Extraction of Audio Features Specific to Speech Production for Multimodal Speaker Detection

Patricia Besson, Vlad Popovici, Jean-Marc Vesin, *Member, IEEE*, Jean-Philippe Thiran, *Senior Member, IEEE*, and Murat Kunt, *Fellow, IEEE*

***Abstract*—A method that exploits an information theoretic framework to extract optimized audio features using video information is presented. A simple measure of mutual information (MI) between the resulting audio and video features allows the detection of the active speaker among different candidates. This method involves the optimization of an MI-based objective function. No approximation is needed to solve this optimization problem, neither for the estimation of the probability density functions (pdfs) of the features, nor for the cost function itself. The pdfs are estimated from the samples using a nonparametric approach. The challenging optimization problem is solved using a global method: the differential evolution algorithm. Two information theoretic optimization criteria are compared and their ability to extract audio features specific to speech production is discussed. Using these specific audio features, candidate video features are then classified as member of the "speaker" or "non-speaker" class, resulting in a speaker detection scheme. As a result, our method achieves a speaker detection rate of 100% on in-house test sequences, and of 85% on most commonly used sequences.**

***Index Terms*—Audio features, differential evolution, multimodal, mutual information, speaker detection, speech.**

I. Introduction

With the increasing capabilities of modern computers, both auditive and visual modalities of the speech signal may be used to improve speaker detection, leading to major improvements in the user-friendliness of man-machine interactions. Just consider, for example, a video-conference system. The most interactive current solution requires an audio engineer and a cameraman so that the speaking person can be emphasized both on audio and video. An intelligent system able to detect the speaker on the basis of sound and image information could focus a moving camera on him/her. Another possible application would be multimedia indexing.

Among the different methods that exploit the information contained in each modality, a few are performing the fusion directly at the feature level, though using such an approach, the detection of the current speaker on an audiovisual sequence can be done with only a camera and a single microphone. Moreover, it has been pointed out, in [1] and [2] for example, that such a fusion can greatly help the classification task: the richer and the more representative the features, the more efficient the classifier.

Some audio-video feature fusion approaches try to directly evaluate the synchronism of the two signals [3], [4], [5]. As suggested in [4], the synchronism is here the perceptive effect of the causal relationship between the two signals. Other methods map first the features onto a subspace where this relationship is enhanced and can therefore be estimated [2], [6], [7]. All the approaches rely on explicit or implicit use of mutual information (MI). An estimation of the features' probability density functions (pdfs) is therefore required. There are two main approaches that may be taken: either a parametric or a nonparametric one. In the first case, the pdfs are assumed to follow a parametric law. Most of the time, a Gaussian distribution is considered, which is not necessarily valid. Fisher in [2], as well as Butz in [1] and [8], estimate the probability density functions directly from the available samples during the feature extraction process through Parzen windowing [9].

The problem addressed in this paper is the detection of the current speaker in a given video sequence with two or more candidates. To this end, audio features specific to speech production are extracted using the information content of both the audio and video signals. Taking a similar approach to Butz and Thiran in [1] and [8], as well as Fisher *et al.* in [10] and [2], the problem is cast in an information theoretic framework to optimize the audio features with respect to the video features. The cost function to be optimized is therefore based on MI, which leads to a highly nonlinear optimization problem. Moreover, an analytical formulation of the gradient of the cost function is difficult to obtain without any parametric approximation of the pdf. For this reason, it is preferable to have a method which does not require such an analytical form of the gradient (gradient-free method). In [2], Fisher and Darell use a second order Taylor approximation of the MI and the Parzen estimator to cast the optimization problem into a convex one and to derive a closed form of the gradient. However, our purpose here is to avoid such an approximation and to directly solve our optimization problem using a suitable optimization method. It turns out that the best solution for solving the optimization problem is differential evolution [11], an evolutionary algorithm.

Once these specific audio features are obtained, the MI between them and the video features of candidate mouth regions is used as a classifier to determine the class of the mouth region ("speaker" or "non-speaker').

Manuscript received February 24, 2006; revised June 12, 2007. This work was supported by the Swiss National Science Foundation through Grant 2000-06-78-59. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. John R. Smith.

P. Besson, J.-M. Vesin, J.-P. Thiran, and M. Kunt are with the Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: patricia.besson@a3.epfl.ch; jean-marc.vesin@epfl.ch; jp.thiran@epfl.ch; murat.kunt@epfl.ch).

V. Popovici is with the Bioinformatics Core Facility, Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland (e-mail: popovici@isb-sib.ch).

Digital Object Identifier 10.1109/TMM.2007.911302

The rest of this paper is organized as follows: first, the use of information theory to extract optimized features for classification problems is presented. Then, the chosen representation for the video and audio signals is described. In the third section, the information theoretic optimization approach is applied to obtain audio features optimized for the specific classification task, regardless the classifier. Different optimization criteria based on MI are defined. The fourth section exposes the optimization problem as well as optimization methods used to solve it. The last part of the paper deals with the experiments and discusses the different optimization criteria used in the feature extraction, the ability of the method to produce audio features specific to speech production, and finally, the performance of the method as a speaker detector.

II. Theoretical Framework

A. Information Theoretic Feature Extraction

In the present work, the detection of the current speaker in an audio-visual sequence is viewed as a classification problem. In this classification problem, O is a binary random variable (r.v.) defined on a sample space Ω_O which models the membership to the "speaker" or "non-speaker" class with respect to an audio-visual source, modelled by a r.v. S defined on another sample space Ω_S.

The goal in a classification process is obviously to minimize the probability of assigning some measurements performed on the original signal to the wrong class. That is, to minimize the classification error probability $P_c = P(\hat{O} \neq O)$. This error probability depends of course on the classifier and on its ability to deal with the particular problem, but it also depends on all the processing steps leading from O to $\hat{O}$ (where $\hat{O}$ is the estimated class).

In the problem at hand, the bimodal source S is not directly accessible but yields two observed signals of different physical nature: audio and video signals A and V. For each of these signals, the unimodal classification process leading from the measurements A - respectively V - to an estimate $\hat{O}_1$ - respectively $\hat{O}_2$ - of the class, can be described through a first order Markov chain [Fig. 1(a)]. Two associated classification error probabilities $P_{c1} = P(\hat{O}_1 \neq O)$ and $P_{c2} = P(\hat{O}_2 \neq O)$ can be defined. They correspond to the probability of committing an error when estimating $\hat{O}_1$ (audio Markov chain), or $\hat{O}_2$ (video Markov chain), from $\hat{S}_A$, or $\hat{S}_V$. However, they are also conditioned by the probability of committing an error when estimating $\hat{S}_A$ or $\hat{S}_V$ from S. These errors are referred to as the estimation error probabilities $P_{e1} = P(\hat{S}_A \neq S)$ and $P_{e2} = P(\hat{S}_V \neq S)$. The estimation of one r.v. from another can be understood as a feature extraction step where some specific information must be recovered from the initial r.v. Therefore, the information theoretic framework for extracting features developed in [10] can be applied.

Using Fano's inequality, it is possible to relate the probability of each error to the conditional entropies $H(S|A)$ or $H(S|V)$ [12]. For the audio Markov chain, this inequality is defined as

$$P_{e1} \geq \frac{H(S|A) - 1}{\log|\Omega_S|} = \frac{H(S) - I(S,A) - 1}{\log|\Omega_S|} \qquad (1)$$

where $H(S)$ and $I(S,A)$ are the entropy of S and the MI between the r.v.'s S and A, respectively, and $|\Omega_S|$ is the cardinality of S. A similar relationship can be established for the video Markov chain, resulting in a lower bound on P_{e2}.

