Extraction of audio features specific to speech in multimodal speaker detection

Patricia Besson*, Vlad Popovici, Jean-Marc Vesin, Jean-Philippe Thiran, Murat Kunt
Signal Processing Institute (ITS)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne
Telephone: +41 21 693 26 01
Fax: +41 693 76 00
Email: patricia.besson@epfl.ch

July 2005

Abstract

We present a method that exploits an information theoretic framework to extract optimized audio features using video information. A simple measure of mutual information (MI) between the resulting audio features and the video ones allows to detect the active speaker among different candidates.

Our method involves the optimization of an MI-based objective function. No approximation is introduced to solve this optimization problem, neither for the estimation of the probability density functions (pdf) of the features, nor for the cost function itself. The pdf are estimated from the samples using a non-parametric approach. As far as the optimization process itself is concerned, three different optimization methods (one local and two global) are compared in this paper. The Differential Evolution algorithm is eventually retained as it outperforms the other methods.

Two information theoretic optimization criteria are compared and their ability to extract audio features specific to speech is discussed. As a result, our method achieves a speaker detection rate of 100% on our test sequences, and of 95% on a most commonly used one.

I. INTRODUCTION

With the increasing capabilities of nowadays computers, both auditive and visual modalities of the speech signal may be used to improve speaker detection, leading to major improvements of the user-friendliness of man-machine interactions. Let us just consider for example a videoconference 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 of interest on the basis of sound and image information could focus a moving camera on her/him.

Among the different methods that exploit the information contained in each modality, a few are performing the fusion directly at the feature level. 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. An estimation of the features' probability density functions (pdf) is therefore required and there are two main approaches that may be taken: either a parametric or a non-parametric one. In the first case, the pdf's 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.

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, the audio features are optimized with respect to the video features. Following Butz and Thiran in [1] and [8], we cast our problem in an information theoretic framework to optimize the audio features with respect to the video features. The objective function to be optimized is therefore based on mutual information,

which turns out to be a highly nonlinear optimization problem. Moreover, an analytical formulation of the gradient of the objective 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 mutual information 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 proper optimization method. Therefore, a local optimization scheme, namely the Powell's method [9], has been tried in a first step. To alleviate the limits encountered with this optimization method, Evolutionary Algorithms methods (Genetic Algorithm in Continuous Space [10] and Differential Evolution [11]) have then been applied and their performance compared and analyzed.

The paper is organized as follows: first the use of information theory to extract optimized features in general unimodal, then multimodal, classification problems is presented. After that, 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 to the classifier. Different optimization criteria based on mutual information are defined. The fourth part exposes the optimization problem as well as the local and the global optimization methods used to solve it. Comparison and analysis of the results obtained with each of the three methods are given. 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, 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 understood as a classification problem.

Following [12] and [1], a general unimodal classification task is formulated as a first order Markov chain process:

$$O \longrightarrow S \xrightarrow[\substack{\textit{feature}\\ \textit{generation}}]{} X \xrightarrow[\substack{\textit{feature}\\ \textit{extraction}}]{} F_X \xrightarrow{\textit{estimation}} \hat{S} \xrightarrow{\textit{classification}} \hat{O}, \tag{1}$$

where O, S, X, F_X and $\hat{S}, \hat{O}$ are random variables (r.v.) . The four first ones stand respectively for the possible classes (defined over the set Ω_O), the physical signal, the observed data, and the feature extracted from the initial feature space, while the two latest are the signal estimated from F_X, defined over Ω_S, and the class estimated from $\hat{S}$. Notice that, as pointed out by the Markov Chain of Eq. (1), the physical signal S itself is not directly observable but through measurements X. It must therefore be estimated from these measurements X or from some features extracted from X, so as to finally being able to estimate the class of this physical phenomenom. Ultimately, the goal in such a classification process is obviously to minimize the probability of assigning the wrong class to the signal. That is, to minimize the classification error probability $P_E = P(\hat{O} \neq O)$ associated to the Markov Chain of Eq. (1).

This error probability depends of course on the classifier and on its ability to deal with the problem at hand, but it also depends on all the processing steps leading from O to $\hat{O}$. In particular, it depends on the estimation process leading from S to $\hat{S}$ and thus on the feature extraction step. The Markov chain of Eq. (1) clearly shows that whatever the classifier, its performance will be poor if the feature F_X extracted from X is bad, resulting in a poor estimation of $\hat{S}$.

Using Fano's inequality, it is possible to relate the probability of committing an error when estimating the discrete r.v. $\hat{S}$ from another r.v. F_X to the conditional entropy $H(S|F_X)$ [13]:

$$P_e \geqslant \frac{H(S|F_X) - 1}{\log |\Omega_S|} = \frac{H(S) - I(S, F_X) - 1}{\log |\Omega_S|}, \tag{2}$$

where $H(S)$ and $I(S, F_X)$ stand respectively for the Shannon's entropy of S and for the Shannon's mutual information between the random variables S and F_X, and $|\Omega_S|$ is the cardinality of S.

The inequality (2) does not allow to directly minimize the error probability P_E. It indicates however that an efficient minimization of P_E is conditioned by an efficient minimization of P_e (called here the estimation error

probability) which is itself conditioned by the minimization of the right hand side of the inequality (2). The minimization of this term implies considering the process leading to the extracted feature F_X, as stated before. This way, the error made in the previous process steps (feature generation and extraction) are taken into account as well and constraints that might improve these stages are introduced. Whatever the classifier used, its input would therefore be fed in with the most significant features from the classification task point of view.

Notice that the estimation of a r.v. from another r.v., such as the estimation of S from F_X can be viewed as a feature extraction step, where the objective is to extract from the initial r.v. the information specific to the classification task to be achieved [12], [1]. Then, the objective is to minimize the conditional entropy $H(S|F_X)$ which corresponds to the information that is present in S and not present in F_X, thus possibly missing for obtaining a good estimate $\hat{O}$ of O from $\hat{S}$. If the mapping is deterministic, this conditional entropy has its minimal possible value.

B. Extension of the information theoretic feature extraction to the multimodal case

Butz *et al.* in [1] have shown that the previous line of reasoning holds in a multimodal case, where two signals have the same physical origin and share some common information.

In particular, in the case of a speaker detection problem, audio and video signals are jointly emitted during the speech production process and these two modalities can be used to constrain the feature extraction step, as it will now be shown.

Let O be a binary random variable which models the membership to the "speaker" or "non-speaker" class with respect to an audio-visual source modelled by the random variable S, defined on Ω_S. Notice that the probability for any prototype to belong to each class is the same, and is equals to $1/|\Omega_O|$, where $|\Omega_O|$ is the cardinality of O. The bimodal source S is not directly accessible but yields two observed signals of different physical nature: the audio and video signals A and V. For each of those 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)), as previously described. Two classification error probabilities with the corresponding two lower bounds can then be derived for each Markov chain. By performing a fusion at the decision or at the classification level, a unique estimate $\hat{O}$ of the class can possibly be obtained at the end of each unimodal signal process.

