Multimed Tools Appl (2016) 75:115–130
DOI 10.1007/s11042-014-2274-x

Audio-visual speaker diarization using fisher linear semi-discriminant analysis

Nikolaos Sarafianos · Theodoros Giannakopoulos · Sergios Petridis

Received: 7 April 2014 / Revised: 21 July 2014 / Accepted: 11 September 2014 /
Published online: 28 September 2014

Abstract Speaker diarization aims to automatically answer the question "who spoke when" given a speech signal. In this work, we have focused on applying the FLsD approach, a semi-supervised version of Fisher Linear Discriminant analysis, both in the audio and the video signals to form a complete multimodal speaker diarization system. Extensive experiments have proven that the FLsD method boosts the performance of the face diarization task (i.e. the task of discovering faces over time given only the visual signal). In addition, we have proven through experimentation that applying the FLsD method for discriminating between faces is also independent of the initial feature space and remains relatively unaffected as the number of faces increases. Finally, a fusion method is proposed that leads to performance improvement in comparison to the best individual modality, which is the audio signal.

Keywords Speaker diarization · FLsD · FLD · Audio-visual fusion

1 Introduction

Speaker diarization is the task that utilizes signal analysis techniques, in order to automatically answer the question "who spoke when" given an audio or a video recording that contains an unknown amount of speech and speakers [4, 30]. This is an important task in multimedia analysis, being used in several applications such as multimedia summarization, speaker recognition and speaker-based retrieval of multimedia. Apart from using only audio information, speaker diarization can also be based on visual features, if such information is available [24, 33] and [18].

N. Sarafianos (✉) · T. Giannakopoulos · S. Petridis
Computational Intelligence Laboratory (CIL), Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos", Athens, Greece
e-mail: nsarafianos@iit.demokritos.gr

T. Giannakopoulos
e-mail: tyiannak@gmail.com

S. Petridis
e-mail: petridis@iit.demokritos.gr

A very important factor of speaker diarization in any modality (video or audio) is the choice of the feature space: it should contain information that only allow differentiating between speakers, and ideally, no other kind of information. The reason is that speaker diarization methods using some type of clustering based on the Euclidean distance, can be misled by the non-speaker discriminative dimensions. A solution to this problem is to apply dimensionality reduction methods to project the initial features to a speaker-relevant subspace, either by using the PCA method [6] or by adopting a supervised rationale [7]. In [20] the Fisher Linear Semi-Discriminant analysis (FLSD) method has been proposed, according to which information from the sequential structure of the audio signal is used as a substitute for unknown speaker labels required by the FLD method.

In [20], FLSD method was applied on audio features. The focus of this work is to extend and test the methodology in the video and the fused audio-visual domains. In particular, the overall goal of the current work is twofold:

- To extend the FLSD approach in the context of visual-based speaker diarization and evaluate its performance.
- To propose a method that fuses the visual and audio-based speaker diarization modules to further boost performance.

2 Fisher linear semi-discriminant analysis for temporal data

In this section, we will briefly describe the Fisher Linear Semi-Discriminant Analysis (FLSD) approach for any type of temporal features. This method was first proposed in [20] in the context of audio-based speaker diarization. FLSD is based on extending the Fisher linear discriminant analysis (FLD) method, which is in general applied in the context of a general classification setting where feature vectors are mapped to particular classes. The following subsection provides a description of the FLD background, while Section 2.2 describes the FLSD approach in general.

2.1 Fisher linear discriminant analysis

The basic rationale behind the FLD approach is to extract linear combinations of features, where the classes' means are far from each other and the variance within each class is small. If $\mathbf{x}$ is a N_x dimensional feature vector, $\mathcal{C} = \{c_k\}$ is the set of class labels, and $\{\mathbf{x}^i \mapsto c^i\}$ is a set of mappings between feature vector samples to classes, the FLD method defines the following matrices:

- the between class scatter matrix

$$S_b = \underset{c\in\mathcal{C}}{\mathcal{E}}[(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^\top], \tag{1}$$

- the average within-class scatter matrix: in an audio or video recording that contains an unknown amount of speech and also an unknown number of speakers

$$S_w = \underset{c\in\mathcal{C}}{\mathcal{E}}\left[\underset{\mathbf{x}^i\mapsto c}{\mathcal{E}}[(\mathbf{x}^i - \mathbf{m}_c)(\mathbf{x}^i - \mathbf{m}_c)^\top]\right], \text{ and} \tag{2}$$

- the total scatter matrix of samples

$$S_m = \underset{\text{all } \mathbf{x}^i}{\mathcal{E}}[(\mathbf{x}^i - \mathbf{m})(\mathbf{x}^i - \mathbf{m})^\top].$$

where

$$\mathbf{m} = \underset{\text{all } \mathbf{x}^i}{\mathcal{E}}[\mathbf{x}^i],$$
$$\mathbf{m}_c = \underset{\mathbf{x}^i \mapsto c}{\mathcal{E}}[\mathbf{x}^i], \forall c \in \mathcal{C}$$

and $\mathcal{E}[\cdot]$ denotes the sample mean. Note that S_m does not depend on the class mappings, while one can easily verify that $\mathrm{S}_m = \mathrm{S}_b + \mathrm{S}_w$.