The inequality (1) does not help us directly to minimize the error probability P_{e1}. It does indicate however that an efficient minimization of P_{e1} is conditioned by the minimization of the right hand side of the inequality. The minimization of this term leads to an optimal source estimate $\hat{S}_A$. Implicitly, the processing steps leading to this optimal are taken into account, as stated before. A similar reasoning can be followed for the video Markov chain.

B. Extension of the Information Theoretic Feature Extraction To the Multimodal Case

A possibility is to firstly obtain optimal estimates $\hat{S}_A$ and $\hat{S}_V$. Then a fusion at the decision or at the classification level can be performed in order to get a unique estimate $\hat{O}$ of the class from both unimodal processes. However, such an approach would not take advantage of the discriminant information offered by the bimodal nature of the source S. A better approach would be to use an extension of the previously described unimodal feature extraction framework to the multimodal case. Such an extension has been proposed by Butz *et al.* in [1] and applied more particularly to image registration. A similar multimodal feature extraction framework is used here. It is extended to speaker detection and the various processing steps are fully justified.

As already stated, the original source S is accessible only through the measurements A and V. But, as mentioned in [2], these two measurements are affected by independent interference sources N_A and N_V. A good estimate of the source should include a feature extraction step which discards this noisy information present in each modality and recovers the information coming from the common source S, thus shared by both modalities. Obviously, such a goal can only be reached by considering the two modalities jointly. Let F_A and F_V be r.v.'s modelling such audio and video features that contain only the information coming from the source S. Since they specifically describe this common source, they are related by their joint pdf $p(F_A, F_V)$. Thus an estimate of the feature related to one modality can be inferred from the other modality with transition probability $p(\hat{F}_V|F_A)$ or $p(\hat{F}_A|F_V)$. These transition probabilities can be obtained by joint probability estimation since $p(\hat{F}_V|F_A) = p(\hat{F}_V, F_A)/p(F_A)$ and $p(\hat{F}_A|F_V) = p(\hat{F}_A, F_V)/p(F_V)$, and if $\hat{F}_V$ and $\hat{F}_A$ are correctly estimated the approximation $p(\hat{F}_A, F_V) \approx p(F_A, \hat{F}_V) \approx p(F_A, F_V)$ can be assumed. These considerations lead to the definition of the classification problem with the two Markov chains shown in Fig. 1(b). Notice that the source estimates associated to each chain are indexed by *AV* or *VA*, to stress that they have been obtained using information present in both modalities, in contrast with the previous case [Fig. 1(a)].

Of course, the estimation error probabilities P_{e1} and P_{e2} and their associated lower bounds are still defined according to inequality (1). However, since each estimation can be viewed as a feature-extraction step where the estimate r.v. is a function of the previous r.v., the data processing inequality for Markov

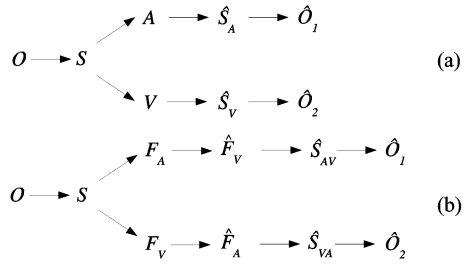


Fig. 1. (a) Graphical representation of the audio and video Markov chains (leading to $\hat{O}_1$ and $\hat{O}_2$, respectively) modelling the two unimodal classification processes associated to each modality. (b) Graphical representation of the related Markov chains modelling the multimodal classification process.

chains can be used to weaken the bounds on the error probabilities. Inequality (1) for the audio and video Markov chains becomes

$$P_{e_1} \geq \frac{H(S) - I(F_A, \hat{F}_V) - 1}{\log |\Omega_S|} \tag{2}$$

$$P_{e_2} \geq \frac{H(S) - I(F_V, \hat{F}_A) - 1}{\log |\Omega_S|}. \tag{3}$$

As previously stated, $p(\hat{F}_A, F_V) \approx p(F_A, \hat{F}_V) \approx p(F_A, F_V)$. Therefore $I(F_A, \hat{F}_V) \approx I(\hat{F}_A, F_V) \approx I(F_A, F_V)$. Introducing this approximation in (2) and (3), a joint lower bound can finally be defined as follows:

$$P_{\{e1,e2\}} \geq \frac{H(S) - I(F_A, F_V) - 1}{\log |\Omega_S|}. \tag{4}$$

Minimizing the lower bound on $P_{\{e_1,e_2\}}$ then amounts to maximizing the MI between the extracted features F_A and F_V corresponding to each modality. The feature sets resulting from the maximization of the MI involved in these equations are expected to compactly describe the relationship between the two modalities. The extraction stage therefore produces optimized features.

However, to get a source estimate with a probability of estimation error close to this bound, a suitable estimator must be found. If F_A and F_V are correctly estimated, they compactly describe the source S. For this last statement to be true, not only the MI $I(F_A, F_V)$ between features extracted from each modality must be increased, but also the conditional entropies $H(F_V|F_A)$ and $H(F_A|F_V)$ must be minimized. Indeed, if the entropies increase, they reduce the inter-feature dependencies. Dividing (4) by the joint entropy $H(F_A, F_V)$, a feature efficiency coefficient [1] can be defined as

$$e(F_A, F_V) = \frac{I(F_A, F_V)}{H(F_A, F_V)} \in [0, 1]. \tag{5}$$

This coefficient defines our estimator. Since $I(F_A, F_V) = H(F_A) + H(F_V) - H(F_A, F_V)$, maximizing $e(F_A, F_V)$ still minimizes the lower bound on the error probability defined in (4) while constraining inter-feature independencies. In other words, the extracted features F_A and F_V will tend to capture just the information related to the common origin of A and V, while discarding the unrelated interferences coming from N_A and N_V: they estimate the source S.

C. Classifier Definition

Applying this framework to extract features, the estimation error probability comes closer to its minimum. However, the classification error probability P_c must still be minimized: this depends on the choice of a suitable classifier. Previous works in the domain have shown that measuring the synchrony between the audio and video measurements is a good way of classifying them as originating from an audio-visual source or not [4]–[6]. In [4] in particular, the authors interpret synchrony as the degree of between audio and video signals. MI also shows good performance in detecting synchronized audio-video sources such as speakers [2]–[4]. Moreover, the feature optimization pre-processing also indicates the MI-based classifier as a good choice. For these reasons, the chosen classifier consists in the evaluation of the MI between candidate audio and video features. The features that exhibit the largest MI are classified as "speaker", while the other ones are labelled as "non-speaker", only one "speaker" class label being authorized per estimation.

Notice that such a classifier has also the advantage of fusing the information at the classification level in a straightforward way, resulting in a unique class estimate $\hat{O}$.

III. Signal Representation

A. Video Representation

When applying this feature extraction framework in the context of speaker detection, the first decision to be made is in the choice of signal representation.

Physiologic evidences point out the motion in the mouth region as a visual evidence for speech. Therefore, the chosen video features are the estimates of the optical flow in the mouth region. In order to have a local pixel-based representation of these video features, the Horn and Schunck's gradient-based algorithm [13] has been chosen. The method is implemented in a two-frame simple forward difference scheme so that the temporal resolution is large enough to capture complex and quickly varying mouth motions. First, a median pre-filtering is used on the raw intensity images to reduce the noise level. The optical flow is computed between each two consecutive frames over a region of $N \times M$ pixels including the lips and the chin of each candidate. These regions are referred to as mouth regions and are estimated over the sequence either from a manual extraction on the first frame, or using a face detector such as the one described in [14]. In order to get reliable pdf estimates without using a very large sample, only the magnitude of the optical flow and the sign of the vertical component are kept, so that the video features are one-dimensional (1-D).

