Fisher Linear Semi-Discriminant Analysis for Speaker Diarization

Theodoros Giannakopoulos and Sergios Petridis

***Abstract*—Given an audio signal with an unknown number of people speaking, speaker diarization aims to automatically answer the question "who spoke when." Crucial to the success of diarization is the distance metric between speech segments, a factor depending on the choice of the feature space: distances should be low for segments of the same speaker and high for segments of different speakers. Starting from an Mel-frequency cepstrum coefficient (MFCC)-based feature space, an algorithm is proposed that finds a Fisher near-optimal linear discriminant subspace, adapted to the particular speakers which exist in the audio signal. The proposed approach relies on a semi-supervised version of Fisher linear discriminant analysis (FLD), leveraging information from the sequential structure of the audio signal as a substitute for unknown speaker labels. The resulting algorithm is completely unsupervised; therefore, the need for speaker labels in the provided or an independent set is dismissed. The eigenvalue perturbation theory is applied in order to provide optimality bounds with respect to FLD, showing the effectiveness of the approach under the assumption that speakers do not significantly modify the characteristics of their voice. A complete diarization system is then proposed, using fuzzy clustering, a non-parametric K-nearest neighbors classifier and a hidden Markov model. The experimental results show a major improvement of speaker diarization accuracy when using the optimal subspace found by the proposed approach with respect to using the initial MFCC feature space or subspaces found by competitive approaches.**

***Index Terms*—Fisher linear discriminant analysis (FLD), speaker diarization.**

I. Introduction

GIVEN an audio signal with an unknown number of people speaking, speaker diarization aims to automatically answer the question "who spoke when." The extraction of such information can help other related tasks, such as audio summarization, speaker recognition and speaker-based retrieval of audio. Earlier work on diarization dates back to 1998 [1] while more recent surveys that cover most aspects of speaker diarization have also become available [2], [3]. Important issues associated with speaker diarization include detecting speaker changes and clustering speech segments into speaker-specific clusters [4], [5]. Modeling the dynamics of conversations using turn-taking patterns and speaker roles has also been recently examined [6], [7].

Manuscript received November 25, 2011; revised February 20, 2012; accepted February 27, 2012. Date of publication March 19, 2012; date of current version May 07, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Man-Wai Mak.

The authors are with the NCSR "DEMOKRITOS," Patriarchou Grigoriou and Neapoleos, Agia Paraskevi, 15310 Greece (e-mail: tyiannak@gmail.com; petridis@iit.demokritos.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2191285

Fundamental to the success of most diarization steps is the choice of the feature space to represent speech. The feature space directly affects all probabilistic distributions used to model speech segments or speakers and therefore popular criteria such as the Bayesian information criterion (BIC) [4] or the Gaussian divergence measure [8] for speaker change detection. The quality of the feature space for speaker diarization depends mostly on two factors. The first one is whether it contains *enough* information to allow differentiating speakers. Though modeling speech with complex features has been proven useful [9], in this study it is assumed that a commonly used Mel-frequency cepstrum coefficients (MFCCs) feature space does contain all required information. The second factor, which is the main focus of this study, is whether the feature space also contains information *irrelevant* to speaker differences, which can be harmful, especially if algorithms do not foresee handling it. The most prominent example is clustering: for all clustering algorithms involving the Euclidean distance, samples in such spaces can be unintentionally clustered according to non-speaker discriminative characteristics, thus failing to define the desired classes. The same holds for applying the BIC criterion, since speech segment models in such spaces may differ not because of speaker characteristics but for other coincidental reasons.

An approach to address irrelevant information is to try to decompose the initial vector space into two orthogonal subspaces, namely the speaker-relevant subspace and the remaining speaker-irrelevant subspace, and then represent speech with projections to the speaker-relevant subspace. A first approach towards this direction is extracting eigenvoices, as proposed by Castaldo *et al.* in [10]. Eigenvoices are extracted using principal component analysis (PCA), which does not require labeled data and therefore is applied naturally to the task of speaker diarization where no *a priori* information for speakers is known. Tsai *et al.* [11] also apply PCA to a feature space based on GMM models adapted to pre-segmented utterances. The downside in both these PCA-based approaches is that the extracted subspace is optimized to capture the overall variance of data, not necessarily the one depending on speaker differences. To compensate, the fishervoices approach is proposed by Chu *et al.* [12], where Fisher linear discriminant analysis (FLD) is used. Also, a variant of fishervoices using a difference distance metric, called spherical discriminant analysis is also proposed by Hao [13]. However, applying FLD requires labeled data and, since these are missing in the setting of speaker diarization, these approaches rely on manually labeled independent sets, therefore targeting a "general" speaker discriminative subspace. Though this approach has its merits, the extracted subspace is optimized to capture the difference between *any* speakers as opposed to the *particular* speakers existing in the target signal. Moreover, its

success depends on the choice of the independent set, while issues such as the difference in recording conditions or the spoken language can lower the compliance of the extracted subspace to the target signal.

The Fisher linear semi-discriminant analysis (FLsD) proposed in this paper, combines the advantages of both PCA and FLD, in that 1) it does not need any manually labeled data and 2) it extracts a subspace that captures differences between the speakers of the target audio signal. The core idea of FLsD is that even if *a priori* information about speakers is missing, there is some related information that can be grasped from the audio signal with little effort. Namely, by leveraging the sequential structure of the signal and making the assumption that neighboring speech samples very likely belong to the same speaker, groups of samples that are tied to the same, though *unknown*, speaker can be formed. Tying samples to the same label is known as a "must-link" constraint in the semi-supervised paradigm [14]. Building on earlier works conducted independently by the authors on speaker clustering [15] and within the context of relevant component analysis (RCA) by Bar-Hillel *et al.* [16], the current study shows that such constraints can be used to obtain an FLD near-optimal discriminative subspace under relaxed assumptions with respect to the sampling of the neighboring regions. It is also shown that the optimal solution of FLsD is the exact optimal solution of the FLD analysis with perturbed input and the magnitude of the perturbation bounds the optimality of the extracted subspace.

An important outcome when applying FLsD is that speech becomes effectively represented by a very small number of features: in our experiments four features were sufficient in order to differentiate between up to five speakers in the test audio signals. This can greatly reduce the complexity of models built in this space, such as GMMs. It also lays the ground for engineering lightweight diarization systems based on simple non-parametric models. In particular, this study shows that by using the off the self Fuzzy C-Means algorithm for clustering speech segments [17], together with the K-Nearest neighbors (K-NN), a high diarization performance is achieved. Also, since the number of samples generated from the test signals are of small or moderate size, both these algorithms are fast.

The novelties of this study can be summarized as follows:

- An approach for near-optimal discriminant subspace extraction (Fisher linear semi-discriminant analysis) is presented. The algorithm is generically defined, in that it can be applied in all cases where must-link constraints between samples are available. In particular, the current study improves over a similar approach proposed in the context of RCA [16] by a) proposing a formal definition using the concept of *class threads* and the *class to class threads mapping*, b) proving optimality bounds with respect to FLD analysis under relaxed class threads sampling assumptions, and c) accounting for both trace-ratio and ratio-trace criteria forms of the FLD criterion.
- The proposed FLsD algorithm is applied in the context of speaker diarization, allowing to detect a small-dimensional feature space relevant to speaker differences. The final features can be considered as "fishervoices," though extraction is done in a completely unsupervised way.
- A complete lightweight speaker diarization system built using FLsD, Fuzzy C-Means, K-NN, and HMM is proposed. Evaluation using the CANAL9 corpus [18] shows that the presented system can reach high diarization accuracy. Comparative experiments show that FLsD has a clear advantage against competitive methods for subspace extraction. It is also demonstrated experimentally that the benefits of extracting the FLsD subspace are complementary to hidden Markov modeling.

