SPEAKER SEGMENTATION USING I-VECTOR IN MEETINGS DOMAIN

Leonardo V. Neri[1] *Hector N. B. Pinheiro*[1]
Tsang Ing Ren[1] *George D. da C. Cavalcanti*[1] *André G. Adami*[2]

[1]Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)
Recife, Brazil

[2] Centro de Ciencias Exatas e da Tecnologia (CCET)
Universidade de Caxias do Sul (UCS)
Caxias do Sul, Brazil

ABSTRACT

In this paper, we propose a speaker segmentation method for meeting audio based on i-vector. The motivation is to utilize the Total Variability (TV) framework as a feature extractor and to exploit the potential of modeling the speaker and channel variabilities for speaker segmentation in meetings. A distance-based segmentation method is designed with the cosine distance. A sliding window with variable length searches for speaker turns, through the distance between the i-vectors extracted from two segments with the same size. The experiments are conducted on the AMI Meeting Corpus, covering several conversation scenarios. For the training data of the UBM and TV matrix, 5 conversations from AMI Meeting Corpus are sampled. Other 10 conversations from AMI Meeting Corpus to compose the test data. The experiments show an improvement in the MDR and FAR curves compared with the FixSlid approach with different distance metrics, and for most of the operating points when compared with the classical BIC based WinGrow. The proposed method has on average a better computational performance, improving in 61.5% compared with the XBIC based FixSlid, and improving in 86.7% compared with the BIC based WinGrow.

***Index Terms*—** speaker diarization, speaker segmentation, i-vector, total variability, meeting conversations

1. INTRODUCTION

Speaker Diarization is the process of splitting an audio stream into homogeneous segments that belongs to each participating speaker. Speaker segmentation is the task of determining the time instants when a transition from one speaker to another occurs, called speaker turns or speaker change points. The main challenge of this task is to form homogeneous speaker segments (only one speaker per segment) [1, 2].

Most of speaker segmentation algorithms relies on a sliding window and a distance metric to evaluate whether two segments belongs to the same speaker. This type of algorithms are categorized as distance-based segmentation [3]. This category assumes no *a priori* information about the speakers in the audio stream. The change points are localized by a sliding window, that can have a fix length (FixSlid approach [4,5]) or a variable length (WinGrow approach [6–8]). Commonly, MFCCs are extracted from the window [2, 9], and a distance between two segments of the features vectors is calculated. A speaker change point is detected when a distance is above a threshold. The Bayesian Information Criterion (BIC) is the most popular distance metric used for this task [2]. Other examples of popular distance metrics are the symmetric Kullback-Leibler (KL2) [4] and the Generalized Likelihood Ratio (GLR) [10].

In speaker recognition, factor analysis methods have proved to be very effective, particularly to telephone speech [11]. These methods are applied to speaker diarization, for telephone conversations [12–15] and broadcast news scenarios [16, 17]. Castaldo et al. [12] introduce the eigenvoices as speaker factor to pre-segment the speech, in a stream segmentation system for multi-speaker telephone conversations. In Desplanques et. al [16], eigenvoices are also proposed for speaker segmentation, in the context of broadcast news, utilizing single-pass and two-pass algorithms. In Shum et. al [14] and Senoussaoui et. al [15], the i-vector as speaker factors is proposed for speaker clustering in telephone conversations.

This work introduces the i-vector for the WinGrow segmentation approach. The i-vectors are extracted through a UBM and a Total Variability (TV) matrix trained for meetings domain. The segmentation relies on detecting change points with a sliding window with variable length. The i-vectors are extracted from two segments with the same length and the cosine distance is calculated between i-vectors. A threshold value is set to decide if there is a speaker change point between the segments boundaries. The UBM and the TV matrix are modeled using the AMI Meeting Corpus [18], sampling meeting audio streams. The experiments are conducted on the AMI Meeting Corpus, with different audio streams than the streams utilized for training of the UBM and TV matrix.

The remainder of this paper is organized as follows. Section 2 presents a brief introduction to the total variability modeling. Section 3 describes the baseline segmentation methods

Thanks to CNPq (446831/2014-0) and Facepe (APQ-0192-1.03/14) agencies for funding.

evaluated in this work. In Section 4, we detail the proposed speaker segmentation method utilizing i-vector. The experimental results are shown in Section 5. Finally, conclusions and discussions in relation to prior work are given in Section 6.

2. TOTAL VARIABILITY MODELING

In Dehak et al. [11], the total variability modeling is introduced, defining a single space containing both speaker and channel variabilities simultaneously. The total variability matrix defines this single space, containing the eigenvectors with the largest eigenvalues of the total variability covariance matrix. It ignores the distinction between the effects of speaker and channel in the Gaussian Mixtures Model (GMM) supervector space. A GMM supervector M is formed by concatenating the means of each mixture component. Given a speaker utterance, the speaker- and channel-dependent supervector M is written as,

$$M = m + Tw \tag{1}$$

where m is the speaker- and channel-independent supervector, taken from the Universal Background Model (UBM) supervector, T is a rectangular matrix of low rank, representing the total variability space, and w is a random vector with standard normal distribution. The vector w is called i-vector and its components are the total factors.

