteur, pages 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 », number 1-3, volume 10, janvier/avril/juillet, 2000, pages 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.* W. VAN DOMMELEN, B. BARRY, editors, in « The Integration of Phonetic Knowledge in Speech Technology », Kluwer Academics, 2002, à paraître.

[5] R. GRIBONVAL. *Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps.* in « IEEE Trans. Signal Proc. », number 5, volume 49, May, 2001, pages 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 », number 13, volume 10, janvier/avril/juillet, 2000, pages 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] L. BENAROYA. *Séparation de plusieurs sources sonores avec un capteur.* thèse de doctorat, Université de Rennes 1, IRISA, Rennes, June, 2003.

Articles in referred journals and book chapters

[10] R. GRIBONVAL, E. BACRY. *Harmonic Decomposition of Audio Signals with Matching Pursuit.* in « IEEE Trans. Signal Proc. », number 1, volume 51, jan, 2003, pages 101–111.

[11] R. GRIBONVAL, M. NIELSEN. *Nonlinear approximation with dictionaries. I. Direct estimates..* in « J. Fourier Anal. and Appl. », 2003, to appear.

[12] R. GRIBONVAL, M. NIELSEN. *On approximation with spline generated framelets.* in « Constr. Approx. », 2003, published online on July 7th, printed version to appear.

[13] R. GRIBONVAL, M. NIELSEN. *Sparse decompositions in unions of bases.* in « IEEE Trans. Inform. Theory », number 12, volume 49, December, 2003, pages 3320–3325.

[14] G. POTAMIANOS, C. NETI, G. GRAVIER, A. GARG, A. W. SENIOR. *Recent advances in the automatic recognition of audio-visual speech.* in « IEEE Proceedings », number 9, volume 91, September, 2003, pages 1306–1326.

Publications in Conferences and Workshops

[15] M. BEN, F. BIMBOT. *D-MAP: a distance normalized MAP estimation of speaker models for automatic speaker verification.* in « IEEE Intl. Conf. on Acoustics, Speech and Signal Processing », 2003.

[16] 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.

[17] 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) », pages 957–961, Nara, Japan, April, 2003.

[18] 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.

[19] 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) », pages 613–616, Hong-Kong, April, 2003.

[20] 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 », pages 71–78, 2003.

[21] 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.

[22] 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) », pages 763–768, Nara, Japan, April, 2003.

[23] R. GRIBONVAL. *Piecewise Linear Separation.* in « Wavelets: Applications in Signal and Image Processing X, Proc. SPIE '03 », volume 5207, M. UNSER, A. ALDROUBI, A. LAINE, editors, San Diego, CA, August, 2003.

[24] R. GRIBONVAL, M. NIELSEN. *Approximation with highly redundant dictionaries.* in « Wavelets: Applications in Signal and Image Processing X, Proc. SPIE '03 », volume 5207, M. UNSER, A. ALDROUBI, A. . LAINE, editors, San Diego, CA, August, 2003.

[25] R. GRIBONVAL, M. NIELSEN. *Sparse Decompositions in "incoherent" dictionaries.* in « Proc. IEEE Intl. Conf. Image Proc. (ICIP'03) », Barcelona, Spain, September, 2003.

[26] C. JUTTEN, R. GRIBONVAL. *L'analyse en composantes indépendantes: un outil puissant pour le traitement de l'information.* in « Proc. GRETSI 2003 », volume I, pages 11, ENST, Paris, France, September, 2003.

[27] 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.

[28] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Audiovisual Integration for Tennis Broadcast Structuring.* in « Conference on Content-Based Multimedia Indexing », pages 421–428, 2003.

[29] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Structuration multimodale d'une vidéo de tennis par modèles de Markov cachés.* in « GRETSI », 2003.

[30] L. MCDONAGH, F. BIMBOT, R. GRIBONVAL. *A granular approach for the analysis of monophonic audio signals.* in « Proc. IEEE Intl. Conf. Acoust. Speech Signal Process (ICASSP'03) », pages 469–472, Hong-Kong, April, 2003.

[31] E. VINCENT, C. FÉVOTTE, R. GRIBONVAL, ET AL.. *Comment évaluer les algorithmes de séparation de sources audio?.* in « Proc. GRETSI 2003 », volume I, pages 27, ENST, Paris, France, September, 2003.