However, such an approach would not take advantage of the discriminant information offered by the bimodal nature of the source S. Indeed, as mentioned by Fisher and coworkers in [2], the two measurements A and V are each one affected by independent interfering sources, denoted here N_A and N_V. The measurements coming from these sources account here for noise since they do not contain any information shared by both modalities. The classification process is then described through two Bayesian networks as shown on Fig. 1.(b). To get a good estimation $\hat{S}$ of the source (and then a good estimation $\hat{O}$ of the class), the classification process should include a step where features F_A and F_V are extracted from A and V respectively. This feature extraction step should try to recover the information present in each modality which originates from the common source S while discarding the noise coming from the interfering sources N_A and N_V. Obviously, such goal can only be reached by considering both modalities all together. The resulting extracted features specifically describe the common source and are therefore related by their joint probability $p(F_A, F_V)$ [1]. Thus, an estimate of the feature related to one modality can be inferred from the other modality. Now, given that such features F_A and F_V can be extracted, this results in carrying out a multimodal classification process described by two first order Markov chains, as shown on Fig. 1.(c), where the transition probabilities for $F_A \longrightarrow \hat{F}_V$ and $F_V \longrightarrow \hat{F}_A$ are obtained by joint probability estimation (since $p(\hat{F}_V|F_A) = p(\hat{F}_V, F_A)/p(F_A)$, and $p(\hat{F}_A|F_V) = p(\hat{F}_A, F_V)/p(F_V)$). Notice that the estimates of the source associated to each chain are indexed by *AV* or *VA*, to stress that these estimates have been obtained using information present in both modalities, in contrast with the previous case (Fig. 1.(a)). Applying the framework described in Sec. II-A for a unimodal classification process to these Markov chains, two estimation error probabilities P_{e_1} and P_{e_2} as well as two corresponding lower bounds can be defined:

$$P_{e_{1,2}} = P(\hat{S}_{AV,VA} \neq S), \tag{3}$$

$$P_{e_{1,2}} \geqslant \frac{H(S) - I(S, \hat{F}_{V,A}) - 1}{\log |\Omega_S|}. \tag{4}$$

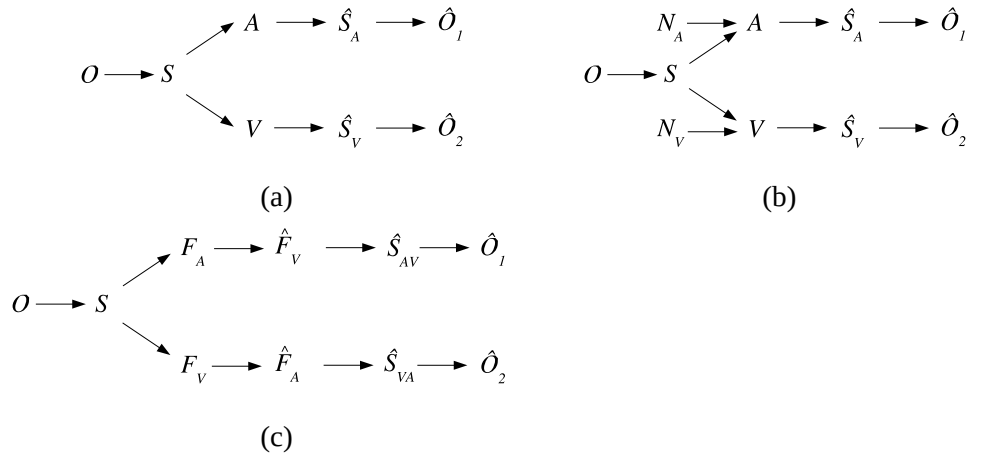

Fig. 1. (a) Graphical representation of the Markov chains modeling the two unimodal classification process associated to each modality; (b) Graphical representation of the Bayesian networks modeling the two unimodal classification process associated to each modality; (c) Graphical representation of the related Markov chains modeling the multimodal classification process.

From the data processing inequality for Markov chain [13], the following inequalities can be stated:

$$I(F_A, \hat{F}_V) \geqslant I(S, \hat{F}_V), \tag{5}$$

and in a similar way, for the second Markov chain:

$$I(F_V, \hat{F}_A) \geqslant I(S, \hat{F}_A). \tag{6}$$

As a result, the bounds on the error probabilities can be weakened [1]:

$$P_{e_1} \geqslant \frac{H(S) - I(F_A, \hat{F}_V) - 1}{\log |\Omega_S|}, \tag{7}$$

$$P_{e_2} \geqslant \frac{H(S) - I(F_V, \hat{F}_A) - 1}{\log |\Omega_S|}. \tag{8}$$

Since the probability densities of $\hat{F}_A$ and F_A, respectively $\hat{F}_V$ and F_V, are both estimated from the same data sequence A, respectively V, it is possible to introduce the following approximations: $I(F_A, \hat{F}_V) \approx I(\hat{F}_A, F_V) \approx I(F_A, F_V)$. Therefore, lower bounds on the estimation error probabilities involving the mutual information between the extracted features can be defined:

$$P_{e1} \geqslant \frac{H(S) - I(F_A, F_V) - 1}{\log |\Omega_S|}, \tag{9}$$

$$P_{e2} \geqslant \frac{H(S) - I(F_A, F_V) - 1}{\log |\Omega_S|}. \tag{10}$$

Because of the symmetry property of mutual information, the bounds of Eqs. (9) and (10) are equivalent and a joint lower bound $P_{\{e_1,e_2\}}$ can finally be defined:

$$P_{\{e_1,e_2\}} \geqslant \frac{H(S) - I(F_A, F_V) - 1}{\log |\Omega_S|}. \tag{11}$$

The cardinality $|\Omega_S|$ of S is supposed to remain fixed during the optimization. Consequently, $H(S)$ remains constant: $H(S) = \log |\Omega_S|$ so that Eq. (11) becomes:

$$P_{\{e_1,e_2\}} \geqslant 1 - \frac{I(F_A, F_V) + 1}{\log |\Omega_S|}. \tag{12}$$

Minimizing the lower bound on $P_{\{e_1,e_2\}}$ comes then eventually to maximizing the mutual information 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 produces therefore optimized features.

However for this last statement to be true, not only the mutual information $I(F_A, F_V)$ 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. Or in other words, the information related to the noise interferences N_A and N_V would be considered in the newly defined features rather than the information coming from the common source S. Dividing Eq. (11) by the joint entropy $H(F_A, F_V)$, a feature efficiency coefficient [1] can be defined:

$$e(F_A, F_V) = \frac{I(F_A, F_V)}{H(F_A, F_V)} \in [0, 1]. \tag{13}$$

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 Eq. (8) while constraining inter-feature independencies. In other words, the extracted features F_A and F_V will tend to specifically capture the information related to the common origin of A and V, discarding the unrelated interference information coming from N_A and N_V.

Applying this framework to extract features, the bound on the estimation error probability is minimized. However, there is no guarantee that this bound is reached during the classification process: 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 [6], [5], [4]. In [4] in particular, the authors interpret synchrony as the degree of mutual information between audio and video signals. Mutual information shows also good performance in detecting synchronized audio-video sources such as speakers [2], [8], [3]. Moreover, the feature optimization pre-processing also indicates MI-based classifier as a good choice. For these reasons, the chosen classifier consists in the evaluation of the MI between candidates audio and video features. The features that exhibit the largest MI are classified as "speaker", while the other one are labeled as "non-speaker", only one "speaker" class label being authorized per estimation.

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

The presented framework combines therefore both feature-level fusion (for the feature optimization) and classifier-level fusion between the two modalities.

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 to choose a representation for the signals.

It has been shown in [8] that the audio signal is more related to the pixel intensity changes than to the raw pixel intensities themselves. 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 [14] 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. Due to the small sample size and the high dimensionality of the video features, we may have difficulties in estimating the pdf (needed in mutual information computation). Thus, only the magnitude of the optical flow and the sign of the vertical component are kept.