3. BASELINE SEGMENTATION METHODS

The proposed method is compared with the classical WinGrow approach [7], utilizing the BIC distance, and the FixSlid approaches, utilizing the distances: GLR [19], KL2 [4] and XBIC [8].

3.1. Window Growing Segmentation Approach

In the Window Growing approach (WinGrow) [7], the multiple change points detection in an audio stream is done with a defined distance metric, a criterion λ, or a threshold value and a sliding window with variable length, to sequentially detect one possible change point inside the window. The length of the window grows N_g vectors if no change is present. To detected one change point, a distance curve is calculated iteratively, splitting the window into two partitions. The length of these partitions is changed for each iteration, the length of the left partition starts with N_{min} vectors and increases over the right partition until the right partition length becomes N_{min}. The distance between partitions is calculated for each iteration and the defined criterion is applied to the curve, to detect the location of a possible speaker change. This approach has a high computation cost if the audio has long homogeneous speech segments and no upper limit is set to the window length. In [7], a upper limit length N_{max} together a slid length N_S parameter shows an efficiently computation performance and no deterioration in the speaker segmentation results. The algorithm presented in [7] is adopted in our method.

3.2. Fixed and Sliding Window Approach

The FixSlid approach detects multiple speaker change points with a defined distance metric, a criterion, or a threshold value and a sliding window with fixed length. The window length N_D is divided in two adjacent segments to calculate a distance value between them. The slide length N_{Slid} is defined to overlap the window and to calculate several distances across the audio stream. An analysis window is defined to find the peaks in the curve that are above the threshold, or are selected as candidates by a defined criterion. The length of the analysis window N_A is defined to evaluate a local set of points under the criterion or threshold and to find a speaker change point.

The criterion proposed in Delacourt et. al [19] is utilized in the baseline method. The absolute differences between a local peak value and the local lowest values on the left and right positions are computed. These differences are compared with the standard deviation of the curve multiplied by an adjustable parameter α. A relevant peak is detected if its value meets this criterion.

4. I-VECTOR BASED SPEAKER SEGMENTATION

In the distance curve computation, the low-level features vectors from the two partitions are modeled by using a single Gaussian density. As the partition length decreases, the voice patterns can become diverse due to the differences in their acoustic contents. The modeling inherently rely on averaging out the phonetic differences between partitions. Therefore, when the spoken material is from the same speaker, the partitions being compared can appear dissimilar, which can lead to an increased number of false alarms.

The i-vector as speaker factors relies on modeling the specific speaker features, considering the variability from its voice, and the channel effect present in the speech acquisition. Given an window analysis, the i-vectors are extracted from two partitions with same size. Then, a distance value is calculated between i-vectors. It is assumed that each partition has sufficient information in the i-vector representation. If the analysed spoken content comes from the same speaker, the extracted i-vectors tends to be similar, even with the variation of the phonetic acoustic information.

The cosine distance is utilized in the segmentation task, to measure the dissimilarity between the i-vectors:

$$cos\_dist(w_1, w_2) = 1 - \frac{\langle w_1, w_2 \rangle}{\|w_1\| \cdot \|w_2\|} \tag{2}$$

this distance has a normalized scale in the range 0 to 2, being easier to set a threshold θ. With no amplitude information,

this range appears independent from the audio application or audio data set.

This approach allows the detection of multiple speaker change points in a continuous speech segment. The analysis window has an initial length of N_{ini} low-level feature vectors. The speaker change candidate is in the middle of the window. The cosine distance is calculated between adjacent segments, corresponding to the left and right middle of the window. If the distance is above the threshold θ, the speaker change candidate is a change point. Then, the window is repositioned to this point and its length is reset to N_{ini}. If no change point is detected, the window length grows N_g features vector and the detection process starts again. If the length reaches the upper limit N_{max}, the window slides N_S features vector until a change point is detected. The algorithm stops when the window reaches the end of the signal. Fig. 1 shows how the proposed i-vector based WinGrow detects multiple speaker change points.

