FEATURE MAPPING FOR SPEAKER DIARIZATION IN NOISY CONDITIONS

Weixin Zhu[1], Wu Guo[1], Guoping Hu[2]

[1]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China
[2]Key Laboratory of Intelligent Speech Technology, Ministry of Public Security, Hefei, China
wxzhu@mail.ustc.edu.cn, guowu@ustc.edu.cn, gphu@iflytek.com

ABSTRACT

Speaker diarization in noisy conditions is addressed in this paper. The regression-based DNN is first adopted to map the noisy acoustic features to the clean features, and then consensus clustering of the original and mapped features is used to fuse the diarization results. The experiments are conducted on the IFLY-DIAR-II database, which is a Chinese talk show database with various noise types, such as music, applause and laughter. Compared to the baseline system using PLP features, a 21.26% relative DER improvement can be achieved using the proposed algorithm.

Index Terms— Speaker diarization, deep neural networks, feature mapping, consensus clustering

1. INTRODUCTION

Speaker diarization relates to the problem of determining "who spoke when". It can be used as an important front-end for audio records that contain more than one active speaker before performing other speech information processing. A typical speaker diarization system consists of four major modules: feature extraction, speaker segmentation, speaker clustering and Viterbi re-segmentation. Short-term acoustic features, such as Perceptual Linear Predictive (PLP) or Mel Frequency Cepstrum Coefficients (MFCC), are the most widely used in speaker diarization. Generally, the traditional approach of speaker segmentation using the Bayesian information criterion (BIC) [1] is adopted, followed by Agglomerative hierarchical clustering (AHC) [2]. A number of approaches exist involving the distance metric of each pair of clusters. When clusters are represented by Gaussian mixture model (GMM), cross likelihood ratio (CLR) [3] and T-Test distance [4] are typically used. More recently, the *i*-vector has become the mainstream method used in the state-of-the-art diarization systems [5, 6], and PLDA [7] or a cosine distance metric are utilized for clustering. After initial clustering, Viterbi re-segmentation is employed to refine initial segmentation boundaries. Furthermore, system fusion approaches are adopted to utilize the complementary information of different features or sub-systems [8，9，10]. These algorithms have greatly improved the diarization performance, with the diarization error rate lower than 1% possible in clean telephone conversational speech [5, 11].

However, the performance of diarization systems will greatly deteriorate under noisy conditions [12，13]. Since the acoustic features represent the speaker's vocal characteristic, we exploited using deep neural networks (DNNs) to obtain the noise robust features to improve the reliability of speaker diarization systems in this paper.

Deep learning techniques have been introduced in speech enhancement. In [14], Xu *et al.* used a regression-based DNN to enhance the noisy speech, and Du *et al.* [15] used the enhanced speech features directly to train acoustic models (GMM-HMM and DNN-HMM) for automatic speech recognition (ASR). In [16], Gao *et al.* mapped the input noisy features to the desired clean acoustic features using a regression DNN and proposed to jointly train a single DNN for both feature mapping and acoustic modeling, which achieved a significant improvement by fusing the enhanced features from different domains. In [17], Wang *et al.* used a similar framework in voice activity detection (VAD). Encouraged by their work, we employ feature mapping to obtain cleaner features in noisy conditions and use the enhanced features for speaker diarization.

The cluster purity of initial speaker models plays an important role in Viterbi re-segmentation. We focus on improving the cluster purity based on consensus clustering [18, 19] in this paper. The regression-based DNN is first used to transform the noise corrupted features into enhanced features, and then the initial clustering results based on the original features and enhanced features are generated. Speech segments belonging to the same cluster in the above-described two clustering results are chosen to train the presumptive speaker initial models, which are used in the following Viterbi re-segmentation.

The remainder of this paper is organized as follows. Section 2 describes the feature mapping based on regression DNN. Section 3 presents the system fusion based on consensus clustering. Analysis of the experiments and results are presented in section 4. Finally, the results are summarized in section 5.