Speakers are observed over a window of T frames. Therefore, a sample of the 1-D r.v. F_V comprises $N \times M \times (T-1)$ observations $\{f_{V,k}\}_{k=1,\ldots,N\times M\times(T-1)}$ which are the optical flow norm values at each spatio-temporal point. These values are normalized for the subsequent optimization (see [15] for details). This approach implicitly considers the observations to be independent, which is obviously a simplification of the real world. Indeed, the neighboring pixels are correlated and cannot be truly independent. This simplification is somewhat mitigated by estimating the pdf with the Parzen window approach [9], as we shall see later.

B. Audio Representation

The audio signal also needs to be represented in a tractable way. This representation should describe salient aspects of the speech signal, while being robust to variations in speaker or acquisition conditions. Mel-cepstrum analysis is one of the methods that fits best these requirements and as such, is widely used in speech-processing research [16], [17]. Accordingly, the speech signal is represented as a set of $T-1$ vectors, each containing P mel-cepstrum coefficients $\{C_t(i)\}_{i=1,\ldots,P}$ with $t=1,\ldots,T-1$ (the first coefficient has been discarded as it pertains to the energy). Notice that the mel-cepstrogram is downsampled to the video feature frame rate to get synchronized audio and video representations.

IV. Extraction of Optimized Speech Audio Features

A. Audio Feature Optimization

In principle, the information theoretic feature extraction discussed in Section II can now be used for audio and video features F_A and F_V. However, the audio representation is P-dimensional thus large samples are required for a correct pdf estimation. The feature extraction framework is then applied so as to decrease the dimensionality of the audio representation while keeping and emphasizing the information related to the video content. Consequently, the 1-D audio features $\{f_{A,t}(\vec{\alpha})\}_{t=1,\ldots,T-1}$ composing the sample of the r.v. F_A, are built as the following linear combination of the P Mel-frequency cepstral coefficients (MFCCs):

$$f_{A,t}(\vec{\alpha})=\sum_{i=1}^{P}\vec{\alpha}(i)\cdot C_t(i)\quad\forall t=1,\ldots,T-1 \tag{6}$$

where the weights $\vec{\alpha}(i)$ are chosen such that $\sum_{i=1}^{P}\vec{\alpha}(i)=1$ and $\vec{\alpha}(i)\geq 0\quad\forall i=1,\ldots,P$. Thus, the $(T-1)P$-dimensional audio observations reduces to $(T-1)$ 1-D observations $f_{A,t}(\vec{\alpha})$. The minimization of the estimation error given by (4) will lead to the optimized vector $\vec{\alpha}$. This optimization therefore requires the availability of the joint probability density function as well as of the marginal densities of the r.v. F_A and F_V. Since the audio features are unknown before the optimization, their distribution is obviously unknown too. To avoid any restrictive assumption, the pdfs are estimated using the nonparametric Parzen windowing approach

$$\hat{f}(y)=\frac{1}{n}\sum_{i=1}^{n}h(y-y_i;\sigma)\quad\forall y\in\Omega_Y \tag{7}$$

where h is a kernel function whose variance is controlled by the parameter σ, n is the sample size, and y an observation of the r.v. Y. A 2-D Gaussian kernel $G(\mu_A,\mu_V,\sigma_A,\sigma_V)$ of mean $[\mu_A,\mu_V]^T$ and diagonal covariance matrix $\mathrm{diag}(\sigma_A;\sigma_V)$ is chosen in our case for its widespread validity. The variances σ_A and σ_V are estimated from the audio and video data respectively, in a robust way, as described in [18]

$$\sigma=\left(\frac{4}{3n}\right)^{1/5}\cdot\frac{\mathrm{median}|y_i-\tilde{\nu}|}{0.6745} \tag{8}$$

where $\tilde{\nu}$ denotes the median of the data points. Since the video data remain the same during the optimization of the audio data, the value for σ_V remains constant for a given set of video features, while σ_A adapts itself to the audio features during the optimization process.

Using the Parzen window to estimate the densities in a nonparametric way yields a better estimate than histogram-based approaches, given the small size of the available samples ($T-1$ for the random variable associated with the audio features).

B. Optimization Criteria

As discussed in Section II, minimizing the estimation error probability is equivalent to maximizing the efficiency coefficient considering the audio and video features over a mouth region. The set of weights to be optimized with respect to the efficiency coefficient criterion (ECC) are defined as

$$\begin{aligned}\vec{\alpha}_{\mathrm{opt}}&=\arg\max_{\vec{\alpha}}\{I(F_V,F_A(\vec{\alpha}))/H(F_A(\vec{\alpha}))\}\\&=\arg\max_{\vec{\alpha}}\{e(F_V,F_A(\vec{\alpha}))\}.\end{aligned} \tag{9}$$

Note that in our case, the normalization term for the MI involves only the audio feature entropy since the video features remain constant during the optimization process.

To verify the necessity of normalizing the MI by the entropy during the optimization, ECC will be compared with a "simple" mutual information criterion (MIC). The set of weights to be optimized is then defined as

$$\vec{\alpha}_{\mathrm{opt}}=\arg\max_{\vec{\alpha}}\{I(F_V,F_A(\vec{\alpha}))\}. \tag{10}$$

Finally, a more constraining criterion is introduced, which takes into account a pair of mouth regions. This criterion, referred to as ΔECC, is the squared difference between the efficiency coefficient computed in each mouth region (referred to as Ω_{M_1} and Ω_{M_2}). This way, the differences between the marginal densities of the video features in each region are taken into account. Moreover, only one optimization is performed for two mouths. If $F_{V_{M_1}}$ and $F_{V_{M_2}}$ denote the random variables associated to regions Ω_{M_1} and Ω_{M_2} respectively, then the optimization problem becomes

$$\vec{\alpha}_{\mathrm{opt}}=\arg\max_{\vec{\alpha}}\{[e(F_{V_{M_1}},F_A(\vec{\alpha}))-e(F_{V_{M_2}},F_A(\vec{\alpha}))]^2\}. \tag{11}$$

V. Optimization Method

A. Definition of the Optimization Problem

The extraction of optimized audio features with respect to our classification task requires finding the real-valued vector $\vec{\alpha}\in\mathbb{R}^P$, that minimizes the chosen cost function $f(\vec{\alpha})$. This function is defined as the negative value of one of the optimization criteria defined in (9), (10), or (11). Moreover, to restrain the set of possible solutions, the P weighting coefficients $\{\alpha_i\}_{i=1,\ldots,P}$ must fulfill the following conditions:

$$0\leq\vec{\alpha}(i)\leq 1\quad\forall i=1,2,\ldots,P \tag{12}$$

$$\sum_{i=1}^{P}\vec{\alpha}(i)=1. \tag{13}$$

This optimization problem is highly nonlinear and gradient-free. Indeed, an analytical formulation of the gradient of the cost

function is difficult to obtain due to the unknown form of the pdf of the extracted audio features. In [2], Fisher and Darell use a second order Taylor approximation of the MI and the Parzen estimator to cast the optimization problem into a convex one and to derive the gradient in an analytical way. However, our purpose here is to avoid such an approximation and to directly solve our optimization problem using a proper optimization method.

Optimization methods can be classified as either local or global. The first category includes steepest gradient descent and gradient descent-based methods such as the Powell's direction set method. They mainly rely on the use of an exact or estimated formulation of the gradient of the cost function to find an optimum. They present the advantage of being fast and easy to use but are very likely to fail to reach the global optimum of the cost function if the latter is not convex.