This paper is structured in five sections. Section II presents a background on FLD, defines FLsD, and explores its optimality. Section III describes how FLsD is applied for speaker diarization and presents the complete system. In Section IV, the proposed system is evaluated based on the CANAL9 corpus and issues that affect its performance are discussed. Finally, Section V concludes the presented work by pointing out particular remarks and sets the grounding for future work.

II. Fisher Linear Semi-Discriminant Analysis

The content of this section applies to a general classification setting where N_x-dimensional feature vectors are mapped to one out of C classes. For example, in the speaker diarization case, the audio signal may be represented by a sequence of N_x-dimensional feature vectors derived from an MFCC short-term audio analysis, each one mapped to a speaker (see Section III-A for the details). The $\mathcal{R}^{N_x}$ space engendered can be considered as the sum of two orthogonal subspaces: a discriminative subspace of dimension N_y and a classification-irrelevant subspace of dimension N_z, such that $N_x = N_y + N_z$. As experiments in Section IV-B show, there are cases where $N_y \ll N_z$, i.e., the biggest part of the original feature space does not contain discriminative information. This may have a negative effect, if one needs to recover the classes based solely on clustering, as is the case in speaker diarization.

Namely, since most clustering algorithms involve the Euclidean distance, samples in the N_x-dimensional space can be unintentionally clustered according to non-class discriminative characteristics, thus failing to define the desired classes. Instead, clustering in the N_y-dimensional discriminant subspace alleviates this problem, since distances in this subspace reflect classes' differences. Therefore, performing clustering using the respective feature vectors projections has significantly more chances to find clusters that correspond to the desired classes.

In this section, a semi-supervised method to extract the N_y-dimensional discriminative subspace is proposed, called Fisher semi-discriminant linear analysis (FLsD). The FLsD method makes use of the Fisher linear discriminant (FLD) analysis discussed in Section II-A. In Section II-B, we propose a generic definition of FLsD, using the concept of class thread and the class-thread to class mapping. Finally, in Section II-C, we proceed by deriving bounds on the optimality of FLsD under non independent and identically distributed (i.i.d.) assumptions for the class threads sampling.

A. Background on Fisher Linear Discriminant Analysis

The basic idea in FLD is to extract linear combinations of features, where the classes' means are far from each other and the variance within each class is small. In the speaker diarization context, classes correspond to speakers. Formally, by letting $\mathbf{x}$ be a N_x-dimensional feature vector, $\mathcal{C} = \{c_k\}$ be the set of classes, and $\{\mathbf{x}^i \mapsto c^i\}$ be a set of mappings between fea-

ture vector samples to classes, we first define the between-class scatter matrix as

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

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. Then, we define the average within-class scatter matrix as

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

and the total scatter matrix of samples as

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

Note that S_m does not depend on the class mappings, while one can easily verify that $S_m = S_b + S_w$.

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 of the following form:

$$\hat{A} = \text{argmax}_{A\in\mathcal{R}^{N_x\times N_y}} r(A, S_1, S_2). \tag{3}$$

where (S_1, S_2) can be any of $\{(S_b, S_w), (S_m, S_w), (S_b, S_m)\}$. Several criteria of this form have been studied, commonly involving determinants or traces of the projected scatter matrices, all of which being extensions of the Fisher criterion for more than two classes [19, Ch. 10]. Notably, the following have received important consideration:

$$r_1 = \text{tr}\left(\frac{A^\top S_1 A}{A^\top S_2 A}\right) \tag{4}$$

and

$$r_2 = \frac{\text{tr}(A^\top S_1 A)}{\text{tr}(A^\top S_2 A)} \tag{5}$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix. Maximizing the trace of the ratio criterion r_1 is probably the most commonly used criterion and amounts to find the (generalized) eigenvectors with largest eigenvalues of the linear matrix pencil (S_1, S_2). Maximizing the ratio of traces criterion r_2 has been first studied by Foley and Sammon [20]. Moreover, Zhang *et al.* [21] have proposed a fast iterative way for obtaining the maximum by solving a sequence of eigenvalue decomposition problems for matrices $(A^\top(S_1 - \theta S_2)A)$, until convergence of the θ parameter. The optimal solution of r_1, respectively r_2, is unique up to any invertible transform, respectively rotation and/or scaling, of matrix $\hat{A}$. Any such matrix can then project the original N_x-dimensional feature vectors to their N_y-dimensional FLD-optimal subspace

$$\mathbf{y} = \hat{A}^\top \mathbf{x}. \tag{6}$$

Assuming that samples are drawn from a probability distribution and the scatter matrices are exact estimates of the corresponding covariance matrices, the FLD matrix is guaranteed to be Bayes-optimal under two conditions: 1) all class-conditional feature vector distributions are Gaussian with the same covariance matrix (homoskedasticity assumption) and 2) the dimension of the resulting subspace is at least as big as the number of classes minus one [22]. However, the wide applicability of FLD shows that it is rather robust under violations of these conditions. Furthermore, increasing the dimension of the extracted subspace may compensate for deviations from the homoscedasticity assumption.

Definition of FLsD

Using the FLD criterion requires to know, for each sample, its mapped class. However, sometimes this information can not be available, at least not entirely. This study considers a less demanding setting, where the requirement to known the samples mapped to each class is reduced to knowing, for each sample, a number of other samples that are mapped 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 neighboring samples, in a relatively small window, most likely belong to the same speaker (see Section III-B for the rationale).

Namely, let each class be composed out of one or more *class threads*, in the sense that all samples mapped to the same class thread v, are also mapped to the same class c. The surjective mapping of class threads to class, denoted by $h(v)$, provides, for each class thread, its corresponding class. Assuming that h is not known, while we do know the mapping of samples to class threads, we can estimate the average *within-class thread* S_w^h and *between-class thread* S_b^h scatter matrices and then apply the FLD criterion using these matrices. It will be shown that, under certain conditions, the subspace found using S_w^h and S_b^h can well approximate the one that would had been found if the mapping with original classes were known.

We define this variant of FLD as Fisher linear semi-discriminant analysis (FLsD). Formally:

Definition 1 (FLsD): Let a set of N_x-dimensional real vectors $\mathcal{X} = \{\mathbf{x}^i\}$, the class set $\mathcal{C} = \{c_k\}$ and the mapping of each observation to an original class $\{\mathbf{x}^i, c^i\}$. Let also a set of *class threads* $\mathcal{V}$, $|\mathcal{V}| > |\mathcal{C}|$, the surjective mapping $h : \mathcal{V} \to C$ that maps each class thread to an original one, and the mapping of the original reference set $\{\mathbf{x}^i, v^i\}$ such that $\forall\, i : h(v^i) = c^i$. Finally, let S_w^h, S_b^h and S_m denote respectively the within-class thread, between-class thread and mixed-class scatter matrix. For any $N_y < N_x$, the matrix found by the optimization

$$\hat{A} = \text{argmax}_{A\in\mathcal{R}^{N_x\times N_y}} r(A, S_1, S_2) \tag{7}$$

where (S_1, S_2) is any of $\{(S_m, S_w^h), (S_b^h, S_w^h), (S_b^h, S_m)\}$ and r given in either (4) or (5), is defined as the optimal $N_x \times N_y$ FLsD matrix.