Given a positive integer $N_y < N_x$, the aim of FLD is to find, among all possible $N_x \times N_y$ full rank matrices A, the matrix that optimizes a criterion which, in most of the cases is the following:

$$r = \mathrm{tr}\left(\frac{\mathrm{A}^\top \mathrm{S}_1}{\mathrm{A}^\top \mathrm{S}_2}\right) \tag{3}$$

where $(\mathrm{S}_1, \mathrm{S}_2)$ can be any of $\{(\mathrm{S}_b, \mathrm{S}_w), (\mathrm{S}_m, \mathrm{S}_w), (\mathrm{S}_b, \mathrm{S}_m)\}$ and $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. Other similar criteria have been proposed in the literature [15, 17] but maximizing the trace of the ratio criterion r is probably the most widely used. It actually amounts to find the eigenvectors with largest eigenvalues of the linear matrix pencil $(\mathrm{S}_1, \mathrm{S}_2)$. The optimal solution resulting from the previously described maximization is unique up to any invertible transform, respectively rotation and/or scaling, of matrix $\hat{\mathrm{A}}$. Then this matrix can project the initial (high dimensional) feature vectors to their N_y-dimensional FLD-optimal subspace:

$$\mathbf{y} = \hat{\mathrm{A}}^\top \mathbf{x} \tag{4}$$

2.2 FLsD

The FLD method is supervised, i.e. it requires to know, for a set of samples their mappings to class labels. However, in many cases (such as speaker diarization) such information is not available. FLsD considers a less demanding setting, according to which the requirement of knowing the samples mapped to each class is reduced to knowing, for each sample, a set of samples that are mapped to the same class. In applications that are associated with temporal data, the samples (feature vectors) are temporally ordered and it may be the case that for each sample, all neighbouring samples, in a relatively small time window, most likely belong to the same class. For example, in the speaker clustering context, we do not know *all* the samples spoken by a speaker beforehand, but one can guess that, for each sample, all neighbouring samples most likely belong to the same speaker.

In [20] the concept of *class threads* was introduced to define the functionality of the FLsD approach. Each (unknown) class is composed out of one or more class threads, therefore all samples of the same class thread are also mapped to the same class. In the context of an unsupervised task, the mapping of class threads to class is not known, while the mapping of samples to class threads can be known. Therefore, we can estimate the average *within-class thread* $\mathrm{S}_w^{\mathrm{h}}$ and *between-class thread* $\mathrm{S}_b^{\mathrm{h}}$ scatter matrices and then apply the FLD criterion using these matrices. It has been proven in [20] that, under certain conditions, the subspace found using $\mathrm{S}_w^{\mathrm{h}}$ and $\mathrm{S}_b^{\mathrm{h}}$ can well approximate the one that would had been found if the mapping with original classes was known.

In the context of audio-based speaker diarization, the class threads are defined as fixed-size speech segments (e.g. 1 second long), as described in Section 3. On the other hand, for the case of visual-based speaker diarization, the class threads will be defined based on video-shots (see Section 4). In the first case, we refer to the respective threads as "speaker threads" while in the case of visual information as "face threads".

3 Audio-based speaker diarization

Audio-based speaker diarization is achieved through FLsD. As a first step, the initial N_x-dimensional feature vectors are generated in a two step methodology, similar to the one in [31]. In particular, the short-term features are first extracted, resulting in $N_x/2$ Mel-frequency Cepstrum Coefficients (MFCCs) for every short-term frame. Selected short-term step and size are 20 ms. Speech Energy has not been included, as it is known to vary importantly between vectors that correspond to the same speaker. It is important to emphasize that the nature of the FLsD method does not require any particular type of initial feature space and therefore any features that convey speaker discriminative information might have been used (e.g. LPC). After the short-term feature sequences are extracted, the means and variances are computed over L subsequent MFCC vectors. In particular, means constitute the first half dimensions of the new vectors and variances the second half. Each of these new vectors describes a *texture window* the duration of which is equal to 1 sec.

Our next step is to obtain the near-optimal speaker discriminative projections of these texture window feature vectors. Since finding the exact FLD optimal subspace would require knowing the speakers of the analysed signal beforehand, we have adopted the FLsD approach, according to which each fixed-size texture segment is assigned a new *speaker thread*. The feature vectors sampled within this segment are used to obtain the speaker-thread mean feature vector and scatter matrix and also to update the overall within-class thread and mixed-class scatter matrices used in the FLD method. Once all the audio signal has been analysed, the scatter matrices are given as arguments to the Fisher criterion to obtain the optimal speaker-discriminative subspace.

Once the audio segments have been represented in the (reduced) speaker-discriminative subspace, the conditional probabilities of speakers given the provided vectors are estimated. Towards this end, a non-parametric discriminative classifier is employed, namely the K-Nearest Neighbour classifier (K-NN). The labels used by K-NN to estimate the speaker probabilities are obtained by applying the Fuzzy C-Mean algorithm [2] on the projected feature vectors, followed by a HMM - based smoothing. Smoothing using HMM allows to improve over the initial clustering speaker labels, by also taking into account the precedent and successive segments. As a final step, successive segments of the same speaker are merged, forming longer speaker-homogeneous segments. The above process is repeated for a range of number of speakers and the Silhouette width criterion [34] is used to decide about the quality of the clustering result in each case and therefore the optimal number of speakers.

