INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team METISS

Modélisation et Expérimentation pour le Traitement des Informations et des Signaux Sonores

Rennes

THEME COG

Table of contents

1. Team

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

Head of project-team
Frédéric Bimbot [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
Gilles Gonon [Temporary Engineer (INRIA)]
Sacha Krstulovic [Temporary Engineer (INRIA)]

Teaching Assistant
Mathieu Ben [until September 2005]

Post-Doc
Daniel Moraru [since May 2005 (CNRS)]

Ph.D. students
Robert Forthofer [CIFRE Grant with TMM, terminated August 2005]
Mikaël Collet [FTR&D-Lannion Funding, 3rd year]
Sylvain Lesage [MENRT Grant, 3rd year]
Alexey Ozerov [FTR&D-Rennes Funding, 3rd year]
Amadou Sall [Regional Grant, 3rd year]
Stéphane Huet [MENRT Grant, 2nd year, also with TEXMEX]
Simon Arberet [CNRS + Region Grant, 1st year (started October 2005)]
Boris Mailhé [ENS Cachan (Bruz), 1st year (started October 2005)]
Ewen Camberlein [FTR&D-Rennes Funding, 3rd year]
Wen Xuan Teng [Telisma Funding, 2nd year]

2. Overall Objectives

2.1. Overall Objectives

The research objectives of the METISS research group are dedicated to audio signal and speech processing and are organised along three axes: speaker characterization, information detection and tracking in audio streams and "advanced" processing of audio signals (in particular, source separation). Some aspects of speech recognition (modeling and decoding) are also addressed so as to reinforce these three principal topics.

The main industrial sectors in relation with the topics of the METISS research group are the telecommunication sector (with voice authentification), the Internet and multi-media sector (with audio indexing), the musical and audio-visual production sector (with audio signal processing), and, marginally, the sector of educational softwares, games and toys.

In addition to the dissemination of our work through publications in conferences and journals, our scientific activity is accompanied with the permanent concern of measuring our progress within the framework of evaluation campaigns, to disseminate software resources which we develop and to share our efforts with other partner laboratories.

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

3. Scientific Foundations

3.1. Introduction

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

Probabilistic approaches offer a general theoretical framework [68] which has yielded considerable progress in various fields of pattern recognition. In speech processing in particular [57], the probabilistic framework indeed provides a solid formalism which makes it possible to formulate various problems of segmentation, detection and classification. Coupled to statistical approaches, the probabilistic paradigm makes it possible to easily adapt relatively generic tools to various applicative contexts, thanks to estimation techniques for training from examples.

A particularily productive family of probabilistic models is the Hidden Markov Model, either in its general form or under some degenerated variants. The stochastic framework makes it possible to rely on well-known algorithms for the estimation of the model parameters (EM algorithms, ML criteria, MAP techniques, ...) and for the search of the best model in the sense of the exact or approximate maximum likelihood (Viterbi decoding or beam search, for example).

In practice, however, the use of the theoretical tools must be accompanied by a number of adjustments to take into account problems occurring in real contexts of use, such as model inaccuracy, the insufficiency (or even the absence) of training data, their poor statistical coverage, etc...

Another focus of the activities of the METISS research group is dedicated to the adaptive representations of signals in redundant systems [71]. The use of criteria of sparsity or entropy (in place of the criterion of least squares) to force the unicity of the solution of a underdetermined system of equations makes it possible to seek an economical representation (exact or approximate) of a signal in a redundant system, which is better able to account for the diversity of structures within an audio signal.

This topic opens a vast field of scientific investigation : sparse decomposition, sparsity criteria, pursuit algorithms, construction of efficient redundant dictionaries, links with the non-linear approximation theory, probabilistic extensions, etc... The potential applicative outcomes are numerous.

This section briefly exposes these various theoretical elements, which constitute the fundamentals of our activities.

3.2. Probabilistic approach

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

For more than a decade, the probabilistic approaches have been used successfully for various tasks in pattern recognition, and more particularly in speech recognition, whether it is for the recognition of isolated words, for the retranscription of continuous speech, for speaker recognition tasks or for language identification. Probabilistic models indeed make it possible to effectively account for various factors of variability occuring in the signal, while easily lending themselves to the definition of metrics between an observation and the model of a sound class (phoneme, word, speaker, etc...).

3.2.1. Probabilistic formalism and modeling

The probabilistic approach for the representation of an (audio) class X relies on the assumption that this class can be described by a probability density function (PDF) $P(.|X)$ which associates a probability $P(Y|X)$ to any observation Y.

In the field of speech processing, the class X can represent a phoneme, a sequence of phonemes, a word from a vocabulary, or a particular speaker, a type of speaker, a language, ... Class X can also correspond to other types of sound objects, for example a family of sounds (word, music, applause), a sound event (a particular noise, a jingle), a sound segment with stationary statistics (on both sides of a rupture), etc.

In the case of audio signals, the observations Y are of an acoustical nature, for example vectors resulting from the analysis of the short-term spectrum of the signal (filter-bank coefficients, cepstrum coefficients, time-frequency principal components, etc.) or any other representation accounting for the information that is required for an efficient separation of the various audio classes considered.

In practice, the PDF P is not accessible to measurement. It is therefore necessary to resort to an approximation $\hat{P}$ of this function, which is usually refered to as the likelihood function. This function can be expressed in the form of a parametric model and the models most used in the field of speech processing (and audio signal) are the Gaussian Model (GM), the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM).

In the rest of this text, we will denote as Λ the set of parameters which define the model under consideration : a mean value and a variance for a GM, p means, variances and weights for a GMM with p Gaussian, q states, q^2 transition probabilities and $p \times q$, means, variances and weights for an HMM with q states the PDF of which being GMMs with p Gaussians. Λ_X will denote the vector of parameters for class X, and in this case, the following notation will be used :

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

Choosing a particular family of models is based on a set of considerations ranging from the general structure of the data, some knowledge on the audio class making it possible to size the model (number of Gaussian p, number of states q, etc.), the speed of calculation of the likelihood function, the number of degrees of freedom of the model compared to the volume of training data available, etc.

