SPEAKER SEGMENTATION USING DEEP SPEAKER VECTORS FOR FAST SPEAKER CHANGE SCENARIOS

Renyu Wang[†] *, Mingliang Gu*[†]*, Lantian Li*[‡]*, Mingxing Xu*[‡]*, Thoms Fang Zheng*[⋆‡]

[†]School of Linguistic Science, Jiangsu Normal University, Xuzhou, 221116, China
[‡]Center for Speech and Language Technologies, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology
Research Institute of Information Technology; Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China
[⋆] Corresponding Author E-mail: fzheng@tsinghua.edu.cn

ABSTRACT

A novel speaker segmentation approach based on deep neural network is proposed and investigated. This approach uses deep speaker vectors (d-vectors) to represent speaker characteristics and to find speaker change points. The d-vector is a kind of frame-level speaker discriminative feature, whose discriminative training process corresponds to the goal of discriminating a speaker change point from a single speaker speech segment in a short time window. Following the traditional metric-based segmentation, each analysis window contains two sub-windows and is shifting along the audio stream to detect speaker change points, where the speaker characteristics are represented by the means of deep speaker vectors for all frames in each window. Experimental investigations conducted in fast speaker change scenarios show that the proposed method can detect speaker change points more quickly and more effectively than the commonly used segmentation methods.

Index Terms— Speaker segmentation, deep neural networks, speaker vector

1. INTRODUCTION

Speaker segmentation is to detect speaker change points in an audio stream and split it into homogeneous segments each with only one speaker ideally. This technique is always used as a pre-processing step for many speech and audio signal processing applications, such as speaker tracking, multi-speaker detection, and speech transcription [1].

In recent years, there are three major categories of speaker segmentation methods: metric-based, model-based, and the hybrid of them. In the metric-based methods, a distance measure needs to be defined firstly, then two adjacent windows are shifted along the audio stream. The distance between two analysis windows is calculated and the boundary between them is detected as a speaker change point if the distance is larger than a predefined threshold. The commonly used distance measures include Bayesian Information Criterion (BIC) [2], Generalized Likelihood Ratio (GLR) [3], Kullback-Leibler Divergence (KL) [4], Support Vector Machine (SVM) [5], and so on. In the model-based segmentation, different speaker models trained from a training set as prior knowledge are used to detect speaker change points when the models identification decision changes from one speaker to another. Typical models are Gaussian mixture models (GMM) [6, 7], eigenvoice-based models [8], hidden Markov models (HMM) [9], etc. The hybrid segmentation combines two of the previously mentioned methods, for example, ELISA [10] is a hybrid of HMM-based method and BIC method. In this paper, we will focus on the metric-based segmentation to solve the problem in the fast speaker change scenarios.

One of the problems in the metric-based segmentation is how to set up the size of analysis window. If the size is too long, there might be more than one speaker change points in the two adjacent windows, which will cause mistakes. Moreover, a large window size will lead to a long time delay in speaker change detection and reduce the accuracy of the final result. If the size is too short, speaker characteristics cannot be extracted accurately, so the distance calculation is often inaccurate and unstable. This problem has a large impact on the metric-based segmentation, especially in fast speaker change scenarios. So if we could know the shortest length of speech segment which can well distinguish two speakers, the metric-based segmentation will be more accurate.

Most of the common-used methods in the metric-based segmentation to discriminate different speakers and detect speaker change points are based on some distance measure assumptions defined by human prior knowledge. Most of these distance measures are based on probabilistic models that require a certain length of speech segment to make the statistical result stable. We hope that this length of speech segment can be shortened in fast speaker change scenarios. Aiming at discriminating two speakers in a shorter time window directly, the effective solution to extracting speaker discriminating characteristics needs to be investigated, and the significant difference between speaker change point and single speaker speech segment in short time needs to be found.

