

Unsupervised Speaker Clustering in a Linear Discriminant Subspace

Theodoris Giannakopoulos, Sergios Petridis
Computational Intelligence Laboratory
NCSR “Demokritos”
Athens, Greece
{tyianak,petridis}@iit.demokritos.gr

Abstract—We present an approach for grouping single-speaker speech segments into speaker-specific clusters. Our approach is based on applying the K-means clustering algorithm to a suitable discriminant subspace, where the euclidean distance reflect speaker differences. A core feature of our approach is approximating speaker-conditional statistics, that are not available, with single-speaker segments statistics, which can be evaluated, thus making possible to apply the LDA algorithm for finding the optimal discriminative subspace, using unlabeled data. To illustrate our method, we present examples of clusters generated by our approach when applied to the ICMLA 2010 Speaker Clustering Challenge datasets.

Keywords—speaker clustering, linear discriminant analysis, K-means

I. INTRODUCTION

Speaker clustering is an important component of a speaker diarization system which aims to group together speech segments produced by the same speaker within an audio stream [1]. Speaker clustering assumes that the speech stream has been previously chopped into homogeneous segments, each one containing speech from a single speaker. We will refer to such segments as single-speaker segments.

A typical approach for speaker diarization systems is to apply hierarchical agglomerative clustering (HAC) to perform speaker clustering after segmentation [2], [3]. In the method present here, we follow a different path, aiming to explore how Linear Discriminant Analysis may be used to improve the results of clustering, by finding optimal speaker-discriminative subspace. LDA is a widely used method that has been used for various tasks, in a supervised context, such as in [4], to distinguish speech from non-speech segments, or in semi-supervised context, such as in [5], for speaker clustering. The main novelty of our approach is deriving an approximation of means and covariance matrices used by the LDA criterion, using information from single-speaker segments, thus making possible the application of LDA in a completely unsupervised framework.

The remainder of the paper has the following structure. Section II presents the steps of the methodology used to find speaker clusters, which includes generating the feature vectors, applying the LDA algorithm to find a suitable speaker-discriminant subspace and applying the K-means algorithm to the subspace, to find the speaker clusters. Section III shows

how to approximate speaker-conditional statistics, based on single-speaker segment information. Finally, Section IV illustrates the results of the approach when applied to the ICMLA 2010 Speaker Clustering Challenge datasets and Section V summarizes the conclusion and discusses future extensions of the approach.

II. METHODOLOGY

In this section, we describe the algorithms upon which our speaker clustering algorithm is based on. Namely, in Section II-A we describe how the feature vectors are generated using short-time analysis and mid-term statistics extraction, in Section II-B we outline the principles of the LDA algorithm used to extract a suitable subspace of the feature space and in Section II-C the K-means clustering algorithm used to group speech segments with respect to the speaker.

A. Generating the feature vectors

The LDA method discussed in Section II-B does not apply directly on speech samples, but on feature vectors extracted from speech segments of specific size. In this section, we describe how these feature vectors are formed.

1) *Single speaker speech segments*: To begin with, we assume that the speech stream has been previously split into speech segments, each one enunciated from a single speaker. This may have been accomplished using for example the BIC criterion [6] or one of its variants [7]. For the purpose of finding an optimal subspace, speech segments, for which there is significant uncertainty regarding whether they stem from one or several speakers, may be removed from the data set used to finding the optimal subspace. In this way, we may safely assume that the probability that a speech segment stems from more than one speaker is practically zero.