3.2.2. Statistical estimation

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

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

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

This approach is generally satisfactory when the number of parameters to be estimated is small w.r.t. the number of training observations. However, in many applicative contexts, other estimation criteria are necessary to guarantee more robustness of the learning process with small quantities of training data. Let us mention in particular the Maximum a Posteriori (MAP) criterion :

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

which relies on a prior probability $p(\Lambda)$ of vector Λ, expressing possible knowledge on the estimated parameter distribution for the class considered. Discriminative training is another alternative to these two criteria, definitely more complex to implement than the ML and MAP criteria.

In addition to the fact that the ML criterion is only one particular case of the MAP criterion (under the assumption of uniform prior probability for Λ), the MAP criterion happens to be experimentally better adapted to small volumes of training data and offers better generalization capabilities of the estimated models (this is measured for example by the improvement of the classification performance and recognition on new data).

Moreover, the same scheme can be used in the framework of incremental adaptation, i.e. for the refinement of the parameters of a model using new data observed for instance, in the course of use of the recognition system. In this case, the value of $p(\Lambda)$ is given by the model before adaptation and the MAP estimate uses the new data to update the model parameters.

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

3.2.3. Likelihood computation and state sequence decoding

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

In addition, when the class model are HMMs, the evaluation of the likelihood requires a decoding step to find the most probable sequence of hidden states. This is done by implementing the Viterbi algorithm, a traditional tool in the field of speech recognition.

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

3.2.4. Bayesian decision

When the task to solve is the classification of an observation into one class among several closed-set possibilities, the decision usually relies on the maximum a posteriori rule :

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

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

In other contexts (for instance, in speaker verification, word-spotting or sound class detection), the problem of classification can be formulated as a binary hypotheses testing problem, consisting in deciding whether the tested observation is more likely to be pertaining to the class X (denoted as hypothesis X) or not pertaining to it (i.e. pertaining to the "non-class", denoted as hypothesis $\overline{X}$). In this case, the decision consists in acceptance or rejection, respectively denoted $\hat{X}$ and $\hat{\overline{X}}$ in the rest of this document.

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

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

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

In practice, however, the Bayesian theory cannot be applied straightforwardly, because the quantities provided by the probabilistic models are not the true PDFs, but only likelihood functions which approximate the true PDFs more or less accurately, depending on the quality of the model of the class.

The rule of optimal decision must then be rewritten :

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

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

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

3.3. Adaptive representations

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

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

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

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

3.3.1. Redundant systems and adaptive representations

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

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

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

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

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

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

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

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

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

3.3.2. Sparsity criteria

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

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

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

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

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

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

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

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

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

3.3.3. Decomposition algorithms

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

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

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

The “Matching Pursuit” approach consists in optimizing incrementally the decomposition of the signal, by searching at each stage the element of the dictionary which has the best correlation with the signal to be decomposed, and then by subtracting from the signal the contribution of this element. This procedure is repeated on the residue thus obtained, until the number of (linearly independent) components is equal to the dimension of the signal. The coefficients α can then be reevaluated on the basis thus obtained. This

greedy algorithm is sub-optimal but it has good properties for what concerns the decrease of the error and the flexibility of its implementation.

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

3.3.4. Dictionary construction

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

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

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

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

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

3.3.5. Signal separation

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

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

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

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

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

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

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

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

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

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

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

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

4. Application Domains

4.1. Introduction

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

4.2. Speaker characterisation

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

The field of speaker characterisation and verification covers a variety of tasks that consist in using a speech signal to determine some information concerning the identity of the speaker who uttered it. Indeed, even though the voice characteristics of a person are not unique [58], many factors (morphological, physiological, psychological, sociological, ...) have an influence on a person's voice. One focus of the METISS group in this domain is speaker verification, i.e the task of accepting or rejecting an identity claim made by the user of a service with access control. We also dedicate some effort to the more general problem of speaker characterisation with two intentions : speaker indexation in the context of information retrieval and speaker selection in the context of speaker recognition.

Speaker recognition and verification has made significant progress with the systematical use of probabilistic models, in particular Hidden Markov Models (for text-dependent applications) and Gaussian Mixture Models (for text-independent applications). As presented in the fundamentals of this report, the current state-of-the-art approaches rely on bayesian decision theory.

However, robustness issues are still pending : when speaker characteristics are learned on small quantities of data, the trained model has very poor performance, because it lacks generalisation capabilities. This problem can partly be overcome by adaptation techniques (following the MAP viewpoint), using either a speaker-independent model as general knowledge, or some structural information, for instance a dependency model between local distributions.

4.2.1. Speaker model and test normalisation

Participants: Mathieu Ben, Frédéric Bimbot, Guillaume Gravier.

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

4.2.2. Scalability and complexity reduction

Participants: Gilles Gonon, Frédéric Bimbot, Rémi Gribonval.

In order to address needs related to the implementation of speaker verification technology on personal devices, specific algorithmic approaches have to be developed to contribute to the scalability, the complexity

reduction and the process distribution. In this context, speaker modelling approaches and classification procedures need to be designed, simulated and tested.

4.2.3. Speaker representation, selection and adaptation

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

METISS also adresses a number of other topics related to speaker characterisation, in particular speaker selection (i.e. how to select a representative subset of speakers from a larger population), speaker representation (namely how to represent a new speaker in reference to a given speaker population) and speaker adaptation for speech recognition.

4.3. Modeling, detecting and structuring information in audio streams

Automatic tools to locate events in audio documents, structure them and browse through them as in textual documents are key issues in order to fully exploit most of the available audio documents (radio and television programmes and broadcasts, conference recordings, etc). In this respect, defining and extracting meaningful characteristics from an audio stream aim at obtaining a more or less structured representation of the document, thus facilitating content-based access or search by similarity. Activities in METISS focus on sound class and event characterisation and tracking in audio documents for a wide variety of features and documents. In particular, speaker detection, tracking, clustering as well as speaker change detection are studied. We also maintain some background activities in speech recognition.

4.3.1. Speaker detection

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

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

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

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

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

4.3.2. Detecting and tracking sound classes and events

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

Participants: Guillaume Gravier, Daniel Moraru, Frédéric Bimbot, 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 multi-modal information

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

Participant: Guillaume Gravier.

