An Analysis of Speaker Diarization Fusion Methods For The First DIHARD Challenge

Bing Yin*, Jun Du*, Lei Sun*, Xueyang Zhang†, Shan He†, Zhenhua Ling*, Guoping Hu† and Wu Guo*
* University of Science and Technology of China
E-mail: bingyin@iflytek.com, jundu@mail.ustc.cn, sunlei17@mail.ustc.edu.cn,zhling@mail.ustc.cn
† iFlytek Research, China
E-mail: shanhe2@iflytek.com, xyzhang12@iflytek.com

***Abstract*—In this paper, we introduce the attempts of our fusion methods during the first DIHARD challenge. To our knowledge, this is the first launch in speaker diarization domain which aims to evaluate the performance of the state-of-the-art system in realistic adverse acoustic environments. Besides speech preprocessing modules including speech denoising and speech activity detection, our attention has been focused on back-end clustering algorithms, especially in system fusion. Consensus clustering is adopt to combine both original speech and denoised speech, for purifying unreliable clusters. Moreover, a score-level fusion is conducted between GMM-UBM-based i-vector and CNN-based i-vector. Finally, our system achieves diarization error rates (DERs) of 36.05% on the evaluation sets, which is the second place in the DIHARD challenge.**

I. Introduction

Speaker diarization task is to segment speaker homogeneous parts given an arbitrary audio recordings. A practical speaker diarization system should work in conditions where no prior information can be used, such as the number of speakers, the dialog styles and environmental scenes. Previously, many studies mainly focused on broadcast news and telephone data, which could not be very representative to realistic application conditions [1]. For this reason, the first DIHARD challenge [2] was proposed where the datasets are well-designed and drawn from a diverse set of challenging domains, in order to explore the benchmark of current state-of-the-art systems. The complexity of data corpora leads people to build systems which should have great capabilities of dealing with noisy speech, reverberations, overlapped speech. Thus, a robust preprocessing system is vital to the final performance. In [3], we have shown that deep learning based denoising method has stronger potentials in coping with realistic noisy environments than traditional enhancement approaches.

Most of current diarization systems use the bottom-up method, which is also known as agglomerative hierarchical clustering (AHC). The recording is first clustered in to smaller segments where each segment ideally comes from only one speaker. Then the most similar segments are merged iteratively until a certain stopping criterion is satisfied. Some metrics like Bayesian Information Criterion (BIC) [4], T-test distance [5], can be used as the distance measure. Recently, i-vector [6], [7] and probabilistic linear discriminant analysis (PLDA) [8], [9] have shown great effectiveness in the field of speaker recognition and speaker diarization. To further enhance the performance, different fusion methods are also explored, including feature-level fusion [10], system output-level fusion [11], and multi-model fusion like audio-visual fusion [12].

The DIHARD challenge proposes two tracks, namely Track1 and Track2. Track1 uses gold speech segmentation while Track2 does diarization from scratch. In this study, we only consider the performance on Track2, because it's the most convictive proof for a real application. First, our proposed diarization system is introduced. Based on that, we describe our novel practices on fusion strategies. In Section III-A, consensus clustering is used to capture complementarity between two systems based on denoised speech and original speech. In Section III-B, a convolutional neural network (CNN) based i-vector is proposed and shows great benefits to traditional i-vector PLDA framework. After all, we evaluate the system on the DIHARD datasets. As conclusion, we discuss some unsolved problems and the future work.

II. The Proposed Diarization System

As in the preprocessing stage of our diarization system, we first adopt a deep-learning based speech denoising model used in [3], in order to mitigate interferences from environmental noises under different recording conditions. Here we expand the diversified simulated training data to more than 400 hours, to make it more robust and stable. Then a deep neural network (DNN) based speech activity detection model is trained on realistic collected data. It can also be observed that speech denoising module can boost the performance of speech activity detection (SAD), especially in reducing the false alarm error.