Recently, deep learning offers a new idea of 'feature learning'. With a deep neural network, task-oriented features can be learned layer by layer from input features. In 2011, deep neural network (DNN) was first used to extract speaker-specific characteristics in speaker segmentation task [11]. However, it was also based on mel-frequency cepstral coefficients (MFCCs), which is a kind of feature elaborately designed by researchers but not designed specifically for tasks of distinguishing speakers. To extract the speaker-discriminative characteristics more thoroughly, we will feed very

This work was supported by Program Granted for Scientific Innovation Research of College Graduate in Jiangsu Province (No. KYLX15_1463), and the National Natural Science Foundation of China under Grant No. 61271389 and No. 61371136, and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

few features to DNN, and expect that it can learn a kind of nonlinear mapping function from the acoustic space to the speaker space. Some studies have shown that it is effective in some speaker verification tasks [12, 13]. Speaker-discriminative features would be more significant in the new space, and it benefits to grasp speaker change point information in speaker segmentation task.

This paper apply d-vectors to the speaker segmentation task with the following contributions:

- We investigate the shortest length of speech segment which can well extract speaker-discriminative feature with our frame-level d-vector approach, and find that even 0.1 seconds (10 frames) length of voice has a certain degree of speaker-discriminative ability.
- We also apply d-vectors to the speaker segmentation task in fast speaker change scenarios, and get more than 26% decrease in false alarm rate (FAR) and more than 21% decrease in miss detection rate (MDR) compared with traditional segmentation methods.

The rest of the paper is organized as follows. Section 2 makes a description of deep speaker vector and its effective speaker-discriminative ability in frame-level. Section 3 illustrates the whole architecture of the proposed segmentation approach using d-vectors. The experiments are presented in Section 4. Conclusions are drawn in Section 5.

2. DEEP SPEAKER VECTORS

It is well-known that DNNs can learn task-oriented features from very raw features layer by layer. This property has been used in ASR (Automatic Speech Recognition) tasks to learn phone-discriminative features [14] and VPR (Voiceprint Recognition) tasks to learn speaker-discriminative features [13]. It has been shown that a well-trained DNN can turn input features into task-oriented features through the DNN structure layer by layer. This feature learning is so powerful that it has defeated the MFCC feature which was elaborately designed by researchers in some tasks.

Fig.1 presents the DNN model used for speaker-discriminative feature learning. Following the work in [13], the input layer involves a window of 40-dimensional filter bank energies (Fbanks). There are 4 hidden layers with each consisting of 200 units. The units of the output layer correspond to the speakers in the training data, and the number is 1,000 in our experiment.

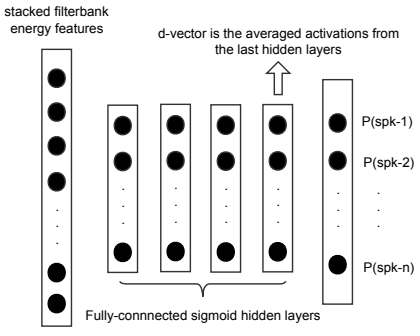


Fig. 1. The DNN model for learning speaker-discriminative features

Once the DNN has been trained successfully, the speaker-discriminative features could be read from any hidden layer. The closer to the output layer, the more speaker-discriminative those features will be. So we extracted features from the last hidden layer as the speaker representation, which is similar to the observation in [12]. There is a necessary underlying hypothesis that the trained DNN, having learned compact nonlinear representations of the speakers in the development set, this may also be able to represent unseen speakers.

Fig.2 shows the distribution of the d-vectors of two speakers on a 10-minutes conversation in the fisher corpus. We visualize d-vectors in the speaker space with PCA dimensionality reduction to 2 dimensions. It can be seen that there exists a distinct nonlinear boundary between most d-vectors of two speakers. That is to say, deep feature has strong discriminability.

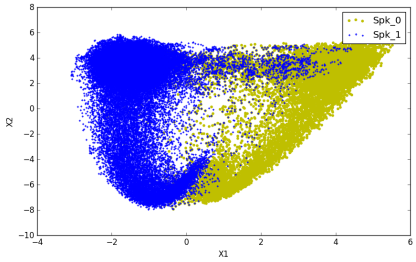


Fig. 2. Plot of the d-vectors of two speakers on a 10-minutes conversation (with PCA dimensionality reduction to 2 dimensions)

3. SEGMENTATION USING D-VECTORS

Since DNNs could learn a nonlinear mapping function from the acoustic space to the speaker space with prior knowledge, it is possible to characterize a speaker using only its d-vectors. According to the training process of DNN, this kind of discriminative feature corresponds to the goal of distinguishing different speakers in the segmentation task.