4 Visual-based speaker diarization

4.1 Overview

Visual-based speaker diarization tries to determine the identity of each speaker using exclusively visual information. Figure 1 presents the flowchart of our visual-based speaker diarization procedure, further detailed in this section. In summary, our approach first extracts video shots and then faces are detected per frame and grouped per shot. For each face, a set of features is extracted and a semi-supervised dimensionality reduction approach, similar to the one described in Section 3, is used to define face-discriminant feature subspaces. A clustering algorithm is then applied on the reduced space and a metric that characterizes the quality of the estimated results is obtained. Finally, a lip movement detection

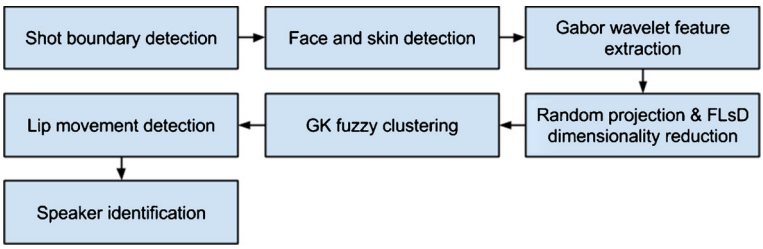

Fig. 1 Flowchart of the visual-based speaker diarization process

technique is used along with a nearest neighbor classifier to extract the final speaker identities. It has to be noted that, since extracting information from multiple faces is not only a difficult task to accomplish but can also lead to misleading results, we chose to deal with frames that contain only one dominant face.

4.2 Shot boundary detection

Video shots can be defined as a sequence of frames that appear to have been continuously captured with the same camera [19]. Here, it will be further assumed that throughout a single video shot, the detected face will belong to the same person. As as result, a video shot meets the requirements of class thread definition for the FLsD approach.

Following that, the first step to find face threads, is to perform video-shot-change detection. We extract the histograms of two successive grayscale frames, in order to exploit the advantages of this representation [29]. We then extract the normalized absolute difference and locate the local maxima of this sequence as proposed by Zhang et al. [38]. The locations of the detected maxima indicate the existence of a shot boundary.

4.3 Face feature extraction

As a next step, we apply the face detection algorithm of Viola and Jones [37] in order to detect faces in each frame. This algorithm is combined with a skin detector which eliminates all bounding boxes that do not correspond to a face. Forysth and Fleck [13] pointed out that the color of human skin has a restricted range of hues and is not deeply saturated. As a result, human's skin has a specific range of values in a color space, but not the same for every color space. The adopted skin detection process takes advantage of the face's hue channel of the HSV color space, by counting the pixels, the hue of which, is either close to zero or close to one. If the number of these (skin-related) pixels is dominant in the respective bounding box (i.e. if it is the majority in the candidate face image) then we proceed to the next step in the processing workflow. If this condition is not satisfied, then the respective bounding box is discarded.

In order to proceed with the feature extraction process we first resize the face image to a fixed size of 100×100 pixels. We then apply a Gabor wavelet face feature extraction technique. We concluded to the use of Gabor wavelets due to their resemblance with the receptive fields of the visual cortex [10] and for the reason that they remain unaffected to changes in illumination and to local distortions caused by the position and the expression of the face [22, 26]. By using two different scales and three different orientations, we balanced the extraction of descriptive features against computational complexity.

4.4 Semi-supervised dimensionality reduction and clustering

Having defined a feature vector, we aim to reduce its dimensions so that the new feature representation is discriminant in terms of faces. First we apply random projection to a lower dimension (e.g. 500 dimensions) as a pre-processing step [14], in order to reduce the computational complexity of the problem. Following that, we perform dimensionality reduction through the FLsD method. For each face-containing shot, FLsD creates a new *face thread*, and the feature vectors that belong in this shot are used to obtain the face-thread mean feature vector and scatter matrix, also updating the overall within-class thread and mixed-class scatter matrices. Once the process is completed, the scatter matrices are given as arguments to the Fisher criterion to obtain the optimal face-discriminative subspace. The assumption used for this purpose was the existence of one face in each shot. Note that the equivalent assumption made in the audio analysis task was that each 1-second segment contains a single speaker.

The Fuzzy C-Means algorithm for clustering speech segments [2], used in this work, requires beforehand the number of faces. Towards that direction, we experimented with a variety of possible number of clusters, computing each time the Silhouette width criterion [34]. The best solution is obtained by maximizing this criterion. After this clustering procedure, we extract, for each shot, the most dominant face label and we assign it to the entire shot. Therefore, each shot is now represented by a unique face label, along with a corresponding probability which is described in the next paragraph (Fig. 2).

Apart from obtaining the cluster (face) labels, we compute the metrics which will be utilized in the fusion process. The first metric, which represents the video probabilities (P_v), is equal to the number of labels in each shot that belong to each cluster divided by the total number of cluster labels within that shot. For example, if face A has been found in 45 out of 60 labels in shot t, the probability $P_v(t, \text{face} = A)$ of the specific cluster will be equal to

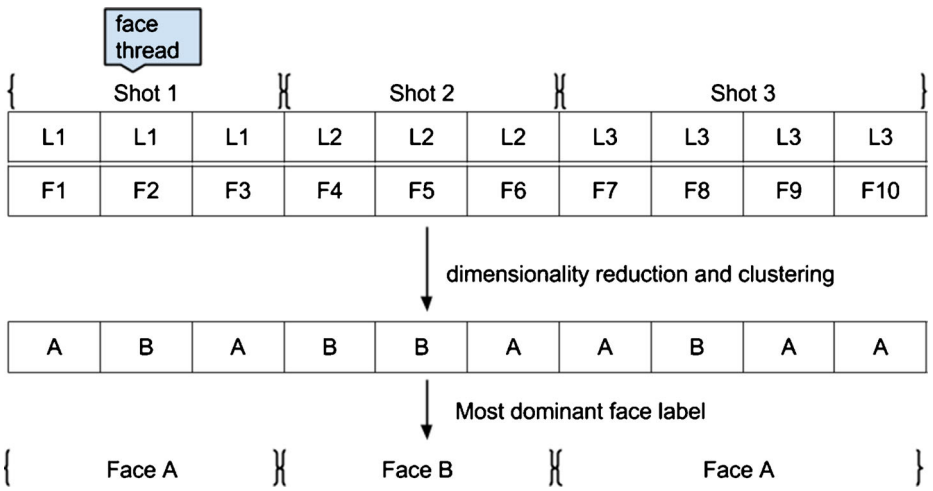

Fig. 2 Dimensionality reduction requires as an input feature vectors (F_i) and their relative labels (L_i) both of which are per frame. The GK fuzzy clustering technique in the reduced feature space along with the obtainment of the most dominant face labels in a shot lead us to the final face labels with a time resolution of one second