Given the valid speech segments, we utilize a two-pass short-long term diarization system. When the segments are relatively short, we use the simple Bayesian information criterion (BIC) as the hypothesis testing metric, to detect speaker turns within an individual segment and then merge different segments which belong to one speaker. After the cluster segments are relatively long (5 seconds in our experiments), i-vector can be used as a more powerful speaker representations [6]. We train the baseline i-vector extractor on the VoxCeleb corpus. The universal background model (UBM) contains 1024 Gaussians and the total variability matrix reduces the dimension to a range between 100 and 400. In this study, with the increase of i-vector dimension, we found the performance

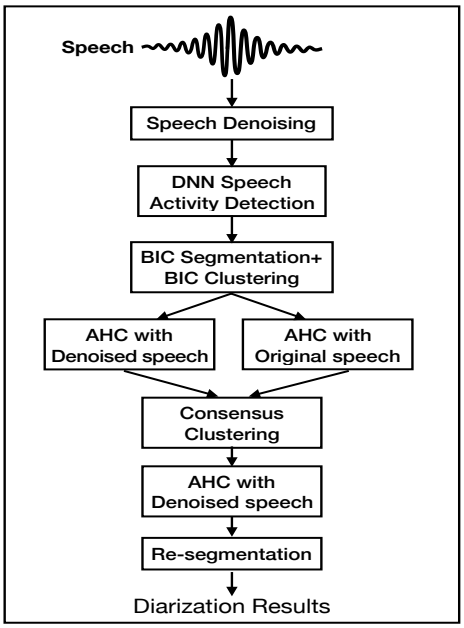

Fig. 1. The framework of system fusion based on consensus clustering.

also improves. So the final i-vector is with 400 dimension and also length-normalized. For agglomerative hierarchical clustering, a PLDA scoring model [8], [9] is trained to measure the similarity between two i-vectors. Finally, a realignment over frames is performed via Viterbi decoding on the specific GMM trained on each speaker.

III. Fusion Strategy

System fusion is a common strategy to improve performance in speaker diarization [11], [10], aiming to catch the complementarity between different systems. In this section, we will present our attempts on seeking for an effective fusion method in adverse acoustic environments.

A. System Fusion Based on Consensus Clustering

In this section, we adopt the consensus clustering method, which aims to remove unreliable clusters via taking consideration of different clustering results from several subsystems [13]. As illustrated in Fig. 1, benefiting from the denoising model, we can run two parallel diarization systems at the beginning, which separately use acoustic features extracted from original speech and denoised speech. As mentioned above, due to the advantage of denoised speech for SAD performance, we use a unified SAD information derived from denoised speech for both sub-systems. Before consensus clustering, we force the two sub-systems to have the same number of clusters.

Firstly, we construct a consensus matrix C with the size $N \times N$, which denotes whether two speech segments belong to the same cluster or not, where N indicates the total number of speech segments. The matrix C is defined as follows [13]:

$$C(i,j) = \frac{\sum_s C_s(i,j)}{S} \tag{1}$$

where S is the total number of diarization systems, and $C_s(i,j)$ denotes :

$$C_s(i,j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ belong to the same cluster} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where s indicates the index of different sub-systems. Thus, the probability matrix of whether the segments i and j should be assigned to the same cluster is represented by $C(i,j)$. Those segments whose corresponding $C(i,j)$ equals to 1, are selected as the reliable parts in the following agglomerative hierarchical clustering. Moreover, we can get more precise speaker models. At the end, the input segments will be assigned to an available cluster through Viterbi decoding.

It is reported in [14] that, consensus clustering can fully utilize the complementary information between original speech and denoised speech, and achieve a significant improvement comparing to each individual system in terms of diarization error rate (DER). However, the generalization ability in adverse acoustic environments of DIHARD challenge is still not examined. In addition, to keep consistent with [14], we also use the T-test distance [5] as the clustering metric, comparing with i-vector PLDA strategy.