The second category refers to algorithms which aim at finding the globally best solution, in the possible presence of multiple local minima. We find in this category stochastic and heuristic methods such as simulated annealing (SA) [19], Tabu Search (TS) [20], or evolutionary algorithms (EAs). These have proven their ability to approach the global optimum of highly nonlinear problems, possibly at a high computational cost. Both SA and TS are more dedicated to solve combinatorial problems. EAs, which include genetic algorithms (GAs), look more suitable for our problem. Such optimization procedures, first introduced by Holland in 1962 [21], are based on natural evolution principles: starting from an initial candidate *population* of *chromosomes* (or sets of parameters to be optimized), operators mimicking the biological ones of *crossover* and *mutation* are used to *select* and *reproduce* fittest solutions, the fitness of a solution being given by a scoring function. Basically, mutation enables the algorithm to explore new regions of the search space by randomly altering some or all *genes* (components) of some chromosomes in the population. On the other hand, crossover reinforce prior successes by recombining parent-chromosomes so as to produce the fittest offspring.

Although the underlying principles are relatively simple, EAs algorithms have proven to be robust and powerful search tools, owing to their remarkable flexibility and adaptability to a given task [22]. As a matter of fact, their tuning relies on a proper selection values for only a few parameters which make them very attractive and easy-to-use. Furthermore, EAs do not try to provide an exact match but an approximation of the optimal solution within an acceptable tolerance, which improve their effectiveness.

B. Multi-Resolution Approach

Whatever the optimization method, a pre-processing of the cost function can be introduced to improve the efficiency of the optimization. Indeed, the MI-based cost functions are *a priori* nonconvex and are very likely to present rugged surfaces. To limit the risk of getting trapped in a local minimum, it is common to smooth the cost function. A trade-off has to be found however between smoothness and loss of information so there is still no guarantee of finding the global optimum. The cost functions require the estimation of the pdfs: using the nonparametric Parzen windowing approach, fine estimates of the distributions are obtained with a small number of observations, but also the cost functions are smoother than what could be expected with histograms. The smoothness of the density estimates and thus the smoothness of the cost functions is controlled by the parameter σ (see Section IV). This parameter must therefore be carefully chosen: if it is too small, the cost functions are likely to be highly irregular, with a negative impact on the optimization algorithm. On the other hand, if it is too large, the loss of information and in particular, the loss of discrimination between the densities can be dramatic and may lead to a wrong solution. The smoothing parameter defined in (8) is a function of the data points y. Therefore it varies during the optimization process as the audio feature data points vary. These audio feature data points tend to evolve so that their distribution tends away from a uniform distribution. Indeed, the optimization process looks for features which maximize the MI, while possibly minimizing the joint entropy between the audio and video features, and the entropy is maximal for r.v. with uniform density. Roughly speaking, the smoothing parameter evolves as follows: at the beginning of the optimization, the audio features are scattered in the space and the smoothing parameter is thus large: the pdf, thus implicitly the objective function, is largely smoothed. Then, as the optimization proceeds, the distribution of the data points tend to concentrate in the sample space and the smoothing parameter decreases: fine structures of the pdf, thus of the objective function, appear. The use of an adaptive smoothing parameter as defined by (8) induces then a multi-resolution approach for solving the optimization problem. Multi-resolution schemes have been shown to perform better in the context of optimization problems involving MI, notably, in image registration problems (see, for example, [23]).

C. Local Optimization: Powell's Direction Set Method

In a first set of experiments, the deterministic Powell's direction set method [24] has been used. This local optimization method is well-suited for problems where no analytical formulation of the gradient is available. It presents the advantage of being fast and easy to use, but, as a local optimization method, it is very likely to fail to reach the global optimum of a nonconvex cost function.

Combining both smoothing and different initial guesses of the solution, good results have been obtained, showing that the proposed approach was able to extract audio features specific to speech production and to detect the current speaking mouth in simple audio-video sequences [25].

However, the solutions found by this method were strongly dependent on the initial conditions, a sign that the cost function still exhibited too many local optima. To ensure that the global optimum is reached, an exhaustive trial of all initial points should be performed; an approach which is, obviously, unfeasible. Consequently, a global optimization strategy turned out to be preferable. Moreover, to be efficient, this global optimization method should fulfill the following requirements.

1) Efficiency for highly nonlinear problems without requiring the cost function to be differentiable or even continuous over the search space.

2) Efficiency with cost functions that present a shallow, rough error surface.
3) Ability to deal with real-valued parameters.
4) Ability to handle the two constraints defined by (12) and (13) in the most efficient way.

D. Differential Evolution (DE)

As previously discussed, evolutionary approaches such as GAs present flexibility and simplicity of use in a challenging context and look therefore suitable to solve the problem at hand.

In order to deal with the four requirements expressed above, the adaptation of the genetic algorithm in the continuous space (GACS) developed in [26] and [27], was firstly used as a global optimization approach. The GACS, first described in [28], is an extension of the original GA scheme which uses real valued parameter vectors instead of bit strings for chromosomes. Thus, the proximity between two points in both the representation and the problem spaces is retained and the third requirement is fulfilled. The adaptation of GACS used speeds up the convergence of the algorithm by requiring the solution domain, or *acceptance domain* [$[0,1]$ for each $\vec{\alpha}_i$ in our case, as indicated by (12)], to be convex. It also relates the genetic operators to the constraints on the solution parameters.

Though better than with Powell, the results can still be improved. A loss of population diversity was observed which caused a notable difference between optima reached from one run to another (premature convergence of the algorithm) due to the difficulty of fixing the parameters as well as the difficulty of reaching a solution when this one was located close to the boundaries of the search space.

The choice of an evolutionary strategy (ES) seemed suitable to alleviate the limits of GACS. As EAs, this kind of approaches first developed by Rechenberg [29] and Schwefel [30], presents the same advantages as GACS and operates according to the same general scenario. But a strategy to exploit the local topology of the search space is considered. Usually, this consists of generating the variance of the perturbation distribution using another *a priori* defined distribution. However, the problem of choosing the right distribution as well as the right variance for this new distribution still remains to be solved.

The differential evolution (DE) approach introduced in 1997 by Storn and Price [11] belongs to this category of EAs. However, rather than applying a perturbation generated by an *a priori* defined distribution, the perturbation in DE corresponds to the difference of chromosomes (termed *vectors* in this context) randomly selected from the population. In this way, the distribution of the perturbation is determined by the distribution of the vectors themselves and no *a priori* defined distribution is required. Since this distribution depends primarily on the response of the population vectors to the topography of the cost function, the biases introduced by DE in the random walk towards the solution match those implicit in the function it is optimizing [31]. In other words, the requirement for an efficient mutation scheme is more closely met: the generated increments move the existing vectors with both suitable displacement magnitude and direction for the given generation.

The exact algorithm used is based on the so-called *DE/rand/1/bin* algorithm [31]. An initial population of N vectors is first generated to lie within the convex acceptance domain, as in the case of GACS optimization, by dividing the search space in Q predefined quantization levels [32]. A perturbed vector $\vec{\alpha'}_{G,i}, i = 1,\ldots,N$ is then generated as a counterpart for each vector $\vec{\alpha}_{G,i}$ of the current population N_G, where G refers to the current generation. This perturbed vector, or child vector, results form the linear combination of three parent vectors $\vec{\alpha}_{G,r_1}, \vec{\alpha}_{G,r_2}, \vec{\alpha}_{G,r_3}$ randomly picked up from the population N_G with $r_1 \neq r_2 \neq r_3 \neq i$ (these conditions ensure the DE mutation to be effective and not to simplify towards a classical crossover scheme [31])

$$\vec{\alpha}'_{G+1,i}(j) = \vec{\alpha}_{G,r_3}(j) + F \cdot (\vec{\alpha}_{G,r_1}(j) - \vec{\alpha}_{G,r_2}(j)) \quad (14)$$

where F is a scaling factor taking value on $[0,2]$. A user-defined crossover probability CR controls the number of child vector element indices j subject to perturbation: P random numbers belonging to $[0,1]$ are generated (i.e., one for each element of the vector under consideration); each time one of these random number is smaller than CR the corresponding vector element index is subject to a perturbation. Thereafter, the child vector differs from its parent by at least one element $(CR = 0)$ and at most, by all of its elements $(CR = 1)$.