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. Speakers are observed over a sequence of T frames resulting in $T-1$ video feature vectors V_t ($t = 1, \ldots, T-1$) where each element of these vectors is an observation of the random variable V. These vectors are normalized for the subsequent optimization (see [15] for details). This approach implicitly considers the observation to be identically independent distributed (i.i.d), which is obviously a simplification of the real world. Indeed, the neighboring pixels are correlated. This simplification is somehow compensated by estimating the pdf with the Parzen window approach [16] (see below).

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 [17], [18]. Finally, the speech signal is represented as a set of $T-1$ vectors $\vec{C}_t$, 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).

IV. Extraction of optimized speech audio features

A. Audio feature optimization

In principle, the information theoretic feature extraction discussed in Sec. II can now be used for audio and video features F_A and V. However, over $T-1$ frames, the dimensionality of the audio features is still too high to be efficiently tractable. Consequently, the one-dimensional (1D) audio features $F_{A,t}(\vec{\alpha})$, associated to the random variable F_A are built as the following linear combination of the P Mel-Frequency Cepstral Coefficients (MFCCs):

$$F_{A,t}(\vec{\alpha}) = \sum_{i=1}^{P} \vec{\alpha}(i) \cdot C_t(i) \quad \forall t = 1, \ldots, T-1, \tag{14}$$

where the weights $\vec{\alpha}(i)$ are chosen such that $\sum_{i=1}^{P} \vec{\alpha}(i) = 1$ and $\vec{\alpha}(i) \geqslant 0 \quad \forall i = 1, \ldots, P$. Thus, the set of $P \cdot (T-1)$ parameters is reduced to $1 \cdot (T-1)$ values $F_{A,t}(\vec{\alpha})$. The minimization of the estimation error given by Eq. (8) will lead to the optimized vector $\vec{\alpha}$. This optimization therefore requires the availability of the joint probability density function (pdf) as well as of the marginal distributions of the r.v. F_A and V. These distributions are obviously unknown. To avoid any restrictive assumption, they are estimated using Parzen windowing:

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} h(y - y_i; \sigma) \quad \forall y \in \Omega_Y, \tag{15}$$

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 2D Gaussian kernel of mean $[\mu_A,\mu_V]^T$ and diagonal covariance matrix diag$(\sigma_A;\sigma_V)$, $G(\mu_A, \mu_V, \sigma_A, \sigma_V)$, 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 [19]:

$$\sigma = \left(\frac{4}{3n}\right)^{1/5} \cdot \frac{\text{median } |y_i - \tilde{\nu}|}{0.6745}, \tag{16}$$

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 will adapt to the audio features during the optimization process.

Using the Parzen window to estimate the densities in a non-parametric way yields a better estimate than histogram-based approaches, given the small number of samples at our disposal ($T-1$ for the random variable associated with the audio features).

B. Optimization criteria

As exposed in Sec. II, minimizing the lower bound on the estimation error 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:

$$\begin{aligned} \vec{\alpha}_{opt} &= \arg\max_{\vec{\alpha}} \{I(V, F_A(\vec{\alpha}))/H(F_A(\vec{\alpha}))\} \\ &= \arg\max_{\vec{\alpha}} \{e(V, F_A(\vec{\alpha}))\}. \end{aligned} \tag{17}$$

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

To verify the necessity of normalizing the mutual information 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:

$$\vec{\alpha}_{opt} = \arg\max_{\vec{\alpha}} \{I(V, F_A(\vec{\alpha}))\}. \tag{18}$$

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 M_1 and 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 V^{M_1} and V^{M_2} denote the random variables associated to regions M_1 and M_2 respectively, then the optimization problem becomes:

$$\vec{\alpha}_{opt} = \arg\max_{\vec{\alpha}} \left\{ [e(V^{M_1}, F_A(\vec{\alpha})) - e(V^{M_2}, F_A(\vec{\alpha}))]^2 \right\}. \tag{19}$$

V. Optimization method

A. Definition of the optimization problem

We saw that the extraction of optimized audio features with respect to our classification task requires to find the real-valued vector $\vec{\alpha} \in \mathbb{R}^P$, that minimizes the chosen objective function $f(\vec{\alpha})$. This objective function is defined as the negative value of one of the optimization criteria defined in Eqs. (17), (18), or (19). Moreover, to restrain the set of possible solutions, the P weighting coefficients $\{\alpha_i\}_{i=1,\ldots,P}$ must observe the following conditions:

$$0 \leq \vec{\alpha}(i) \leq 1 \quad \forall i = 1, 2, \ldots, P \,, \tag{20}$$

$$\sum_{i=1}^{P} \vec{\alpha}(i) = 1. \tag{21}$$

This optimization problem is highly nonlinear and gradient-free. Indeed, an analytical formulation of the gradient of the objective 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 mutual information 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 to be 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) [20], Tabu Search (TS) [21], 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 [22], 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 enable 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 fittest offsprings.

Although the underlying principles are relatively simples, EAs algorithms have proven to be robust and powerful search tools, owing to their remarkable flexibility and adaptability to a given task [23]. As a matter of fact, their tuning relies on a proper selection of only a few parameter values 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. Local Optimization: Powell's direction set method

In a first set of experiments, we have used the deterministic Powell's direction set method [9]. This optimization algorithm is well-suited for problems where no analytical formulation of the gradient is available. It finds the minimum of a multidimensional cost function by solving sequences of one-dimensional minimizations (using for example the one-dimensional Brent's optimization method) along N linearly independent, mutually conjugate set of directions. This method belongs however to the category of local optimization methods: if the surface of the cost function is not smooth and exhibits several local optima, the ability of the algorithm to reach the global optimum relies on a judicious initial guess of the solution.

Indeed, our MI-based objective functions are *a priori* non-convex 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 find however between smoothness and loss of information so there is still no guarantee of finding the global optimum. The objective functions require the estimation of the pdf: using the non-parametric Parzen windowing approach, we do not only obtain fine estimates of the distributions with a small number of samples, but also smoother objective functions than what could be expected with histograms. The smoothness of the density estimates and thus the smoothness of the objective functions is controlled by the parameter σ (see Sec. IV). This parameter must therefore be carefully chosen: if it is too small, the objective 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. Therefore we have introduced an adaptive scheme for the estimation of the audio feature density function, in which σ is varied at each iteration (Eq. (16)). Roughly speaking, the smoothing parameter evolves as follow: at the beginning of the optimization, the audio features are scattered in the space and the smoothing parameter is thus large. Then, as the optimization proceeds, the sample tends toward a unimodal distribution and the smoothing parameter decreases. Therefore, the optimization problem is solved using a multi-resolution scheme. Such an approach has been shown to perform better in the context of optimization problems involving mutual information, notably, in image registration problems (see for example [24]).

Combining both smoothing and different initial trials, we obtained good results, showing that our framework was able to extract audio features specific to speech. The mutual information measured thereafter between the extracted audio features and the video features of different mouth regions indicated the current speaking mouth in simple audio-video sequences [25].

However, the solutions found by this method were strongly dependent on the initial conditions, showing that the objective function still exhibited too many local optima. Therefore the method was not performing at its best level. To ensure the global optimum to be 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 objective functions that present a flat, rough error surface;
3) Ability to deal with real-valued parameters;
4) Ability to handle the two constraints defined by Eqs. (20, 21) in the most efficient way;

C. Global optimization: Genetic Algorithm in Continuous Space (GACS)

An evolutionary approach such as genetic algorithm (GA) answers the two first demands previously defined while presenting flexibility and simplicity of use in a challenging context. Conventional GAs however have difficulties to handle the third and fourth requirements because they encode the solutions under the form of quantized and binarized representations (the *chromosomes*). Hence, working with real-valued parameters requires additional bits in chromosome representation to improve the precision, increasing the computational cost. Moreover, the crossover is likely to produce out-of-range values. Thus a validity test is required, decreasing the efficiency of the process. Finally, possible links between different solution parameters are ignored during crossover, slowing down once again the convergence process [26].