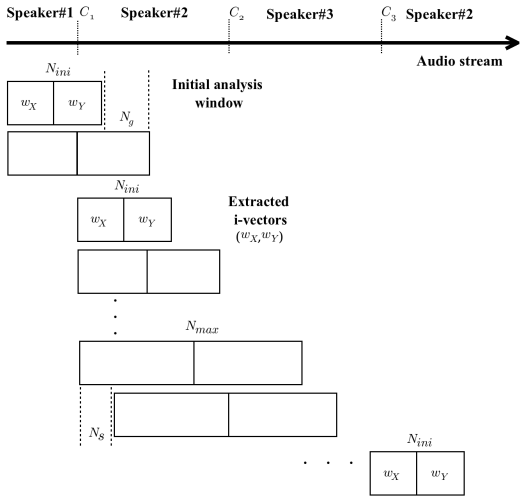

Fig. 1. Multiple speaker change points detection utilizing an i-vector based WinGrow approach. The cosine distance is calculated between i-vectors extracted from two window partitions with same length.

5. EXPERIMENTS

5.1. Data

The AMI Meeting Corpus [18] is an open access corpora with 100 hours of meeting recordings. For the experiments, we obtained 15 conversation having different speaker for all the samples. We divide the conversations into training and test data sets. In the training data, 5 conversations are sampled for both UBM and TV matrix modeling. The total of speakers for all the conversations is 10, and total duration for all conversations is approximately 4 hours. The evaluation data set is composed by 10 conversations. The conversations have between 4 and 5 speakers. The total duration of all conversations is approximately 6 hours and 24 minutes. There is a total of 6916 speaker change points in these conversations.

5.2. Speech Activity Detection

Before parametrization, a Speech Activity Detection (SAD) method based on energy is applied to extract all non speech segments longer than 0.25 second. In addition, speech segments shorter than 0.5 second are also discarded.

5.3. Evaluation measures

Given the inconsistencies in the labelling process and the uncertainty of when a speaker change precisely occur, a non-scoring collar around every reference speaker change is defined in evaluating the performance of the systems. In this work, a 500 ms collar is used to account for such issues.

Two types of errors are computed: False Alarm Rate (FAR) and Miss Detection Rate (MDR) [2, 9]. The MDR refers to the percentage of reference speaker changes that were not detected by the system. The FAR refers to the percentage of detected speaker changes that are not present in the audio.

The MDR and FAR from different operating points are used to compute the Receive Operation Characteristics (ROC) curve. The Area Under the Curve (AUC) is also used as a evaluation measure. In our experiments, we choose two error rates to compute the ROC curve, therefore, the best AUC is the most closely to 0.

5.4. Baseline Methods Configuration

For all baseline methods we extract 12 MFCCs from each cluster, with 32 ms frame length for every 10 ms.

For the BIC based WinGrow the following configuration is set: N_{ini}, N_g, N_{max} and N_S are set to 2, 1, 10 and 2 seconds, respectively. N_{min} is set to 20 frames of MFCC. We evaluate 10 values of λ, in a range from 0.5 to 5.5 in intervals of 0.5.

For the FixSlid approaches, the following configuration is set: N_D, N_{Slid} and N_A are set to 2, 0.2 and 2 seconds, respectively. We evaluate 10 values of α, in a range from 0.1 to 1.9 in intervals of 0.1.

5.5. Proposed Method Configuration

From each speaker cluster in the training data we extract 12 MFCCs, with a frame length of 32 ms for every 10 ms. The UBM is trained with all features vectors. The number of Gaussian components is set to 8. The matrix T is trained considering the variability intra-cluster. The rank of both T and the i-vector is set to 100.

Table 1. The Area Under the Curve (AUC) for the ROC curves of each segmentation method in the test data set.

Methods	AUC
WinGrow i-vector (COSINE)	**46.57**
WinGrow MFCC (BIC)	47.41
FixSlid MFCC (KL2)	52.07
FixSlid MFCC (GLR)	52.01
FixSlid MFCC (XBIC)	52.29

Table 2. Computational performances of the evaluated methods, measured in seconds. The mean and standard deviation (dev) are computed for the processing time for 10 runs of experiments.

Methods	mean (secs)	dev
WinGrow i-vector (COSINE)	**2.64**	**2.44**
WinGrow MFCC (BIC)	18.46	16.31
FixSlid MFCC (KL2)	16.18	20.29
FixSlid MFCC (GLR)	10.21	12.91
FixSlid MFCC (XBIC)	6.85	8.85

For the proposed approach, the WinGrow configuration follows the baseline using the BIC based WinGrow. We evaluate 10 values of θ, in a range from 0.1 to 1.9 in intervals of 0.1.

5.6. Speaker Segmentation Results

Fig. 2 shows the ROC curves for all evaluated methods. The points of the curve are obtained varying the adjustable parameters or threshold value. Table 1 shows the AUC for the ROC curves presented in Fig. 2. The ROC curves show that the proposed method outperforms the FixSlid approaches evaluated, in both MDR and FAR. The proposed method outperforms the BIC based WinGrow in some operating points. Table 1, the proposed method yields the lowest AUC.