Applied to the sound track of a video, detecting and tracking audio events, as mentioned in the previous section, can provide useful information about the video structure. Such information is by definition only partial and can seldom be exploited by itself for multimedia document structuring or abstracting. To achieve these goals, partial information from the various media must be combined. By nature, pieces of information extracted from different media 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 [59], [60] [77].

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.

4.3.5. Music modeling

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

Participants: Amadou Sall, Frédéric Bimbot.

Music pieces constitue a large part of the vast family of audio data for which the design of description and search techniques remain a challenge. But while there exist some well-established formats for synthetic music (such as MIDI), there is still no efficient approach that provide a compact, searchable representation of music recordings.

In this context, the METISS research group dedicates some investigative efforts in high level modeling of music content along two tracks. The first one is the acoustic modeling of music recordings by deformable probabilistic sound objects so as to represent variants of a same note as several realisation of a common underlying process. The second track is music language modeling, i.e. the symbolic modeling of combinations and sequences of notes by statistical models, such as n-grams.

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

4.4. Advanced audio signal processing

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

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

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

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

4.4.1. Audio source separation

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

The general problem of “source separation” consists in recovering a set of unknown sources from the observation of one or several of their mixtures, which may correspond to as many microphones. In the special case of *speaker separation*, the problem is to recover two speech signals contributed by two separate speakers that are recorded on the same media. The former issue can be extended to *channel separation*, which deals with the problem of isolating various simultaneous components in an audio recording (speech, music, singing voice, individual instruments, etc.). In the case of *noise removal*, one tries to isolate the “meaningful” signal,

holding relevant information, from parasite noise. It can even be appropriate to view audio compression as a special case of source separation, one source being the compressed signal, the other being the residue of the compression process. The former examples illustrate how the general source separation problem spans many different problems and implies many foreseeable applications.

While in some cases –such as multichannel audio recording and processing– the source separation problem arises with a number of mixtures which is at least the number of unknown sources, the research on audio source separation within the METISS project-team rather focusses on the so-called under-determined case. More precisely, we consider the cases of one sensor (mono recording) for two or more sources, or two sensors (stereo recording) for $n > 2$ sources.

4.4.2. Audio signal analysis and decomposition

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

The norms within the MPEG family, notably MPEG-4, introduce several sound description and transmission formats, with the notion of a "score", *i.e.* a high-level MIDI-like description, and an "orchestra", *i.e.* a set of "instruments" describing sonic textures. These formats promise to deliver very low bitrate coding, together with indexing and navigation facilities. However, it remains a challenge to design methods for transforming an arbitrary existing audio recording into a representation by such formats.

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

5. Software

5.1. SPro+AudioSeg : audio signal processing, segmentation and classification toolkit

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

Participants: Guillaume Gravier, Mathieu Ben, Daniel Moraru.

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 license. It is used by several other french laboratories working in the field of speech processing.

In the framework of our activities on audio indexing and speaker recognition, audioseg, a toolkit for the segmentation of audio streams 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. It was also extensively used for various work and developments, in particular for the detection of audio events in video sound tracks.
Contact : guillaume.gravier@irisa.fr

5.2. Sirocco : a speech recognition search engine

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 [62] based on the algorithm described in [75]. 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 [56].

We are also currently using Sirocco as the heart of a broadcast news indexing system, in collaboration with the Computer Science Dept. at ENST Paris, in combination with the know-how of METISS in terms of segmentation into sound classes and into speakers.

This broadcast news indexing system (called IRENE) has been evaluated in the framework of the national evaluation campaign ESTER (Broadcast News Rich Transcription System Evaluation).

Contact : guillaume.gravier@irisa.fr

5.3. MPTK: the Matching Pursuit Toolkit

Participants: Rémi Gribonval, Sacha Krstulovic.

The Matching Pursuit ToolKit (MPTK) is a fast and flexible implementation of the Matching Pursuit algorithm for sparse decomposition of monophonic as well as multichannel (audio) signals. MPTK is written in C++ and runs on Windows, MacOS and Unix platforms. It is distributed under a free software license model (GNU General Public License) and comprises a library, some standalone command line utilities and scripts to plot the results under Matlab.

MPTK has been entirely developped within the METISS group mainly to overcome limitations of existing Matching Pursuit implementations in terms of ease of maintainability, memory footage or computation speed. One of the aims is to be able to process in reasonable time large audio files to explore the new possibilities which Matching Pursuit can offer in speech signal processing. With the new implementation, it is now possible indeed to process a one hour audio signal in as little as twenty minutes.

METISS efforts this year have been targeted at designing the new flexible Matching Pursuit API which allows easy additions of new classes of dictionaries that can be used in the Matching Pursuit framework. Current dictionaries include Gabor, Harmonic and Dirac atoms, and Chirps are currenlty being developped. A description of the various algorithms implemented in MPTK can be found in [67], [65], [66], [63].

Contact : remi.gribonval@irisa.fr

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

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

Participant: Rémi Gribonval.

BSS_EVAL is a MATLAB toolbox to compute performance measures in (blind) source separation within an evaluation framework where the original sources are available as ground truth. BSS_EVAL has been developped in collaboration with C. Févotte and E. Vincent with the support of the French GdR-ISIS/CNRS Workgroup “Resources for Audio Source Separation”. The measures implemented in BSS_EVAL are based on a decomposition of each estimated source into four contributions corresponding to the target source, interferences of unwanted sources, remaining additive noise and artifacts such as "musical noise". They are valid for all usual types of signals, such as real-valued audio or biomedical signals or complex-valued subbands of these signals. A more detailed description of the BSS_EVAL methodology as well as a reference manual can be found in [27], [53] and [64].

Contact : remi.gribonval@irisa.fr

Relevant links : http://www.irisa.fr/metiss/bss_eval.

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

Participant: Rémi Gribonval.

BSS_ORACLE is a MATLAB toolbox to evaluate the best performance achievable by a class of source separation algorithms in an evaluation framework where the true reference sources are known. BSS_ORACLE has been developped in collaboration with E. Vincent. A description of the BSS_ORACLE methodology as well as a reference manual can be found in [52] and online at http://www.irisa.fr/metiss/bss_oracle.

Other relevant references include [27], [53] and [64].

Contact : remi.gribonval@irisa.fr

Relevant links : http://www.irisa.fr/metiss/bss_oracle, http://www.irisa.fr/metiss/bss_eval.