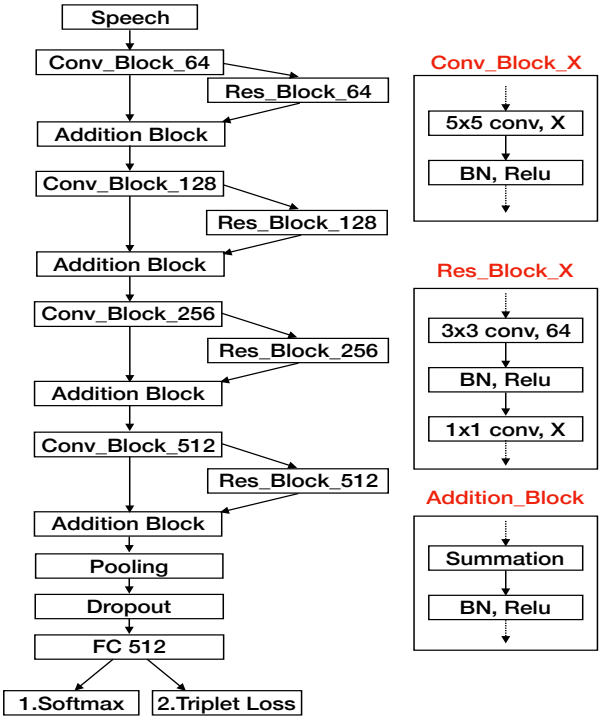

Fig. 2. The architecture of CNN-based i-vector extractor.

B. Scoring Fusion Based on Different Embeddings

Traditional i-vector extractor based on GMM-UBM works well in most scenes. But the modeling capability of UBM is relatively limited when we have large amounts of data [15]. Due to the powerful modeling ability of deep networks, an end-to-end residual convolutional neural network (CNN) based i-vector extractor [16] was used as a supplementary. The CNN architecture is shown in Fig. 2, where BN denotes the batch normalization layer, Conv_Block_X denotes the convolutional layer with X feature maps, Res_Block_X denotes the residual layer with X feature maps. For the input layer, 512 frames of 64 dimensional filterbank features which belong to the same speaker are packed together. For the output layer, a 512 dimensional vector is generated as the identity vector of the specific speaker. During the first stage of training, we pre-train the network by predicting the speaker identity using softmax loss. Then triplet loss [17] is used as the second stage training criterion. Similarities between different CNN i-vectors are measured by cosine score. In the testing stage, a complete segment will be split into a sequence of 512-frame windows. Each window can get a separate CNN-based i-vecotr, then the final embedding vector takes average of them.

Conventionally, scoring fusion is conducted on PLDA scores which are produced from several individual diarization systems. However, we found the complementarity between systems of different scoring methods is stronger than it between systems using the same scoring strategy. Hence, we explore fusing the traditional PLDA score of baseline system and the cosine score of CNN-based i-vector system via a weighted sum with coefficients. The new score function is shown as follows:

$$\text{Score} = \text{PLDA}(\text{Uvec}_i, \text{Uvec}_j) + w * \cos(\text{Cvec}_i, \text{Cvec}_j) \quad (3)$$

where PLDA calculates the score between UBM i-vectors of segment i and j (Uvec_i and Uvec_j), while cos calculates the cosine score between CNN i-vectors of segment i and j (Cvec_i and Cvec_j). Due to the dimension distinction between these two kinds of score methods, the coefficient w is the key weighting parameter to tune the balance between those two scores. In DIHARD development dataset, we go through the diarization process of our baseline system before PLDA AHC. The distribution of PLDA score is illustrated in Fig. 4. It can be observed that most scores are in the range of [-60, 60], and roughly follow a Gaussian distribution. Therefore, we set w to 60.

IV. Experiments And Results

In this section, we present our experiments on the first DIHARD challenge datasets, including both development set and evaluation set. We use the DER to evaluate the performance of a diarization system, which is defined by the evaluation campaigns organized by NIST. Note that, there is no unscored collars during the evaluation. Moreover, multiple speakers in overlap speech segments are all counted. In this study, we only focus on Track2 which means a complete system from scratch without any prior information.

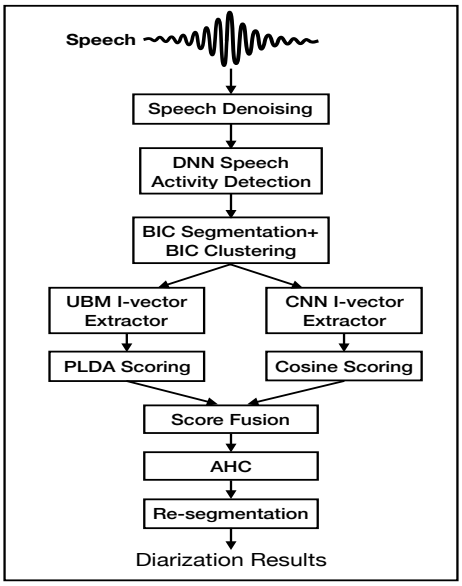

Fig. 3. The scoring fusion based on different embeddings.