The genetic algorithm in the continuous space (GACS), an extension of the original GA scheme first described in [10] and [27], alleviates these limitations by using the real valued parameter vectors instead of bit strings of

chromosomes. This floating point representation presents the obvious advantage of retaining the proximity between two points in both the representation and the problem spaces. The final requirement (4^{th}) still has to be fulfilled, namely efficient handling of the constraints defined by Eqs. (20) and (21).

The adaptation of GACS developed in [28] and [29], relates the genetic operators to the constraints on the solution parameters. It also speeds up the convergence of the algorithm by requiring the solution domain to be convex (the *acceptance domain*). This domain is problem-dependent and has to be defined accordingly. At generation $t+1$, a mutated vector $\vec{\alpha}_{t+1,k}$ (with $k = 1, \ldots, N$) is then generated from a chromosome $\vec{\alpha}_{t,k}$ selected from the old population at generation t, by performing the following addition:

$$\vec{\alpha}_{t+1,k}(i) = \vec{\alpha}_{t,k}(i) + \epsilon, \tag{22}$$

where ϵ, the increment, is a zero-mean Gaussian perturbation which is applied to one element i of the chromosome vector that will mutate, with i randomly selected in the set $\{1, \ldots, P\}$. This scheme has shown to be more efficient in our case than mutating all the elements of the given chromosome vector at once.

For the mutation to be effective, that is, to eventually lead to improvement in the future populations by permitting the exploration of new regions of the search space, the variance of the Gaussian perturbation must be adequately chosen. A suitable value can be defined based on the acceptance domain for each element (i.e. $[0, 1]$ in our case, as indicated by Eq. (20)), as a certain fraction of this range. Note that it is necessary to check if the mutated gene still belongs to its acceptance domain. If it is not the case, the mutation is rejected. The role of crossover is to reinforce the prior successes by merging the good characteristics of two chromosomes using a linear combination of candidates. To ensure that the recombined chromosome $\vec{\alpha}_{t+1,k_1}$ belongs to the above-defined acceptance domain, the crossover operator is defined as follow:

$$\vec{\alpha}_{t+1,k_3}(i) = \lambda \cdot \vec{\alpha}_{t,k_1}(i) + (1-\lambda) \cdot \vec{\alpha}_{t,k_2}(i), \tag{23}$$

where $\vec{\alpha}_{t,k_1}$ and $\vec{\alpha}_{t,k_2}$ refer to two parent chromosome vectors at generation t, λ and i are randomly selected in the set $\{1, \ldots, P\}$. Since λ remains fix for each crossover operation, the search space is convex. Then the new chromosome vector $\vec{\alpha}_{t+1,k_3}$ is guaranteed to be valid if $\vec{\alpha}_{t,k_1}$ and $\vec{\alpha}_{t,k_2}$ are valid as well.

Finally, to ensure that the constraint defined by Eq. (21) is satisfied, all the chromosomes of the new generation are normalized. This implies *de facto* that each gene of each chromosome (excluding the replicated best one) in the new population is finally modified at the end of the iteration.

The specific evolution strategy implemented for the application addressed here is an extension of the scheme given in [28] and [29]. It is presented in Fig. (2) and can be summarized as follow:

1) Generate an initial population of N chromosomes (with N odd number) within the convex acceptance domain. Instead of randomly distribute the initial chromosome vectors in the search space, they are regularly placed in the acceptance domain according to a user-defined number of quantization levels Q [30].
2) Rank the chromosomes according to the evaluation (fitness) function, given by one of Eqs. (17), (18), (19). Reproduction is performed by keeping unchanged the best one for the next generation.
3) The remaining chromosomes then compete in pairs. Local pair-competitions for crossover are performed between a mutated and a crossovered chromosome of the previous generation. Crossover, using Eq. (23) is then applied to the winners of these local competitions until $(N-1)/2$ new chromosomes are generated and included in the next generation. Contrary to global competition, these local competitions allow the algorithm to preserve genetic diversity in the succeeding generations.
4) Complete the next generation by mutation of the best ranked chromosome $(N-1)/2$ times, using Eq. (22). These chromosomes combined with $(N-1)/2$ new chromosomes produced by crossover and the best ranked chromosome form the N chromosomes for the next generation. If the new chromosomes do not lie in the acceptance domain, reject the mutation.
5) Normalize the new parameters vectors such that the sum of the vector elements equals 1.
6) A stagnation of the best (reproduced) chromosome over a certain number of generations (typically 10 in our case) may indicate that the algorithm has reached a local extremum. To avoid such a situation, all chromosomes but the best one are in this case reset to random values.
7) Steps 2 to 6 are reiterated until the pre-defined maximum number of generations has been reached.

This evolution strategy is guaranteed not to diverge since the best chromosome is retained for the succeeding generations. Thus the GACS behaves at least like a random search process in a bounded search space. Note that

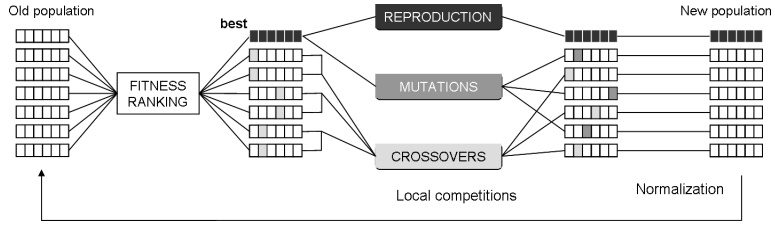

Fig. 2. Population renewal policy in GACS: reproduction, mutations and crossover (based on [29]).

unlike conventional optimization methods where the decrease of cost function over successive iterations can be used as the criterion to terminate the process, it is more difficult to assess the convergence in GACS since the stagnation in the cost function does not necessarily mean that the optimum is reached. Good results have been obtained using GACS. In particular, the optima reached were quite better than those obtained with Powell's method. However, the choice of an appropriate value for the parameters, especially the number of generations and the variance of the Gaussian perturbation still is a weak point. The latter has to be relatively high for the algorithm not to get stucked in local minima (i.e. to efficiently explore the search space). However, the highest the variance, the more likely a mutated parameter to fall outside the acceptance domain. As a result, the number of rejected mutations is too high for the population to preserve its diversity along the generations. Therefore, the mutation operator is not much more efficient with a high perturbation variance than with a small one. On some runs, only crossover maintains the evolution process active. Moreover, our solutions have proven to be sometimes very close to the boundaries of the search space. However, it is unlikely to approach the boundaries of the acceptance domain. As a result, a lost of the population diversity is observed which caused a noticeable difference between optima reached from one run to another.

What is needed is a scheme where the mutations applied are small for some parameters, and larger for others, allowing a better exploration of all the search space, including the region close to the boundaries, *i.e.* the perturbations need to adapt to the population evolution.