6. New Results

6.1. Speaker characterisation

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

6.1.1. Speaker characterisation in the model space

Participants: Mathieu Ben, Frédéric Bimbot, Guillaume Gravier.

In speaker recognition, Bayesian adaptation of Gaussian Mixture Models (GMM) [78] with the Maximum A Posteriori (MAP) criterion has 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... [3]. In the framework of his PhD [1], 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).

Furthermore, the relation between likelihood ratio scores and the Euclidean distance between GMM parameters can be exploited for efficient score computation, avoiding explicit likelihood computation. Experiments in speaker verification on the NIST 2005 Speaker Recognition corpus and in speaker tracking on the ESTER Broadcast News corpus demonstrated that the computation time can be reduced by up to 75% without any decrease in performance [28].

6.1.2. Relative speaker information and related metrics

Participants: Mikaël Collet, Frédéric Bimbot.

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

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

In previous work, several deterministic metrics (euclidean, angular and correlation) were investigated and evaluated for the comparison of speakers in the anchor space [79], [72] [32]. More recently, a probabilistic approach based on a speaker-dependent Gaussian modeling of the SCV was proposed [33] and yielded considerable improvement of the anchor speaker approach, making it competitive with respect to conventional GMMs.

6.1.3. Optimizing the speaker coverage of a speech database

Participants: Sacha Krstulovic, Mathieu Ben, Frédéric Bimbot.

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

In this context, we have proposed a general database design method aiming at optimizing the contents of new speech databases by focusing the data collection on a selection of speakers chosen for its good coverage of the voice space. Such databases would be better adapted to the development of recent speech processing methods, such as those based on multi-models (e.g adaptation of speech recognition with specialized models, speaker recognition with anchor models, speech synthsis by unit selection, etc.). Such developments require indeed a much larger quantity of data per speaker than the traditional databases can offer [61]. Nevertheless, the increase in the collection cost for such newer and larger databases should be limited as much as possible, while preserving a good coverage of the speaker variability.

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

The resulting methodology proposes to operate a selection by optimizing a quality criterion defined in a variety of speaker similarity modeling frameworks [31]. The selection can be operated and validated with respect to a unique similarity criterion, using classical clustering methods such as hierarchical or k-median clustering, or it can be operated and validated across several speaker similarity criteria, thanks to a newly developed clustering method that we called Focal Speakers selection [43]. In this framework, four different speaker similarity criteria have been tested, and three different speaker clustering algorithms have been compared. The outcome of this work has been used for the final specification of the list of speakers to be recorded in the NEOLOGOS database.

A manuscript detailing the methodology and the results of this speaker-driven database design was submitted to an international journal.

6.1.4. Improved CART trees for fast speaker verification

Participants: Gilles Gonon, Rémi Gribonval, Frédéric Bimbot.

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

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

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

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

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

These improvements applied to the classical classification and regression trees (CART) method in a speaker verification system allow to reduce more than 10 times the complexity of the LLR function computation. Considering a baseline state of the art system with an equal error rate (EER) of 11.6% on the NIST 2005 evaluation, a previous CART method provided typical EER ranging between 19% and 22% while the proposed improvements decrease the EER to 13.7% [38].

This work was carried out in the framework of a feasibility study concerning security requirements for a "Trusted Personal Device" within the Inspired IST Project [54].

6.2. Audio indexation and information extraction

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

6.2.1. Speaker tracking and turn segmentation

Participants: Daniel Moraru, Mathieu Ben, Guillaume Gravier.

In various applications, there is a need to do both speaker turn detection and speaker tracking. So far, these tasks were implemented separately though it seems obvious that they are related. Indeed, grouping together segments from the same speaker before doing speaker tracking should help take better decision based on a larger amount of data. On the opposite side, using prior information on speakers can benefit to speaker turn segmentation, for example by grouping together segments that were identified as belonging to the same known speaker.

We investigated the possible interactions between speaker tracking and speaker turn detection in the framework of broadcast news indexing. We observed that speaker tracking did not benefit from steps of clustering, mostly because the speaker verification system is robust enough to the limited amount of data. On the other hand, the performance of speaker turn detection algorithm can be greatly improved by the knowledge of some of the speakers by applying a speaker tracking algorithm before clustering segments. It was observed that performance are unaffected if none of the known speakers are present [48].

Future works include the use of transcripts for speaker turn detection and tracking. For example, word usage, knowledge of the phonetic content of a segment, are valuable information for speaker related tasks that we plan to explore.

6.2.2. Part of speech tagging for multiple hypothesis speech transcription rescoring

Participants: Stéphane Huet, Guillaume Gravier.

In the framework of the Ph.D. of Stéphane Huet (joint PhD with the Tex-Mex team), we are studying the relations and interactions between natural language processing (NLP) techniques and automatic speech recognition (ASR) techniques.

The first step of NLP techniques often consists of a part-of-speech (POS) tagger. The aim of the POS tagger is to tag each word of a text with morphological and syntactical information concerning the gender, the number and the grammatical function (article, noun, verb, etc.). POS taggers are primarily designed to work on written texts free of grammatical errors and with punctuation marks. We therefore investigated the robustness of various POS taggers on ASR transcriptions which typically contain errors and do not include punctuation. A qualitative analysis of various taggers, either rule-based or statistical, showed that POS taggers are robust to transcription errors and loose punctuation, since they are mostly based on local decisions.

This robustness makes the use of POS tagger practical to improve the transcription. Indeed, one of the most common sources of errors in automatic transcription of French speech is the agreement in gender and number,

mostly due to the mute 'e' and 's' in the French language. To deal with this problem, we investigated the use of part-of-speech (POS) taggers to rescore N-best lists. A statistical POS tagger was used to provide a score for the sequence of tags in a sentence. N-best transcriptions were then rescored based on the POS tagger score combined with the acoustic and linguistic ones. This first approach proved effective in correcting some errors but introduced new ones, actually increasing the error rate. We are currently investigating other ways to use POS tags to improve speech transcription.

This work was carried out in collaboration with Pascale Sébillot (IRISA, Tex-Mex).