Both the perturbed and the original populations are evaluated by the cost function and pair competitions are performed between child and parent vectors (so the population size remains constant). At the end of one iteration, a new population eventually emerges, composed by the winners of each local competition.

The constraints defined in (12) and (13) still hold. Therefore, the validity of each vector of the perturbed, or child, population should be verified before starting the decision process. If the element j of a child vector i does not belong to the acceptance domain, it is replaced by the mean between its pre-mutation value and the bound that is violated [31]. This scheme is more efficient than the simple rejection adopted with GACS. Indeed, it allows the bounds to be asymptotically approached, thus covering efficiently the whole search space. To handle the second constraint ((13)), a simple normalization is performed on each child vector, as it was done with GACS.

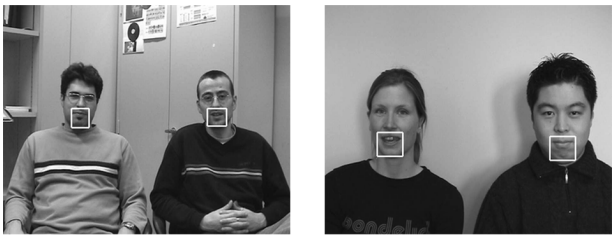
A good introduction to DE as well as some rules to tune the parameters in an adequate way can be found in [33] and [31].

Both the generation of the perturbation increment using the population itself instead of a predefined probability density function and the handling of the out-of-range values allow the DE algorithm to achieve outstanding performance in the context of our problem.

VI. Audiovisual Speaker Detection Results

A. Experimental Protocol

Two different sets of test sequences have been used. The first set of sequences is part of an in-house data set containing five audio-video sequences of duration 4 s (labelled $1,2,\ldots,5$), each shot in PAL format (25 frames/second (fps), 44.1-kHz stereo sound). In each sequence, two individuals are present, only one of them speaks during the entire sequence. Notice however that both are referred to as "speakers", since either



a b

Fig. 2. Typical frame extracted from the in-house test sequences. White rectangles delimits the extracted mouth regions. (a) Frame extracted from sequence 5. (b) Frame extracted from the third sequence.

one of them may have uttered the recorded audio. These sequences are of increasing complexity, the fifth being the most challenging with the nonspeaking individual moving randomly his head and lips. Frames extracted from two sequences are shown as an example in Fig. 2. These sequences, shot under controlled conditions, are used to test the theoretical points developed in this paper. For that purpose, the mouth regions are extracted based on manual localization initialized on the first frame. The duration of the sequences allows us to obtain a data sample set large enough to correctly estimate the pdfs. It also allows the speakers to remain still enough for the initial mouth localization to be valid throughout the sequence.

The second set of sequences is part of the CUAVE database [34]. The 11 two-speaker sequences considered, *g11* to *g22*,[1] are shot in the NTSC standard (29.97 fps, 44.1 kHz stereo sound). On these sequences each speaker utters in turn two series of digits. The final seconds of the video clips, where both speakers read simultaneously different digit strings, have not been used since signal separation is not in our goal. The first seconds present challenging properties, making the detection task difficult: in some sequences the nonspeaking person moves his lips and chin, sometimes even formulating the words without sounding them. A tradeoff has to be found between the sample size required for correctly estimating the pdf, and a detection window offering enough flexibility to correctly deal with the speaker changes. Thus, the optimization is done over a $2s$ temporal window, shifted in one second steps over the whole sequence to make decisions once per second. The mouth regions are tracked along the sequence using the face detector described in [14].

For both sequence sets, the $N \times M$ mouth regions are extracted, with N and M varying between 22 and 57 pixels, depending on speakers' characteristics and acquisition conditions. Thus the video feature set (video sample) is composed of the $N \times M \times (T-1)$ values of the optical flow norm at each pixel location (T being the number of video frames within the analysis window, i.e., $T = 100$ or $T = 60$ frames).

From the audio signal, $P = 12$ mel-cepstrum coefficients are computed using 30 ms Hamming windows [16], [17].

Considering each mouth region and its associated video features, the MFCCs are projected on a new 1-D subspace as defined in Section IV. As a result of the optimization, two sets of weights are obtained (one for each mouth region). They give the optimal linear combination of mel-cepstrum coefficients with respect to the optimization criterion (either ECC or MIC). Let us denote them $\vec{\alpha}^{\text{opt}}_{M_1}$ and $\vec{\alpha}^{\text{opt}}_{M_2}$, where the indices M_1 and M_2 indicate whether these weights result from the optimization performed on the first or second mouth region. Two corresponding audio feature sets derive from these weight sets: $F^{\text{opt}}_{A_{M_1}}$ and $F^{\text{opt}}_{A_{M_2}}$.

Two pairs of MI values can then be evaluated between the audio features and the video features in each mouth region. If $F_{V_{M_1}}$ denotes the video features of the first mouth region and $F_{V_{M_2}}$ those of the second, the two pairs of MI are given by

$$\{I(F_{V_{M_1}}, F^{\text{opt}}_{A_{M_1}}), I(F_{V_{M_2}}, F^{\text{opt}}_{A_{M_1}})\} \tag{15}$$

$$\{I(F_{V_{M_1}}, F^{\text{opt}}_{A_{M_2}}), I(F_{V_{M_2}}, F^{\text{opt}}_{A_{M_2}})\}. \tag{16}$$

First, a comparison of both MIC and ECC criteria is performed on the in-house sequences. As a result, ECC turned out to be indeed more discriminative than MIC. Therefore, ECC alone is then used on the same sequences to analyze the ability of the method to extract audio features specific to speech production and to perform speaker detection. Finally, the discussion of the results leads to the definition of a more efficient criterion ΔECC given by (11) whose performance on both sequence sets are presented and discussed in Section VI-D.

B. Comparison of Optimization Criteria MIC and ECC

The initial hypothesis is that ECC is more effective than the simpler MIC. The first set of experiments, carried out on the in-house sequence set, aims at testing this hypothesis. Therefore, the knowledge of the active mouth region is introduced *a priori* so that the optimization is only performed on this region, with each of the optimization criteria successively. Using the resulting audio feature sets, the difference of MI between the speaking mouth region and the nonspeaking one, normalized by the speaking mouth region MI (i.e., the *normalized difference of MI*), is measured for each of the five test sequences. Table I presents the results (ΔI_{MIC} and ΔI_{ECC} refer to the normalized difference of MI measured between the speaking and the nonspeaking mouth regions when using optimization criterion MIC and ECC, respectively). Two observations can be made from these results. Firstly, the MI is always greater in the active mouth region, regardless the optimization criterion used, confirming that our scheme permits the detection of the current speaker. Secondly, we see that in four cases out of five, the ECC criterion leads to larger difference between MI in the two regions. This indicates that the use of the ECC criterion gives rise to more discriminative features. Consequently, normalizing the MI by the entropy during the optimization allows to extract more specific information than using simply the MI alone, as stated in Section IV.