D. Differential Evolution (DE)

To overcome the problems encountered with GACS, the Differential Evolution approach (DE) introduced in 1997 by Storn and Price [11] has been used. As an evolutionary algorithm, it presents the same advantages than GACS and operates according to the same general scenario. The core difference between the two methods lies in the way the perturbation is generated. Rather than applying a perturbation generated by an *a priori* defined distribution as in the case of GACS, the perturbation in DE corresponds to the difference of chromosomes (rather called *vectors* in this context) randomly selected from the population. This way, the distribution of the perturbation is determined by the distribution of the vectors themselves. Since this distribution depends primarily on the response of the population vectors to the objective function topography, 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 value and direction for the given generation.

The exact algorithm we used is based on the so-called *DE/rand/1/bin* algorithm [31]. Its pseudo-code, including the modifications for handling the constraints, is given on Algorithm 1. Let us describe here more in detail the different steps of the DE algorithm. 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 [30]. 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 in 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]). The user-defined crossover probability CR controls the number of child vector element indices 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 inferior 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 ($CR = 1$). Lines 2 to 11 of the Algorithm 1 sum up these operations.

Both the perturbed and the original populations are evaluated by the objective 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 decision process is described in lines 8 to 11 of the algorithm.

The constraints defined in Eqs. (20, 21) still hold. Therefore, the validity of each vector of the perturbed, or child, population has to 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] (lines 12 to 19 of the algorithm, where $\alpha^{(lo)}(j)$ and $\alpha^{(hi)}(j)$ refer respectively to the lowest and highest bounds defined for the j^{th} parameter - that is, 0 and 1 in our case). This scheme is more efficient than the simple rejection adopted with GACS. Indeed, it allows to asymptotically approach the bounds, thus covering efficiently the whole search space. To handle the second constraint (Eq. 21), a simple normalization is performed on each child vector, as it was done with GACS (lines 20-21 of the algorithm).

A good introduction to DE as well as some rules to tune the parameters in an adequate way can be found in [32] and [31].

Algorithm 1: DE/rand/1/bin with modification for handling the constraints given by Eqs. (20, 21). Based on [31]

Input: $P, G_{max}, N \geq 4, F$ (scaling factor) $\in [0,2], CR \in [0,1], \vec{\alpha}^{(lo)}, \vec{\alpha}^{(hi)}$.
Initialize: initialization of the population;
$i = \{1, 2, \ldots, N\}$, $j = \{1, 2, \ldots, P\}$, $G = 0$;
while $G < G_{max}$ **do**
  for $i = 1, \ldots, N$ **do**
    Mutate and recombine:
    randomly select r_1,r_2,$r_3 \in \{1, 2, \ldots, N\}$, s.t. $r_1 \neq r_2 \neq r_3 \neq i$;
    $j_{rand} \in \{1, 2, \ldots, P\}$, randomly selected once each i;
    for $j = 1, \ldots, P$ **do**
      s=rand([0,1))
      if $s < CR \vee j = j_{rand}$ **then**
        $\vec{\alpha}'_{G+1,i}(j) = \vec{\alpha}_{G,r_3}(j) + F \cdot (\vec{\alpha}_{G,r_1}(j) - \vec{\alpha}_{G,r_2}(j))$
      else
        $\vec{\alpha}'_{G+1,i}(j) = \vec{\alpha}_{G,i}(j)$
      Check validity:
      if $\vec{\alpha}'_{G,i}(j) < \vec{\alpha}^{(lo)}(j)$ **then**
        $\vec{\alpha}'_{G+1,i}(j) = (\vec{\alpha}_{G,i}(j) + \vec{\alpha}^{(lo)}(j))/2$
      else
        if $\vec{\alpha}'_{G,i}(j) > \vec{\alpha}^{(hi)}(j)$ **then**
          $\vec{\alpha}'_{G+1,i}(j) = (\vec{\alpha}_{G,i}(j) + \vec{\alpha}^{(hi)}(j))/2$
        else
          $\vec{\alpha}'_{i,G+1}(j)$
    Normalize:
    $\vec{\alpha}'_{i,G+1} = \vec{\alpha}'_{G+1,i} / \sum_{k=1}^{P} \vec{\alpha}'_{G+1,i}(k)$
    Select:
    if $f(\vec{\alpha}'_{G+1,i}) \leq f(\vec{\alpha}_{G,i})$ **then**
      $\vec{\alpha}_{G+1,i} = \vec{\alpha}'_{G+1,i}$
    else
      $\vec{\alpha}_{G+1,i} = \vec{\alpha}_{G,i}$
  $G = G + 1$

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

E. Comparison of the optimization methods

The performances of the three different optimization methods are compared, while using them to minimize the objective function corresponding to *ECC* (Eq. (17)). For these tests, a simple audio-video sequence involving a single speaker - thus a single mouth region - has been used. A frame of this test sequence is shown, as an example, on Fig. 3. More details about the sequence are given in the next section, where the main results on speaker detection are presented (this one-speaker sequence presents the same characteristic than the two-speaker ones presented hereinafter).

For both GACS and DE algorithms, different tests have first been performed so as to tune the parameters properly. Notice that the implementation of the DE algorithm has been based on Storn's public domain version software [33]. As far as concerned GACS, a choice of $Q = 5$ quantization levels (resulting in a population of 125 chromosomes) combined with 400 generations and a perturbation variance σ fixed to 0.1, gave good results. DE

Fig. 3. Frame example of the test sequence used to perform the comparison between the different optimization methods. The white rectangular box delimites the extracted mouth region.

algorithm achieved good performance with $Q = 5$ quantization levels, 500 generations, a scaling factor $F = 0.5$ and a crossover probability CR equals to 1.

Once determined these optimal parameters, different runs have been performed with GACS and DE algorithms, whereas different initial conditions (*i.e.* different initial solution guesses) have been tried for Powell's method. Table I summarizes the results obtained with each method. Obviously, much better minimization is obtained using the global optimization schemes instead of the local one (Powell's). A finer analysis of the results in Tab. I reveals that DE reaches the best solution and in a more stable way. Indeed, the standard deviation of the solutions is much smaller in the case of DE than in the case of the other two methods, giving us more confidence in the results.

As another feature of this better behavior of the DE algorithm with respect to Powell's, it can be observed on Fig. 4 that the weight values obtained at the end of each runs are more scattered in the parameter space with the latter. This indicates two things. First, the objective function is highly irregular and exhibit plenty of local minima. Secondly, the values close to the global optimum are clustered in the solution space. These two characteristics of the cost function make Powell's method inadequate.

While the high variation of the solutions found with Powell's method is not a surprise (as it is very sensitive to initial conditions), the instability of GACS solution seems intriguing. However, this is less surprising when we analyze the evolution of the algorithm towards the solution: the degeneration of the population combined with the less systematic exploration of the solution space (especially the boundaries) make GACS solutions to be very different from run to run. On Figs. 4.(a) and 4.(b), the evolution of GACS and DE over different runs has been plotted.

Another issue in using GACS and DE algorithms is the stopping criterion. One simple way consists in running the algorithm for an *a priori* defined number of iterations. However, the number of iterations needed to reach a good approximation of the global minimum depends on the data and the inherent randomness of the algorithm. Thus this approach is unsteady. A more suitable criterion should be based on the analysis of the algorithm's evolution towards the global optimum. One may choose to stop if, during a number of iterations, the solution is not improved significantly. Even from this perspective, DE seems more convenient: from Figs. 4.(a) and 4.(b), it is clear that GACS exhibits long generations with no changes in the solution, possibly followed by slight improvements. This means that it is very hard to find a suitable stopping criterion for GACS, as we may always get an improvement after a long period of stagnation of the solution. So an early termination has the chance to leave the solution far from the best achievable one.