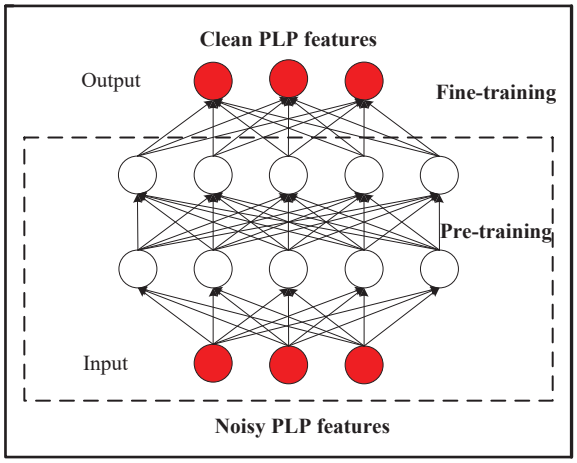

Figure 1. DNN for Feature mapping

2. FEATURE MAPPING

Regression-based DNN is used for feature mapping [16, 17]; the DNN architecture is shown in Figure 1. This regression DNN performs as a highly non-linear regression function to obtain clean speech features from noisy speech features. To improve the continuity of estimated clean features, the acoustic context information along both time axis (with multiple neighboring frames) and frequency axis (with full frequency bins) is utilized by DNN. As for training a more generalized DNN model, a large amount of pairs of noisy and clean speech data is required. Since it is difficult and expensive to collect so much noisy training data from real scenarios, the noisy training data is produced by corrupting the clean speech data with different types of noise at various signal-to-noise-ratio (SNR) levels. The training process is similar to classification DNN, consisting of unsupervised pre-training and supervised fine-tuning. A minor difference exists between regression DNN and classification DNN in fine-tuning. As for regression DNN, we aim at minimizing mean squared error (MMSE) between the DNN output and the reference clean speech features, as follows:

$$E = \frac{1}{N}\sum_{n=1}^{N} \| \hat{x}_{n-\tau}^{n+\tau}(y_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b}) - x_{n-\tau}^{n+\tau} \|_2^2 + \lambda \| \boldsymbol{W} \|_2^2 \quad (1)$$

where $\hat{x}_{n-\tau}^{n+\tau}$ and $x_{n-\tau}^{n+\tau}$ are the $D(2\tau+1)$ dimensional vectors of estimated and reference clean features for frame n, respectively, N represents the mini-batch size, and $y_{n-\tau}^{n+\tau}$ is a $D(2\tau+1)$ dimensional vector of input noisy feature where the window size of context is $2\tau+1$. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote the weight and bias parameters to be learned. λ is the weight decay coefficient to avoid over-fitting. A mini-batch stochastic gradient descent algorithm is used to learn the model parameters. It is noted that the DNN output contains the same number of frames as the input, and all the input and output features are normalized with a global mean and variance of the noisy features of the training set. For convenience, the DNN output features are labelled as enhanced features in the rest of this paper.

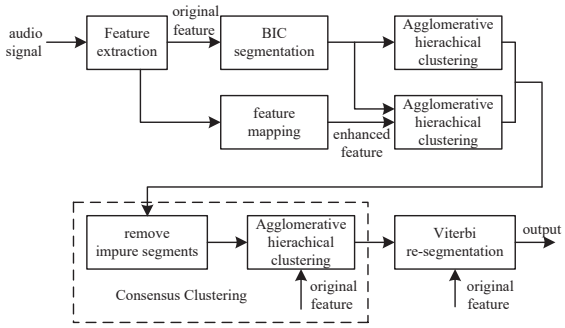

Figure 2: Framework of diarization system

3. SYSTEM FUSION BASED ON CONSENSUS CLUSTERING

The main purpose of consensus clustering is to access the stability of the discovered clusters and discard the unstable clusters [18]. When it is used for cluster purification, it can remove the incorrect assignments of speech segments in speaker clusters before cluster modeling.

The framework of our speaker diarization system is shown in Figure 2. Two sub-systems are adopted in consensus clustering with the same diarization algorithm, where original and enhanced features are used separately. These two sub-systems employ the same initial segments from BIC segmentation with original features.

In our experiments, the consensus clustering has two steps. The first step is to remove the impure speaker segments to obtain more purified clusters, and then the pure speaker segments are used to train initial speaker probability models. Generally speaking, the number of purified clusters is more than that of real speaker after the first step, so agglomerative hierarchical clustering is performed at the second step using the original features.

At the first step, consensus matrix [18] $\boldsymbol{M}$ is used to remove the impure speaker segments. Consensus matrix $\boldsymbol{M}$ with the size $N \times N$ denotes whether two speech segments belong to the same cluster or not, where N is the total number of speech segments in a recording. Consensus matrix $\boldsymbol{M}$ is defined as follows:

$$\boldsymbol{M}(i,j) = \frac{\sum_h \boldsymbol{M}_h(i,j)}{H} \quad (2)$$

where H denotes the total number of diarization systems, and $\boldsymbol{M}_h(i,j)$ is defined as follows:

$$\boldsymbol{M}_h(i,j) = \begin{cases} 1 & \textit{if segments } i \textit{ and } j \textit{ belong to the same cluster} \\ 0 & \textit{otherwise} \end{cases} \quad (3)$$

where subscript h denotes the index of the sub-system used for consensus clustering. $\boldsymbol{M}(i,j)$ represents the probability that segments i and j are assigned to the same cluster. If $\boldsymbol{M}(i,j)$ is equal to 1, the confidence of segments i and j

in the same cluster is high and they are chosen in the following agglomerative hierarchical clustering. Furthermore, we discard the consensus clusters whose duration is shorter than 5 seconds in the experiments.