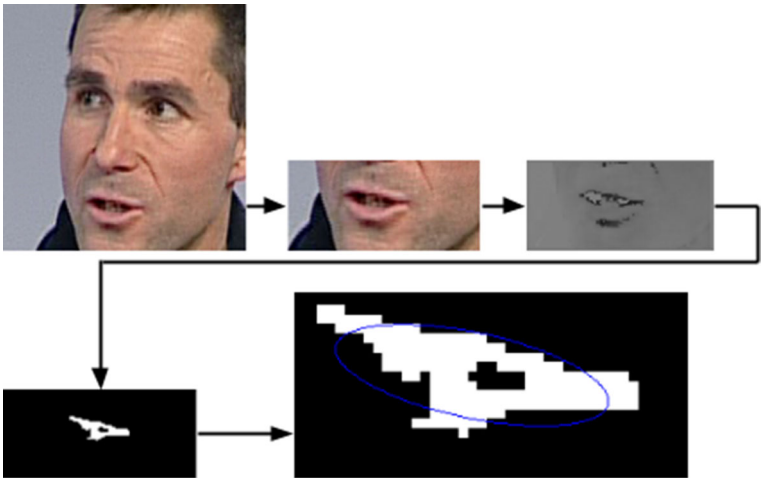

Fig. 3 Representative example of the aforementioned lip detection method

0.75. Moreover, we performed simple linear regression using the ratio of the average inter-cluster distance divided by the average outer-cluster distance as an explanatory variable. The coefficients of the estimated regression line were determined by conducting experiments on a subset of the data and varied depending on the estimated number of clusters. The outcome of this procedure is a cluster-based weight Cl_w, which along with the estimated video probabilities are used in the fusion process, as described in Section 5.3. Cl_w is therefore estimated for each separate input video sequence as described above.

4.5 Visual-based speaker extraction

Until this point, the algorithm provides an answer to the question "which face is shown and when". In order to move from the extracted face labels to the respective speaker labels, we further use moving lip information, obtained using the following lip movement detection algorithm (Fig. 3).

1. The lower part of the detected face is isolated by cropping the mouth region from the initial face image.
2. We apply the lip detection technique proposed by Soetedjo et al. [28] in order to transform the mouth region, from the RGB color space to a grayscale image. Its pixel values correspond to a confidence level that the respective pixels belong to the lips region. A simple heuristic post-processing technique is applied to remove the pixels that correspond to the internal region of an open mouth.
3. We transform the mouth image into binary, by thresholding the 10 % of its highest values, to obtain the brightest regions, i.e. the regions where the confidence of lips is high.
4. A 3×3 median filter is applied to remove noise.
5. Aiming to model the shape of the lips during speech, we use the extracted lip pixels of the previous step to construct an ellipse which is commonly used for that purpose [3, 9].

6. We compute a distance-based metric to measure abrupt changes in two successive detected ellipses (i.e., two successive estimated lip regions). In particular, this metric is defined as follows:

$$M = \text{mean}\left(\frac{|\alpha_{t+1} - \alpha_t|}{\alpha_{t+1}}, \frac{|\beta_{t+1} - \beta_t|}{\beta_{t+1}}\right)$$

where α_t, β_t are the two semi-axes of symmetry of the ellipse at time t. Finally, to transform this metric from frame to time resolution we obtain its average every one second.

As a next step, we threshold the lip movement metric described above, in order to keep the face labels (which now correspond to speaker labels) for which we can grant a confident answer that there is a lip movement. For the remaining segments, which (a) do not contain a face or (b) contain a face which does not correspond to moving-lips, we assign to each one of them, the label of the most neighboring speaker. In this way we obtain a complete sequence of speaker labels.

5 Fusion

5.1 Overview

Expecting fusion of audio and video information to improve overall performance is justified with information theoretic arguments as follows. Let the audio and video observation random variables be denoted as X_a and X_v respectively, while the (ground truth) reference person be denoted as Y and let these random variables take independently values at each considered time step t. Then, for the mutual information I between these variables, it holds that

$$I(X_a, X_b; Y) \geq I(X_a; Y)$$

and

$$I(X_a, X_b; Y) \geq I(X_b; Y)$$

i.e. the mutual information with the target variable can only increase if we take into consideration both audio and video observations together. This is direct consequence of the data processing inequality principle, stating that, for any deterministic function f, $I(X; Y) \geq I(f(x); Y)$ [8], i.e. the mutual information can only be decreased when applying a function to the variable. Considering that the audio variable is obtained by applying a (deterministic) projection function on the joined audio-video space, the two inequalities above follow directly. Moreover, these inequalities are expected to hold in the strong case, since X_a and X_v are expected to be independent to a great extent: the sound of one's voice is not a-priori correlated with one's look. In other words, two persons with similar voice may differ in how the look and vice-versa. This raises the upper bound of performance of an approach taking both modalities into consideration. Note, however, that this is a theoretical bound: a fusion method may show decreased performance if it fails to cope with the increased join dimension space created by the two modalities, with the finite sample space and/or with time alignment of modalities. Engineering a method that manages to boost the performance may be proven to be challenging task.