Note that S_m is used to refer to both mixed-class and mixed-class thread scatter matrices, which are equal, since there is no involvement of the class or the class-thread mapping in their definition.

Fig. 1 shows a toy example in a 2-D space, with two classes composed by three class threads each. The FLD projection (dashed line in the figure) has been evaluated using the mapping to the original classes, $\{\mathbf{x}^i, c^i\}$. Evaluation of the optimal FLsD projection (solid line in the figure) uses the class threads instead, $\{\mathbf{x}^i, v^i\}$, not the mapping to the original classes, neither

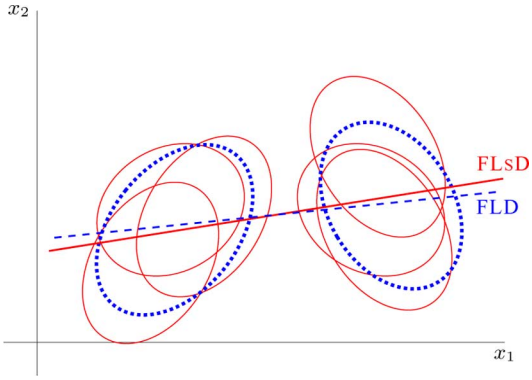

Fig. 1. FLsD example in two dimensions with two classes and six class threads. Solid (respect. dashed) ellipses correspond to the contours of two class threads (resp. classes) variances. The projection found by FLsD (solid line) closely approximates the one found by FLD (dashed line).

h. Notice that, in this example, the FLsD projection is a close approximation to the FLD projection.

B. Optimality of FLsD

A first result on the FLsD optimality has been derived in the context of RCA [16]. Namely, using the terminology define in this study, it has been shown that if each class thread has been i.i.d. sampled from the samples of the corresponding class, then the variance of the within-class thread scatter matrix rapidly converges to the variance of the within-class scatter matrix. In particular, for the minimum of class threads of size 2, the variance of S_w^h is only twice as of S_w. This is important, since it guarantees that, under the i.i.d. assumption, FLsD optimal solution will be very close to the FLD one, assuming that all class threads are well sampled.

The current study strengthens these results, by deriving bounds upon the optimality of FLsD, with respect to FLD, in the weakest case where the i.i.d. assumption does not hold. Our analysis is based on eigenvalue perturbation theory. Namely, in realistic situations, samples that are known to be in the same class thread may also depend in other ways. It follows that the means and covariance matrices of the classes threads cannot be assumed to be identical to the ones of the corresponding classes. This has a clear impact on the within-class and between-class matrices estimated using class threads, which the Fisher criteria depends on. To quantify the difference of these matrices, we define the class thread scatter matrix perturbation based on the theorem as follows:

Definition 2 (Class Thread Perturbation Matrix): Given a class thread v, let $\mathbf{d}_v$ be the difference of the sample mean of the corresponding class $h(v)$ to the sample mean of the class thread v

$$\mathbf{d}_v = \mathbf{m}_{h(v)} - \mathbf{m}_v.$$

Then, the class thread perturbation matrix S_h is defined as the covariance matrix of these differences over all class threads:

$$S_h = \underset{v\in\mathcal{V}}{\mathcal{E}}\left[\mathbf{d}_v\mathbf{d}_v{}^\top\right].$$

The term perturbation is justified by the following lemma:

Lemma 1: For any class thread mapping h,

$$S_w^h = S_w + S_h$$

and

$$S_b^h = S_b - S_h$$

i.e., S_h is an additive matrix perturbation to the within-class, respectively between-class, scatter matrix.

The proof is given in Appendix A. The class thread perturbation matrix summarizes how the class thread means are scattered around their respective class means. In essence, the class-thread scatter matrices can be thought of as perturbed versions of the respective class scatter matrices. By using class threads, we increase the within-class scatter and reduce, by an equal amount, the between-class scatter.

Note that the scatter matrix perturbation depends only on differences of the first-order statistics of class threads. Differences on second order statistics, manifested as differences of the elliptic shapes of class threads against the ones of the corresponding classes, do not further increase the perturbation. This is due to the fact the Fisher criterion considers only averages of scatter matrices, therefore neglecting *a priori* the differences the scatter matrices of either classes or class threads have between them. As a result, bounds on first-order statistics differences imply a bound on the matrix perturbation, as shown in the following Lemma:

Lemma 2: Let $\forall\, v \in \mathcal{V} : \|\mathbf{m}_{h(v)} - \mathbf{m}_v\|_2 \leq \delta$ for some $\delta > 0$. Then

$$\|S_h\|_2 \leq \delta^2$$

where $\|\cdot\|_2$ denotes the spectral matrix norm.

Proof: Using the definition of S_h, $\|S_h\|_2 = \left\|\mathcal{E}_v\left[\mathbf{d}_v\mathbf{d}_v{}^\top\right]\right\|_2 \leq \mathcal{E}_v\|\mathbf{d}_v\mathbf{d}_v{}^\top\|_2 = \mathcal{E}_v\|\mathbf{d}_v\|_2^2 \leq \delta^2$. ■

In words, if we can guarantee that the means of all class threads do not deviate more than a small δ from the corresponding class means, then the magnitude of the perturbation matrix does not exceed δ^2. In some sense, δ quantifies a bound on how wrong the assumption that the class threads have been i.i.d. sampled from the corresponding classes is.

The solutions of both the trace of ratio criteria and ratio of trace criteria for FLD rely on solving eigenvalue and generalized eigenvalue problems. Therefore, to see how the FLsD solution approximates the FLD one, we have to examine how the eigenvalues and eigenvectors of FLsD relate to the one of FLD. To simplify the discussion, we present the bounds based on the $(S_1 = S_m, S_2 = S_w)$ case. Note however that the same bounds also apply to all other combinations, since these achieve exactly the same solutions.

We start by examining the r_2 criterion defined in (5), since it comes down to solving a simple eigenvalue problem. It has been shown in [21] that the optimal solution of r_2 equals the optimal solution of $(A^\top(S_1 - \theta S_2)A)$ for a particular choice of the θ parameter. Therefore, we just need to show the following:

Theorem 1: Let $\forall\, v \in \mathcal{V} : \|\mathbf{m}_{h(v)} - \mathbf{m}_v\|_2 \leq \delta$ for some $\delta > 0$. Let also $\lambda_1 < \lambda_2 < \cdots \lambda_{N_x}$ and $\lambda_1^t < \lambda_2^t < \cdots \lambda_{N_x}^t$ be the eigenvalues of the matrices $(S_w - \theta S_m)$ and

$(S_w^h - \theta S_m)$, respectively. Then, for any θ, the difference of the corresponding eigenvalues is bounded as

$$|\lambda_i - \lambda_i^t| \leq \delta^2, \quad i = 1 : N_x.$$

Proof: Matrix S_m is not perturbed, and therefore the θS_m terms are canceled out. So the difference between these matrices is still the symmetric matrix $S_h = S_w - S_w^h$, whose norm, by Lemma 1, is bounded by δ^2. The proof then follows by applying Corollary 8.1.6 of [23, p. 396]. ∎