At the end of this procedure, the speech input has been converted to a list of single-speaker speech segments of varying duration, indexed by the order of their enunciation.

$$\mathbf{S} = \{\mathbf{s}_c\}, \quad c = 1 \dots C$$

2) *Short-time analysis*: For each such speech segment \mathbf{s}_n , a short term analysis is then performed, resulting in 12 mel-frequency cepstrum coefficients (MFCC) and an energy feature for each window considered. In the experiment presented,

short-term window size has been set to 25msec while short-term window step to 10msec. Depending on the duration of the speech segment \mathbf{s}_c , a specific number M_c of feature vectors are generated:

$$\mathbf{s}_c \rightarrow \{\mathbf{f}_{c,m}\}, \quad \mathbf{f}_{c,m} \in \mathbb{R}^{D_1}, m = 1 \dots M_c, D_1 = 13$$

3) *Mid-term statistics*: The sequence of 13-MFCC feature vectors corresponding to a single-speaker segment is then given as input to a second-level analysis, with a specific window size W_m and step 1, aiming to capture mid-term characteristics of the speech signal. In the experiments presented, window size has been set to 30 (which corresponds to 300msec of speech) and window step to 1 (which corresponds to 10msec of speech). For each mid-term window, two statistics for each MFCC are calculated: the mean and the standard deviation.

$$\mathbf{x}_{c,m}(i) = \frac{1}{W_m} \sum_{l=i \dots i+W_m} \mathbf{f}_{c,m}(l), \quad i = 1 \dots 13$$

$$\mathbf{x}_{c,m}(i+13) = \frac{1}{W_m} \sum_{l=i \dots i+W_m} (\mathbf{f}_{c,m}(l) - \mathbf{x}_{c,m}(i))^2$$

As a result, at the end of mid-term analysis, the complete speech input is represented as a list of sequences of 26-dimensional feature vectors, each element of the list being indexed by the order of the particular single-speaker segment enunciation:

$$\begin{aligned} \mathbf{S} &\rightarrow \{\mathbf{s}_c\}, \quad c = 1 \dots C, \\ \mathbf{x}_{c,m} &\rightarrow \{\mathbf{x}_{c,m}\}, \quad \mathbf{x}_{c,m} \in \mathbb{R}^D, m = 1 \dots M_c, D = 26 \end{aligned} \quad (1)$$

B. Linear Discriminant Analysis

Reducing the dimension of the observation space, by finding a suitable linear subspace in which the class separability is optimally maintained, has been commonly used as a pre-processing step in pattern recognition systems [8], since a reduced feature space dimension may lead to better classifier training with improved generalization ability. The benefits of this approach have been demonstrated even when combined with very competent classifier models, such as support vector machines [9]. Here, we show that it may also be used in a unsupervised framework.

In particular, Linear Discriminant Analysis (LDA), also known as generalized Fisher criterion, has been used in many application fields, such as face recognition [10], [11], or document classification [12]. Several extensions and variations of the basic LDA algorithm have been developed concerning either implementation and robustness issues [13]–[15] or deviations from the model assumptions [16], [17].

The basic idea behind LDA is to extract a subspace in which the classes means are far from each other, whereas the within class covariance matrices (i.e. class conditional covariance matrices) are small. In our context, classes roughly correspond to speakers, though details are given in Section III.

To formulate the LDA criterion, we need the following definitions:

- the feature vector of MFCC-based coefficients, \mathbf{x}
- a class, $k, k = 1 \dots K$,

- the overall mean of vectors \mathbf{x} , $\mathbf{m} = \mathbb{E}[\mathbf{x}]$,
- the mean of all \mathbf{x} attributed to class k , $\mathbf{m}_k = \mathbb{E}_{\mathbf{x}|k}[\mathbf{x}]$,
- the covariance matrix of vectors \mathbf{x} :

$$\Sigma = \mathbb{E}_{\mathbf{x}}[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top],$$

- the covariance matrix of vectors \mathbf{x} attributed to class k :

$$\Sigma_k = \mathbb{E}_{\mathbf{x}|k}[(\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^\top], \text{ and}$$

- the average of class-conditional covariance matrices Σ_k :

$$\bar{\Sigma} = \mathbb{E}_k[\Sigma_k]. \quad (2)$$

Then, given a positive integer m , lower than the original feature vector size n , the LDA criterion requires to find among all possible $n \times m$ full rank matrices \mathbf{A} , the matrix defined by

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmax}} \frac{|\mathbf{A}^\top \Sigma \mathbf{A}|}{|\mathbf{A}^\top \bar{\Sigma} \mathbf{A}|}. \quad (3)$$

The criterion involves actually the *sample estimates* of means, covariance matrices and class probabilities, while it can be expressed in several other equivalent forms (see [18]). The LDA-optimal matrix can be found by solving a generalized eigenvalue problem. In particular, one has to form a matrix using the m eigenvectors of $\bar{\Sigma}^{-1}\Sigma$ which correspond the m greater eigenvalues.