Fusing the audio and visual modalities in the speaker diarization task can be achieved either in an early stage, where the audio and visual features are combined to form a large feature dimensionality, or in a late manner. A related class of works [16, 18, 32] investigates

the problem of audio-visual fusion in the feature space for the speaker diarization problem. Garau & Bourlard [18] refer to a number of multimodal features in order to investigate an initialization approach whereas Friedland et al. [16] produce a combined log-likelihood of the features of the two individual modalities. Another related class of works [24, 33] is that of late audio-visual fusion (i.e. mapping of the corresponding labels). Noulas et al. [24] create a joint audio-visual space that results in the composition of two generative sets (whether or not the visible person corresponds to a speaker) for each person model.

In this work, we have opted not to apply fusion in the feature vector formation stage, since combining the audio and visual features would result in very high dimensional spaces. In addition, the FLsD approach is not applied exactly in the same way for the two modalities. This is due to the fact that different time resolutions are used for generating the samples used by the FLsD dimensionality reduction step.

On the other hand, fusion of audio-visual information has been done in two distinct steps, summarized as video-guides-audio, audio and video together. In particular, at a "pre-fusion" step (see Section 5.2), the video shot limits extracted from the visual module are used to exclude respective audio signal areas from the audio-based process. Then, the core fusion process is executed by combining the audio and video labels to extract the final fused speaker labels, as described in Section 5.3.

5.2 Using video shot information to improve audio-based diarization

As a pre-processing fusion step we exploit the accuracy of the shot boundary detection method by using the extracted shot limits (extracted based on the visual information), in order to improve the FLsD results applied on the audio module. Firstly, we isolate and exclude from the dimensonality reduction process of the audio problem, the time indices that correspond to a shot boundary in the video module. This is done due to the fact that the probability of multiple speakers is high near a shot boundary. Therefore, the probability that the FLsD method is fed with non-homogeneous data (which may lead to instability) is also high near the shot boundaries.

At a second step, we use the shots' ID in the FLsD process of the audio segments. In particular, the " speaker threads" described in Section 3 are defined based on the video shots, instead of the fixed-size one-second texture audio segments. So, information taken from visual analysis, the video shot, is used to improve on the one-second assumption somehow arbitrarily chosen for audio. Since video shots last more than one second, (audio based) speaker thread statistics defined on video shots will be closer to the speaker class ones, and therefore allow the FLsD approach to derive an improved feature subspace.

5.3 Core fusion

This subsection describes the core fusion process, i.e., how to obtain the fusion labels from the respective audio and video individual decisions. In other words, our purpose in this submodule is to estimate a *mapping* between the labels of audio and video and then to combine the (mapped) labels in order to extract the final decisions.

Given the number of estimated speakers in the video (N_v) and in the audio (N_a), we construct a table (T) of size equal to $N_v \times N_a$, the contents of which are filled in as follows:

$$T(m,n) = \sum_{\substack{i:v_i=m \\ j:a_j=n}} \frac{P_v(i,v_i) + P_a(j,a_j)}{2}$$

Where $m = 1, \ldots, N_v$, $n = 1, \ldots, N_a$ and v_i, a_i are the video and audio labels respectively. Once we have completed this procedure, we apply the Hungarian method [21] to the resulting matrix T in order to derive a confusion matrix (CM) which maps the labels extracted from the video module with those extracted from the audio. In other words, apart from the confusion matrix, each video label is mapped to a unique audio label using the mapping function $L(i)$. For example, $L(2) = 1$ means that the second visual label is mapped to the first audio label.

Having taken into consideration that the audio module demonstrates better performance measures, we chose to attribute less significance to the Cl_w of the video module. Thus, fusion probabilities are obtained for the t^{th} second from the following formula:

$$P_f(t, c) = P_a(t, c) \cdot \left(1 - \frac{Cl_w}{3}\right) + P_v(t, L(c)) \cdot \frac{Cl_w}{3}$$

where c refers to the respective cluster. Then, we derive the labels which display higher fusion probability and by applying the Viterbi algorithm, the most probable speaker path is obtained.

6 Experimental results

6.1 Dataset and performance measures

For the evaluation of the proposed algorithm, we used the publicly available Canal9 corpus [36]. The corpus consists of 70 debate recordings where participants do not act, but are actually engaged in unprompted, and often vivid, conversations. There are 190 unique participants in total, 165 of which are men and 25 are women, where each one participates in a maximum of three different debates. The number of participants in each debate varies from 3 to 5 (including the moderator) and as a result we experimented with numerous different face poses and orientations of each participant. Manual speaker segmentation is also provided and has been used as ground truth for the evaluation of all experiments.

In addition, for the evaluation of our "Face Diarization" approach described in the following subsection, we used a subset of the Canal9 corpus consisting of 25 videos which were manually annotated according to which face was shown and when. The duration of the videos varied from 20 to 90 seconds, whereas the number of faces was equally distributed between 2 and 5. Thus, our subset ended up with 40 different faces which were sufficient for the execution of our experiments.

The evaluation for our diarization technique was performed using the diarization accuracy rate (DAR), the average cluster purity (ACP) and the average speaker purity (ASP) measures, all of which are defined in [20].

6.2 Evaluation of the face diarization approach

In order to measure the ability of the video module to discriminate between different faces, we conducted a variety of experiments regarding the initial feature space, the dimensionality reduction method and the final dimensions of the FLsD. In order to evaluate the whole algorithm regardless of the adopted feature methodology, we have also computed performance measures when the pixel values of the detected faces were directly used as features, instead of the adopted Gabor features described in Section 4.3. Our interest here is to measure the influence of the initial feature space to the overall method.