C. Performance Using ECC

From the first set of experiments, we may conclude that ECC is more suitable as an optimization criterion for active speaker detection. This is why in the following we will focus only on its use and analyze its properties in detail. The purpose of the experiments described here is to assess the ability of our algorithm to

[1] *g18* has been discarded as it exhibits strong noise due to the compression.

TABLE I
NORMALIZED DIFFERENCE OF MI MEASURED IN EACH MOUTH REGION FOR EACH OF THE FIVE IN-HOUSE TEST SEQUENCES, CONSIDERING THE AUDIO FEATURES EXTRACTED WITH OPTIMIZATION CRITERION MIC OR ECC, ON THE SPEAKING MOUTH REGION

Sequence	1	2	3	4	5
ΔI_{MIC}	73.54 %	76.18 %	91.67 %	69.64 %	52.13 %
ΔI_{ECC}	76.00 %	76.73 %	90.93 %	76.29 %	69.72 %

extract audio features specific to speech production and to perform speaker detection. The tests are carried out on the in-house sequences.

The capacity of the proposed method to act as a speaker detector is shown first. In contrast to the experiments described in Section VI-B, no *a priori* knowledge of the active speaker is assumed. Then the technique described in Section VI-A is applied. Recall that the optimization is performed on each of the mouth regions (M_1 and M_2) and the MI between two pairs of audio and video features is measured as stated by (15) and (16). If the approach is correct, the highest MI value should be measured between the video features of the speaking mouth and the audio features resulting from the optimization on the active speaker. The values of MI are plotted in Fig. 3. We note that for all sequences (including the challenging sequence 5), the MI measured on mouth M_1 with $\vec{\alpha}_{\text{opt}}$ optimized on this same region is always strikingly greater than all the other 3. Indeed, in all these sequences, M_1 is the speaking mouth, which gives 100% correct detections, a rather encouraging result.

Another issue that it is necessary to investigate is whether the features extracted from audio are specific to speech production. For this, the difference between the normalized MI computed on mouth regions and the corresponding audio is measured as follows:

$$\Delta I_{M_1} = \frac{\max_i \left(I\left(F_{V_{M_i}} F^{\text{opt}}_{A_{M_1}}\right)\right) - \min_i \left(I\left(F_{V_{M_i}}, F^{\text{opt}}_{A_{M_1}}\right)\right)}{\max_i \left(I\left(F_{V_{M_i}}, F^{\text{opt}}_{A_{M_1}}\right)\right)} \tag{17}$$

$$\Delta I_{M_2} = \frac{\max_i \left(I\left(F_{V_{M_i}}, F^{\text{opt}}_{A_{M_2}}\right)\right) - \min_i \left(I\left(F_{V_{M_i}}, F^{\text{opt}}_{A_{M_2}}\right)\right)}{\max_i \left(I\left(F_{V_{M_i}}, F^{\text{opt}}_{A_{M_2}}\right)\right)} \tag{18}$$

where $i = 1, 2$.

The results are listed in Table II. It can be seen that $\Delta I_{M_1} > \Delta I_{M_2}$ and $\Delta I_{M_1} > 0$ for all the sequences. But ΔI_{M_2} is sometimes negative. In other words, when the audio features are obtained on the nonspeaking mouth region, the difference of MI is sometimes favoring the nonspeaking mouth (sequences 2, 4, and 5). So when optimizing on the nonspeaking region, the features extracted cannot (and are not expected to) reflect any underlying relationship between audio and video. This result also appears in Fig. 3, since the MI measured between $F_{V_{M_1}}$ and $F_{A_{M_2}}$ is always smaller than the one measured between $F_{V_{M_1}}$ and $F_{A_{M_1}}$. Therefore, the audio features can be said to be specific to speech production.

D. Results Obtained With ΔECC on the In-House Data Set

Two optimizations were performed previously to decide who is the current speaker. They are now combined in a single optimization problem, which aims at maximizing the discrepancy between the two mouth regions. For this, the ΔECC, given by (11), is used. The result of the optimization is a vector $\vec{\alpha}_{\text{opt}}$ which generates a single audio feature vector. The latter is expected to maximize thereafter the MI with the video features of the active mouth region. This new detection approach has firstly been tested on the five in-house test sequences. Results are summarized in Table III. The normalized difference of MI is always in favor of the active speaker, i.e., the correct speaking mouth region is always indicated. It is also interesting to note that the difference of MI here is greater than what was obtained with the previous ECC optimization scheme (Table I). This stresses the benefit of using the video content related to each mouth region during the optimization.

E. Results Obtained with ΔECC on the CUAVE Database

To validate the results obtained with this simple detection scheme using ΔECC, experiments on the CUAVE database have been performed. Recall that a two second analysis window is shifted in one second steps over a given sequence. Due to the resulting overlap between the windows, the evaluation is restricted to the second half of each detection window, except for the very first window one. The results are then evaluated based on the experimental framework described in [35]: the ground truth for the evaluation window takes the label that mainly occurs over these 30 frames (*speaker 1* or *speaker 2*). Since our detector is not tuned to detect a silent state, the silent frames are not considered. The results are listed in Table IV. As a comparison, the average rate of correct detections over the 11 sequences when using a simple motion-based detector (the highest power value of the video features indicating the speaking mouth) is 60%. These results indicate that the use of both audio and video information significantly improves the detection.

It is interesting to compare these results to those presented by Nock *et al.* in [3]. They compute the MI at each pixel location, considering the difference of pixel intensity as video features, and MFCCs as audio features. In a first stage, the highest total MI value in the left and right of the image is assumed to indicate the current speaker (76% of correct detections in such a scheme). In a second experiment, the highest concentration of MI value in a $N \times M$ region indicates the speaking mouth. It is classified as correct if this region falls within a $K \times K$ pixels square centered around the true speaking mouth (K being equal to 200). They obtain 70% of correct detection with this detection scheme, which is the most directly comparable to ours since we also limit the MI measure to mouth regions. Our results are significantly better, thus the optimization of the audio features as presented in this work leads to better classification results. A comparative study of the classification results obtained with and without introducing the feature extraction step prior to the classification has been performed in [36]. As a result, the performance of the classification process is increased when audio features at the input of the classifier are an optimized linear combination of MFCCs instead of a simple average of the same MFCCs.

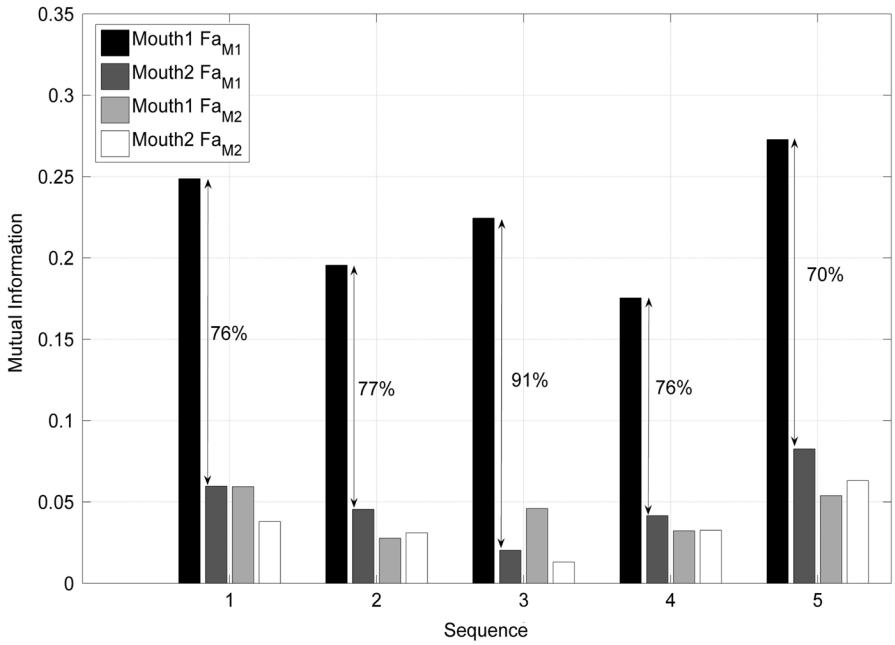


Fig. 3. MI measured between the M_1 or M_2 mouth region features and the audio features$F_{A_{M_1}}$ and $F_{A_{M_2}}$ obtained with optimization on mouth region M_1 or M_2 [(15) and (16)] for the five in-house sequences. The normalized difference of MI between the best value found and the corresponding value found in the opposite mouth is indicated.