6.2.3. Audio and audio-visual structuring of sports programmes

Participants: Guillaume Gravier, Robert Forthofer, Frédéric Bimbot.

The problem of (sport) videos analysis has so far been seen mainly from the image point of view. Based on our previous work on the extraction of audio information, we investigated how the latter can be combined with visual information in order to structure automatically sports video.
Previous work by Ewa Kijak, based on HMMs, demonstrated the potential of the Markovian formalism to integrate multimodal (sound and image) information [70], [69] 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.

To overcome this problem, we investigated segment models (SM) 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 [76]. In this case, a state corresponds to a semantic event with its own duration, modeled at the state level, and with 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. We investigated various ways of modeling the audio stream in this framework. In the HMM framework, audio information resulted from the segmentation of the audio track into broad sound classes (speech, applause, ball hits, etc.) and were incorporated as a shot descriptor, thus loosing the dynamic of such information.

With segment models, we were able to model independently the audio events and to take into account their dynamic. Since broad class segmentation is error prone, we also investigated direct modeling of low-level audio features to get rid of the audio segmentation step. We also explored the incorporation of scores (extracted from incrustations) in the segment model framework.

All the segment model-based approaches yielded significant improvement over the conventional HMM approach. Experimental results also showed that SMs on top of the broad sound class segmentation outperform SMs with low-level cepstral audio features, the latter using data that are too variable. However, we demonstrated that the use of a simple audio frame pre-classifier can result in the same level of performance as with a full broad sound class segmentation [36], [34].

The work on tennis videos is a joint work with Tex-Mex and is the prime focus of the PhD of Manolis Delakis, co-directed by Patrick Gros (Tex-Mex), Pascale Sébillot (Tex-Mex) and Guillaume Gravier (Metiss).

6.2.4. Statistical models of music

Keywords: *musical description, statistical models*.

Participants: Amadou Sall, Frédéric Bimbot.

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

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

In practice, we consider a corpus of MIDI files on which we learn co-occurences of concurrent and consecutive notes, and we use these statistics to cluster music pieces into classes of models and to measure predictability of notes within a class of models. Preliminary results have shown promising results that are currently being consolidated. Bayesian networks will be investigated.

At the longer term, the model is intended to be used in complement to source separation and acoustic decoding, to form a consistent framework embedding signal processing techniques, acoustic knowledge sources and music rules modeling.

6.3. Source separation

6.3.1. Source separation using multichannel Matching Pursuit

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

Participants: Sylvain Lesage, Sacha Krstulovic, Rémi Gribonval.

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

The principle of the methods for the sources estimation is to use the space localization of the sources to discriminate them. To do this, we assume that the signals can be sparsely decomposed on a dictionary, e.g. local cosines or wavelets. This is done using the Matching Pursuit algorithm. Two methods are proposed. The first one consists in decomposing the channels by Matching Pursuit, choosing at each step one atom, and substracting it on all the channels, with the corresponding value on each channel. A direction, corresponding to the vector of these values, is then associated to each atom. A clustering of all these directions is done in a second step to assign each atom to the nearest source. Each source is then reconstructed by adding the atoms assigned to it.

The second approach is to include the knowledge of the sources directions in the Matching Pursuit process. The channels are then decomposed on multichannel atoms, constituted of a monochannel waveform, present on all the channels with a different strength, the ratio of these strengths being the direction of one source. By this way, the atoms are directly assigned to the sources.

These methods perform similarly to DUET and Bofill-Zibulevski's algorithm, two reference methods, for source separation of audio data. The second method is directly transposable for convolutive mixtures if the filters are known, which is a work in progress. Moreover, using adapted dictionaries (learnt from training data) for these methods instead of analytically designed atoms (such as Gabor atoms) is current work.

This work has been presented in [46].

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

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

Participants: Simon Arberet, Rémi Gribonval, Frédéric Bimbot.

One main problem in sound source separation is the estimation of the number of mixed sources. Another issue is the estimation of the source directions in a multisensor mixture.

In complement to the separation methods based on Matching Pursuit, which we developed and evaluated assuming the mixing matrix is known, we have proposed a new robust method to estimate both the number

of audio sources and the mixing directions in a linear instantaneous mixture, even with more sources than sensors.

Our method is based on a multiscale Short Time Fourier Transform (STFT), and relies on the assumption that at some (unknown) scales and time-frequency points, only one source contributes to the mixture. Such points provide estimates of the corresponding directions. Our main contribution is a new method to detect points where this assumption is valid, along with a confidence measure. We also propose a new clustering algorithm called DEMIX to estimate the number of sources and their directions.

In contrast to DUET or other similar sparsity-based algorithms, which rely on a global scatter plot, our algorithm exploits the new confidence measure to weight the influence of each time-frequency point in the estimated directions. The proposed DEMIX algorithm is inspired from work by Deville, based on a confidence measure using time-frequency local persistence of the activity/inactivity of each source. The performance of DEMIX is assessed for counting the sources and estimating the mixing directions on stereophonic mixtures.

In our experiments, DEMIX yields better experimental results than those obtained by K-means and ELBG clustering algorithms to estimate source directions. Moreover DEMIX is, to our knowledge, the only algorithm, to count the number of sources. This work is currently submitted for publication.

6.3.3. Single channel source separation

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

Participants: Alexey Ozerov, Rémi Gribonval, Frédéric Bimbot.

The problem of one microphone source separation applied to singing voice extraction is considered. An approach based on a priori Gaussian Mixture Models of two sources is used. Instead of using general source models (i.e., models learned on sources issued from recordings different from those to be separated) we propose to use adapted models (i.e., models with characteristics mapped to those of the mixed sources). Assuming that processed recording is segmented into vocal and non-vocal parts, music model is learned on the non-vocal parts and the general voice model is adapted on the vocal parts. For voice model adaptation we introduce two constrained adaptation techniques : filter adaptation and Power Spectral Density (PSD) gains adaptation. Joint filter and PSD gains adaptation are also possible and give the best performance. Finally, we show that our singing voice extraction system can be also used for singing voice pitch estimation in polyphonic music.

This work has been published in [50], [51]. It was done in collaboration with FTR&D (Pierrick Philippe).

6.3.4. Evaluation of source separation algorithms

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

Participant: Rémi Gribonval.

