IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

The Intelligent Voice System for the IberSPEECH-RTVE 2018 Speaker Diarization Challenge

Abbas Khosravani, Cornelius Glackin, Nazim Dugan, Gérard Chollet, Nigel Cannings

Intelligent Voice Limited, St Clare House, 30-33 Minories, EC3N 1BP, London, UK

Abstract

This paper describes the Intelligent Voice (IV) speaker diarization system for IberSPEECH-RTVE 2018 speaker diarization challenge. We developed a new speaker diarization built on the success of deep neural network based speaker embeddings in speaker verification systems. In contrary to acoustic features such as MFCCs, deep neural network embeddings are much better at discerning speaker identities especially for speech acquired without constraint on recording equipment and environment. We perform spectral clustering on our proposed CNN-LSTM-based speaker embeddings to find homogeneous segments and generate speaker log likelihood for each frame. A HMM is then used to refine the speaker posterior probabilities through limiting the probability of switching between speakers when changing frames. We present results obtained on the development set (dev2) as well as the evaluation set provided by the challenge.

Index Terms: speaker diarization, CNN, LSTM, IberSPEECH-RTV, speaker embedding

1. Introduction

Speaker diarization is the task of marking speaker change points in a speech audio and categorizing segments of speech bounded by these change points according to the speaker identity. Speaker diarization which is also referred to as who speaks when, is an important front-end processing for a wide variety of applications including information retrieval and automatic speech recognition (ASR).

The speaker diarization performance largely varies by the application scenario including the broadcast news audio, interview speech, meetings audio (with distance microphone), pathological speech, child speech, and conversational telephone speeches. Each domain presents a different quality of the recordings (bandwidth, microphones, noise), types of background sources, number of speakers, duration of speaker turns, as well as the style of the speech. Apart from its similarity to speaker identification in classifying homogeneous segments as spoken by the same or different speakers, each domain presents unique challenges to speaker diarization. Since there is no prior knowledge regarding the speakers involved in an audio speech, the number of speakers as well as the approximate speaker change points need also to be addressed. Moreover, speech segments may be of very short duration, making i-vectors as the most common speaker representation not an appropriate option [1].

The most widely used approach in speaker diarization involves: (1) speech segmentation, where speech activity detection (SAD) or speaker change point detection is used to find rough boundaries around each speaker's regions of speech; (2) segment clustering, where same speaker segments as well as the number of speakers will be determined; and finally (3) re-segmentation, where the boundaries are further refined to produce the diarization results. The first stage intends to divide the speech into short segments of a few seconds with a single dominant speaker. Thus, the quality of the speech activity detection plays an important role in the performance [2, 3]. Using word boundaries generated by an Automatic Speech Recognition (ASR) system to produce homogeneous short segments has been proposed in [4]. Neural-network based approached has also been investigated in [5]. The basic principle of detecting speaker turn is to use a short duration window of a few seconds and use a similarity measure to decide whether a speaker change has occurred or not. This window will then slides frame by frame over the entire audio to mark all the potential speaker change points. This window needs to be long enough to include speaker identifying information (usually 1-2 seconds long). The most common measures for change detection includes, Bayesian Information Criterion (BIC) [6], local Gaussian Divergence [7] and more recently deep neural networks [8, 9, 10].

In the clustering stage, features extracted from same speaker segments should ideally go to one cluster. These features could be the basic MFCCs, speaker factors [11], i-vectors [12, 13], or more recently d-vectors [14]. For many years, *i*-vector based systems have been the dominating approach in speaker verification [15]. However, thanks to the recent progress of deep neural networks, neural network based audio embeddings (*d*-vector) could significantly outperforms previously state-of-the-art techniques based on i-vectors, especially, on short segments [16, 17, 18, 19, 20]. The LSTM recurrent neural network [21] has been incorporated to produce speaker embeddings for the task of speaker verification [19] and later for speaker diarization [14]. In this work, we explore a similar text-independent *d*-vector based approach to speaker diarization. The proposed speaker embeddings using deep neural networks at the frame level is used for the task of speaker diarization.

The predominant approach for clustering is Agglomerative Hierarchical Clustering (AHC) [6] where a stopping criteria could be used to determine the number of clusters. Other popular clustering approaches include Gaussian Mixture Models (GMMs) [22], mean-shift clustering [23] and spectral clustering [13, 22, 14]. We found the effectiveness of spectral clustering algorithm which relies on analyzing the eigen-structure of an affinity matrix in our proposed diarization framework.