Definitely, the behavior of DE is preferable as we have steeper changes in the current solution and an early stop is not so dramatic from the perspective of the quality of the solution. All these considerations justify our choice of optimization algorithm for all subsequent experiments: we will use DE in its form given by Fig. 1 for our study of different speaker detection criteria.

VI. AUDIOVISUAL SPEAKER DETECTION RESULTS

A. Experimental protocol

A number of experiments have been performed on a home-grown dataset containing five audio-video sequences of duration 4s (labeled $1, 2, \ldots, 5$), each shot in PAL standard (25 frames/second (fps), 48kHz stereo sound). In

	Best Value	Mean Value	Standard Deviation
Powell	-0.0213	-0.0183	0.0047
GACS	-0.0695	-0.0619	0.0052
DE	-0.0788	-0.0774	0.0017

TABLE I

VALUES OF THE OBJECTIVE FUNCTION CORRESPONDING TO *ECC* FOR DIFFERENT RUNS USING POWELL'S, GACS, AND DE APPROACHES. ALL THE RUNS WERE PERFORMED UNDER THE SAME CONDITIONS (EXCEPT FOR POWELL WHERE DIFFERENT INITIAL CONDITIONS WERE TRIED) ON THE SAME AUDIO-VIDEO SEQUENCE.

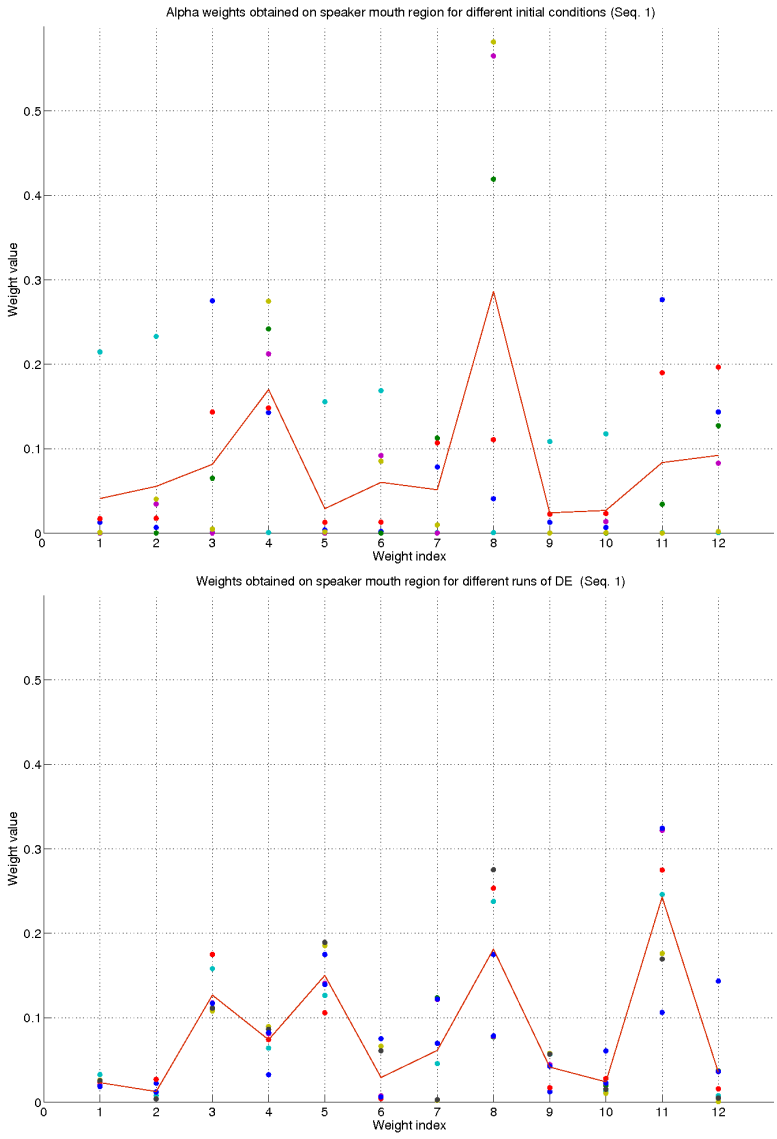

Fig. 4. Values of the MFCCs linear combination obtained on a given sequence with Powell's optimization algorithm with different initial guesses (top) and different runs of DE (bottom). The continuous line connects the mean values of the weights obtained.

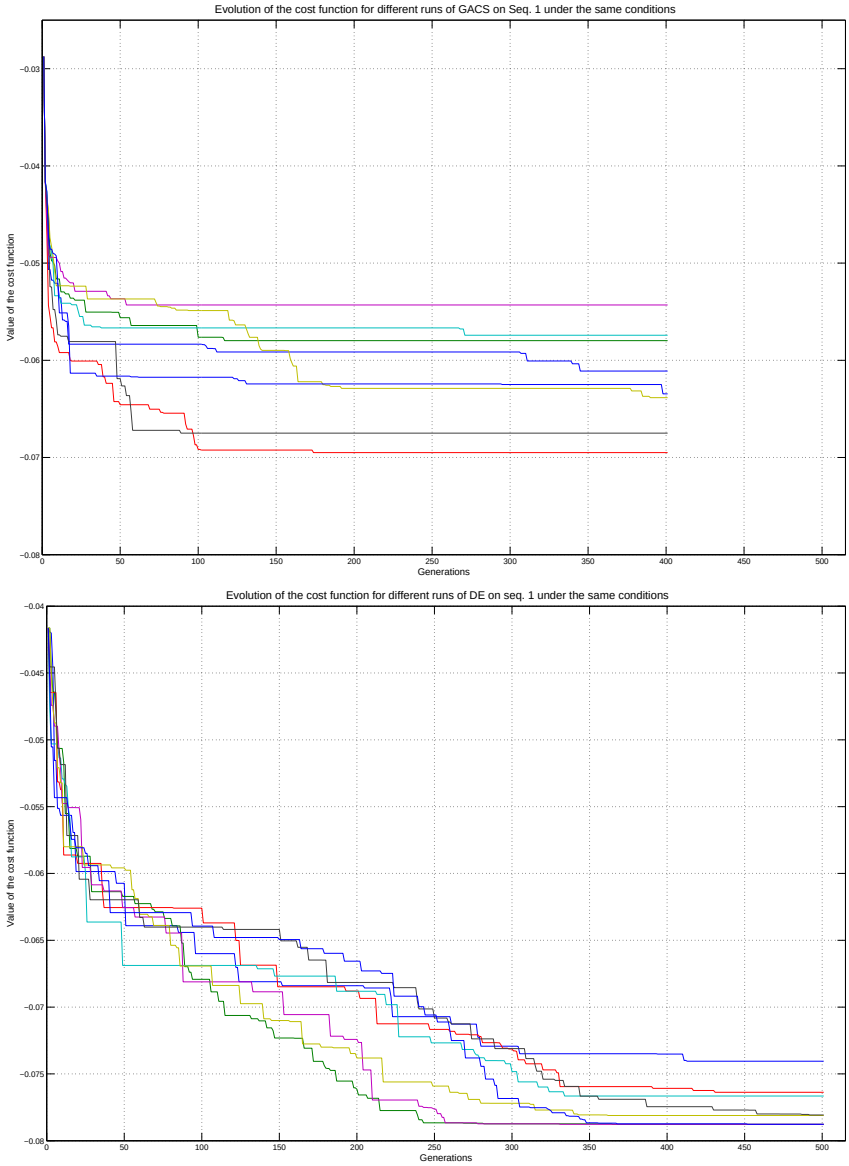

Fig. 5. Evolution of the cost function -*ECC* for different runs obtained with GACS (top) and DE (bottom) on a given audio-video sequence.

each sequence, two individuals are present. Both of them are referred to as ”speakers”, since any of them may have utter the recorded audio. Note, however, that only one is speaking at a time. These sequences are of increasing complexity, the fifth being the most challenging with the non-speaking individual moving randomly his head and lips.