Source separation of under-determined and/or convolutive mixtures is a difficult problem that has been addressed by many algorithms which may include parametric source models, mixing models, linear or nonlinear separation systems, etc. Their separation performance is usually limited by several factors including badly designed source models or local maxima of the function to be optimized. But also, performance may be limited by constraints on the estimate, such as the length of the demixing filters or the number of frequency bins of the time-frequency masks. The best possible source that can be estimated under these constraints (in the ideal case where source models and optimization algorithms are perfect) is called an oracle estimator of the source. In order to study the performance of some families of source separation algorithms in an evaluation framework where the reference sources are available, we expressed and implemented oracle estimators for two classes (stationary filtering separation algorithms and time-frequency masking separation algorithms) and studied their performance on a few audio mixture examples.

This work has been published in [52]. It was done in collaboration with Emmanuel Vincent (Queen Mary University).

6.4. Sparse decompositions: theory and algorithms

6.4.1. Learning of shift-invariant atoms (MoTIF algorithm)

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

Participants: Sylvain Lesage, Boris Mailhé, Rémi Gribonval.

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

This algorithm, called MoTIF for Matching of Translation Invariant Features, learns the generating functions iteratively, from a set of learning signals. Each step is an alternate routine : we begin with an initial function, then we find the location, in each learning signal, where this function is the most present. From these located patches, the function is updated by a least-square approximation (Principal Component Analysis). This mechanism is done iteratively. It is monotonic and converges in a finite number of iterations.

The estimation of the next functions needs the addition of a constraint that helps the atoms to be as decorrelated as possible with the previous ones. This way, no atom is selected multiple times. Note that with this constraint, if there were two close underlying patterns in the signal, the algorithm would not retrieve both.

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

This work can be found in [41]. It was done in collaboration with Philippe Jost and Pierre Vandergheynst (EPFL, Lausanne).

6.4.2. The Matching Pursuit Toolkit : Matching Pursuit made tractable

Keywords: *Matching Pursuit, sparsity*.

Participants: Sacha Krstulovic, Rémi Gribonval.

Matching Pursuit (MP) aims at finding sparse decompositions of signals over redundant bases of elementary waveforms. Traditionally, MP has been considered too slow an algorithm to be applied to real-life problems with high-dimensional signals. Indeed, in terms of floating points operations, its typical numerical implementations have a complexity of $\mathcal{O}(N^2)$ and are associated with impractical runtimes. In this work, we propose a new architecture which exploits the structure shared by many redundant MP dictionaries, and thus decreases its complexity to $\mathcal{O}(N \log N)$. This architecture is implemented in a new software toolkit, called MPTK (the Matching Pursuit Toolkit), which is able to reach, e.g., $0.25\times$real time for a typical MP analysis scenario applied to a 1 hour long audio track. This substantial acceleration makes it possible, from now on, to explore and apply MP in the framework of real-life, high-dimensional data processing problems.

This work is currently submitted for publication and the corresponding software is distributed at http://mptk.gforge.inria.fr

6.4.3. Structured sound decomposition with Matching Pursuit

Keywords: *Matching Pursuit, chirplets, harmonic structures, sparsity*.

Participants: Sacha Krstulovic, Rémi Gribonval.

In the framework of audio signal analysis, it is desirable to obtain sparse representations that are able to reflect the harmonic structures, e.g., issued from musical instruments. In this work, we compare two approaches which introduce some explicit models of harmonic features into the Matching Pursuit analysis framework. The first approach is the Harmonic Matching Pursuit (HMP), where the harmonic structures are

modeled by sets of harmonically related Gabor atoms which are directly optimized in the analysis loop. The second approach, called Meta-Molecular Matching Pursuit (M3P), is based on the a posteriori agglomeration of elementary features coming from a Short Time Fourier Transform. We discuss the pros and cons of each method through experiments involving different audio signals, and conclude on possible approaches for combining the two methods. This work is published in [44].

By definition, the Matching Pursuit algorithm with constant (or "flat") Gabor atoms provides a coarse estimate of frequency modulated sinusoids in music and voice signals. Chirped Gabor atoms, closer to the nature of these signals, would fit them in a finer and sparser way. Though a method for the direct analytic estimation of chirped Gabor atoms has been proposed in the past [65], we propose an alternative method where the chirp factor and scale parameter are estimated through a regression over an iteratively selected chain of small-scale atoms defined by a Short Time Fourier Transform. This new technique suits the Matching Pursuit framework, and is compared with a "flat atoms" version of the algorithm. The influence of various frequency interpolation techniques over the sparsity of the resulting representation is also studied.

This work is published in [47]. It was done in collaboration with Pierre Leveau and Laurent Daudet.

6.4.4. A simple test to check the optimality of a sparse signal approximation

Participant: Rémi Gribonval.

Approximating a signal or an image with a sparse linear expansion from an overcomplete dictionary of atoms is an extremely useful tool to solve many signal processing problems. Finding the sparsest approximation of a signal from an arbitrary dictionary is a NP-hard problem. Despite of this, several algorithms have been proposed that provide sub-optimal solutions. However, it is generally difficult to know how close the computed solution is to being "optimal", and whether another algorithm could provide a better result. In this work, we provide a simple test to check whether the output of a sparse approximation algorithm 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. As a by-product of our theorems, we obtain results on the identifiability of sparse overcomplete models in the presence of noise, for a fairly large class of sparse priors.

This work will appear in [19]. It was done in collaboration with Rosa Figueras and Pierre Vandergheynst, EPFL.

6.4.5. Beyond sparsity : recovering structured representations

Keywords: *basis pursuit*, *matching pursuit*, *recovery analysis*, *sparse decomposition*.

Participant: Rémi Gribonval.

In a series of recent results, several authors have shown that both L_1 minimization (Basis Pursuit) and greedy algorithms (Matching Pursuit) can successfully recover a sparse representation of a signal provided that it is sparse enough, that is to say if its support (which indicates where are located the nonzero coefficients) is of sufficiently small size. In this paper we define more general identifiable structures that support signals that can be recovered exactly by L_1 minimization and greedy algorithms. In addition, we obtain that if the output of an arbitrary decomposition algorithm is supported on an identifiable structure, then the representation is optimal within the class of signals supported by the structure. As an application of the theoretical results, we give a detailed study of a family of multichannel dictionaries with a special structure (corresponding to the representation problem $X = \mathbf{A}C\Phi$) often used in, e.g., under-determined source separation problems or in multichannel signal processing. The theoretical results obtained in this framework have served as an inspiration for new source separation algorithms.

This work will appear in [21]. It was done in collaboration with Morten Nielsen, Aalborg University.

6.4.6. An adaptive computational strategy for optimal sparse signal approximation

Keywords: *basis pursuit*, *matching pursuit*, *recovery analysis*, *sparse decomposition*.

Participant: Rémi Gribonval.

Sparse approximation using redundant signal dictionaries is most useful for compressing, denoising or separating mixtures of signals, images and other high-dimensional data, but has the reputation of being a computationally intensive, almost intractable task. Yet, algorithms as simple as thresholding sometimes provide solutions that are surprisingly close to those obtained with more time hungry techniques such as Matching Pursuit, Basis Pursuit or FOCUSS. We propose an adaptive computational strategy to compute, with as simple techniques as possible, nearly as good sparse approximations as with substantially more complex ones. The strategy is based on a low cost *a priori* prediction of the behaviour of complex algorithms which serves as an *a posteriori* test of the "success" of simple algorithms. We discuss the tradeoff between the accuracy and the computational cost of the test with numerical examples, and we provide simple tests for the thresholding, Matching Pursuit and Basis Pursuit algorithms. Preliminary proposals for FOCUSS type algorithms are also discussed. The proposed tests – which rely on combined properties of the signal dictionary and the computed sparse approximation – are more accurate and only slightly more complex than known tests based on the (cumulative) coherence of the dictionary and the number of terms of the sparse approximation. To our knowledge, our results tests also provide, for the first time, practical tools that are applicable in real settings, for many standard dictionaries that are not necessarily "quasi-incoherent".

This work was done in collaboration with Morten Nielsen, Aalborg Univ. and Pierre Vandergheynst, EPFL. A paper is in preparation.

7. Contracts and Grants with Industry

7.1. Initiatives funded by the French Network RNRT

7.1.1. 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, Daniel Moraru, Stéphane Huet.

This project entitled "Multimodal description for automatic structuring of TV streams" started in Oct. 2004 and is funded by the ACI Masse de Données. The partners are the METISS and Tex-Mex groups at IRISA and the DCA 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, Texmex) 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 was focused on constrained architectures and algorithm for biometry.

IRISA's participation terminated as planned in July 2005.

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) took place at IRISA, 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 2005, METISS has been participating for the 8th consecutive year to the NIST evaluation, with a system based on the ELISA platform, and obtained well-positioned performances. [74].

This year, a consolidated version of the ELISA platform has been finalised 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

Rémi Gribonval was a member of the Programme Committee for the GRETSI 2005 Workshop on Speaker Recognition, in Louvain-La-Neuve, 6-9 September, 2005.

Rémi Gribonval was the Local Chairman for the workshop SPARS'05 (Signal Processing with Adaptive Sparse Structured Representations) held in Rennes, November 16-18 2005. The workshop was organised in coordination with the MathSTIC initiative "Sparse and structured approximations in audio signal processing".

Frédéric Bimbot and Guillaume Gravier are chairing and organising the Journées d'Etudes sur la Parole 2006, which will be held in Dinard, June 12-16, 2006.

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

9.2. Leadership within scientific community

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

Guillaume Gravier has been the coordinator, on behalf of AFCP, for the ESTER action on the evaluation of enriched transcription systems for broadcast news [9].

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."

Rémi Gribonval has participated to the CNRS expert committee "methods in signal and image processing".

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.

10. Bibliography

Major publications by the team in recent years

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

[2] L. BENAROYA. *Séparation de plusieurs sources sonores avec un seul microphone*, Ph. D. Thesis, Université de Rennes 1, IRISA, Rennes, June 2003.

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

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

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

[6] 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.

[7] J.-F. BONASTRE, F. BIMBOT, L.-J. BOË, J. CAMPBELL, D. REYNOLDS, I. MAGRIN-CHAGNOLLEAU. *Person authentication by voice : a need for caution*, in "Proceedings Eurospeech, Genève, Suisse", 2003.

[8] 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.

[9] 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.

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

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

[12] R. GRIBONVAL, M. NIELSEN. *Sparse decomposition in unions of bases*, in "IEEE Trans. Inf. Th.", vol. 49, n° 12, Décember 2003, p. 3320-3325.

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

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

[15] 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

[16] L. M. DONAGH. *Modèles granulaires pour les signaux sonores : contributions théoriques et expérimentales*, Thèse de doctorat, Université de Rennes 1, IRISA, Rennes, April 2005.

Articles in refereed journals and book chapters

[17] L. BENAROYA, F. BIMBOT, R. GRIBONVAL. *Audio source separation with a single sensor*, in "IEEE Trans. Speech and Audio Processing", to appear, January 2006.

[18] G. GRAVIER, F. YVON, B. JACOB, F. BIMBOT. *Introducing contextual transcription rules in large vocabulary speech recognition*, in "The integration of phonetic knowledge in speech technology", W. J. BARRY, W. A. V. DOMMELEN (editors). , Text, Speech and Language Technology, vol. 25, chap. 6, Springer, 2005, p. 87–106.

[19] R. GRIBONVAL, R. FIGUERAS, P. VANDERGHEYNST. *A simple test to check the optimality of a sparse signal approximation*, in "EURASIP Signal Processing Journal, Special Issue on Sparse Approximations in Signal and Image Processing", accepted, 2006.

[20] R. GRIBONVAL. *From Projection Pursuit and CART to Adaptive Discriminant Analysis ?*, in "IEEE Trans. Neural Networks", vol. 16, n° 3, May 2005, p. 522–532.

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

[22] R. GRIBONVAL, M. NIELSEN. *Nonlinear Approximation with Dictionaries. II. Inverse Estimates*, in "Constructive Approximations", accepted, 2006.

[23] R. GRIBONVAL, M. NIELSEN. *Sparse Approximations in Signal and Image Processing - Editorial*, in "EURASIP Signal Processing Journal, Special Issue on Sparse Approximations in Signal and Image Processing", accepted, 2006.

[24] R. GRIBONVAL, P. VANDERGHEYNST. *On the exponential convergence of Matching Pursuit in quasi-incoherent dictionaries*, in "IEEE Trans. Information Theory", to appear, January 2006.

[25] P. GROS, M. DELAKIS, G. GRAVIER. *Multimedia indexing: the multimedia challenge*, in "ERCIM News", vol. 62, July 2005, p. 11–12.

[26] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Audiovisual integration for tennis broadcast structuring*, in "Multimedia Tools and Application", 2005.

[27] E. VINCENT, C. FÉVOTTE, R. GRIBONVAL. *Performance measurement in Blind Audio Source Separation*, in "IEEE Trans. Speech and Audio Processing", accepted, 2006.

Publications in Conferences and Workshops

[28] M. BEN, F. BIMBOT, G. GRAVIER. *A model space framework for efficient speaker detection*, in "Proc. Interspeech'05 (Eurospeech, Lisbonne)", September 2005, p. 3061–3064.

[29] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Nonlinear Approximation with Bi-framelets*, in "Proc. Approximation Theory XIth Conf.", 2005.

[30] L. BORUP, R. GRIBONVAL, M. NIELSEN. *Nonlinear Approximation with Redundant Dictionaries*, in "Proc. IEEE-ICASSP (International Conference on Acoustics, Speech and Signal Processing)", vol. IV, 2005, p. 261–264.

[31] D. CHARLET, S. KRSTULOVIC, F. BIMBOT, O. BOËFFARD, E. AL.. *Neologos : an optimized database for the development of new speech processing algorithms*, in "Proc. Interspeech'05 (Eurospeech, Lisbonne)", September 2005, p. 1549–1552.

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

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

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

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

[36] M. DELAKIS, P. GROS, G. GRAVIER. *Multimodal Segmental-Based Modeling of Tennis Video Broadcasts*, in "Meeting of the MUSCLE Network of Excellence", April 2005.

[37] S. GALLIANO, E. GEOFFROIS, D. MOSTEFA, K. CHOUKRI, J.-F. BONASTRE, G. GRAVIER. *The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News*, in "European Conference on Speech Communication and Technology", 2005.

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

[39] G. GRAVIER, F. YVON, M. BEN. *IRENE, le système IRISA – ENST d'indexation d'émissions radiophoniques*, in "Atelier ESTER Phase II", 2005.

[40] R. GRIBONVAL, R. FIGUERAS, P. VANDERGHEYNST. *A simple test to check the optimality of sparse*

signal approximations, in "Proc. IEEE-ICASSP (International Conference on Acoustics, Speech and Signal Processing)", vol. V, 2005, p. 717–720.

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

[42] S. KRSTULOVIC, M. BEN, G. GONON, D. MORARU, G. GRAVIER, R. GRIBONVAL, F. BIMBOT. *The IRISA/METISS systems for NIST SRE 2005*, in "Proc. of the NIST 2005 Speaker Recognition Workshop", 2005.

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

[44] S. KRSTULOVIC, R. GRIBONVAL, P. LEVEAU, L. DAUDET. *A comparison of two extensions of the Matching Pursuit algorithm for the harmonic decomposition of sounds*, in "Proc. WASPAA", 2005.

[45] S. LESAGE, R. GRIBONVAL, F. BIMBOT, L. BENAROYA. *Learning Unions of Orthonormal Bases with Thresholded Singular Value Decomposition*, in "Proc. IEEE-ICASSP (International Conference on Acoustics, Speech and Signal Processing)", vol. V, 2005, p. 293–296.

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

[47] P. LEVEAU, L. DAUDET, S. KRSTULOVIC, R. GRIBONVAL. *Model-Based Matching Pursuit - Estimation Of Chirp Factors And Scale Of Gabor Atoms with Iterative Extension*, in "SPARS, Rennes", 2005.

[48] D. MORARU, M. BEN, G. GRAVIER. *Experiments on speaker tracking and segmentation in radio broadcast news*, in "European Conference on Speech Communication and Technology", 2005.

[49] X. NATUREL, G. GRAVIER, P. GROS. *Etiquetage automatique de programmes de télévision*, in "Compression et Représentation des Signaux Audiovisuels (CORESA)", 2005.

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

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

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

Internal Reports

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

[54] G. GONON, F. BIMBOT, E. AL.. *Security requirements for TPD (Deliverable) – Chapter 8 : Enhanced User Authentication / Biometry for TPD*, Technical report, nº D8, Inspired Consortium, IST-2003-507894, June 2005.

[55] R. GRIBONVAL, M. NIELSEN. *Beyond sparsity : recovering structured representations by L1 minimization and greedy algorithms. Application to the analysis of sparse underdetermined ICA*, Technical report, nº 1684, IRISA, Rennes (France), January 2005, http://www.irisa.fr/bibli/publi/pi/2005/1684/1684.html.

Bibliography in notes

[56] 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.

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

[58] 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.

[59] H. BOURLARD, S. DUPONT, C. RIS. *Multi-stream speech recognition*, Research Report, nº RR 96-07, IDIAP, Dec. 1996.

[60] S. DUPONT, J. LUETTIN. *Audio-Visual Speech Modeling for Continuous Speech Recognition*, in "IEEE Trans. on Multimedia", vol. 2, nº 3, September 2000, p. 141–151.

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

[62] 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.

[63] R. GRIBONVAL, E. BACRY. *Harmonic Decomposition of Audio Signals with Matching Pursuit*, in "IEEE Trans. Signal Proc.", vol. 51, nº 1, jan 2003.

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

[65] R. GRIBONVAL. *Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps*, in "IEEE Trans. Signal Proc.", vol. 49, nº 5, May 2001, p. 994-1001.

[66] 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.

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

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

[69] 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.

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

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

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

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

[74] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *The 2005 NIST Speaker Recognition Evaluation*, 2005, http://www.nist.gov/speech/tests/spk/2005/.

[75] 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.

[76] M. OSTENDORF. *From HMMs to Segment Models*, in "Automatic Speech and Speaker Recognition - Advanced Topics", chap. 8, Kluwer Academic Publishers, 1996.

[77] 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, nº 9, September 2003, p. 1306–1326.

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

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