Table 2 presents the average computational performance for 10 runs of experiments in the test data set for each evaluated method. The proposed method have the best average computation performance among the baseline methods, with an improvement of 61.5% compared to the FixSlid approach with the XBIC distance, and in 86.7% compared to the BIC based WinGrow.

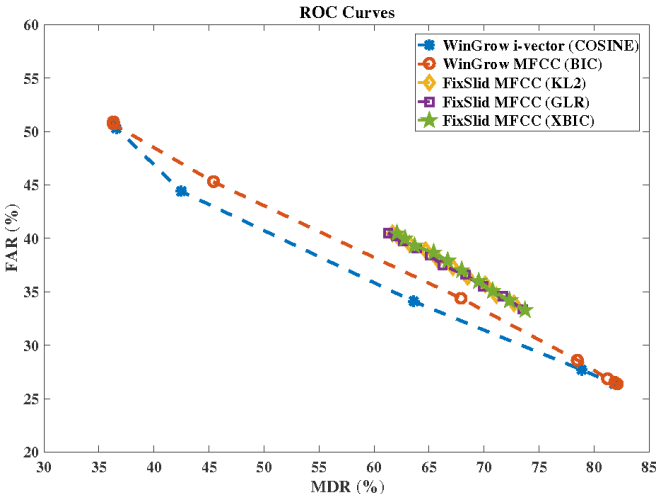

Fig. 2. The MDR x FAR curves of the evaluated methods. Each point is obtained varying the threshold value or the adjustable parameters.

6. CONCLUSION

This work presented an i-vector based WinGrow approach for speaker segmentation in meeting audio, using a cosine distance. In the experiments using the AMI Meeting Corpus, the i-vector and the cosine distance demonstrate a higher discrimination capacity among different speakers compared with the Gaussian distributions modeled from MFCC using the distances BIC, GLR, KL2 and XBIC. The results showed that the proposed method obtains better MDR and FAR for most of operating points compared with the classical BIC based WinGrow, and outperforms the classical FixSlid approaches with the distances GLR, KL2 and XBIC. The AUC of the proposed segmentation method is less than all baseline methods. The proposed method has a superior average computational performance.

As future work, we will investigate the results with more data for the UBM and T matrix modeling, and varying the rank of both T and i-vector. More experiments are needed in different application domains, such as telephone conversations and broadcast news.

7. ACKNOWLEDGEMENTS

This research are sponsored by the Conselho Nacional de Pesquisa (CNPq).

8. REFERENCES

[1] N Evans, S Bozonnet, Dong Wang, C Fredouille, and R Troncy, "A Comparative Study of Bottom-Up and Top-Down Approaches to Speaker Diarization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 20, no. 2, pp. 382–392, 2012.

[2] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 15, 2012.

[3] Shih Sian Cheng, Hsin Min Wang, and Hsin Chia 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.

[4] Matthew A Siegler, Uday Jain, Bhiksha Raj, and Richard M Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Chantilly, 1997, pp. 97–99.

[5] Lie Lu and HongJiang Zhang, "Real-Time Unsupervised Speaker Change Detection," in *International Conference on Pattern Recognition*, Quebec, 2002, pp. 358–361.

[6] S Chen and P Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, 1998, number 6, pp. 67–72.

[7] M Cettolo and M Vescovi, "Efficient audio segmentation algorithms based on the BIC," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong-Kong, 2003, vol. 6, pp. 537–540.

[8] X Anguera and J Hernando, "Xbic: Real-time cross probabilities measure for speaker segmentation," *ICSI - Berkeley Technical Report*, vol. 1, no. 1, pp. 05–08, 2005.

[9] Margarita Kotti, Vassiliki Moschou, and Constantine Kotropoulos, "Speaker segmentation and clustering," *Signal Processing*, vol. 88, no. 5, pp. 1091–1124, 2008.

[10] Perrine Delacourt and Christian Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing.," *Speech Communication*, vol. 32, no. 1-2, pp. 111–126, 2000.

[11] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4133–4136.

[13] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[14] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas A. Reynolds, and James R. Glass, "Exploiting intra-conversation variability for speaker diarization," in *INTERSPEECH*, 2011.

[15] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 1, pp. 217–227, 2014.

[16] Brecht Desplanques, Kris Demuynck, and Jean-Pierre Martens, "Factor analysis for speaker segmentation and improved speaker diarization," in *INTERSPEECH*, 2015, pp. 3081–3085.

[17] Jan Silovsky and Jan Prazak, "Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4193–4196.

[18] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner, "The ami meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, Berlin, Heidelberg, 2006, pp. 28–39, Springer-Verlag.

[19] Perrine Delacourt, David Kryze, and Christian Wellekens, "Detection of speaker changes in an audio document," in *European Conference Spoken Language Processing*, Budapeste, 1999, pp. 1195–1198, ISCA.