Indeed, results indicate that there is negligible difference in the face diarization accuracy rate of the proposed method between the two initial feature spaces (raw pixel values and Gabor features). In particular, Fig. 4 confirms that (a) if FLsD is not adopted, the face diarization process experiences better results when Gabor features are used instead of simple pixel values, (b) the FLsD approach is independent of the initial feature space (similar DAR for pixel values and Gabor features) and (c) the DAR of the proposed method (including FLsD) remains unaffected as the number of faces increases. This is not the case when the Diarization process is applied without the FLsD dimensionality reduction method regardless of the type of features. In other words, in the later case, both pixel values and Gabor features, lead to performance that declines as the number of faces increases. This is another proof of robustness of the FLsD approach. Regarding the optimal number of FLsD dimensions, the projection of the feature space to four dimensions demonstrated the best performance, although three dimensions were usually sufficient enough to solve a simple Face Diarization problem.

Furthermore, Fig. 5 demonstrates a comparison between the Diarization Accuracy Rates of each method for the Face Diarization problem. Note that PCA refers to the standard statistical procedure [6] without any use of thin classes as in FLsD. Additionally, the difference between the optimal FLsD and the non-optimal described in Section 2.2 is that the first obtains the class threads from the ground truth in order to provide us with an upper boundary for the evaluation of our approach.

6.3 Evaluation of the FLsD approach for speaker diarization

Tables 1 and 2 show the performance indices of the diarization system for the CANAL9 corpus in the FLsD subspace. The last rows refer to the combination of the pre-fusion method with the core fusion technique.

In both cases, the final system that contains both fusion steps leads to almost 2 % DAR increase respective to the performance of the best individual modality (i.e. the audio-based

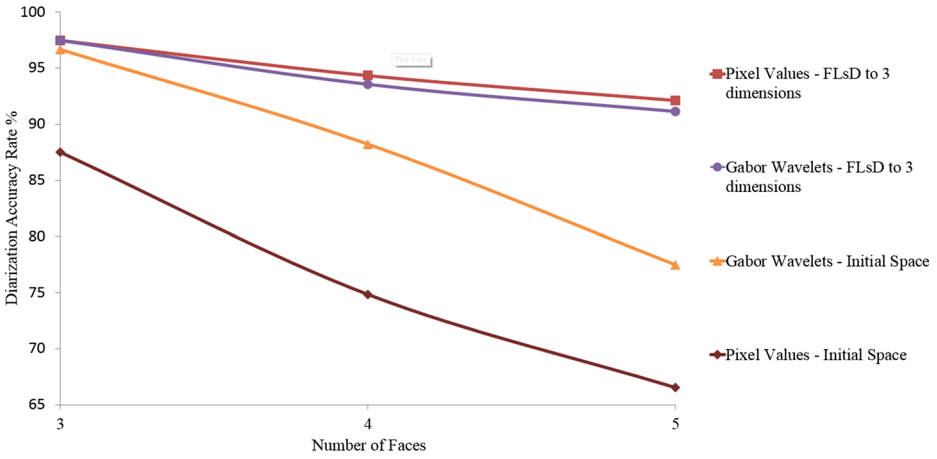

Fig. 4 Comparison of the feature extraction techniques for both the initial space and the projection of our feature space to 3 dimensions through the FLsD method. Significance of the FLsD subspace in function of the number of faces and evidence that our approach is independent of the initial feature space

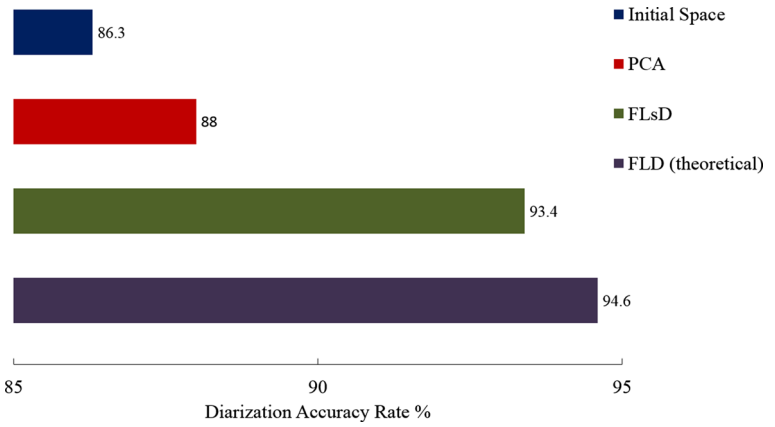

Fig. 5 Comparison of the proposed approach with other methods for subspace extraction

method). Both the ACP and the ASP measures demonstrate far better results when comparison is made between the audio and the video modalities and also exhibit a considerable increase when we proceed to the fusion step.

Providing a detailed comparison in terms of performance measures for the multimodal speaker diarization task is not trivial, since it involves a wide variety of data acquisition setups, scenarios and context. In this work, we have focused on political debates since they are at the same time simple and realistic. The approach described in [27] also adopted the Canal9 political debate benchmark to evaluate their methods, leading to an overall accuracy of 78.6 % for the audio domain, while the fused output was 83.2 % accurate. In [11] the audiovisual clustering process led to a performance rate of around 80 %, evaluated on the Canal9 dataset, however restricted in cases where a single person in foreground is also speaking (which cover less than 90 % of the whole data streams). Vallet et al. [33] perform speaker diarization on a TV talk show dataset which is similar to the Canal9 corpus, however no details on the performance boost obtained by the fusion procedure are provided.