TABLE II
NORMALIZED DIFFERENCE OF MI MEASURED BETWEEN THE M_1 AND M_2 MOUTH REGIONS WITH THE AUDIO FEATURES OBTAINED WITH OPTIMIZATION ON MOUTH REGIONS $M_1(I_{M_1})$ AND $M_2(I_{M_2})$. THE OPTIMIZATION CRITERION USED IN BOTH CASE IS ECC

Sequence	1	2	3	4	5
ΔI_{M_1}	76.00 %	76.73 %	90.93 %	76.29 %	69.72 %
ΔI_{M_2}	36.09%	-11.66	71.65 %	-0.66%	-17.28 %

TABLE III
NORMALIZED DIFFERENCE OF MI MEASURED BETWEEN THE SPEAKING AND THE NONSPEAKING MOUTH REGIONS WITH THE AUDIO FEATURES OBTAINED USING ΔECC AS COST FUNCTION (TESTS PERFORMED ON THE IN-HOUSE DATABASE)

Sequence	1	2	3	4	5
ΔI	84.23%	86.27%	95.55%	80.9%	76.15%

TABLE IV
RESULTS ON THE CUAVE SEQUENCES, USING THE EVALUATION FRAMEWORK GIVEN IN [35] WITH EVALUATION ON THE LAST SECOND OF EACH DETECTION WINDOW (SILENT WINDOWS ARE NOT CONSIDERED)

Sequence number	Correct detection rate (in %)
g11	69
g12	82
g13	53
g14	95
g15	89
g16	83
g17	100
g19	86
g20	92
g21	90
g22	95
Mean	85

VII. CONCLUSION

We have presented a method that exploits the common content of speech audio and video signals to detect the active speaker among different candidates. This method uses the information theoretic framework in a similar way to that in [1] and [2] to derive optimized audio features with respect to the video ones. No assumption is made about the distributions of the features. They are rather estimated from the samples. Moreover, no approximation of the MI-based cost functions is used but the optimization is performed in a straightforward manner using a global method, the Differential Evolution algorithm.

A study of two optimization criteria that can be used in this information theoretic framework has been carried out. Results shown that the best performing criterion (namely, ECC) is able to extract audio features that are specifically related to the speaker video features. Using only these extracted features, the algorithm performs detection of the current speaker with 100% correct detection on five in-house test sequences.

In order to optimize the detection in the case of two-people sequences, a third optimization criterion (ΔECC) has been introduced and tested on the same sequence set as before as well as on the more widely used CUAVE database [34]. This criterion aims at simplifying the detection scheme, as well as improving the audio feature specificity by taking advantage of the video information related to both mouth regions. Indeed, the resulting audio features have been shown to be even more specific than with the previous optimization criterion. A number of experiments have therefore been carried out on 11 sequences of the CUAVE database to assess and compare the performance of this ΔECC-based detection method to the results presented in [3]. In the latter, MFCCs are used as audio features, without any optimization. The better results achieved by our method show that optimizing the features improves the classifier performance.

Only two potential speakers are present in these test sequences but the method can easily be extended to sequences containing more speaker candidates using ECC as the optimization criterion. These speakers should remain face on to the camera. However, it is not a problem if they move, provided the mouth detector is able to deal with moving faces. In its actual form, the computation time does not allow the algorithm to be used in real-time applications. An example application would be multimedia indexing.

Future work aims at extending this optimization process to the video features, and introducing a silent state detection.

ACKNOWLEDGMENT

The authors would like to thank Prof. P. Vandergheynst, Prof. P. Frossard, Prof. A.C. Davison, Dr. T. Butz, Dr. X. Bresson, G. Monaci, and P. Berlin for fruitful discussions.

REFERENCES

[1] T. Butz and J.-P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Process.*, vol. 85, pp. 875–902, 2005.

[2] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 406–413, Jun. 2004.

[3] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, Urbana, IL, Jul. 2003, pp. 488–499.

[4] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Proc. NIPS*, Denver, CO, 1999, vol. 12, pp. 813–819.

[5] G. Monaci, O. D. Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," in *Proc. IEEE Int. Conf. Image Processing (ICIP'05)*, Geneva, Italy, Sep. 2005, pp. 814–820.

[6] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronisation of video facial images and audio tracks," in *Proc. NIPS*, 2001, vol. 13.

[7] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proc. ICA*, Nara, Japan, Apr. 2003, pp. 709–714.

[8] T. Butz and J.-P. Thiran, "Feature space mutual information in speech-video sequences," in *Proc. ICME*, Lausanne, Switzerland, 2002, vol. 2, pp. 361–364.

[9] E. Parzen, "On estimation of a probability density function and mode," *Ann. Mathemat. Statist.*, vol. 33, pp. 1065–1076, 1962.

[10] J. W. Fisher and J. C. Principe, "A methodology for information theoretic feature extraction," in *Proc. Int. Joint Conf. Neural Networks*, Anchorage, AK, May 1998, vol. 3, pp. 1712–1716, ser. IEEE World Congress on Computational Intelligence.

[11] R. Storn and K. Price, "Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces," *J. Global Optimiz.*, vol. 11, pp. 341–359, 1997.

[12] T. M. Cover and J. A. Thomas, , D. L. Schilling, Ed., *Elements of Information Theory*. New York: Wiley, 1991.

[13] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, 1981.

[14] J. Meynet, V. Popovici, and J.-P. Thiran, Face Detection with Mixtures of Boosted Discriminant Features, EPFL, 1015 Ecublens, Tech. Rep. 2005-35, Nov. 2005.

[15] P. Besson and M. Kunt, Information theoretic optimization of audio features for multimodal speaker detection, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, EPFL-ITS Tech. Rep. 08/2005, Feb. 2005.

[16] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York: Wiley, 2000.

[17] J. W. Picone, "Signal modeling techniques in speech recognition," in *Proceedings of the IEEE*, Sept. 1993, vol. 81, no. 9.

[18] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press, 1997.

[19] S. Kirkpatrick, C. D. Gelatt, and J. M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.

[20] F. Glover, "Future paths for integer programming and links to artificial intelligence," *Comput. and Oper. Res.*, vol. 13, no. 5, pp. 533–549, 1986.

[21] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor: Univ. of Michigan Press, 1975.

[22] T. Spalek, P. Pietrzyk, and Z. Sojka, "Application of the genetic algorithm joint with the Powell method to nonlinear least-squares fitting of powder EPR spectra," *J. Chem. Inf. Model.*, vol. 45, pp. 18–29, 2005.

[23] A. A. Cole-Rhodes, K. L. Johnson, J. LeMoigne, and I. Zavorin, "Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient," *IEEE Trans. Image Process.*, vol. 12, no. 12, pp. 1495–1511, Dec. 2003.

[24] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 1992.

[25] P. Besson, M. Kunt, T. Butz, and J.-P. Thiran, "A multimodal approach to extract optimized audio features for speaker detection," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, Sep. 2005.