The segmentation algorithm used in our technique is summarized as follows. First, Fbank features need to be extracted when an audio stream comes after pre-processing. Second, Fbank features for each frame need to be fed to DNN to generate the d-vector sequence of an audio stream. We calculate the distance between two adjacent windows of a fixed size for the d-vector sequence. $D(t)$ the distance between two neighboring windows at frame t, is computed as the cosine distance of the means for deep speaker vectors in each analysis window. After two windows slide from left to right along the whole d-vector sequence of the audio stream with d frames shift, a curve of distance scores can be obtained as shown in Fig.4 (b).

There are two assumptions.

H_0: if the speakers of two neighboring windows are identical, the distance score of means for d-vectors between the two analysis windows is large.

H_1: if the speakers of two neighboring windows are different, the distance score of means for d-vectors between the two analysis windows is small.

Based on the above assumptions, if the distance score between two windows is smaller, it is more highly possible that there is a speaker change point across the boundary between these two windows. However, the problem is that how small the distance score

should be when there is a real speaker change point. So, in the last step, we detect peaks from the distance score curve with a threshold to find speaker change points. The whole segmentation architecture is shown in Fig.3.

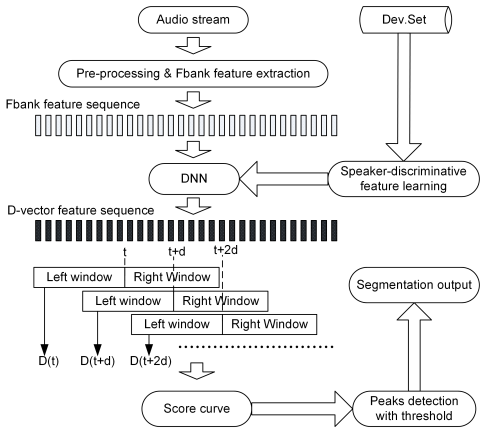


Fig. 3. The d-vector segmentation architecture

Fig.4 shows the differences of distance curves between a traditional segmentation method (BIC-based distance measure selected) and the d-vector segmentation approach. From the distance score curves, we can see that the d-vector segmentation is more precise than the BIC segmentation, and it makes a more detailed description of the trend of changes in score curve. Moreover, with the d-vector segmentation approach, the distance scores change more significantly when a real speaker change occurs, it is more beneficial for peaks detection and choosing a suitable threshold to detect real speaker change point.

4. EXPERIMENTS

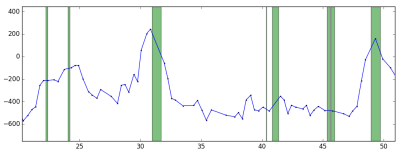
4.1. Database and experimental set-up

We randomly chose 1,000 speakers (with gender balanced; and each speaking segment length more than 10 minutes) from the fisher corpus for deep neural network training. Data used for segmentation experiments were also selected from the fisher corpus. It contains 100 fast switching telephone conversations, and each conversation has about 10 minutes and 100-200 change points (totally 16 hours and; 20,180 change points). The histogram of the speech segment durations based on the transcriptions is shown in Fig.5.

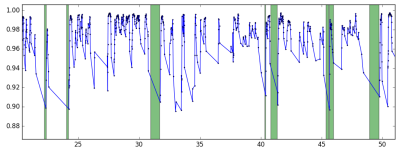
Feature extraction was performed on a 20ms frame width with 10ms shift. The pre-emphasis coefficient was 0.97 and the hamming windowing was applied to each frame. An energy-based Voice Activity Detection (VAD) was performed to remove the silence regions of each speech signal. In the DNN training step, the 40-dimensional Fbanks were extracted with speaker target corresponding to the unit of the output layer. In the segmentation step, the same size Fbanks were extracted and fed to the deep neural network to extract deep speaker vectors. Means of deep speaker vectors were calculated in each window to generate the speaker characteristics vectors.