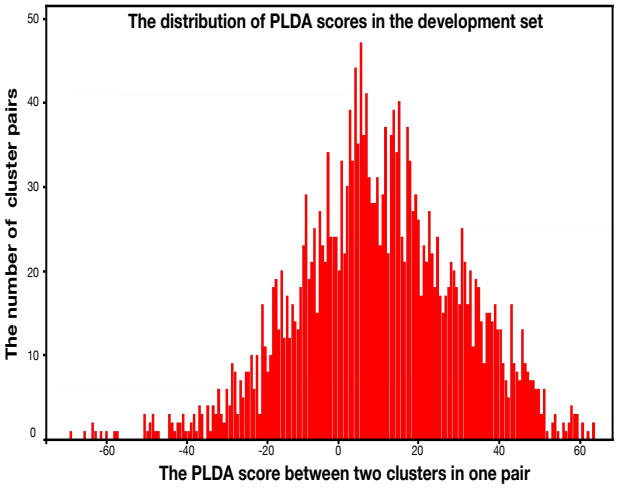

Fig. 4. The distribution of PLDA scores across the development dataset.

A. Training Data

For the denoising model, we maintain the model architecture presented in [3]. Beyond using English speech corpus WSJ0 [18], we also add a 50-hour Chinese speech corpus from 863 Program to increase the diversity of clean speech data. 115 noise types are adopted with clean data to generate pairs of clean and noisy utterances. Total size of the synthesized training data is about 400 hours. For SAD training, 600-hour home-made realistic speech data in iFlytek was used. Human annotations on each speech segment are set as the learning target.

Speaking of i-vector extractor, VoxCleb corpus which contains over 100,000 utterances for 1,251 celebrities [19], is used

for UBM and PLDA training. Besides, due to the huge demand of large scale training data for training deep networks, we use another home-made corpus in iFlytek to train the CNN-based i-vector extractor. It contains realistic recordings from more than 30,000 persons.

B. Evaluation Data

The evaluation data for diarization performance is taken from the first DIHARD challenge. The very challenging corpora is selected from a diverse set of challenging domains, including child language acquisition recordings, clinical interviews, far-field dialogs, web videos and so on. The development set and evaluation set contains 19 hours and 21 hours respectively. More details of the data can refer to [2], [20], [21].

TABLE I
EXPERIMENTS OF CONSENSUS CLUSTERING WITH DIFFERENT METRICS.

DER(%)	Devlopment Set				Evaluation Set			
System	Miss	FA	SpkrErr	Overall	Miss	FA	SpkrErr	Overall
PLDA	16.50	6.00	7.60	30.10	18.09	5.40	13.35	36.84
+Consensus	16.50	6.00	8.89	31.39	18.09	5.40	14.81	38.30
T-test	16.50	6.00	8.60	31.10	18.09	5.40	14.67	38.16
+Consensus	16.50	6.00	8.20	30.70	18.09	5.40	13.95	37.44

C. Results

The baseline system is constructed with UBM i-vector extractor and PLDA scoring model which are both trained with denoised speech. It's denoted as PLDA in Table I. For consensus clustering, original speech and denoised speech are simultaneously utilized together to purify the clusters before agglomerative hierarchical clustering, each uses its own i-vector extractor. In this section, both PLDA score and T-test distance are investigated as clustering metric. As shown in Table I, the miss and false alarm rates are the same since all systems adopt the SAD results extracted from denoised speech. Note that, our systems can not tackle with overlap speech segments, where all those segments can only be distributed to one character. That's the reason why miss error rate is relatively high. Specifically, the single PLDA system using the denoised data yields the best performance. Using PLDA as the clustering metric, consensus clustering obtains no performance gain, but an obvious decline. On the contrary, T-test distance system can benefit by consensus clustering with 0.4 and 0.72 reductions of speaker error in development set and evaluation set. Comparing with more significant improvements in [14], the reason can be categorized as follows: First, most of the diarization data here is noisy, where the denoising performance is of great importance; Second, our denoising model architecture is much more advanced and robust. Hence, a single system using the well-behaved denoising model can achieve best performance with i-vector/PLDA clustering method, original speech with consensus clustering brings no additional gain.