Note also that the LDA subspace engendered is assured to be optimal under two conditions: first, that the class-conditional distributions are Gaussian with the same covariance matrix and, second, that the dimension of the subspace engendered is at least as big as the number of classes minus one (see [19]). Nevertheless, the wide applicability of the criterion shows that it is rather robust under violations of these conditions.

C. The K-means algorithm

K-means is a widely used clustering algorithm formulated based on a formal objective function. Namely, given a set of N samples in \mathbb{R}^D , and an integer K , the objective is to determine a set of K points in \mathbb{R}^D , the *centers*, so as to minimize the mean squared distance from each data point to its nearest center. By letting \mathbf{o}_i be the i -th center and $S_k = \{\mathbf{x}_j\}$ the set of points closest to the k -th center, the optimal partition of points $S = \{S_k\}$ is defined as:

$$\hat{S} = \underset{S}{\operatorname{argmin}} \sum_{k=1}^K \sum_{\mathbf{x}_j \in S_k} \|\mathbf{x}_j - \mathbf{o}_k\|_2$$

A common way to find a solution is to use the Lloyd's algorithm, whose underlying principle is that the optimal placement of a center is at the centroid of the associated cluster. Lloyd's algorithm can be formulated as an expectation-maximization algorithm as follows.

- start k samples are selected randomly as centers of the clusters
- e-step Each sample is assigned to the cluster with the closest center

m-step The means of the clusters are calculated and replace the centers found at the previous iteration
 stop If the new means do not differ from the old ones, stop iterating.

Though the above algorithm always converges (unless samples are exactly equidistant from two centers), the result is not necessarily the global minimum, but a local one that depends on the initial selection of centers. Nevertheless, in our context, a global optimum has significant chances of being found given that the dimension of the space in which clustering takes place is probably very low.

III. OPTIMAL DISCRIMINANT SUBSPACES

A. Extracting suitable subspaces

A basic assumption made by our algorithm is that the the *original* D -dimensional feature space contains enough information to discriminate among speakers. This assumption most probably holds for the MFCC/energy-based feature vector that our experiments depend on. In other words, we expect that, the euclidean distance between two feature vectors that correspond to different speakers, all other things being equal (e.g. same phone), would be greater than two feature vectors that correspond to the same speaker, again all other things being equal.

However, it also holds that the same feature space contains information to discriminate between other speech qualities, such as phones, speech energy or pitch. Therefore, when attempting to cluster samples in this feature space, using the euclidean distance, the resulting clusters may correspond to various similarities and will not necessarily group together samples from the same speaker.

On the other hand, assuming that feature vectors stemming from different speakers are linearly separable, then, as discussed in Section II-B, a suitable low dimension linear subspace would be enough to discriminate between speakers. In particular, for two speaker this would be a line. Relaxing the assumption with respect to linear separability, we may still expect that a low dimensionality subspace will contain all speaker-discriminative information.

It then follows that, by construction, an optimal such speaker-discriminant subspace will not allow as to discriminate between speech qualities that are irrelevant to speaker discrimination. Therefore, performing clustering in this subspace has significantly more chances to group segments with respect to the speaker.

The remaining issue is, of course, finding the speaker-discriminant subspace.

B. Speaker-specific segments as classes

At first sight, the problem we are addressing falls into the *unsupervised* pattern recognition tasks, since there is no available information with respect to which speech segment belongs to which speaker. In particular, what is missing from the LDA criterion of Eq. (3) is the average speaker-conditional matrix $\bar{\Sigma}$ as defined in Eq. (2). However, by closer examination of the available data, as summarized in Eq. (1), one may see

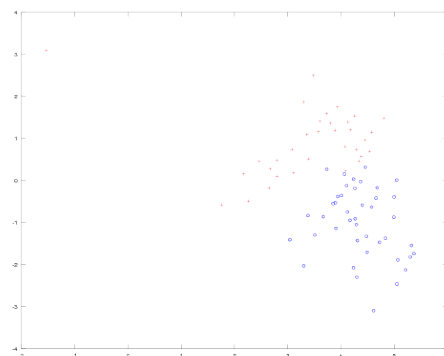


Fig. 1. Clusters found for the 4th dataset, depicted in the LDA-optimal two-dimensional space

that there is some information that may be used to evaluate $\bar{\Sigma}$ approximatively.

Namely, we do know that, for a particular c , all M_c feature vectors belong to the same speaker. Let the mean of these feature vectors be

$$\mathbf{m}_c = E_{\mathbf{x}|c}[\mathbf{x}], \quad \forall c \in \{1 \dots C\}$$

and their covariance matrix

$$\Sigma_c = E_{\mathbf{x}|c}[(\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^\top], \quad \forall c \in \{1 \dots C\} \quad (4)$$

Now, for a particular speaker k , there is a subset of \mathbf{S} , say \mathbf{S}_k , that contains all its speech segments:

$$\mathbf{S}_k = \{\mathbf{s}_c\} \subset \mathbf{S}, \quad c \in C_k \subset \{1, \dots, C\}.$$

One may easily see that, the mean of the feature vectors that belong to any of these \mathbf{s}_c , i.e. all feature vectors corresponding to speaker k , can be exactly evaluated as the mean of all \mathbf{m}_c for speech segments of the same speaker:

$$\mathbf{m}_k = E_{c \in C_k}[\mathbf{m}_c], \quad \forall k = \{1 \dots K\}.$$

Going one step further, we may reasonably argue that, under the assumption that each one of the \mathbf{s}_c is *approximately* i.i.d distributed with all others segments of the same speaker, then

$$\mathbf{m}_c \simeq \mathbf{m}_k, \quad \forall c \in C_k \quad (5)$$

i.e. the mean feature vector within each speech segment \mathbf{m}_c is approximately equal to the mean feature vector of all speech segments of the same speaker \mathbf{m}_k .

This allow as to approximate the speaker-conditional covariance matrix as

$$\begin{aligned} \Sigma_k &= E_{\mathbf{x}|k}[(\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^\top] \\ &= E_{c \in C_k} [E_{\mathbf{x}|c}[(\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^\top]] \\ &\simeq E_{c \in C_k} [E_{\mathbf{x}|c}[(\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^\top]] \\ &\simeq E_{c \in C_k} [\Sigma_c] \end{aligned} \quad (6)$$

and, last, their average (see Eq. 2) as

$$\begin{aligned} \bar{\Sigma} &= E_k [\Sigma_k] \\ &\simeq E_k [E_{c \in C_k} [\Sigma_c]] = E_c [\Sigma_c] \end{aligned} \quad (7)$$

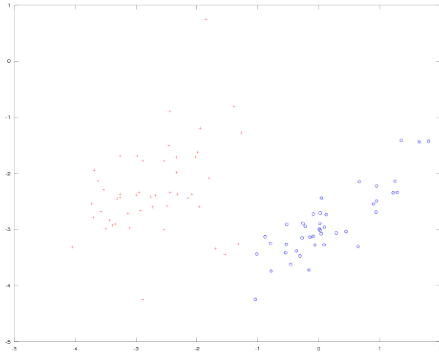


Fig. 2. Clusters found for the 4th dataset, depicted in the LDA-optimal two-dimensional space

In other words, we may approximate the average speaker covariance matrix by the average single-speaker segment covariance matrix, which can be readily evaluated from the available data.