4.2. Differentiated performance between the traditional distance measures and the d-vector approach

The first experiment investigated the performance of the d-vector differentiated performance, and compared it with the three traditional



(a) BIC: change points often detected around the local maximum values



(b) d-vector: change points often detected around the local minimum values

Fig. 4. Window distance score curves of d-vector and BIC segmentation (the green segment represents real speaker change segment)

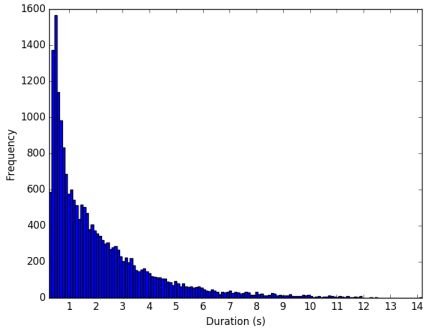


Fig. 5. Histogram of the speech segment durations based on the transcriptions

distance measures. In this experiment, we only selected 20 speakers' speech from the fisher corpus (8 males and 12 females). Five types of short-length segments (from 0.1 to 2 seconds per person) were extracted from each person's voice with each length of segments containing 20 cases. In all length of segments, we computed the distance scores of same speakers and different speakers on any 2 test cases, and the equal error rate (EER) was used as a relative standard, the same as in speaker verification task. Experimental set-up was the same as defined in section 4.1. The distinguishing abilities of different distance measures and d-vector approach are shown in Table 1. We cannot calculate the BIC distance in 0.1s speech length because we get singular matrix if feature dimensions smaller than number of frames in an analysis window.

The experimental result shows that traditional distance measures can only work in a certain speech length, while the d-vector has a great speaker-discriminative ability for short speech segment length even as short as 0.1 seconds (10 frames). As a result, d-vector approach is more suitable to grasp speaker-discriminative characteristics in a very short time window, and may be beneficial for segmentation task in fast speaker change scenarios.

Table 1. Performances in EER for different distance measures

Speech Length (s)	BIC	GLR	KL2	d-vector
0.10	-	49.39%	48.45%	19.61%
0.50	38.51%	39.52%	44.18%	10.44%
1.00	26.86%	27.47%	38.78%	8.16%
1.50	20.00%	21.02%	36.47%	6.94%
2.00	15.71%	15.97%	34.74%	5.00%

4.3. Analysis window size selection

In this experiment, we selected 10 conversations each of which contains around 150 speaker change points to examine the effect of window size in our d-vector segmentation approach. Because of the effective distinguishing ability of deep speaker vector, and the fast speaker change scenarios, we investigated multiple analysis window sizes (ranging from 0.01s to 1s) with 0.01s window shift.

FAR and MDR are used to evaluate the performance of the segmentation algorithm, which is defined as

$$FAR = FA/(ASC + FA)$$

$$MDR = MD/ASC$$

where FA denotes the number of false alarms, MD denotes the number of miss detections, and ASC denotes the actual number of speaker change points. We gave a 0.3s tolerance between the reference speaker change point and its nearest putative change point.

As shown in Fig.6, we took averages of two types of error rate for total 10 conversations. We can conclude that 0.05 to 0.1 seconds were the most effective window sizes for our fast speaker change segmentation task. A too short analysis window cannot extract enough speaker distinguishing features, and a too long analysis window may contain more than one speaker change points in two adjacent windows. The problem has been introduced in the first part of this paper and this is also a validation experiment from the side.

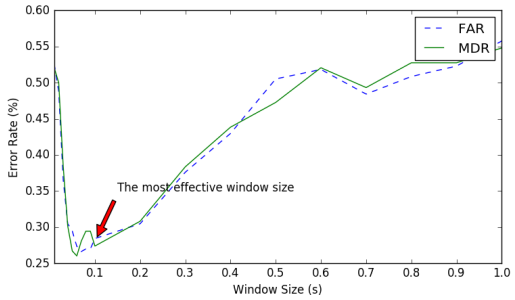


Fig. 6. The effect of window size with d-vector segmentation

4.4. The superior performance of d-vector segmentation

In this experiment, three traditional methods (BIC, GLR, KL2) were chosen as the baseline systems. We compared our proposed d-vector segmentation approach with them in all 16-hours conversations. The 0.1-seconds window size that achieves a considerable performance in section 4.3 was applied.

Due to the accuracy of speakers' distinguishing characteristics and the stability of distance calculation in a short time window, the proposed d-vector segmentation approach obtained a more substantial performance, which achieved a more than 26% decrease in FAR and a more than 21% decrease in MDR compared with the traditional ones. The DET curves for two types of error rates between traditional methods and the d-vector approach are shown in Fig.7.