A similar result holds for the generalized eigenvalue problem related to r_1 criterion defined in (4) by (S_w, S_m). Note that the perturbation applies only to S_w. Namely:

Theorem 2: Let $\forall\, v \in \mathcal{V}: \quad \|\mathbf{m}_{h(v)} - \mathbf{m}_v\|_2 \leq \delta$ for some $\delta > 0$. Also, let $\lambda_1 < \lambda_2 < \cdots \lambda_{N_x}$ and $\lambda_1^t < \lambda_2^t < \cdots \lambda_{N_x}^t$ be the generalized eigenvalues of the pencils (S_w, S_m) and (S_w^h, S_m), respectively. Then, the difference of corresponding generalized eigenvalues is bounded as

$$|\lambda_i - \lambda_i^t| \leq \frac{\delta^2}{\lambda_{\min}(S)}, \quad i = 1 : N_x.$$

Proof: The proof is a direct consequence of Theorem 2.1 in [24], using the result of Lemma 2. ∎

Similar results obtained for eigenvalues also hold for eigenvectors, in the sense that the distance of subspaces engendered by the FLD and FLsD eigenvectors are bounded by differences on eigenvalues and therefore on δ^2. In particular:

Theorem 3: Let $\mathbf{a}_i$ be an eigenvector of $(S_w - \theta S_m)$ with eigenvalue λ_i. Let also $|\lambda_i - \lambda_j| \leq \delta_\lambda, \forall\, j$. Then there exists an eigenvector $\mathbf{a}_1^t$ of $(S_w^h - \theta S_m)$ such that

$$\operatorname{dist}(\operatorname{span}\{\mathbf{a}_i\}, \operatorname{span}\{\mathbf{a}_i^t\}) = \sqrt{1 - \left(\mathbf{a}_i^\top \mathbf{a}_i^t\right)} \leq \frac{4\delta^2}{\delta_\lambda}.$$

Proof: The proof is on the same line of the one provided for Lemma 1, using Theorem 8.1.12 of [23, p. 399]. ∎

In essence, Theorems 1–3 point out that we should expect a near-optimal behavior of the FLsD criterion when using either the r_1 or the r_2 criterion. The FLsD criterion converges to the optimal one, as δ^2 goes to zero. This means that we can seek for an approximate discriminative subspace using class threads, instead of the classes, as long as the first-order statistics over the classes' threads do not differ much from the first-order statistics over their corresponding original ones.

III. Speaker Diarization Using FLsD

In Section II, we described how the Fisher criterion can be used within a semi-supervised framework, to obtain a discriminative linear subspace. In this section, a speaker diarization system is described that applies the FLsD criterion in a completely unsupervised way, leveraging information from the sequential structure of the audio signal. Section III-A describes how the original feature vector is generated. Section III-B shows how the FLsD criterion is applied, using class threads that correspond to single-speaker segments. Finally, Section III-C describes the complete diarization approach in the projected space, using a combination of fuzzy clustering, a K-NN classifier and HMM smoothing.

A. Generating the Initial Feature Vectors

To generate the initial N_x-dimensional feature vectors a two step methodology is followed, similar to the one in [25] and [26]. Initially, a short-time analysis is conducted, resulting in $N_{x/2}$ MFCCs for every w_s of audio signal:

$$\left\{\mathbf{o}[n] \in \mathcal{R}^{N_{x/2}}\right\}, n \in \left[1 \ldots \frac{T}{w_s}\right]$$

where T is the duration of the audio signal and T/w_s is the number of w_s-sized non-overlapping windows. The values used in our experiments are $N_x = 24$ for the number of coefficients and $w_s = 20$ ms for the analysis window and analysis step, which have been experimentally found to give good results. Note that Delta MFCC were not found to be useful, and therefore have not been included in the original feature space. Also, neither energy has been included, since it has been found to vary importantly between speaker threads of the same speaker, thereby deteriorating FLsD performance (see also Assumption 1b in Section III-B). In general, any features that convey speaker discriminative information might have been used (e.g., LPC). On the same spirit, analysis that tends to discard speaker discriminative information, such as as low-order PLP, should be avoided.

Subsequently, by evaluating the means and variances over L subsequent MFCC vectors $\mathbf{o}[n]$, N_x-dimensional vectors $\mathbf{x}[n]$ are created. Means constitute the first half dimensions of the vectors: $x_i[n] = (1/L)\sum_{m=n}^{n+L} o_i[m]$, $i = [0 \ldots N_{x/2} - 1]$, and variances the second half: $x_i[n] = (1/L)\sum_{m=n}^{n+L} \left(o_{i-N_{x/2}}[m] - x_{i-N_{x/2}}[n]\right)^2$, $i = [N_{x/2} \ldots N_x]$.

Following the terminology in [25], each $\mathbf{x}[n]$ describes a *texture window* of duration $w_l = L \cdot w_s$. Texture windows are used for all tasks described later, including non-overlapping speech detection, clustering, HMM smoothing and FLsD-based subspace extraction. Their length has been set to $L_1 = 50$ ($w_l = 1000$ ms) for all tasks, except for FLsD, where a smaller value $L_2 = 20$ ($w_l = 400$ ms) is used, to allow the feature vector of the texture window to vary within a speaker thread, as discussed in Section III-B. These two lengths result, respectfully, to two sequences of features vectors, namely $\mathbf{x}_1[n]$ and $\mathbf{x}_2[n]$.

B. Applying FLsD

This section describes how to obtain near-optimal speaker discriminative projections of the MFCC-based vectors $\mathbf{x}[n]$. Finding the *exact* FLD optimal subspace would require knowing the speakers of the analysed signal beforehand, or to resort to an independent set to extract a general speaker-discriminative subspace, as suggested in [12] where the term *fishervoice* is probably first used. Instead, here we use FLsD, together with the following assumption:

Assumption 1: There exist a duration bound L and a distance bound δ, such that:

a) any window $|w_l| < L$ contains single speaker speech;
b) the distance of the first order statistics obtained from any such window to the overall first order statistics of the corresponding speaker is bound by δ.

In FLsD terminology, Assumption 1a) guarantees that the speech sampled within windows of such duration, e.g., 1 s, define a *speaker thread*. In most practical cases, Assumption 1 a may hold only approximatively, since speech from different speakers sometimes overlap. Nevertheless, we may enforce this assumption by preprocessing the audio signal and removing all segments with overlapping speech. Even if after this step some

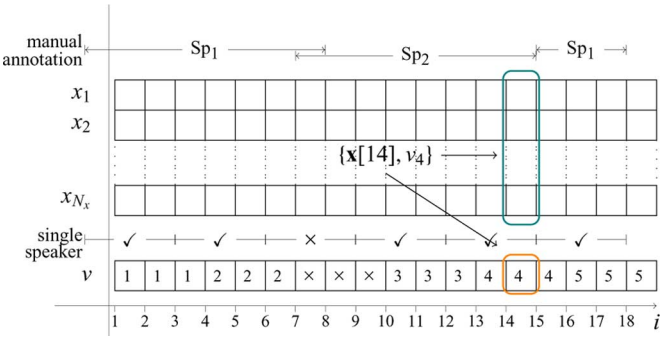

Fig. 2. Using fixed-size speech segments as class threads for the FLsD algorithm. As an example, the feature vector of the 14th texture window is mapped to the 4th speaker thread. The manual annotation (e.g., "Sp2") is not used during FLsD. Non-speech and overlapping/mixed speech segments, detected automatically, are not used while seeking the FLsD optimal subspace.

segments with overlapping speech still exist, all quantities in FLsD are evaluated as averages and therefore the outcome will not be seriously affected. On the other hand, Assumption 1b) essentially points out that, the voice characteristics of a speaker should not significantly change, which is a quite reasonable assumption to make for a diarization system. Using the results of Section II-C, we are ensured that by using FLsD, a near-optimal speaker voice discriminative subspace will be obtained.