[26] P. Schroeter, J.-M. Vesin, T. Langenberger, and R. Meuli, "Robust parameter estimation of intensity distributions for brain magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 17, no. 2, pp. 172–186, Apr. 1998.

[27] V. Vaerman, "Multi-dimensional object modeling with application to medical image coding," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, 1999.

[28] X. Qi and F. Palmieri, "Theoretical analysis of evolutionary algorithms with an infinite population size in continuous space, Part I: Basic properties of selection and mutation. Part II: Analysis of the diversification role of the crossover," *IEEE Trans. Neural Netw.*, vol. 5, no. 1, pp. 102–129, Jan. 1994.

[29] I. Rechenberg, Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution 1973, Stuttgart: Frommann-Holzboog.

[30] H.-P. Schwefel, *Numerical Optimization of Computer Models*. Chichester, U.K.: Wiley, 1981.

[31] K. V. Price, *New Ideas in Optimization*. New York: McGraw-Hill, 1999, ch. 6, pp. 79–108, An Introduction to Differential Evolution.

[32] Y.-W. Leung and Y. Wang, "An orthogonal genetic algorithm with quantization for global numerical optimization," *IEEE Trans. Evol. Comput.*, vol. 5, no. 1, pp. 41–53, Feb. 2001.

[33] R. Joshi and A. C. Sanderson, "Minimal representation multisensor fusion using differential evolution," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 29, no. 1, pp. 63–76, 1999.

[34] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, May 2002, vol. 2, pp. 2017–2020.

[35] P. Besson, G. Monaci, P. Vandergheynst, and M. Kunt, Experimental evalutation framework for speaker detection on the CUAVE database, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, Tech. Rep. TR-ITS- 2006.003, Jan. 2006.

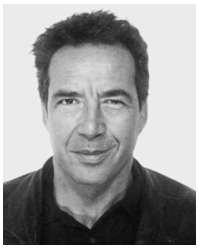
[36] P. Besson and M. Kunt, "Hypothesis testing as a performance evaluation method for multimodal speaker detection," in *Proc. 2nd Int. Workshop on Biosignal Processing and Classification (BPC2006), ICINCO*, Setubal, Portugal, 2006, pp. 106–115.

Patricia Besson was born in Charenton-le-Pont, France, in July 1977. She received the M.Sc. degree in biomedical engineering from the University of Lyon (UCBL), Lyon, France, in June 2001 and the Ph.D. degree from the Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in June 2007.

She spent the 2000–2001 academic year as an exchange student at the University of Montreal (UdM), Montreal, QC, Canada. In June 2002, she joined the ITS, EPFL, where she started a thesis on the multimodal detection of the speaker in audiovisual sequences under the supervision of Prof. Murat Kunt. In October 2007, she will integrate the Laboratory "Motion and Perception" at the National Center for Scientific Research (CNRS), Marseille, France as a Postdoctoral Researcher. Her research will focus on the modelization of the multisensorial perception and control of self-orientation in space by humans.

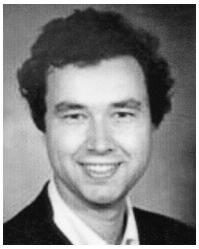
Vlad Popovici received the Eng. and M.Sc. degrees in computer science from the Technical University of Cluj-Napoca, Romania, in 1998 and 1999, respectively, and the Ph.D. degree from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2004.

He is Researcher with the Swiss Institute of Bioinformatics (SIB), working on different, theoretical, and applied aspects of statistical pattern recognition. Before joining SIB, he was with the Digital Media Institute, Tampere University of Technology, as Research Scientist, and with the Signal Processing Institute, EPFL as an Assistant Researcher and later as Postdoctoral Researcher. His current research focusses on sparse classifiers andmultiple clsissifier systems, and their applications to life sciences.



Jean-Marc Vesin (M'98) graduated from the Ecole Nationale Superieure d'Ingenieurs Electricians de Grenoble (ENSIEG), Grenoble, France, in 1980. He received the M.Sc. degree from Laval University, Quebec City, QC, Canada, in 1984, where he spent four years on research projects. He received the Ph.D. degree from the Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), in 1992.

He was previously involved in research projects at Laval University and later spent two years working in industry. He is currently in charge of the activities in 1-D signal processing at ITS, EPFL. His main interests are biomedical signal processing, adaptive signal modelling and analysis, and the applications of genetic algorithms in signal processing. He has authored or co-authored more than 30 publications in renowned peer-reviewed journals, as well as several book chapters.



Jean-Philippe Thiran (M'93–SM'05) was born in Namur, Belgium, in 1970. He received the Elect.Eng. and Ph.D. degrees from the Universite catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 1993 and 1997, respectively.

He joined the Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in February 1998 as a Senior Lecturer. Since January 2004, he has been an Assistant Professor, responsible for the Image Analysis Group. His current scientific interests include image segmentation, prior knowledge integration in image analysis, partial differential equations and variational methods in image analysis, multimodal signal processing, medical image analysis, including multimodal image registration, segmentation, computer-assisted surgery, and diffusion MRI. He is author or co-author of two book chapters, 51 journal papers, and some 90 peer-reviewed papers published in proceedings of international conferences. He holds four international patents.

Dr. Thiran was Co-Editor-in-Chief of *Signal Processing* (published by Elsevier Science) from 2001 to 2005. He is currently an Associate Editor of the *International Journal of Image and Video Processing* (published by Hindawi), and member of the Editorial Board of *Signal, Image and Video Processing* (published by Springer). He will be the General Chairman of the 2008 European Signal Processing Conference (EUSIPCO 2008).

Murat Kunt (SM'70–M'74–SM'80–F'86) was born in Ankara, Turkey, in 1945. He received the M.S. degree in physics and the Ph.D. degree in electrical engineering, both from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1969 and 1974, respectively.

From 1974 to 1976, he was a Visiting Scientist at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, where he developed compression techniques for X-ray images and electronic image files. In 1976, he returned to the EPFL where he is presently a Professor of Electrical Engineering and Director of the Signal Processing Institute (ITS), one of the largest at EPFL. He conducts teaching and research in digital signal and image processing with applications to modeling, coding, pattern recognition, scene analysis, industrial developments and biomedical engineering. His laboratory participates in a large number of European projects under various programmes such as Esprit, Eureka, Race, HCM, Commett and Cost. He is the author or the co-author of more than 200 research papers and 15 books, and holds seven patents. He supervised more than 60 Ph.D. students, some of them being today university professors. He consults for governmental offices, including the French General Assembly.

Dr. Kunt has been the Editor-in-Chief of *Signal Processing* for 28 years and is a founding member of EURASIP, the European Association for Signal Processing. He is now the Editor-in-Chief of *Signal, Images and Video Processing* (Springer) and serves as a Chairman and/or a member of the Scientific Committees of several international conferences and on the editorial boards of the PROCEEDINGS OF THE IEEE, *Pattern Recognition Letters*, and *Traitement du Signal*. He was the Co-Chairman of the first European Signal Processing Conference held in Lausanne in 1980 and the General Chairman of the International Image Processing Conference (ICIP'96) held in Lausanne in 1996. He was the President of the Swiss Association for Pattern Recognition from its creation until 1997. He received the Gold Medal of EURASIP for meritorious services, the IEEE Acoustics, Speech, and Signal Processing's Technical Achievement Award, the IEEE Third Millennium Medal, an honorary doctorate from the Catholic University of Louvain, the Technical Achievement Award of EURASIP, and the Imaging Scientist of the Year Award of the IS&T and SPIE in 1983, 1997, 2000, 2001, and 2003, respectively.