In the second step, clusters obtained from the first step are served as initial clusters, and the original features are used in agglomerative hierarchical clustering. By using consensus clustering, many impure segments are removed and we can obtain more precise speaker model which is used for segment re-assignment at the Viterbi decoding step.

4. EXPERIMENTS AND RESULT ANALYSIS

4.1. Experiment data

The experiments carried out on the IFLY-DIAR-II database, which is drawn from Chinese talk shows, and the sample rate is 16 kHz. The duration of the recordings in the IFLY-DIAR-II database vary from 20 minutes to one hour. The number of speakers in each recording ranges from 2 to 9, and there are generally one host and several guests. The speaking style is spontaneous and causal, and short conversation turns and overlapped speech are often encountered. Furthermore, the speech is corrupted by music, laughter, applause, or other noises.

The training set contains 171 recordings (86 hours), the development set consists of 90 conversations (47 hours), and the test set contains 367 audio files (193 hours).

4.2. Performance evaluation metric

We use the diarization error rate (DER) [20] to measure the performance of the speaker diarization systems. DER can be expressed as:

$$DER = \frac{T_{FA} + T_{miss} + T_{SpkErr}}{T_{speech}} \quad (4)$$

where $T_{FA}, T_{miss}, T_{SpkErr}$ and T_{speech} are the duration of silence wrongly classified as speech, the duration of speech misclassified as silence, the duration of speech wrongly classified to other speakers, and the total duration of speech. We use oracle speech activity detection (SAD) for the following experiments. The oracle SADs are derived from the reference human transcriptions, so only speaker confusion error is attributed to DER.

In addition to DER, the average cluster purity rate (ACPR) [19] is adopted to evaluate the cluster purity before performing Viterbi re-segmentation as follows.

$$ACPR = \frac{T_{PurSeg}}{T_{speech}} \quad (5)$$

where T_{PurSeg} is the total duration of pure speaker segments in all clusters.

4.3. Regression DNN training

The IFLYTEK-HIFICM database is used for regression DNN training, and is composed of 16880 clean Mandarin utterances. Additive white Gaussian noise (AWGN), music, applause and laughter are used as noise signals, which are added to these clean utterances at five different SNR levels from 0dB to 20dB with an increment of 5dB. As for the DNN training, the input layer is a context window of 11 frames of 39-dimensional PLP feature with delta and acceleration coefficients. The DNN architecture was 429-2048-2048-429, the mini-batch size N is set to 128, and the regularization weighting coefficient λ is 1e-5.

4.4. Speaker diarization system description

The state-of-the-art speaker diarization algorithms are adopted in our system building. Since the oracle SADs are used to mark the speech segments, VAD algorithm is not applied. Oracle SADs and the Bayesian information criterion (BIC) based change point detector [1] are first used to partition the recording into short segments, and then an agglomerative hierarchical clustering (AHC) algorithm is performed. At this step, the T-test distance [4] is adopted for clustering; the clustering process stops when a stopping threshold is met. After the clustering, several iterations of Viterbi segmentation and models retraining are performed.

In our experiment, the duration of each segment is relatively long in the audio files of the talk show, and the well-known i-vector/PLDA method is not as good as the method based on T-test distance; thus, we adopted the T-test distance at AHC step in this paper.

4.5. Experimental results and analysis

4.5.1 Experiment using different acoustic features

First, we compare the diarization performance with three different acoustic features: PLP, MFCC and enhanced PLP (denoted as EnPLP). Table 1 presents the results on IFLY-DIAR-II database. The enhanced PLP can achieve 7.32% relative DER reduction and 1.74% absolute ACPR improvement compared to PLP, and can achieve 7.79% relative DER reduction and 1.82% absolute ACPR improvement compared to MFCC. These results demonstrate the effectiveness of the enhanced speech features obtained by regression DNN. We can get clearer speaker's personal characteristics from the enhanced speech features, which can improve the diarization performance.

Table 1. Experimental results for different features without consensus clustering in the IFLY-DIAR-II test set.