First, each mouth region is manually extracted from each of the 100 frames of a sequence, resulting in two regions of $N \times M$ pixels, where N and M vary between 22 and 33 pixels, depending on speakers' characteristics and acquisition conditions. Thus the video feature set (video sample) is composed of the $N \times M \times 99$ values of the optical flow norm at each pixel location.

From the audio signal, 12 mel-cepstrum coefficients are computed using 23.22ms Hamming windows [17], [18].

Considering each mouth region and its associated video features, the MFCCs are projected on a new 1D subspace as defined in Sec. 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 $\vec{\alpha}^{opt}_{M_1}$ and $\vec{\alpha}^{opt}_{M_2}$, where the indices M_1 and M_2 indicate whether these weights result from the optimization performed on the first mouth region or on the second one respectively. Two corresponding audio feature sets derive from these weight sets: $F^{opt}_{A_{M_1}}$ and $F^{opt}_{A_{M_2}}$.

Two pairs of mutual information values can then be evaluated between these audio features and the video features

a b

Fig. 6. Typical frame extracted from the test sequences. White rectangles delimite the extracted mouth regions. (a) Frame extracted from sequence 5; (b) frame extracted from the third sequence.

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 %

TABLE II

NORMALIZED DIFFERENCE OF MUTUAL INFORMATION MEASURED IN EACH MOUTH REGION FOR EACH OF THE FIVE TEST SEQUENCES, CONSIDERING THE AUDIO FEATURES EXTRACTED WITH OPTIMIZATION CRITERION *MIC* OR *ECC*, ON THE SPEAKING MOUTH REGION.

in each mouth region. If V_{M_1} denote the video features of the first mouth region and V_{M_2} those of the second one, the two pairs of mutual information are given by:

$$\{I(V_{M_1}, F^{opt}_{A_{M_1}}),\ I(V_{M_2}, F^{opt}_{A_{M_1}})\}, \tag{24}$$

$$\{I(V_{M_1}, F^{opt}_{A_{M_2}}),\ I(V_{M_2}, F^{opt}_{A_{M_2}})\}. \tag{25}$$

First, a comparison of both *MIC* and *ECC* criteria is performed. As a result, *ECC* turned out to be indeed more discriminative than *MIC*. Therefore, *ECC* only is then used to analyze the ability of the method to extract specific audio features and to perform speaker detection. Finally, the discussion of the results leads to the definition of the more efficient criterion ΔECC given by Eq. (19) whose performances are presented and discussed.

B. Comparison of optimization criteria MIC and ECC

The initial hypothesis is that *ECC* is more efficient that the simpler *MIC* and the first set of experiments 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 normalized difference of mutual information between the speaking mouth region and the non-speaking one for each of the five test sequences is measured. Table II presents the results (ΔI_{MIC} and ΔI_{ECC} refer to the normalized difference of mutual information measured between the speaking and the non-speaking mouth regions when using optimization criterion *MIC* and *ECC* respectively). Two observations can be made from these results. Firstly, the mutual information 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 4 cases out of 5, the *ECC* criterion led to larger difference between MI in the two regions. This indicates that using the *ECC* criterion gives rise to more discriminative features. Consequently, normalizing the mutual information by the entropy during the optimization leads to extract more specific information than using simply the mutual information alone, as stated in sec. IV.

C. Performances using ECC

The first set of experiments leads to the conclusion that *ECC* is a 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 in detail its

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 MUTUAL INFORMATION 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*.

properties. The purpose of the experiments described here is to assess the ability of our algorithm to extract audio features specific to speech and to perform speaker detection using these features.

The capacity of the proposed method to act as a speaker detector is shown first. In contrast with the experiments described in Sec. VI-B, none *a priori* knowledge of the active speaker is assumed. Then the technique described in sec. VI-A is applied: the optimization is performed on each of the mouth regions (M_1 and M_2) and the mutual information between two pairs of audio and video features is measured as stated by Eqs. (24, 25). 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 computed as described above are plotted in Fig. 7. We note that for all sequences (including the challenging seq. 5), the MI measured on mouth M_1 with $\vec{\alpha}_{opt}$ 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% of correct detections. Therefore the proposed method performs well as a speaker detector.

Another issue necessary to investigate is the specificity of the features extracted from audio with respect to video. For this, the difference between the normalized mutual information computed on mouth regions and the corresponding audio is measured as follow:

$$\Delta I_{M_1} = \frac{\max_{i\in\{1,2\}}(I(V_{M_i}, F^{opt}_{A_{M_1}})) - \min_{i\in\{1,2\}}(I(V_{M_i}, F^{opt}_{A_{M_1}}))}{\max_{i\in\{1,2\}}(I(V_{M_i}, F^{opt}_{A_{M_1}}))}, \tag{26}$$

$$\Delta I_{M_2} = \frac{\max_{i\in\{1,2\}}(I(V_{M_i}, F^{opt}_{A_{M_2}})) - \min_{i\in\{1,2\}}(I(V_{M_i}, F^{opt}_{A_{M_2}}))}{\max_{i\in\{1,2\}}(I(V_{M_i}, F^{opt}_{A_{M_2}}))}, \tag{27}$$

The results are listed in Table III. It can be seen that $\Delta I_{M_1} > \Delta I_{M_2}$ and $\Delta I_{M_1} > 0$ for all the sequences. On the other hand, ΔI_{M_2} is sometimes negative. In other words, when the audio features used for the measurement of mutual information have been obtained on the non-speaking mouth region, the difference of MI is sometimes favoring the non-speaking mouth (sequences 2, 4 and 5). So when optimizing on the non-speaking region, the features extracted cannot (and are not expected to) reflect any underlying relationship between audio and video. This result also appeared on Fig. 7, since the mutual information measured between V_{M_1} and $F_{A_{M_2}}$ is always smaller than the one measured between V_{M_1} and $F_{A_{M_1}}$. Therefore the audio features thus extracted are specific to speech.

D. Results obtained with ΔECC

Two optimizations were performed previously to decide who is the current speaker. Now, the two optimizations will be combined in a single one, which aims at maximizing the discrepancy between the two mouth regions. For this, the ΔECC, given by Eq. (19), will be used. The result of the optimization is a vector $\vec{\alpha}_{opt}$ which generates a single audio feature vector. It is expected to maximize thereafter the mutual information with the video features of the active mouth region. This new detection approach has been tested on the same five test sequences than before. Results are summarized in Table IV. The normalized difference of mutual information 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 mutual information is here greater than what was obtained with the previous *ECC* optimization scheme (Tab. II). This stresses the benefit of using the video content related to each mouth region during the optimization.