In Table II, the single CNN i-vector system using cosine score, denoted as Cosine in Section III-B, does not achieve the best performance. This is partly because the data used in training CNN network is all Chinese speech while the most test recordings are English. But the complementarity between PLDA score and cosine score is captured via scoring fusion presented in Section III-B. Comparing to the single PLDA scoring, the fusion method obtains relative SpkrErr reductions of 9.2% in development set and 2.1% in evaluation set, respectively. Finally using the fusion system, we achieve the second place on the evaluation set of DIHARD challenge.

TABLE II
EXPERIMENTS OF SCORING FUSION WITH DIFFERENT EMBEDDINGS.

DER(%)	Devlopment Set				Evaluation Set			
System	Miss	FA	SpkrErr	Overall	Miss	FA	SpkrErr	Overall
PLDA	16.50	6.00	7.60	30.10	18.09	5.40	13.35	36.84
Cosine	16.50	6.00	9.10	31.60	18.09	5.40	16.06	39.55
Fusion	16.50	6.00	6.90	29.40	18.09	5.40	12.56	36.05

V. DISSUSION AND FUTURE WORK

Speaker diarization in adverse acoustic environments still remains challenging, the state-of-the-art performance is barely satisfactory. The diarization errors can be mainly attributed into several types. First, the overlapped segments are difficult to detect and correctly classified to corresponding speakers. Usually, those overlapped segments will cluster together as a new speaker. After the statistical analysis, around 8.5% speech segments on the development set are overlapped parts, while it is 9.56% on the evaluation set. Second, background noises can greatly hurt the overall diarization performance, due to its effects on both SAD results and clustering process. For examples, some constantly emerging noises have strong similarity, such as car whistles, wind blows, microphone gratings and so on. Although our denoising model removes most of non-speech interferences, it's still a trade-off problem between noise reduction and speech reservation.

As conclusion, we have presented our attempts on system fusion in a realistic speaker diarization task. On condition that the denoised speech system performs well enough, consensus clustering of denoised speech system and original speech system is helpless. Furthermore, a score fusion between PLDA score and cosine score is more implementable and effective. In the future, we aim to improve the diarization performance by investigating the overlap detection and speech separation.

VI. ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, and MOE-Microsoft Key Laboratory of USTC, and Huawei Noah's Ark Lab.

REFERENCES

[1] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J. Bonastre, "The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–373.

[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," in *https://zenodo.org/record/1199638*, 2018.
[3] L. Sun, J. Du *et al.*, "A novel lstm-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *ICASSP*, 2018.
[4] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, 1978.
[5] T. H. Nguyen, E. S. Chng, and H.-Z. Li, "T-test distance and clustering criterion for speaker diarization," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.
[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
[7] S. Madikeri, I. Himawan, P. Motlicek, and M. Ferras, "Integrating online i-vector extractor with information bottleneck based speaker diarization system," Idiap, Tech. Rep., 2015.
[8] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.
[9] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," *CRIM, Montreal, Technical Report*, 2008.
[10] A. Friedland, B. Vinyals, C. Huang, and D. Muller, "Fusing short term and long term features for improved speaker diarization," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4077–4080.
[11] S. Bozonnet, N. Evans, X. Anguera, O. Vinyals, G. Friedland, and C. Fredouille, "System output combination for improved speaker diarization," in *Interspeech 2010, September 26-30, Makuhari, Japan*, 2010, pp. Interspeech–2010.
[12] I. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, 2017.
[13] T. L. Nwe, H. Sun, H. Li, and S. Rahardja, "Speaker diarization in meeting audio," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4073–4076.
[14] W.-X. Zhu, W. Guo, and G.-P. Hu, "Feature mapping for speaker diarization in noisy conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5445–5449.
[15] Y. Xu, I. McLoughlin, Y. Song, and K. Wu, "Improved i-vector representation for speaker diarization," *Circuits, Systems, and Signal Processing*, vol. 35, no. 9, pp. 3393–3404, 2016.
[16] 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.
[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
[18] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.
[19] A. Nagrani, J. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.
[20] E. Bergelson, "Bergelson Seedlings HomeBank Corpus," doi:10.21415/T5PK6D.
[21] N. Ryant, "DIHARD Corpus," dLinguistic Data Consortium.