Feature	DER(%)	ACPR(%)
PLP	9.83	84.44
MFCC	9.89	84.36
EnPLP	9.11	86.18

4.5.2 Experiment with consensus clustering

We further analyze the diarization performance based on consensus clustering. Although consensus clustering can be carried out on more than two sub-systems, we only compare the pairwise combinations of the above-mentioned three acoustic features in this paper. The results of consensus clustering based on different combination are listed in Table 2. MFCC-EnPLP denotes the fusion of MFCC and EnPLP based on consensus clustering; similarly, PLP-EnPLP represents the fusion of PLP and EnPLP, and PLP-MFCC is the fusion of PLP and MFCC. We can see that the difference between MFCC-EnPLP and PLP-EnPLP is small for both DER and ACPR. Compared to PLP-MFCC, PLP-EnPLP can achieve 9.15% relative DER reduction and 2.23% absolute ACPR improvement. The results in Table 1 and Table 2 show that PLP-EnPLP can achieve 21.26% relative DER reduction and 6.97% absolute ACPR improvement compared to the original PLP feature.

Table 2. Results of consensus clustering based on different acoustic feature combinations.

SYSTEM	DER(%)	ACPR(%)
MFCC-EnPLP	7.82	91.29
PLP-EnPLP	7.74	91.41
PLP-MFCC	8.52	89.18

To sum up, the use of enhanced features can increase the robustness of speaker diarization systems in noisy conditions. Moreover, the enhanced features can add complementary information to the original short-term features.

5. CONCLUSIONS

Speech is always corrupted by all types of noise in practical applications; such noise will degrade the performance of speaker diarization system. In this paper, we utilize a regression DNN to map noisy features to clean features and find the enhanced features can improve the purity of the speaker diarization clusters. Furthermore, the results of consensus clustering in the IFLY-DIAR-II test set reported a 21.26% relative DER improvement compared to the PLP baseline system, and these experiments confirm that the enhanced PLP features can add complementary information to the original PLP features in noisy conditions.

6. ACKNOWLEDGEMENT

This work was partially funded by The National Key Research and Development Program of China (Grant No.2016YFB100130300) and Natural Science Foundation of Anhui Province (Grant No.1408085MKL78).

7. REFERENCES

[1] Chen S, Gopalakrishnan P. Speaker, environment and channel change detection and clustering via the bayesian information criterion[C]//Proc. DARPA Broadcast News Transcription and Understanding Workshop. 1998, 8: 127-132.

[2] Han K J, Narayanan S S. A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system[C]//INTERSPEECH. 2007: 1853-1856.

[3] Zhu X, Barras C, Meignier S, et al. Combining speaker identification and BIC for speaker diarization[C]//INTERSPEECH. 2005, 5: 2441-2444.

[4] Nguyen T H, Chng E, Li H. T-test distance and clustering criterion for speaker diarization[C]//INTERSPEECH. 2008: 36-39.

[5] Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. A., & Glass, J. R. (2011, August). Exploiting Intra-Conversation Variability for Speaker Diarization. In INTERSPEECH (Vol. 11, pp. 945-948).

[6] Zhu W, Pelecanos J. Online speaker diarization using adapted i-vector transforms[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5045-5049.

[7] Sell G, Garcia-Romero D. Speaker diarization with PLDA i-vector scoring and unsupervised calibration[C]//Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014: 413-417.

[8] Zewoudie A W, Luque J, Pericás H, et al. Using voice-quality measurements with prosodic and spectral features for speaker diarization[C]//INTERSPEECH 2015: 16th Annual Conference of the International Speech Communication Association: Dresden, Germany: September 6-10, 2015. International Speech Communication Association (ISCA), 2015: 3100-3104.

[9] Sarria-Paja M, Senoussaoui M, O'Shaughnessy D, et al. Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 5480-5484.

[10] Planet S, Iriondo I. Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition[C]//Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on. IEEE, 2012: 1-6.

[11] Xu Y, McLoughlin I, Song Y, et al. Improved i-vector representation for speaker diarization[J]. Circuits, Systems, and Signal Processing, 2015: 1-12.

[12] Zelenák M, Schulz H, Hernando J. Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2012, 2012(1): 1-9.

[13] Gupta V, Boulianne G, Kenny P, et al. Speaker diarization of French broadcast news[C]//2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008: 4365-4368.

[14] Xu Y, Du J, Dai L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68.

[15] Du J, Wang Q, Gao T, et al. Robust speech recognition with speech enhanced deep neural networks[C]//INTERSPEECH. 2014: 616-620.

[16] Gao T, Du J, Dai L R, et al. Joint training of front-end and back-end deep neural networks for robust speech recognition[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4375-4379.

[17] Wang Q, Du J, Bao X, et al. A universal VAD based on jointly trained deep neural networks[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[18] Nwe T L, Sun H, Li H, et al. Speaker diarization in meeting audio[C]//2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009: 4073-4076.

[19] Nwe T L, Sun H, Ma B, et al. Speaker clustering and cluster purification methods for RT07 and RT09 evaluation meeting data[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(2): 461-473.02.

[20] The 2009 (RT09) Rich Transcription Meeting Recognition EvaluationPlan,[Online].Available:http://itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf