Investigation of Segmentation in i-Vector based Speaker Diarization of Telephone Speech

Zbyněk Zajíc[1], Marie Kunešová[1,2], and Vlasta Radová[1,2]

University of West Bohemia, Faculty of Applied Sciences,
[1]NTIS - New Technologies for the Information Society and [2]Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic, www.zcu.cz
zzajic@ntis.zcu.cz, mkunes@kky.zcu.cz,radova@kky.zcu.cz

Abstract. The goal of this paper is to evaluate the contribution of speaker change detection (SCD) to the performance of a speaker diarization system in the telephone domain. We compare the overall performance of an i-vector based system using both SCD-based segmentation and a naive constant length segmentation with overlapping segments. The diarization system performs K-means clustering of i-vectors which represent the individual segments, followed by a resegmentation step. Experiments were done on the English part of the CallHome corpus. The final results indicate that the use of speaker change detection is beneficial, but the differences between the two segmentation approaches are diminished by the use of resegmentation.

Keywords: speaker diarization, speaker change detection, i-vector, segmentation

1 Introduction

Speaker diarization is defined as the task of categorizing different speech sources in an unlabeled conversation. Or in other words, determining "Who spoke when", typically without any prior information regarding the number and identities of the speakers.

The majority of diarization systems follow one of two basic approaches. The most common approach consists of the segmentation of the input signal, followed by the merging of the segments into clusters corresponding to the individual speakers [1, 2]. The alternative is to combine the segmentation and clustering steps into a single iterative process [3, 4].

In systems which have a standalone segmentation step, speaker change detection (SCD) is often applied to this purpose, as it allows to obtain segments which ideally contain only the speech of a single speaker (e.g. [1]). However, due to some of the common obstacles typically present in spontaneous telephone conversation, namely very short speaker turns and frequent overlapping speech, diarization systems aimed at telephone speech often omit the SCD process and use a simple constant length segmentation of areas of speech found by a speech activity detector (e.g. [2, 5]).

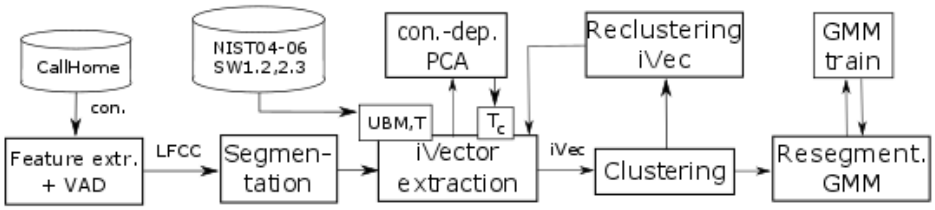

Fig. 1. Diagram of the diarization process.

In this paper, we compare the two segmentation approaches on telephone data from the CallHome corpus [15]. Our goal is to determine whether the SCD approach offers any improvement under such conditions.For this purpose, we implement an i-vector based speaker diarization system. The use of i-vectors in speaker diarization has become increasingly popular in recent years [2, 5], following their success in speaker recognition tasks [6, 7].

This paper is organized as follows: The i-vector based speaker diarization system is described in Section 2. In Section 3, two approaches to segmentation are introduced: segmentation with constant length segments and segmentation based on SCD. The i-vector extraction is explained in Section 4, clustering using K-means in Section 5 and the resegmentation step is described in Section 6. The comparison of the efficiency of the two proposed segmentation approaches is presented in Section 7.

2 Speaker Diarization System

Our speaker diarization system is based on the use of i-vectors to represent segments of speech, as introduced in [8]. The diarization process starts with the extraction of acoustic features from the conversation and the identification of the regions of speech by a voice activity detector. Following this, the non-speech regions are discarded and the rest is split into short segments, using SCD-based or constant length segmentation. In the next step, a single i-vector is extracted from each segment and the i-vectors are clustered using cosine distance in order to determine which parts of the signal were produced by the same speaker. Finally, the system iteratively performs resegmentation using a similar i-vector based clustering process, followed by a single iteration using GMMs to refine the final results. A diagram of our diarization system can be seen in Figure 1 and the main steps are described in detail in the following sections.

3 Segmentation

The purpose of the segmentation step of a speaker diarization system is to divide an audio recording into short segments, so that they can be subsequently

merged into clusters corresponding to the individual speakers. The length of the segments should be enough to allow the extraction of speaker-identifying information, in our case represented by an i-vector, while limiting the risk of a speaker change being present within the segment, as may happen in longer segments, depending on the used method. In the following subsections, we describe the two segmentation approaches which were considered.

3.1 Constant Length Segments

The naive approach to segmentation is to simply split the speech regions into short segments of fixed length. The main issue with this simple method is that the segment boundaries do not correspond in any way to the speaker change points and so many of the segments may contain the speech of more than one speaker. For this reason, it is preferable to use very short segments. On the other hand, a certain minimal duration is required for i-vector extraction. Typically, this is selected as 1-2 seconds of speech. As in [2], segment overlap is used to increase the amount of information contained in a single i-vector while retaining the same precision of the segmentation.

3.2 Speaker Change Detection

The standard approach to speaker change detection consists of applying a pair of sliding windows on the signal and computing the distance between their contents. Speaker changes are then found at the boundary between the two windows, at the points in which the distance achieves a significant local maximum. An example of this approach can be found in [1].

Commonly used distance metrics include the Bayesian Information Criterion (BIC), Generalized Likelihood Ratio (GLR) and Kullback-Leibler divergence.

In our system, we use a GLR-based segmentation. In order to obtain segments of consistent length, comparable to the constant length approach described in Section 3.1, we use a two-step algorithm which incorporates a fixed minimum and maximum segment length.

In the first step of the segmentation, we identify a smaller number of the most likely speaker change points by performing standard GLR-based speaker change detection using two neighboring sliding windows of 2 s with a step size of 0.1 s.

The distance between two windows X_i and X_j is calculated as

$$d(X_i, X_j) = -\log GLR(X_i, X_j)\,, \tag{1}$$

where $GLR(X_i, X_j)$ is the generalized likelihood ratio, which is defined as

$$GLR(i, j) = \frac{L(X_i \cup X_j | M)}{L(X_i | M_i) \cdot L(X_j | M_j)} \tag{2}$$

and is used to express whether X_i and X_j are better represented by a single model M or two different ones, M_i and M_j [9]. In our system, M, M_i and M_j are

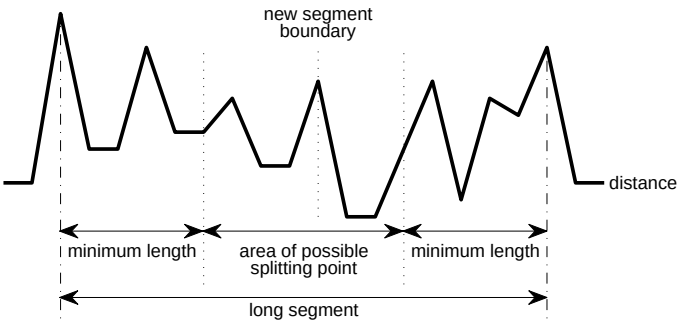

Fig. 2. The process of splitting longer segments.

single Gaussians with full covariance matrices, estimated from the corresponding data.

Likely speaker changes are identified as the locations of significant local maxima of the distances. For this purpose, we calculate the prominence of individual peaks in the distances and select those with values exceeding a threshold.

Peak prominence measures how much a given peak stands out within the signal and is calculated as follows: on each side of the peak, find the minimum of the signal that lies in the area between the peak and either the nearest higher point or the edge of the signal. The prominence of the peak is given as the difference between the value of the peak and the higher of the two minima.

The second step of the segmentation consists of further splitting any segments which are longer than the maximum allowed length. The point where a long segment is split is found in the following manner:

First, the system identifies an area where a split can occur, such that neither of the resulting new segments would be shorter than the minimum allowed length. If there are any peaks within this smaller area, the one with the highest prominence (as calculated during the first step of the segmentation) is selected as the new segment boundary. If no peaks are present, the segment is cut at the edge of the area, at the point where the distance is highest. Figure 2 illustrates this process.

4 Segment description

For each segment of parametrized conversation the supervector of statistics is accumulated. Subsequently, an i-vector is extracted from the supervector.

4.1 Statistics Extracted on GMM

For each segment of a parametrized conversation the supervector of statistics is accumulated. Supervector of statistics contains the first and zeroth statistical moments of speakers' data related to a Universal Background Model (UBM) based on GMM. This idea has origins in the speaker adaptation process [10], where these statistics are used as a descriptor of a new speaker.

First, a GMM trained on a huge amount of data from different speakers is used as a UBM and consists of a set of parameters $\boldsymbol{\lambda}_{\text{UBM}} = \{\omega_m, \boldsymbol{\mu}_m, \boldsymbol{C}_m\}_{m=1}^{M}$, where M is the number of Gaussians in the UBM, ω_m, $\boldsymbol{\mu}_m$, $\boldsymbol{C}_m$ are the weight, mean and covariance of the m^{th} Gaussian, respectively. In our case, the covariance matrix $\boldsymbol{C}_m$ is diagonal with vector σ_m on diagonal. Let $\boldsymbol{O} = \{\boldsymbol{o}_t\}_{t=1}^{T}$ be the set of T feature vectors $\boldsymbol{o}_t$ of dimension D of one segment of conversation, and

$$\gamma_m(\boldsymbol{o}_t) = \frac{\omega_m \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_m, \boldsymbol{C}_m)}{\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_m, \boldsymbol{C}_m)} \tag{3}$$

be the posterior probability of m^{th} Gaussian given a feature vector $\boldsymbol{o}_t$. The soft count of the m^{th} Gaussian (zeroth statistical moments of feature vectors) is $n_m = \sum_{t=1}^{T} \gamma_m(\boldsymbol{o}_t)$ and the sum of the first statistical moments of feature vectors with respect to the m^{th} Gaussian is $\boldsymbol{b}_m = \sum_{t=1}^{T} \gamma_m(\boldsymbol{o}_t)\boldsymbol{o}_t$. The speaker's supervector for given data $\boldsymbol{O}$ is a concatenation of the zeroth and first statistical moments of $\boldsymbol{O}$.

4.2 i-Vectors

For i-vectors extraction the Factor Analysis (FA) approach [11] (or extended Joint Factor Analysis (JFA) [12] to handle more sessions of each speaker) is used for dimensionality reduction of the supervector of statistics. The generative i-vector model has the form

$$\boldsymbol{\psi} = \boldsymbol{m}_0 + \boldsymbol{T}\boldsymbol{w} + \boldsymbol{\epsilon}, \;\; \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{T}$ (of size $D \times D_w$) is called the total variability space matrix, $\boldsymbol{w}$ is the segment's i-vector of dimension D_w having standard Gaussian distribution, $\boldsymbol{m}_0$ is the mean vector of $\boldsymbol{\psi}$, however often the UBM's mean supervector $\boldsymbol{m}_0$ is taken instead as an approximation, and $\boldsymbol{\epsilon}$ is some residual noise with a diagonal covariance $\boldsymbol{\Sigma}$ constructed from covariance matrices $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m$ of the UBM ordered on the diagonal of $\boldsymbol{\Sigma}$. The i-vectors are also length-normalised [7]. Details about training of total variability space matrix $\boldsymbol{T}$ can be seen in [13] or [14].

Because of the differences between each conversation (and the similarity in one conversation), we also compute a conversation dependent PCA transformation, which further reduces the dimensionality of the i-vector $\boldsymbol{w}$. The dimension of the PCA latent space is dependent on the parameter p, the ratio of eigenvalue mass [8] (in our case $p = 0.5$).

5 Clustering

The clustering of all segments is used for determining which segments are produced by the same speaker. Since our data only includes conversations with 2 speakers, we use K-means clustering into 2 clusters, based on cosine distance [8] of two i-vectors:

$$dist(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{\boldsymbol{w}_1^T \boldsymbol{w}_2}{\|\boldsymbol{w}_1\| \cdot \|\boldsymbol{w}_2\|}, \tag{5}$$

where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are these i-vectors.

6 Resegmentation

After clustering the segments, the new i-vector of each cluster is computed (only from data of each cluster) and resegmentation is made to get better results. This process is repeated iteratively until the clusters consist of the same segments as in previous iteration (or the maximum number of iterations is reached). After the i-vector resegmentation, data (in the form of acoustic features) belonging to each cluster are used to train the Gaussian Mixture Model (GMM) of this cluster. The whole conversation is then resegmented frame by frame according to the likelihood of each GMM.

7 Experiments

In this paper, we try to answer the question of whether segmentation by SCD can improve the performance of an i-vector based speaker diarization system compared to the use of a naive segmentation with constant length segments. The experiment was carried out on telephone conversations from the English part of CallHome corpus [15], where only two speaker conversations were selected (so the clustering can be limited to two clusters), this is 109 conversation each with about 10 min duration in a single telephone channel sampled at 8 kHz.

The feature extraction was based on Linear Frequency Cepstral Coefficients (LFCCs), Hamming window of length 25 ms with 10 ms shift of the window. There are 25 triangular filter banks which are spread linearly across the frequency spectrum, and 20 LFCCs were extracted. Delta coefficients were added leading to a 40-dimensional feature vector. Instead of the voice activity detector, the reference annotation about missed speech was used.

For naive segmentation, a 2 second window with 1 second of overlap was used. For segmentation by SCD, the length of the segments was set to 4 seconds maximum and 0.1 second minimum.

The i-vector extraction system was trained using the following corpora: NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, Switchboard 1 Release 2 and Switchboard 2 Phase 3. The number of Gaussians in the UBM was set to 512. The latent dimension (dimension of i-vectors) in the FA total variability space matrix $\boldsymbol{T}$ in the i-vector extraction was set to 400. Finally, the dimension of

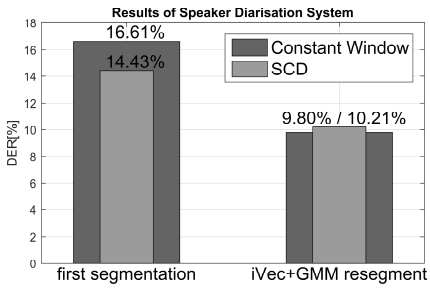

Fig. 3. Comparison of the system using SCD-based segmentation and constant window segmentation, before and after resegmentation. Results are given as given as DER[%].

the final i-vector was reduced by conversation dependent PCA with the ratio of eigenvalue mass $p = 0.5$.

In the resegmentation, the maximum iteration was set to 1000. The GMMs consisted of 1024 components and were trained by adaptation from a UBM.

7.1 Results

For evaluation, the Diarization Error Rate (DER) was used as described and used by NIST in the RT evaluations [16], with 250 ms tolerance around the reference boundaries. DER combines all types of error (missed speech, mislabeled non-speech, incorrect speaker cluster), but with correct information about the silence from the reference annotation, DER represents only the error in speaker cluster. The results are shown in Figure 3.

The experimental results of two approaches to the segmentation for speaker diarization task indicate, that the segmentation based on SCD brings better information for further clustering. However, the following iterations of resegmentation reduce the impact of inaccurate segmentation, making the final differences between systems with or without SCD negligible.

8 Conclusions

In this work, we compared two approaches to segmentation in an i-vector based speaker diarization system. The SCD segmentation method is based on finding the precise boundaries where the speaker is changing. On the other hand, the segmentation with constant length divides a conversation into short segments and relies on clustering and further resegmentation to refine the boundaries. The experimental results of these two approaches show that the SCD approach offers significantly better performance in the clustering stage, but the differences are diminished by the resegmentation. Therefore the naive segmentation is a sufficient approach for the speaker diarization system based on i-vectors.

Acknowledgments. The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506 and by the grant of the University of West Bohemia, project No. SGS-2016-039. Access to computing and storage facilities (CESNET LM2015042) is greatly appreciated.

References

1. Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., Meignier, S.: An open-source state-of-the-art toolbox for broadcast news diarization. Tech. rep., Idiap (2013)
2. Sell, G., Garcia-Romero, D.: Speaker Diarization with PLDA I-vector Scoring and Unsupervised Calibration. In: IEEE Spoken Language Technology Workshop, pp. 413–417 (2014)
3. Fredouille, C., Bozonnet, S., Evans, N.: The lia-eurecom rt 09 speaker diarization system. In: RT09, NIST Rich Transcription Workshop (2009)
4. Shum, S.H., Dehak, N., Dehak, R., Glass, J.R.: Unsupervised methods for speaker diarization: An integrated and iterative approach. Audio, Speech, and Language Processing, IEEE Transactions on 21(10), pp. 2015–2028 (2013)
5. Senoussaoui, M., Kenny, P., Stafylakis, T., Dumouchel, P.: A study of the cosine distance-based mean shift for telephone speech diarization. Audio, Speech, and Language Processing, IEEE/ACM Transactions on 22(1), pp. 217–227 (2014)
6. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on 19(4), pp. 788–798 (2011)
7. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of I-vector Length Normalization in Speaker Recognition Systems. In: Interspeech 2011. pp. pp. 249–252, Florence (2011)
8. Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D., Glass, J.: Exploiting intra-conversation variability for speaker diarization. In: INTERSPEECH. pp. 945–948, August (2011)
9. Gish, H., Siu, M.H., Rohlicek, R.: Segregation of speakers for speech recognition and speaker identification. In: ICASSP, pp. 873–876 (1991)
10. Zajíc, Z., Machlica, L., Müller, L.: Initialization of fMLLR with Sufficient Statistics from Similar Speakers. Lecture Notes in Computer Science 6836, pp. 187–194 (2011)
11. Kenny, P., Dumouchel, P.: Experiments in Speaker Verification using Factor Analysis Likelihood Ratios. In: Odyssey - Speaker and Language Recognition Workshop. pp. 219–226, Toledo (2004)
12. Kenny, P.: Joint factor analysis of speaker and session variability: Theory and algorithms. Tech. rep. (2006)
13. Machlica, L., Zajíc, Z.: Factor Analysis and Nuisance Attribute Projection Revisited. In: Interspeech 2012. pp. 1570–1573, Portland (2012)
14. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A Study of Interspeaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing 16(5), pp. 980–988 (2008)
15. Canavan, A., Graff, D., Zipperlen, G.: CALLHOME American English Speech LDC97S42. In: LDC Catalog. Philadelphia: Linguistic Data Consortium (1997),
16. Fiscus, J.G., Radde, N., Garofolo, J.S., Le, A., Ajot, J., Laprun, C.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. Machine Learning for Multimodal Interaction 4299, pp. 309–322 (2006)