In the final stage, the initial rough boundaries are then refined. This is usually done using Viterbi algorithm at the acoustic feature level. For each cluster a Gaussian Mixture Model (GMM) is estimated to calculate posterior probabilities of speakers at the frame level and the process usually iterates until convergence. The Viterbi re-segmentation on raw MFCC features was found to be effective [11]. Re-segmentation in a factor analysis subspace has also been investigated in [24, 11]

10.21437/IberSPEECH.2018-48

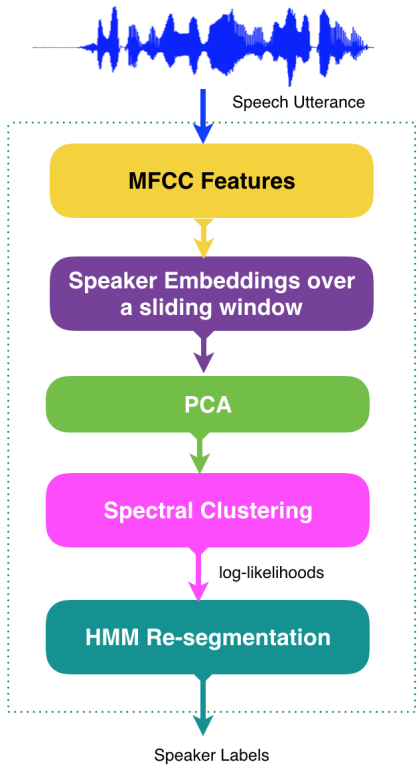

Figure 1: *System diagram for the proposed diarization system.*

where Variational Bayes (VB) system [25] proved to be the most effective. This approach implicitly performs a soft speaker clustering in a way which avoids making premature hard decisions. An extension to this was proposed at the 2013 Center for Language and Speech Processing (CLSP) Summer Workshop at Johns Hopkins University, where temporal continuity was modeled by an HMM. This extension will constrain speaker transitions and defines the speaker posterior probabilities [24]. In this paper, we intend to perform diarization with the help of deep neural network speaker embeddings. The log likelihoods for frame-level speaker embedding is estimated using a spectral clustering algorithm. In another way, HMM will serve as the re-segmentation for the spectral clustering algorithm.

The remainder of this paper is organized as follows: We outline the specific architecture in our proposed diarization system in Section 2 which includes CNN-LSTM based speaker embedding extraction, clustering and re-segmentation. Experimental results on the development set of the IberSPEECH-RTVE 2018 speaker diarization challenge are presented in Section 3. We will then conclude the paper in Section 5.

2. Diarization Framework

In this section, we review the key sub-tasks used to build current speaker diarization system based on DNN embeddings. The flowchart of our diarization system is shown in Fig. 1.

2.1. Acoustic Features

For speech parameterization we used 20-dimensional Mel-Frequency Cepstral Coefficients (MFCCs). These features are extracted at 8kHz sample frequency using Kaldi toolkit with 25 ms frame length and 10 ms overlap. For each utterance, the features are centered using a short-term (3s window) cepstral mean and variance normalization (ST-CMVN).

2.2. Speaker Embeddings

The i-vector based systems have been the dominating approach for both speaker verification and diarization applications. However, with the recent success of deep neural networks, a lot of efforts have been made into learning fixed-dimensional speaker embeddings (d-vectors) using an end-to-end network architecture that could be more effective relative to i-vectors on short segments [26, 20, 19, 27]. We employed a generalized end-to-end model using a convolutional neural networks (CNNs) and LSTM recurrent neural network. The network architecture is shown in Fig 2. It consists of two CNN layers and two dense layers at frame level followed by a bi-directional LSTM and two dense layers at utterance level. The LSTM layer maps a sequence of input feature vectors into an embedding vector. The output of the LSTM layer is then followed by two more dense layer and a length-normalization layer to produce a fixed dimensional representation for the input segment. Training is based on processing a large number of utterances in the form of a batch that contains N speakers and M utterances each. Each utterance could be of arbitrary duration. But to train the network in batch, they need to be of the same duration. We used variable length speech segments ranging from 10-20 seconds and construct batches with 30 speakers, each having 10 different segments. In the loss layer, a generalized end-to-end (GE2E) loss [19] builds a similarity matrix that is defined based on the cosine similarity between each pair of input utterances. During the training, we want the embedding of each utterance to be similar to the centroid of all that speaker's embeddings, while at the same time, far from other speakers' embeddings. A detailed description of GE2E training can be found at [19].

During evaluation, for every test utterance we apply a sliding window of 5 seconds (500 frames) with 80 percent overlap (we would have an speaker embedding every 100 frames). We compute the d-vector for each window. No speech activity detection has been used for this processing. Finally, a principle component analysis (PCA) is incorporated to reduce the dimensions of the resulting length-normalized embeddings (we used 8 dimension in our experiments) so as to be ready for clustering.