To validate the results obtained with this simplest ΔECC detection scheme, experiments on a sequence of the CUAVE speech corpus [34] have been performed. This is a speaker-independent corpus of multiple speakers with

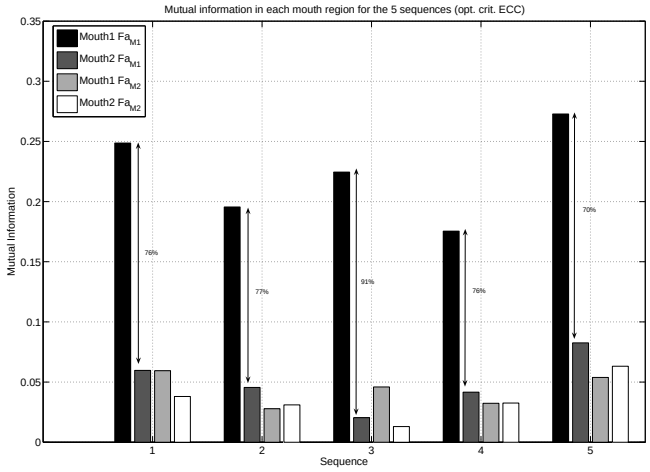

Fig. 7. Mutual information measured between the M_1 or M_2 mouth region features and the audio features obtained with optimization on mouth region M_1 or M_2 (Eqs (24, 25). The normalized difference of mutual information between the best value found and the best value found in the opposite mouth is indicated.

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

TABLE IV

NORMALIZED DIFFERENCE OF MUTUAL INFORMATION MEASURED BETWEEN THE SPEAKING AND THE NON-SPEAKING MOUTH REGIONS WITH THE AUDIO FEATURES OBTAINED USING ΔECC AS OBJECTIVE FUNCTION. TESTS HAVE BEEN PERFORMED ON THE FIVE TEST SEQUENCES DESCRIBED IN SEC. VI-A.

audio-video sequences of either single or two speakers. In the sequence considered (referred to as *g22* in [34]), a male and a female speakers are present.

The first 22 seconds of the clip have been used, where one person speaks at a time. These first seconds present challenging properties, making the detection task uneasy: the left speaker in the sequence often moves his lips so as to formulate without sounding the words. Since the frame rate of these sequences is 30fps, we have considered $2s$ long temporal windows instead of 4s as in the previous tests. This analysis window has been shifted each second over the whole sequence, so that 21 optimizations and mutual information measures have eventually been achieved (optimization criterion is still ΔECC). To compare the results with the one of Nock et col. [3], the same evaluation protocol has been used: the output of our detector has been compared with the groundtruth for the central frame of each detection window. This gives 3 wrong detection points out of 21. Notice however that evaluating the detection at the middle points of the analysis window somehow implies that the detection requires information on the future state to perform well. This is not the case for the proposed method. If the detection is rather evaluated at the last frame of each analysis windows, only one false detection occurs out of the 21 evaluation points (95% of good detections).

VII. CONCLUSIONS

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 exposed in [1] 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 objective functions is used but the optimization is performed in a straightforward manner using a global method. A comparison of the performance of three optimization methods (one local and two global) has been carried out, showing that the intrinsic properties of the Differential Evolution algorithm make it the best choice for our problem.

A study of two optimization criteria that can be used in this information theoretic framework has been carried out. Results have shown that the most 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% of good detections on 5 test sequences. Only two potential speakers are present in these test sequences but the detection method involving *ECC* can easily be extended to sequences containing more speaker candidates.

To optimize the detection in the case of two-people sequences, a third optimization criterion (ΔECC) has finally been introduced and tested on the same test sequence set as before. This criterion aimed 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.

Finally a number of tests have been carried out on a sequence of the CUAVE database [34] to assess and compare the performance of the ΔECC-based method to the state-of-the-art. Results are comparable to those obtained by Nock et al. in [3] for this particular sequence. Future work will include performing an extensive comparison of the proposed method with other published results, using the whole CUAVE corpus.

Acknowledgements

The authors would like to thank Dr. X. Bresson, E. Frejinger and M. Themans for fruitful discussions, as well as the Swiss National Found which supports this work.

REFERENCES

[1] T. Butz and J.-P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Processing*, vol. 85, pp. 875–902, 2005.
[2] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 406–413, June 2004.
[3] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proceedings of the International Conference on Image and Video Retrieval (CIVR)*, Urbana, IL, USA, July 2003, pp. 488–499.
[4] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Proc. of NIPS*, vol. 12, Denver, CO, USA, 1999, pp. 813–819.
[5] G. Monaci, O. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, Genova, Italy, September 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. of NIPS*, vol. 13, 2001.
[7] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proceedings of ICA*, Nara, Japan, April 2003, pp. 709–714.
[8] T. Butz and J.-P. Thiran, "Feature space mutual information in speech-video sequences," in *Proceedings of ICME*, vol. 2, Lausanne, Switzerland, 2002, pp. 361–364.
[9] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, 1992.
[10] 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 Transaction on Neural Networks*, vol. 5, no. 1, pp. 102–129, 1994.
[11] R. Storn and K. Price, "Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, pp. 341–359, 1997.
[12] J. W. Fisher III and J. C. Principe, "A methodology for information theoretic feature extraction," in *Proceedings of International Joint Conference on Neural Networks*, ser. IEEE World Congress on Computational Intelligence, vol. 3, Anchorage, Alaska, May 1998, pp. 1712 – 1716.
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, D. L. Schilling, Ed. John Wiley & Sons, 1991.
[14] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.
[15] P. Besson and M. Kunt, "Information theoretic optimization of audio features for multimodal speaker detection," EPFL, Lausanne, Switzerland, EPFL-ITS Technical Report 08/2005, February 2005.
[16] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.
[17] B. Gold and N. Morgan, *Speech and audio signal processing*. John Wiley & sons, Inc, 2000.
[18] J. W. Picone, "Signal modeling techniques in speech recognition," in *Proceedings of the IEEE*, vol. 81, no. 9, Sept. 1993.
[19] A. W. Bowman and A. Azzalini, *Applied smoothing techniques for data analysis*. Oxford science publications, 1997.
[20] S. Kirkpatrick, C. D. Gelatt, and J. M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, may 1983.
[21] G. Fred, "Future paths for integer programming and links to artificial intelligence," *Comput. and Ops. Res.*, vol. 13, no. 5, pp. 533–549, 1986.
[22] J. H. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.
[23] T. Spalek, P. Pietrzyk, and Z. Sojka, "Application of the genetic algorithm joint with thePowell method to nonlinear least-squares fitting of powder EPR spectra," *J. Chem. Inf. Model.*, vol. 45, pp. 18–29, 2005.

[24] 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 Transaction on Image Processing*, vol. 12, no. 12, pp. 1495–1511, December 2003.

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

[26] R. He and P. A. Narayana, "Global optimization of mutual information: application to three-dimensional retrospective registration of magnetic resonance images," *Computerized Medical Imaging and Graphics*, vol. 26, pp. 277–292, 2002.

[27] X. Qi and F. Palmieri, "General properties of genetic algorithms in the euclidean space with adaptive mutation and crossover." Department of Electrical and System Engineering U-157, University of Connecticut, Storrs, Tech. Rep. CT 06269-3157, 1992.

[28] P. Schroeter, J.-M. Vesin, T. Langenberger, and R. Meuli, "Robust parameter estimation of intensity distributions for brain magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 17, no. 2, pp. 172–186, april 1998.

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

[30] Y.-W. Leung and Y. Wang, "An orthogonal genetic algorithm with quantization for global numerical optimization," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 1, pp. 41–53, february 2001.

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

[32] R. Joshi and A. C. Sanderson, "Minimal representation multisensor fusion using differential evolution," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 29, no. 1, pp. 63–76, 1999.

[33] R. Storn, "Differential evolution homepage [online]." Available: http://www.icsi.berkeley.edu/ storn/code.html. [Online]. Available: http://www.icsi.berkeley.edu/ storn/code.html

[34] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1189–1201, 2002.