Based on the above, we propose an algorithm that incrementally evaluates S_w^h through a long-term analysis of the audio signal. The process is exemplified for a sample audio signal in Fig. 2 while the steps are analytically described in Algorithm 1. Namely, the algorithm proceeds by sequentially analyzing fixed-size segments of duration w_l. Each segment is first checked to contain non-interrupted non-overlapping speech using a combination of classifiers. For every such segment, a new speaker thread is created, and the feature vectors sampled within this segment are used to obtain the speaker-thread mean feature vector and scatter matrix, also updating the overall within-class thread and mixed-class scatter matrices. Once all the audio signal is considered, the scatter matrices are given as arguments to the Fisher criterion to obtain the optimal speaker-discriminative subspace.

The size w_l of the analysis segments is tuned with respect to two opposite goals. On one hand, it needs to be large enough to contain an important number of texture windows, so that statistics within each segment are robustly calculated (i.e., speaker threads are well sampled). On the other hand, it must be short enough, to maintain low the probability that a segment includes speech from more than one speakers. In our experiments, we have used $w_l = 1$ s, resulting in 30 different samples for each speaker thread. Alternatively, one could consider longer variable size single-speaker segments, using for instance the BIC criterion, though this would entail the risk of compromising the validity of results, in case of speaker change detection failure.

As shown in Algorithm 1, to guarantee that speaker threads do correspond to single speakers, each segment is checked to contain non-overlapping speech. This is done in two steps, namely speech/non-speech classification, followed by rejection of segments that contain overlapping speech. Our approach for discarding non-speech segments has been to employ a general-purpose speech/non-speech K-NN classifier trained on $\sim$1000 audio segments, recorded and manually labeled

Algorithm 1: FLsD fishervoices

Input: $\mathbf{x}_1[n], n \in 1 \dots (T/w_s - L_1)$ // non-overlapping speech detection

Input: $\mathbf{x}_2[n], n \in 1 \dots (T/w_s - L_2)$ // used for the scatter matrices

Parameter: N_y // subspace dimension

Output: $\hat{\mathrm{A}}_{N_x \times N_y}$ // the optimal FLsD matrix

$n \leftarrow 1$ // initialise the analysis window sequence index

$v \leftarrow 1$ // initialise the class thread index

$\mathbf{m} \leftarrow \mathbf{0}_{N_x}, \mathrm{S}_m \leftarrow 0_{N_x \times N_x}, \mathrm{S}_w^h \leftarrow 0_{N_x \times N_x}$ // initialisation

while $n < \frac{T}{w_s} - L_1$ **do**

  if `isNonOverlappingSpeech` $\mathbf{x}_1[n]$ **then**

    $R \leftarrow [n \dots n + L_1 - L_2)$ // range of texture windows

    $\mathbf{m}_c \leftarrow \frac{1}{|R|} \sum_{k \in R} \mathbf{x}_2[k]$, // class thread mean

    $\mathrm{S}_c \leftarrow \frac{1}{|R|} \sum_{k \in R} \mathbf{x}_2[k] \mathbf{x}_2[k]'$ // class thread cov. mat.

    $\mathrm{S}_w^h \leftarrow \mathrm{S}_w^h + \frac{w_l}{T} \left(\mathrm{S}_c - \mathbf{m}_c \mathbf{m}_c^\top \right)$ // within-class

    $\mathbf{m} \leftarrow \mathbf{m} + \frac{w_l}{T} \mathbf{m}_c$, $\mathrm{S}_m \leftarrow \mathrm{S}_m + \frac{w_l}{T} \mathrm{S}_c$ // mixed-class

    $v \leftarrow v + 1$ // advance the speaker-thread index

  $n \leftarrow n + L_1$ // advance the analysis window

$\mathrm{S}_m \leftarrow \mathrm{S}_m - \mathbf{m}\mathbf{m}^\top$ // evaluate mixed-class scatter matrix

$\hat{\mathrm{A}} = \mathrm{argmax}_{\mathrm{A}_{N_x \times N_y}} r(\mathrm{A}, \mathrm{S}_m, \mathrm{S}_w^h)$ // apply the Fisher criterion

from several audio sources, e.g., online radio broadcasts, video sharing sites, so it can be considered as practically independent from any corpus. Note, however, that completely unsupervised approaches do exit for this task [27].

Likewise, classifying the speech as overlapping versus single-speaker can also be done in a completely unsupervised way, though this may be more complicated [28]. For practical considerations, we have built a non-overlapping versus overlapping speech classifier based on few audio files from the CANAL9 corpus. Hence, strictly speaking, the non-overlapping speech detection we have been experimented with uses some supervised information. However, we have found that this component is not essential for the success of our approach, and removing it altogether results in only $\sim 1\%$ drop in diarization accuracy. This is attributed to the fact that the overlapping speech duration is small compared to the total speech duration and therefore does not significantly alter the estimation of the within-class thread covariance matrix.

C. Obtaining Speakers and Speaker Segments

In previous sections, we have described the FLsD approach in order to find the speaker-discriminative subspace. Even though we have been based on speaker threads, no actual speakers have been detected. This section shows how the proposed approach can be integrated into a diarization system that allows to obtain speakers and speaker segments. Before going into the details, it is important to point out that this is fairly independent to the FLsD algorithm. Therefore, other approaches could be plugged in to obtain the complete diarization system, by simply using the projected feature vectors instead of the original ones.

We implicitly assume that speakers are represented by probability distributions hosted in the N_y-dimensional speaker-discriminative subspace defined by matrix $\hat{A}$. However, the mainstream approach of estimating the speaker probability distributions, which is a hard task even at the low-dimensional projected feature vector space, may actually be avoided. This is because the diarization goal is not to create generative models for speakers but to decide the most likely speaker for each speech

segment. Therefore, we directly estimate the conditional probabilities of speakers given the projected vectors. In particular, we employ a non-parametric discriminative classifier, namely the K-Nearest Neighbor classifier (K-NN). Diarization is then equivalent to classifying projected feature vectors to the most likely speaker, in the K-NN sense. Using K-NN is also very convenient, since it does not require any training and the reference sets in this setting are small.

The labels used by K-NN to estimate the speaker probabilities are obtained by clustering the projected feature vectors of the fixed-duration speech segments, 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.

Overall, the following steps are applied:

Step 1) For each non-overlapping window of duration w_l, a feature vector $\mathbf{x}_1$ is generated (Section III-A) and subsequently projected to the precomputed FLsD subspace (Section III-B) resulting to vector $\mathbf{y} = \hat{A}^\top \mathbf{x}_1$.