Table 2. Performance comparison among the d-vector approach and traditional methods

Methods	FAR	MDR
BIC	52.92%	51.35%
GLR	52.07%	54.01%
KL2	51.58%	60.69%
d-vector	39.00%	40.15%

However, our proposed approach in Table 2 cannot reach the best performance as in Fig.6. This is mainly because the threshold is conversation dependent and it is hard to find a global optimum threshold suitable for all given conversations.

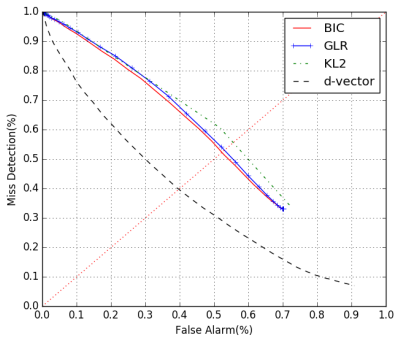


Fig. 7. DET curves comparison between three traditional approaches and d-vector based segmentation

5. CONCLUSION

In this paper, we propose a novel speaker segmentation approach based on deep speaker vectors. To deal with the problem of distinguishing speaker change points and single speaker speech segment in a very short time window, the d-vector approach has the following advantages compared with traditional distance measures: 1) Speaker representations in traditional distance measures are 'descriptive', and they are represented by constructing probability distributions. The d-vector is 'discriminative', which represents the speaker by removing speaker irrelevant variance, this matches the discriminative goal in speaker segmentation task directly. 2) Traditional distance measures require a certain length of voice to make the statistical result stable. However, the d-vector is a 'local' description which can be inferred from 'each' frame. This means that d-vector is more superior with short time window tasks. The experiment shows that with our d-vector approach, speaker characteristics can be extracted in only 0.1 seconds (10 frames) length of voice to distinguish different speakers, and in the case of fast speaker change scenarios, it got more than 26% decrease in FAR and more than 21% decrease in MDR compared with the traditional segmentation methods.

Future work will include improving the current cosine distance measure between d-vectors, as well as trying other transformations for the raw distance score. Another plan is to investigate the automatic threshold selection method to reduce the impact on conversation dependency.

6. REFERENCES

[1] Sue E Tranter and Douglas A Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] S. Chen, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. Darpa Broadcast News Transcription and Understanding Workshop*, 2000, pp. 127–132.

[3] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *ICASSP, International Conference*, 1991, pp. 873–876.

[4] Matthew Siegler, Uday Jain, Bhiksha Raj, Stern, and Richard, "Automatic segmentation, classification and clustering of broadcast news audio," *Proc Darpa Speech Recognition Workshop*, pp. 97–99, 1997.

[5] B. Fergani, M. Davy, and A. Houacine, "Speaker diarization using one-class support vector machines," *Speech Communication*, vol. 50, no. 5, pp. 355–365, 2008.

[6] I. Magrin-Chagnolleau, A. E. Rosenberg, and S. Parthasarathy, "Detection of target speakers in audio databases," in *icassp*, 1999, pp. 821–824.

[7] Amit S. Malegaonkar, Aladdin M. Ariyaeeinia, and Perasiriyan Sivakumaran, "Efficient speaker change detection using adapted gaussian mixture models," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 6, pp. 1859–1869, 2007.

[8] F. Castaldo, D. Colibro, E. Dalmasso, and P. Laface, "Stream-based speaker segmentation using speaker factors and eigenvoices," pp. 4133–4136, 2008.

[9] Sylvain Meignier, Jean Francois Bonastre, and Sylvain Meignier, "E-hmm approach for learning and adapting sound models for speaker indexing," *A Speaker Odyssey*, 2001.

[10] D. Moraru, S. Meignier, C. Fredouille, and L. Besacier, "The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation," 2004, pp. I–373–6 vol.1.

[11] Ke Chen and A. Salman, "Learning speaker-specific characteristics with a deep neural architecture," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1744–1756, 2011.

[12] E. Variani, Xin Lei, E. Mcdermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," 2014, pp. 4052–4056.

[13] Lantian Li, Dong Wang, Zhiyong Zhang, and Thomas Fang Zheng, "Deep speaker vectors for semi text-independent speaker verification," *Computer Science*, 2015.

[14] Jinyu Li, Dong Yu, Jui Ting Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm," in *Spoken Language Technology Workshop*, 2012, pp. 131 – 136.