[32] E. VINCENT, C. FÉVOTTE, R. GRIBONVAL, X. RODET, E. LE CARPENTIER, L. BENAROYA, A. RÖBEL, F. BIMBOT. *A Tentative Typology of Audio Source Separation Tasks.* in « Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation (ICA2003) », pages 715–720, Nara, Japan, April, 2003.

Internal Reports

[33] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Bi-framelet systems with few vanishing moments characterize Besov spaces.* Technical report, number R-2003-18, Dept of Math. Sciences, Aalborg University, November, 2003, submitted to Appl. Comp. Harmonic Anal. in October 2003.

[34] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Tight wavelet frames in Lebesgue and Sobolev spaces.* Technical report, number R-2003-05, Aalborg Univ., Dept of Math., March, 2003, submitted to J. Function Spaces.

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

[36] R. GRIBONVAL, M. NIELSEN. *On a problem of Gröchenig about nonlinear approximation with localized frames.* Technical report, number R-2003-19, Dept of Math. Sciences, Aalborg University, November, 2003, submitted to J. Fourier Anal. and Appl. in October 2003.

Bibliography in notes

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

[38] H. BOURLARD, S. DUPONT, C. RIS. *Multi-stream speech recognition.* Research Report, number RR 96-07, IDIAP, Dec., 1996.

[39] C. K. CHUI, W. HE, J. STÖCKLER. *Compactly supported tight and sibling frames with maximum vanishing moments.* in « Appl. Comput. Harmon. Anal. », number 3, volume 13, 2002, pages 224–262.

[40] I. DAUBECHIES, B. HAN, A. RON, Z. SHEN. *Framelets: MRA-based constructions of wavelet frames.* in « Applied and Computational Harmonic Analysis », number 1, volume 14, 2003, pages 1–46.

[41] S. DUPONT, J. LUETTIN. *Audio-Visual Speech Modeling for Continuous Speech Recognition.* in « IEEE Trans. on Multimedia », number 3, volume 2, September, 2000, pages 141–151.

[42] J. FOOTE, M. COOPER. *Media Segmentation using Self-Similarity Decomposition.* in « Proc. SPIE Storage and Retrieval for Multimedia Databases , Vol. 5021 », pages 167-75, January, 2003.

[43] 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 », number 2, volume 2, April, 1994.

[44] 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.

[45] 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.

[46] G. GRAVIER, F. YVON, B. JACOB, F. BIMBOT. *Sirocco, un système ouvert de reconnaissance de la parole.* in « Journées d'étude sur la parole », pages 273-276, Nancy, June, 2002.

[47] R. GRIBONVAL, E. BACRY. *Harmonic Decomposition of Audio Signals with Matching Pursuit.* in « IEEE Trans. Signal Proc. », number 1, volume 51, jan, 2003.

[48] R. GRIBONVAL. *Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps.* in « IEEE Trans. Signal Proc. », number 5, volume 49, May, 2001, pages 994-1001.

[49] 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.

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

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

[52] H. HERMANSKY, N. MORGAN. *RASTA processing of speech.* in « IEEE Trans. on Speech and Audio », number 4, volume 2, 1994, pages 578–589.

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

[54] L. F. LAMEL, J.-L. GAUVAIN, M. ESKÉNAZI. *BREF, a large vocabulary spoken corpus for French.* in « Proc. European Conf. on Speech Processing (EUROSPEECH'91) », pages 505–508, 1991.

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

[56] S. ORTMANNS, H. NEY. *A word graph algorithm for large vocabulary continuous speech recognition.* in « Computer Speech and Language », volume 11, 1997, pages 43-72.

[57] G. PEETERS, X. RODET. *SINOLA: A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum.* in « Proc. Int. Computer Music Conf. (ICMC'99) », ICMC, Beijing, 1999.

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

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

[60] L. VILLEMOES. *Nonlinear Approximation with Walsh Atoms.* in « Proceedings of "Surface Fitting and Multiresolution Methods", Chamonix 1996 », Vanderbilt University Press, A. LE M'EHAUT'E, C. RABUT,

L. SCHUMAKER, editors, pages 329–336, 1997.

[61] A. ZILS, F. PACHET. *Musical Mosaicing.* in « Proceedings of DAFX '01 », University of Limerick, December, 2001.

[62] ACTION JEUNES CHERCHEURS DU GDR ISIS (CNRS). *Ressources pour la séparation de signaux audiophoniques.* 2002-2003, http://www.ircam.fr/anasyn/ISIS/.

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