Step 2) The set of all projections $\mathbf{y}$, independently to their sequential index, are partitioned using the Fuzzy C-Means clustering algorithm described in [17]. By considering each cluster to be a speaker, a speaker label is assigned to each $\mathbf{y}$. Moreover, for each $\mathbf{y}$, a speaker probability is estimated as the ratio of feature vectors attributed to the given speaker that are among the K closest to the given $\mathbf{y}$. In our experiments, K has been set as the 10% of the sample set.

Step 3) Using the time sequence of the estimated speaker labels, the speakers transition matrix as well as the prior probabilities of speakers are evaluated. Together with the K-NN, these define an HMM model with states as many as speakers. A particular non-speech state is also considered, and the corresponding prior and transition probabilities are evaluated based on the speech/non-speech classification results. Then, by applying the Viterbi algorithm, the most probable speaker path is obtained.

Step 4) Through HMM smoothing, some segments end up with having a speaker label different from the one proposed by the clustering algorithm. It follows that the K-NN estimates of the speaker conditional distributions are modified and hence Step 3 can be repeatedly applied to further improve the results. It has been experimentally found that this process converges within a few iterations. Note that this iterative scheme would have not been so easy to follow, if we have used models needing training, such as generative models, since these should be reestimated in each iteration.

Step 5) Successive segments of the same speaker are merged, forming longer speaker-homogeneous segments.

The fuzzy clustering algorithm that has been used requires knowing the number of speakers beforehand. Since this information is typically not available, Steps 2 to 4 are applied for a range of number of speakers N_s and the silhouette width criterion [29] is used to decide about the quality of the clustering result in each case and therefore the optimal number of speakers.

IV. Experimental Results

A. Data and Performance Measures

To evaluate the proposed algorithm, we have used the audio part of the publicly available Canal 9 corpus [18]. The corpus consists of 70 debate recordings where participants do not act, but are actually engaged in spontaneous, and often vivid, conversations. In total, there are 190 unique participants, 165 men and 25 women, where each one participates in a maximum of three different debates. Manual diarization is also provided and has been used as the ground truth for the evaluation. A basic audio preprocessing has been conducted, by downsampling it to 16 kHz and removing music segments typically occurring at the beginning and end of recordings.

We evaluated our approach for diarization based on the diarization accuracy rate (DAR), defined as the ratio of *correctly* clustered single-speech segments duration to the *total* duration of non-overlapping speech. DAR is based on the optimal one-to-one mapping of the cluster labels with the true speaker labels. This is achieved by applying the Hungarian method to the resulting confusion matrix between clusters and speakers. DAR is closely related to the unit-complement of the diarization error rate (DER), which is commonly used in diarization evaluation [30]. Their difference is that DER is evaluated using the exact boundaries, as opposed to the fixed w_l segments, and that DAR is evaluated only on single-speaker speech segments, ignoring both non-speech and overlapping speech segments. We have experimentally found that for all methods compared for the CANAL9 corpus, DAR is overestimated by $\sim 2.5\%$ with respect to $(1 - \mathrm{DER})$.

In addition, we have considered a number of clustering evaluation metrics. In particular, we have used the average cluster purity (ACP) and average speaker purity (ASP) measures, defined, respectively, as $\mathrm{ACP} = (1/N) \sum_{i=1}^{N_c} \max_{j=1 \ldots N_s} n_{ij}$ and $\mathrm{ASP} = (1/N) \sum_{j=1}^{N_s} \max_{i=1 \ldots N_c} n_{ij}$ where N is the total number of segments, N_s is the total number of speakers, N_c is the total number of detected clusters, and n_{ij} is the total number of segments classified in cluster i and spoken by speaker j [31]. ACP focuses on the frequency of the most common speaker into each cluster, while ASP on the frequency of the most common detected speaker (cluster) within each speaker class. An alternative definition of the purity metric proposed in [1] gives essentially the same results.

Finally, we have also considered the normalized mutual information max $(\mathrm{NMI}_{\max})$ measure as suggested in [32], which is used to compare two partitions over the same data. It is bounded in $[0, 1]$, equalling 1 when the two partitions are identical, and 0 when they are independent, i.e., they share no information about each other. Other variants of other mutual information measures proposed in [32] gave similar results.

B. Evaluation of the FLsD Approach for Diarization

Table I shows the performance indexes of the diarization system for the CANAL 9 corpus in the original space and the FLsD subspace. The original space is formed using the MFCC features described in Section III-A. The ratio-trace criterion (4) for FLsD has been preferred, instead of the trace-ratio (5), since it achieved slightly better performance. Clearly, the FLsD subspace proposed here improves significantly the DAR,

TABLE I
COMPARISON OF THE DIARIZATION PERFORMANCE ON THE CANAL 9 CORPUS, WITH AND WITHOUT THE FLSD SUBSPACE STEP. A BULLET INDICATES STRONG ($a = 0.1\%$) STATISTICAL SIGNIFICANCE

Performance measure %	Feature space		
	Original	FLsD	
DAR	54.4±16.5	86.3±10.6	•
ACP	46.7±16.9	83.5± 9.8	•
ASP	90.0± 6.9	85.6± 7.9	
NMI_{max}	30.4±17.1	74.8±13.0	•

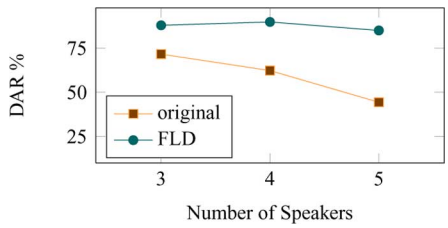

Fig. 3. Significance of the FLsD subspace in function of the number of speakers. Each DAR score has been evaluated as an average of DAR scores on CANAL9 recordings with the given number of speakers.

ACP, and NMI_{max} indexes. In particular, the DAR is improved by $\sim 32\%$. All improvements have been found to be strongly statistically significant (paired t-test with $a = 0.1\%$).

On the other hand, the ASP rate is decreased, though the difference has not been found to be statistical significant even in a more tolerate setting (paired t-test with $a = 1\%$). This loss may be attributed to the fact that clustering in the original feature space tended to overestimate the number of speakers. Therefore, clusters become small and the chance that samples of a speaker are in the same cluster becomes larger. Also, it is important to point out that even though a variance in performance is notable with and without the FLsD step, diarization in the FLsD subspace reduces standard deviation of the DAR by 6%. Thus, the proposed FLsD step makes diarization performance more predictable.

1) Number of Speakers: Fig. 3 shows how the gain in performance varies with different number of speakers. Even though the number of speakers in the CANAL 9 corpus varies only between 3 and 5, the curves allow to drive some conclusions. Namely, one may see that the system performance without FLsD deteriorates as the number of speakers increases, while it remains almost unaffected using the FLsD subspace. Thus, it seems that the FLsD subspace makes diarization more resilient to increased number of speakers.

2) Comparative Study: To see the relative merits of the proposed approach, Fig. 4 shows a comparison of the DAR indexes evaluated using competitive methods for reduced size feature space. Namely, comparison is done against the following:

- The subspace obtained using PCA, as in [10].
- The subspace obtained using FLD together with a manually labeled independent set, as in [12]. Regarding the independent sets, we have used the following corpora: the CHAINS corpus [33], the UMICH corpus [34] and the complete CANAL-9 corpus. Note that test audio signals speakers are among the 190 speakers of CANAL-9 corpus, so the later is not a truly independent set. However, we use it to see the merits of the this approach when the independent set shares the same recoding conditions.
- Finally, we also compare against the FLD upper bound of FLsD, obtained by using the actual speaker labels. This is a purely theoretical comparison, since these are not available but provides an upper bound on the performance of our system using FLsD as well as an index of how justified is the approximation of the within-class by the within-class thread scatter matrix.

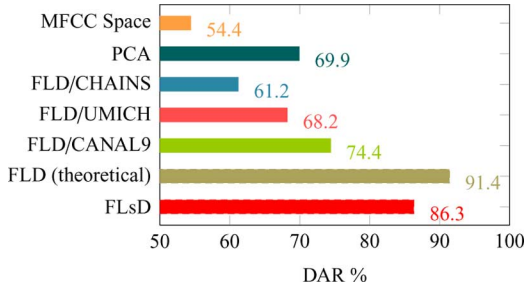

Fig. 4. Comparison of the proposed approach with other approaches for subspace extraction. The bars depicts the DAR rate obtained by each approach.

To allow a fair comparison, the same diarization steps described in Section III-C have been used in all experiments, except for the subspace extraction approach, where all competitive methods have been used in place of the proposed FLsD method. The reported results correspond to the optimal subspace size for each method, obtained through exhaustive search. By looking at the performance indexes, we readily conclude that the proposed approach has a clear benefit with respect to both the FLD with independent sets approaches and the PCA approach. In particular, there is a performance gain of more than $\sim 12\%$ even when using the same corpus as an independent set. Regarding the theoretical FLD upper bound, we notice that not knowing the exact classes results in $\sim 5.1\%$ loss in diarization accuracy, i.e., the proposed approximation has some non-negligible cost. Nevertheless, this cost is rather small when compared to not using the FLsD at all, since there is a $\sim 32\%$ gain when using the proposed approximation.

As a side note, PCA achieved better performance compared to FLD trained with completely independent sets of speakers (CHAINS and UMICH). This somehow contradicts the findings reported in [12] and [13]. A possible explanation is that the independent set of speakers, in these studies, has been obtained from the same reference corpus and therefore has similar overall recording conditions with the testing sets. In any case, the FLsD approach has a clear advantage, since it does no depend on an independent set.

C. Parameters Affecting the Performance

1) Size of the FLsD Extracted Subspace: The system performance depends on the size of the FLsD subspace extracted. Too few dimensions may not be enough to discriminate among speakers, while too many may add noise to the feature space, deteriorating the resulting clusters. Fig. 5 shows how the gain in performance varies with a different number of dimensions. Best performance is achieved for $N_x = 4$, which is what is used in all experiments for FLsD reported here.

2) HMM Smoothing Step: Theoretically, we expect the benefits of using the HMM step to be complementary to the benefits

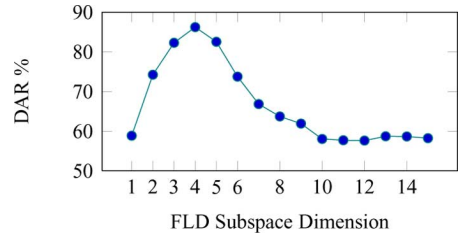

Fig. 5. Significance of the subspace dimension. The curve shows DAR as a function of the subspace dimension. Maximum DAR is obtained for $N_x = 4$.

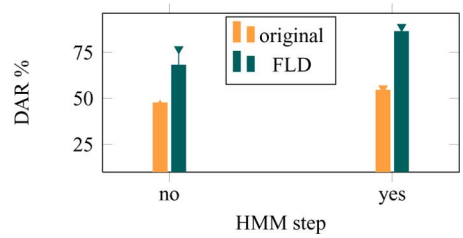

Fig. 6. Significance of the HMM step in the original and the FLD subspaces. Inverted triangle indicate the additional gain obtained when the number of speakers is given beforehand.

of using the FLsD subspace. FLsD focuses on the static representation of speech segments while the HMM tries to capture the dynamics of the conversation. This is confirmed by the experiment results shown in Fig. 6. Note however that the gain of using both FLsD and HMM is more than the sum of gains of using each one alone. Since the FLsD step precedes the HMM step, this may attributed to the fact that the HMM has more chances to correctly smooth the output when the initial probabilities are better estimated.

3) Relevant Components: The RCA method [16] proposes to scale the features with respect to how meaningful they are for differentiating the classes. Our experiments show that the results obtained simply by scaling are significantly worse than reducing the dimension, while reducing the dimension and scaling the dimensions drop DAR performance by 4%. For this reason, scaling of features has not been further considered.

V. Discussion

On a concluding note, this study has shown that the relevance of the feature space has a great influence over the diarization accuracy. The use of the proposed approach allowed to start with 54% DAR in the original MFCC-based feature space, and end with 86%, achieved by removing the speaker-irrelevant feature space. We conjecture that the performance can be further boosted, by considering other richer features to complement the initial feature space, such as prosodic [35] or amplitude modulation spectral features [36]. A second conclusion is that information that seems insignificant may be proven useful when utilized correctly. In the current study, the fact that speakers do not alternate with a very high frequency allowed to consider groups of samples which can be very likely attributed to the same, though unknown, speaker. It has been theoretically shown that considering these groups as class threads and applying FLD analysis, results in near-optimal subspaces. It has also been shown that performance using this subspace is superior to the one obtained with PCA or FLD with manually annotated independent sets. A further remark is that reducing the feature space to very few dimensions, gives a new perspective to speaker diarization. We have been able to engineer a fast system using the off the self algorithms Fuzzy C-Means and K-NN instead of specifically tailored methods. Modeling directly the conditional probabilities of speakers using a zero-cost training classifier (K-NN) also allowed us to automatically adapt probabilities after the HMM step. To give an index of the performance, diarization is $5\times$ faster than real time, using MATLAB on a Intel SU7300 and 4 GB RAM.

Comparison with the theoretical FLD upper-bound shows that there is room for improvement. First, though detection of overlapping speech has not significantly affected the performance, when speakers overlap more with each other, this factor can be an issue. A way to address this issue is to further use FLsD to extract an optimal subspace to differentiate between overlapping and non-overlapping speech. The idea in [28] can be used in this context to define class threads. Furthermore, it has been experimentally found that the benefits of using FLsD are complementary to the benefits of HMM for modeling the dynamics of speakers. Therefore, a path to further boost the diarization performance is to incorporate more elaborate ways of modeling speaker turn-taking behavior and roles [6]. Finally, extending a single-pass version of the proposed approach will eventually allow real-time speaker diarization. This entails tracking the optimal speaker relevant subspace simultaneously with clustering into speakers, and add new speakers on the fly. In this perspective, extending the recent approach in [37] to work in an adaptive FLsD subspace seems a promising direction of research.

Appendix

Proof of Lemma 1: Consider first the samples of a particular class thread. Their covariance matrix assuming that the mean of the respective class is unknown, respectively, known, is $\mathcal{E}_{\mathbf{x}\mapsto\mathbf{v}}\left[\mathbf{x}\mathbf{x}^\top\right] - \mathbf{m}_v\mathbf{m}_v^\top$, respectively, $\mathcal{E}_{\mathbf{x}\mapsto\mathbf{v}}\left[\mathbf{x}\mathbf{x}^\top\right] - \mathbf{m}_{h(v)}\mathbf{m}_{h(v)}^\top$. By averaging their difference over all threads, we obtain

$$S_h = S_w - S_w^h = \mathcal{E}_{v\in\mathcal{V}}\left[\mathbf{m}_v\mathbf{m}_v^\top - \mathbf{m}_{h(v)}\mathbf{m}_{h(v)}^\top\right]$$

where the samples correlation matrices have been cancelled out. Now, letting $\mathbf{d}_v = \mathbf{m}_{h(v)} - \mathbf{m}_v$, this difference is further rewritten as follows:

$$\begin{aligned} S_h &= \mathop{\mathcal{E}}_{v}\left[\left(\mathbf{m}_{h(v)} - \mathbf{d}_v\right)\left(\mathbf{m}_{h(v)} - \mathbf{d}_v\right)^\top - \mathbf{m}_{h(v)}\mathbf{m}_{h(v)}^\top\right] \\ &= \mathop{\mathcal{E}}_{v\in\mathcal{V}}\left[\mathbf{d}_v\mathbf{d}_v^\top - \mathbf{m}_{h(v)}\mathbf{d}_v^\top - \mathbf{d}_v\mathbf{m}_{h(v)}^\top\right] \\ &= \mathop{\mathcal{E}}_{v\in\mathcal{V}}\left[\mathbf{d}_v\mathbf{d}_v^\top\right] + \\ &\quad + \mathop{\mathcal{E}}_{c\in\mathcal{C}}\left[\mathbf{m}_c\left(\mathop{\mathcal{E}}_{v:h(v)=c}\mathbf{d}_v\right)^\top + \left(\mathop{\mathcal{E}}_{v:h(v)=c}\mathbf{d}_v\right)\mathbf{m}_c^\top\right]. \end{aligned}$$

The class threads difference from the means for every class c cancel out, i.e., $\forall\, c\in\mathcal{C} : \mathcal{E}_{v:h(v)=c}\mathbf{d}_v = \mathbf{0}$. Consequently,

$$S_h = \mathop{\mathcal{E}}_{v\in\mathcal{V}}\left[\mathbf{d}_v\mathbf{d}_v^\top\right].$$

Finally, $S_m = S_w + S_b = S_w^h + S_b^h$ through which we readily obtain that $S_b^h - S_b = S_h$.

REFERENCES

[1] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proc. ICASSP*, 1998, vol. 2, pp. 757–760.

[2] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[3] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Process.*, vol. 88, pp. 1091–1124, 2008.

[4] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. ICASSP*, 2005, vol. 5, pp. V-953–V-953.

[5] D. M. S. Meignier, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Comput. Speech Lang.*, vol. 20, pp. 303–330, 2006.

[6] A. Vinciarelli, "Capturing order in social interactions [social sciences]," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 133–152, Sep. 2009.

[7] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, vol. 27, pp. 1775–1787, 2009.

[8] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Improving speaker diarization," in *Proc. Rich Transcript. Workshop (RT-04)*, 2004.

[9] H.-P. Shen, J.-F. Yeh, and C.-H. Wu, "Speaker clustering using decision tree-based phone cluster models with multi-space probability distributions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1289–1300, Jul. 2011.

[10] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proc. ICASSP*, 2008, pp. 4133–4136.

[11] W. Tsai, S. Cheng, and H. Wang, "Automatic speaker clustering using a voice characteristic reference space and maximum purity estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1461–1474, May 2007.

[12] S. M. Chu, H. Tang, and T. S. Huang, "Fishervoice and semi-supervised speaker clustering," *Proc. ICASSP*, pp. 4089–4092, 2009.

[13] H. Tang, S. M. Chu, and T. S. Huang, "Spherical discriminant analysis in semi-supervised speaker clustering," in *Proc. NAACL '09, Companion volume: Short Papers*, Morristown, NJ, 2009, pp. 57–60.

[14] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006, vol. 2.

[15] T. Giannakopoulos and S. Petridis, "Unsupervised speaker clustering in a linear discriminant subspace," in *Proc. 9th Int. Conf. Mach. Learn. Applicat. (ICMLA)*, 2010, pp. 1005–1009.

[16] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *JMLR*, vol. 6, pp. 937–937, 2005.

[17] R. Babuka, P. Van der Veen, and U. Kaymak, "Improved covariance estimation for Gustafson-Kessel clustering," in *Proc. FUZZ-IEEE'02*, 2002, vol. 2, pp. 1081–1085.

[18] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "CANAL9: A database of political debates for analysis of social interactions," in *Proc. Affective Comput. Intell. Interact. Workshops*, 2009, pp. 1–4.

[19] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Boston, MA: Academic Press, 1990.

[20] D. Foley and J. Sammon Jr., "An optimal set of discriminant vectors," *IEEE Trans. Comput.*, vol. 100, no. 3, pp. 281–289, Mar. 1975.

[21] L.-H. Zhang, L.-Z. Liao, and M. Ng, "Superlinear convergence of a general algorithm for the generalized Foley–Sammon discriminant analysis," *J. Optimiz. Theory Applicat.*, pp. 1–13, 2011.

[22] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recogn.*, vol. 37, pp. 857–874, 2004.

[23] G. Golub and C. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1996, vol. 3.

[24] Y. Nakatsukasa, "Absolute and relative weyl theorems for generalized eigenvalue problems," *Linear Algebra Its Applicat.*, vol. 432, pp. 242–248, 2010.

[25] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[26] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 174–186, Jan. 2009.

[27] H. Maganti, P. Motlicek, and D. Gatica-Perez, "Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms," in *Proc. ICASSP*, 2007, vol. 4, pp. IV-1037–IV-1040.

[28] M. Huijbregts, D. Leeuwen, and F. Jong, "Speech overlap detection in a two-pass speaker diarization system," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1063–1066.

[29] L. Vendramin, R. Campello, and E. Hruschka, "On the comparison of relative clustering validity criteria," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 733–744.

[30] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*. New York: Springer, 2008, vol. 4625, pp. 373–389.

[31] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Inf. Retrieval*, vol. 12, pp. 461–486, 2009.

[32] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *JMLR*, pp. 2837–2854, 2010.

[33] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: Characterizing individual speakers," in *Proc. SPECOM*, 2006, pp. 431–435.

[34] A. Reda, S. Panjwani, and E. Cutrell, "Hyke: A low-cost remote attendance tracking system for developing regions," in *Proc. 5th ACM Workshop Netw. Syst. Develop. Regions*, 2011, pp. 15–20.

[35] M. Kockmann, L. Ferrer, L. Burget, E. Shriberg, and J. Cernocky, "Recent progress in prosodic speaker verification," in *Proc. ICASSP*, 2011, pp. 4556–4559.

[36] J. Bach, J. Anemüller, and B. Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features," *Speech Commun.*, 2010.

[37] L. Ren, L. Du, L. Carin, and D. Dunson, "Logistic stick-breaking process," *JMLR*, vol. 12, pp. 203–239, 2011.

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 main research interests are pattern recognition and multimedia analysis.

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, Athens, Greece, in 1996, the M.Sc. degree in pattern recognition from UPMC, Paris, France, in 1997 and the Ph.D. degree from the Department of Informatics and Telecommunications, University of Athens, Athens, Greece.

He is a Research Associate at the Institute of Informatics and Telecommunication, NCSR "Demokritos," Athens. His interests include machine learning and multimedia analysis.