2.3. Clustering

We employed a spectral clustering which is able to handle unknown cluster shapes. It is based on analyzing the eigenstructure of an affinity matrix. A more detailed analysis of the algorithm is presented in [28]. We used an Euclidean distance measure to form a nearest neighbor affinity matrix on the frame-level embeddings. To estimate the number of clusters a simple heuristic based on the eigenvalues of the affinity matrix is used [13]. To mitigate the computational complexity of the spectral clustering, especially when the number of frames are too large, we can employ sampling at a specific rate.

To estimate the number of speakers, a simple heuristic based on the Calinski & Harabasz criterion [29] has been incorporated. We cluster d-vectors using K-means clustering algorithm using different value of cluster number and choose the one that maximize Calinski & Harabasz score.

2.4. Re-segmentation

The clustering algorithms are typically followed by a re-segmentation algorithm that refines the speaker transition

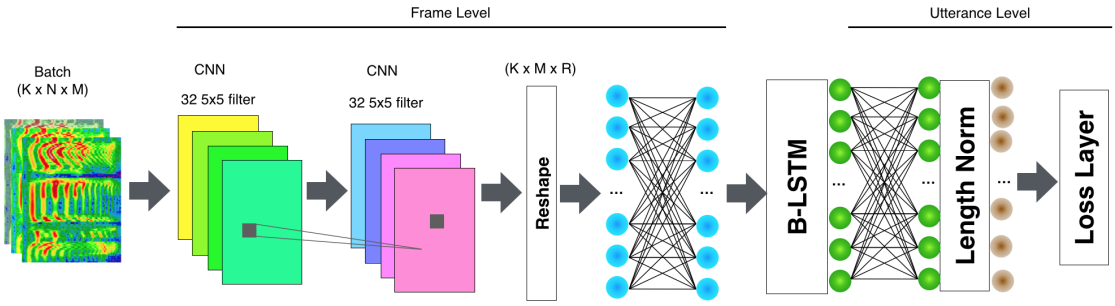

Figure 2: *Deep neural network architecture used to extract speaker embeddings.*

boundaries. This could be either in the feature space like MFCC or in the factor analysis subspace [24]. Speaker diarization in factor analysis space allows us to take advantages of speaker specific information. By contrast, lower-level acoustic features such as MFCCs are not quite as good for discerning speaker identities, but can only provide sufficient temporal resolution to witness local speaker changes. The proposed framework for diarization provides a stronger speaker representation at the frame level. As a result, when combined with a HMM to refine the speaker posterior probabilities through limiting the speaker transitions [24], the system is able to detect speaker change points. The speaker log likelihoods for the HMM are computed by the spectral clustering algorithm as described in section 2.3.

Table 1: *DER (%) on the development data (dev2) as well as evaluation data of the IberSPEECH-RTVE 2018 speaker diarization challenge.*

	DER(%)	Err(%)	FA(%)	Miss(%)
dev2	15.96	10.5	3.6	1.8
eval	30.96	25.2	4.8	0.9

3. Experiments

3.1. Training Data

Switchboard corpora (LDC2001S13, LDC2002S06, LDC2004S07, LDC98S75, LDC99S79) and NIST SRE 2004-2010 which consists of conversational telephone and microphone speech data at 8kHz sample frequency from around 5k speakers were used for training the system. Augmentation increases the amount and diversity of the existing training data. Our strategy employs additive noises and reverberation. Reverberation involves convolving room impulse responses (RIR) with audio. We use the simulated RIRs described in [30], and the reverberation itself is performed with the multi-condition training tools in the Kaldi ASpIRE recipe [31]. For additive noise, we use the MUSAN dataset, which consists of over 900 noises, 42 hours of music from various genres and 60 hours of speech from twelve languages [32]. Both MUSAN and the RIR datasets are freely available from http://www.openslr.org. We use a 3-fold augmentation that combines the original "clean" training list with two augmented copies [33]. To augment a recording, we choose between one of the following randomly:

- *babble*: Three to seven speakers are randomly picked from MUSAN speech, summed together, then added to the original signal (13-20dB SNR).
- *music*: A single music file is randomly selected from MUSAN, trimmed or repeated as necessary to match duration, and added to the original signal (5-15dB SNR).
- *noise*: MUSAN noises are added at one second intervals throughout the recording (0-15dB SNR).
- *reverb*: The training recording is artificially reverberated via convolution with simulated RIRs.

3.2. Performance Metrics

We measured performance with Diarization Error Rate (DER), the standard metric for diarization. It is measured as the total percentage of reference speaker time that is not correctly attributed to a speaker. More concretely, DER is defined as:

$$DER = \frac{FA + MISS + ERROR}{TOTAL} \quad (1)$$

where FA is the total system speaker time not attributed to a reference speaker, $MISS$ is the total reference speaker time not attributed to a system speaker, and $ERROR$ is the total reference speaker time attributed to the wrong speaker. Like the traditional conventions used in evaluating diarization performance [11], a forgiveness collar of 0.25 seconds will be applied before and after each reference boundary prior to scoring. The DER is reported based on the NIST RT Diarization evaluations [34].

4. Results

The IberSPEECH-RTVE 2018 Speaker Diarization is a new challenge in the ALBAYZIN evaluation series. This evaluation consists of segmenting broadcast audio documents according to different speakers and linking those segments which originate from the same speaker. We used two Intel Xeon CPU (E5-2670 @ 2.60GHz and 8 cores), 64G of DDR3 memory, 400G disk storage and an NVIDIA TITAN X GPU (12G of memory) to train the network. Keras API with tensorflow backend has been used for system development. Training takes almost a week to process around half a million segments of 10-20 seconds long. To process a single 20 minute recording the system execution times is around 7 seconds. We report the performance of our proposed diarization framework on the development set (dev2) using the provided speaker marks and also the result of the submitted system on the evaluation set in Table 1. Our system was trained on publicly accessible data which totally differ from both the development and evaluation data (open-set condition). The results indicate the effectiveness of the proposed approach on challenging domains.

5. Conclusion

The IberSPEECH-RTVE 2018 Speaker Diarization has proven to be a highly challenging contest especially in the detection of the number of speakers and dealing with background noise. We have presented our system and reported the results on the development set as well as the evaluation set of the challenge. We found deep neural network embeddings much better at discerning speaker identities especially for speech acquired without constraint on recording equipment and environment. Our strategy to employ additive noises and reverberation for data augmentation plays an important role in the success of our system on challenging domain. We will perform research on the evaluation set once the labels are released to gain insights on the real effects of the approaches presented in the paper.

6. Acknowledgement

The research leading to the results presented in this paper has been (partially) granted by the EU H2020 research and innovation program under grant number 769872.

7. References

[1] A. Khosravani and M. M. Homayounpour, "Nonparametrically trained plda for short duration i-vector speaker verification," *Computer Speech & Language*, vol. 52, pp. 105–122, 2018.

[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," 2018. [Online]. Available: https://zenodo.org/record/1199638

[3] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando, "Hybrid speech/non-speech detector applied to speaker diarization of meetings," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*. IEEE, 2006, pp. 1–6.

[4] D. Dimitriadis and P. Fousek, "Developing on-line speaker diarization system," in *Proc. Interspeech*, 2017, pp. 2739–2743.

[5] S. Thomas, G. Saon, M. Van Segbroeck, and S. S. Narayanan, "Improvements to the ibm speech activity detection system for the darpa rats program," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4500–4504.

[6] S. Chen, P. Gopalakrishnan *et al.*, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA broadcast news transcription and understanding workshop*, vol. 8. Virginia, USA, 1998, pp. 127–132.

[7] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, vol. 1997, 1997.

[8] V. Gupta, "Speaker change point detection using deep neural nets," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4420–4424.

[9] R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast tv using bidirectional long short-term memory networks," in *Proc. Interspeech 2017*, 2017, pp. 3827–3831.

[10] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4945–4949.

[11] 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.

[12] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 413–417.

[13] S. Shum, N. Dehak, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[14] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," *arXiv preprint arXiv:1710.10468*, 2017.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[16] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech 2017*, pp. 999–1003, 2017.

[17] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *Proc. of Interspeech*, 2017.

[18] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.

[19] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *arXiv preprint arXiv:1710.10467*, 2017.

[20] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matejka, and L. Burget, "End-to-end dnn based speaker recognition inspired by i-vector and plda," *arXiv preprint arXiv:1710.02369*, 2017.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[23] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 1, pp. 217–227, 2014.

[24] G. Sell and D. Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4794–4798.

[25] F. Valente and C. Wellekens, "Variational bayesian methods for audio indexing," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 307–319.

[26] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4930–4934.

[27] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[28] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[29] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[30] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE*, 2017, pp. 5220–5224.

[31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[32] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[33] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Submitted to ICASSP*, 2018.

[34] "The 2009 (rt-09) rich transcription meeting recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, accessed on June 2, 2016.