In the context of a totally different experimental setup, various speaker diarization approaches have focused on meeting room-related scenaria. However, this is a rather different application domain since it requires to take into consideration more complex parameters, e.g. multiple cameras, speaker positions in the room, microphone arrays and even equipment topology. Noulas et al. [24] used IDIAP A [23] and Edinburgh [24] meeting datasets to achieve DAR of 67 % and 80 % respectively for the audio modality and 84 % and 89 % for the fused modality. Furthermore, Friedland et al. [16] used the whole AMI dataset [5] to achieve Diarization Accuracy Rates of 67.9 % for the audio modality and 74.7 % for the multimodal problem. Tranter [30] evaluated his multimodal approach on the RT-04F [12]

Table 1 Results of the evaluation process when the number of speakers in the video is not known beforehand

Performance measure %	DAR	ACP	ASP
Video	70.7 ± 8.7	72.7 ± 8.4	74.6 ± 7.5
Audio	84.3 ± 12.3	86.9 ± 11.2	89.9 ± 8.3
Core fusion	85.0 ± 12.4	86.8 ± 11.2	91.9 ± 6.4
Pre+Core fusion	86.1 ± 11.6	87.0 ± 11.3	92.1 ± 5.5

Table 2 Results of the evaluation process when the number of speakers in the video is provided

Performance measure %	DAR	ACP	ASP
Video	71.4 ± 8.9	73.0 ± 8.0	73.8 ± 7.3
Audio	87.6 ± 9.5	88.9 ± 7.5	89.9 ± 6.7
Core fusion	88.0 ± 8.4	89.5 ± 4.8	90.9 ± 3.4
Pre+Core fusion	89.0 ± 7.7	89.8 ± 6.4	91.6 ± 5.6

meeting corpus and reported results ranging from 73.1 % - 87.0 %. For the particular case of the audio modality, more detailed corresponding comparisons can be found in [1, 20] and [25].

Finally, aiming to give an index of the performance, it has to be mentioned that the whole diarization process (audio, video and fusion) is $1.65\times$ faster than real time when the number of speakers is known beforehand and $1.35\times$ when the number of speakers in the stream is unknown. More than half of the consuming time is allocated to the processing of the video frames and to the feature extraction stage. All the experiments were carried out using MATLAB 2013a on an Intel Core i7-3770 and 8 GB RAM.

7 Conclusions and future work

We have presented a multimodal method of speaker diarization based on clustering sequences of features in a reduced space that stems from the application of the Fisher Linear Semi-Discriminant Analysis method both in the audio and visual domains. We have extended the results obtained from [20] not only by proposing a way to apply the FLsD method in the visual-based face diarization problem but also by demonstrating a fusion approach that combines the results of both modalities to boost the overall performance. Extensive experimental evaluations lead to the following main conclusions:

- The FLsD approach, when applied on the task of face diarization leads to improved performance which is also independent of the initial feature space and remains relatively unaffected as the number of faces increases.
- The proposed fusion approach for the task of speaker diarization leads to better results compared to the best individual modality, i.e. audio.

Our future reseach will mainly focus on improving the fusion process. In particular, a promising direction of further research could be the addition of an extra step to the aforementioned fusion technique in a late rationale by using the audio labels which, in general, demonstrate more accurate performance measures, in order to guide the process of the video clustering. This could be achieved by detecting a speaker obtained from the audio module who demonstrates the largest mismatch in the respective video labels so as to use the FLD method to estimate a 1-D feature space that discriminates between the two classes. The vector extracted, would be concatenated to the previous FLsD vector leading to a new transformation to the reduced subspace. Moreover, to further improve the diarization performance, additional research could be focused on utilizing techniques of modeling the turn-taking behavior of each speaker as well as his role in the conversation in a more representative and realistic manner [35].

References

1. Anguera Miro X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O (2012) Speaker diarization: a review of recent research. IEEE Trans Audio Speech Lang Process 20(2):356–370
2. Babuka R, Van der Veen P, Kaymak U (2002) Improved covariance estimation for gustafson-kessel clustering. In: FUZZ-IEEE'02, vol 2. IEEE, pp 1081–1085
3. Barnard M, Holden EJ, Owens R (2002) Lip tracking using pattern matching snakes. In: Proceedings of the 5th Asian conference on computer vision
4. Barras C, Zhu X, Meignier S, Gauvain J (2006) Multistage speaker diarization of broadcast news. IEEE Trans Audio Speech Lang Process 14(5):1505–1512
5. Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, Kadlec J, Karaiskos V, Kraaij W, Kronenthal M et al (2006) The ami meeting corpus: a pre-announcement. In: Machine learning for multimodal interaction. Springer, pp 28–39
6. Castaldo F, Colibro D, Dalmasso E, Laface P, Vair C (2008) Stream-based speaker segmentation using speaker factors and eigenvoices. In: IEEE international conference on coustics, speech and signal processing. ICASSP 2008. IEEE, pp 4133–4136
7. Chu SM, Tang H, Huang TS (2009) Fishervoice and semi-supervised speaker clustering. ICASSP:4089–4092
8. Cover TM, Thomas JA (1991) Elements of information theory. Wiley
9. Dalka P, Czyzewski A (2010) Human-computer interface based on visual lip movement and gesture recognition. IJCSA 7(3):124–139
10. Daugman JG et al (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Opt Soc Amer J A: Optics Image Sci 2(7):1160–1169
11. Dielmann A (2010) Unsupervised detection of multimodal clusters in edited recordings. In: 2010 IEEE international workshop on multimedia signal processing (MMSP). IEEE, pp 177–182
12. Fiscus J, Garofolo J, Le A, Martin A, Pallett D, Przybocki M, Sanders G (2004) Results of the fall 2004 stt and mde evaluation. In: RT-04F workshop
13. Fleck MM, Forsyth DA, Bregler C (1996) Finding naked people. In: Computer vision ECCV'96. Springer, pp 593–602
14. Fodor IK (2002) A survey of dimension reduction techniques. Tech. rep., Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory
15. Foley D, Sammon J (1975) An optimal set of discriminant vectors. IEEE Trans Comput 100:281–289
16. Friedland G, Hung H, Yeo C (2009) Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In: IEEE international conference on acoustics, speech and signal processing. ICASSP 2009. IEEE, pp 4069–4072
17. Fukunaga K (1990) Introduction to statistical pattern recognition. Academic Press Limited, Boston
18. Garau G, Bourlard H (2010) Using audio and visual cues for speaker diarisation initialisation. In: 2010 IEEE international conference on acoustics speech and signal processing (ICASSP). IEEE, pp 4942–4945
19. Gargi U, Kasturi R, Strayer SH (2000) Performance characterization of video-shot-change detection methods. IEEE Trans Circ Syst Video Technol 10(1):1–13
20. Giannakopoulos T, Petridis S (2012) Fisher linear semi-discriminant analysis for speaker diarization. IEEE Trans Audio Speech Lang Process 20(7):1913–1922
21. Kuhn HW (1955) The hungarian method for the assignment problem. Naval Res Logist Q 2(1-2):83–97
22. Liu C, Wechsler H (2002) Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans Image Process 11(4):467–476
23. Moore D (2002) The idiap smart meeting room
24. Noulas A, Englebienne G, Krose BJ (2012) Multimodal speaker diarization. IEEE Trans Pattern Anal Mach Intell 34(1):79–93
25. Pardo JM, Anguera X, Wooters C (2007) Speaker diarization for multiple-distant-microphone meetings using several sources of information. IEEE Trans Comput 56(9):1212–1224
26. Schiele B, Crowley JL (2000) Recognition without correspondence using multidimensional receptive field histograms. Int J Comput Vis 36(1):31–50

27. Seichepine N, Essid S, Févotte C, Cappe O (2013) Soft nonnegative matrix co-factorizationwith application to multimodal speaker diarization. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3537–3541
28. Soetedjo A, Yamada K (2008) Skin color segmentation using coarse-to-fine region on normalized rgb chromaticity diagram for face detection. IEICE Trans Inf Syst 91(10):2493–2502
29. Swain MJ, Ballard DH (1991) Color indexing. Int J Comput Vis 7(1):11–32
30. Tranter SE, Reynolds DA (2006) An overview of automatic speaker diarization systems. IEEE Trans Audio Speech Lang Process 14(5):1557–1565
31. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans ASP 10:293–302. doi:10.1109/TSA.2002.800560
32. Vajaria H, Islam T, Sarkar S, Sankar R, Kasturi R (2006) Audio segmentation and speaker localization in meeting videos. In: 18th international conference on pattern recognition. ICPR 2006, vol 2. IEEE, pp 1150–1153
33. Vallet F, Essid S, Carrive J (2013) A multimodal approach to speaker diarization on tv talk-shows. IEEE Trans Multimedia 15(3):509–520
34. Vendramin L, Campello R, Hruschka E (2009) On the comparison of relative clustering validity criteria. In: SIAM international conference on data mining, pp 733–744
35. Vinciarelli A (2009) Capturing order in social interactions [social sciences]. IEEE Signal Process Mag 26(5):133–152
36. Vinciarelli A, Dielmann A, Favre S, Salamin H (2009) Canal9: a database of political debates for analysis of social interactions. In: 3rd International conference on affective computing and intelligent interaction and workshops. ACII 2009. IEEE, pp 1–4
37. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol 1. IEEE, pp I–511
38. Zhang H, Kankanhalli A, Smoliar SW (1993) Automatic partitioning of full-motion video. Multimedia Syst 1(1):10–28

Nikolaos Sarafianos was born in Athens, Greece in 1990. He received the Diploma Degree in Electrical and Computer Engineering from the National Technical University of Athens in 2013. He is currently a research assistant at the Institute of Informatics and Telecommunication of the NCSR "Demokritos" Athens. His interests are pattern recognition, computer vision and multimedia analysis.

Theodoros Giannakopoulos was born in Athens, Greece, in 1980. He received the Degree in Informatics and Telecommunications from the University of Athens (UOA), Athens, Greece, in 2002, the M.Sc. (Honors) Diploma in signal and image processing from the University of Patras, Patras, Greece, in 2004 and the Ph.D. degree in the department of Informatics and Telecommunications, UOA, in 2009. He is currently a Research Associate in the Institute of Informatics and Telecommunication, NCSR Demokritos. His research interests are pattern recognition and multimedia analysis. He is the coauthor of more than 20 publications and the coauthor of a book titled "Introduction to Audio Analysis: A MATLAB Approach".

Sergios Petridis was born in Athens, Greece in 1973. He received the Diploma Degree in Electrical and Computer Engineering from the National Technical University of Athens in 1996, the M.Sc in Pattern Recognition from UPMC, Paris in 1997 and a Ph.D. at the Department of Informatics and Telecommunications of UOA, Greece. He is a research associate at the Institute of Informatics and Telecommunication of the NCSR "Demokritos", Athens. His interests include machine learning and multimedia analysis.