Note that the quality of the approximation of Eq. (7) depends only on the validity of assumption of Eq. (5). In practice, this means that the average speaker characteristics do not change significantly from segment to segment, for each particular speaker. Also, one has to note the risk that the sample estimate of means for each segment, \mathbf{m}_c , will be more biased than the ones that would have been evaluated for each speaker, \mathbf{m}_k , since the size of the respective sample set may be considerably smaller.

IV. EXPERIMENTS

The experiments done comply to the *first task* of the ICMLA 2010 Speaker Clustering Challenge, as described in http://www.icmla-conference.org/icmla10/CFP_Challenge1_files/CFP_Challenge1.html. Namely, 7 datasets are provided, each one of them comprised of speech coming from two different speakers. The aim of the proposed method is to identify two clusters within each dataset.

As an illustration of the results of the proposed method, Figures 1 and 2 show how clusters are represented in the first two dimensions of the LDA-optimal subspace found.

V. CONCLUSION

We have presented a methodology that allows grouping single-speaker speech segments into speaker-specific clusters. The main novelty of the methodology is approximating the unknown average speaker conditional covariance matrix of feature vectors with the average single-speaker-segment conditional covariance matrix, that can be evaluated from unlabeled data. Plugging these covariances matrix into the LDA criterion allowed us to perform clustering in a low-dimensional subspace where the euclidean distance has improved chances of reflecting speaker differences.

As a future direction, we work on integrating the dimensionality reduction step with the clustering step, in order to allow for more robust estimation of means and covariance matrices.

ACKNOWLEDGMENT

This paper has been supported by the EU, in the context of the CASAM project (Contract number FP7-217061, Web site: www.casam-project.eu).

REFERENCES

- [1] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization," *IEEE Trans. Audio, Speech and Language Eng.*, vol. 14, no. 5, pp. 1557–1565, 2006.
- [2] X. Zhu, C. Barras, L. Lamel, and J. Gauvain, "Speaker diarization: From broadcast news to lectures," *Machine Learning for Multimodal Interaction*, pp. 396–406, 2006.
- [3] S. Meignier, D. Moraru, C. Fredouille, J. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 303–330, 2006.
- [4] A. Martin, D. Charlet, and L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC," in *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, vol. 1. Citeseer, 2001.
- [5] H. Tang, S. M. Chu, and T. S. Huang, "Spherical discriminant analysis in semi-supervised speaker clustering," in *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 57–60.
- [6] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Citeseer, 1998, pp. 127–132.
- [7] S. Cheng, H. Wang, and H. Fu, "BIC-based Speaker Segmentation Using Divide-and-Conquer Strategies with Application to Speaker Diarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 1, pp. 141–157, 2010.
- [8] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, Jan 2000.
- [9] J. Kittler, "A framework for classifier fusion: Is it still needed," in *Proceedings of the Joint SPR&SPRR Workshop*, 2000, pp. 45–56.
- [10] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data – with application to face recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, Oct 2001.
- [11] Y.-D. Guo, T.-T. Shu, J.-Y. Yang, and S.-J. Li, "Feature extraction method based on the generalised fisher discriminant criterion and facial recognition," *Pattern Analysis & Applications*, vol. 4, pp. 61–66, 2001.
- [12] K. Torkkola, "Linear discriminant analysis in document classification," in *IEEE ICDM*, 2001.
- [13] D. M. Hawkins and G. J. McLachlan, "High breakdown lda," *Journal of the American Statistical Association*, vol. 92, no. 437, pp. 136–143, mar 1997.
- [14] J. Yang and J.-Y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognition*, vol. 36, p. 563, 2003.
- [15] D. Xu, J. C. Principe, and H.-C. Wu, "Generalized eigen-decomposition with an on-line local algorithm," *IEEE Signal Processing Letters*, vol. 5, pp. 298–301, 1999.
- [16] T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society Series B-Methodological*, vol. 58, no. 1, pp. 155–176, 1996.
- [17] R. Lotlikar and R. Kothari, "Fractional step dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 623–627, jun 2000.
- [18] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press Limited, Boston, MA, 1990.
- [19] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York, 1992.