Interspeech 2018
2-6 September 2018, Hyderabad

Speaker Diarization with Enhancing Speech for the First DIHARD Challenge

Lei Sun[1], Jun Du[1], Chao Jiang[2], Xueyang Zhang[2], Shan He[2], Bing Yin[2], and Chin-Hui Lee[3]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]iFlytek Research, Hefei, Anhui, P. R. China
[3]Georgia Institute of Technology, Atlanta, GA. USA

sunlei17@mail.ustc.edu.cn, jundu@ustc.edu.cn, chaojiang2@iflytek.com,
xyzhang12@iflytek.com, shanhe2@iflytek.com, bingyin@iflytek.com, chl@ece.gatech.edu

Abstract

We design a novel speaker diarization system for the first DIHARD challenge by integrating several important modules of speech denoising, speech activity detection (SAD), i-vector design, and scoring strategy. One main contribution is the proposed long short-term memory (LSTM) based speech denoising model. By fully utilizing the diversified simulated training data and advanced network architecture using progressive multitask learning with dense structure, the denoising model demonstrates the strong generalization capability to realistic noisy environments. The enhanced speech can boost the performance for the subsequent SAD, segmentation and clustering. To the best of our knowledge, this is the first time we show significant improvements of deep learning based single-channel speech enhancement over state-of-the-art diarization systems in highly mismatch conditions. For the design of i-vector extraction, we adopt a residual convolutional neural network trained on large dataset including more than 30,000 people. Finally, by score fusion of different i-vectors based on all these techniques, our systems yield diarization error rates (DERs) of 24.56% and 36.05% on the evaluation sets of Track1 and Track2, which are both in the second place among 14 and 11 participating teams, respectively.

Index Terms: speaker diarization, speech denoising, speech activity detection, i-vector, DIHARD challenge

1. Introduction

Speaker diarization is a task to segment an audio recording into speaker homogeneous regions without any prior information including the number of speakers [1], the dialog styles, environmental scenes and so on. Good speaker diarization results can be very beneficial to several speech areas, such as transcription of dialogues, dominant speaker detection, speech indexing, and meeting summary [2]. All these domains are extremely significant to promote and popularize the practical use of speech technology in daily life.

A conventional speaker diarization algorithm can be roughly divided into two main components: speaker segmentation and clustering. Depending on the difference of sequential order between these two components, most of state-of-the-art speaker diarization systems fall into two categories: the bottom-up and the top-down approaches [3]. The bottom-up method, also known as agglomerative hierarchical clustering (AHC) [4], first cuts the whole speech recording into smaller segments where each segment ideally comes from only one speaker. The closet segments selected by some distance metrics like Bayesian information criterion (BIC) [5], are merged iteratively until a certain stopping criterion is satisfied. On the contrary, the top-down approach will successively split speech segments to new clusters until reaching the number the speakers. In general, bottom-up approaches are far more popular than top-down ones. Recently, i-vector has shown great effectiveness in the field of speaker recognition [6, 7]. It is natural to introduce i-vector to speaker diarization as a more powerful feature to enhance speaker specific information. Moreover, a probabilistic linear discriminant analysis (PLDA) scoring fuction [8, 9] is learned to discriminate whether two i-vectors are from the same person.

Apart from the abovementioned diarization process, a practical speaker diarization system should also include the preprocessing stage [3], which involves speech denoising, multichannel acoustic beamforming and speech activity detection. The background noises, reverberations and other interferences in real scenes, can greatly hurt the overall diarization performance. Thus the accumulated error during the whole process becomes uncontrolled and untraceable. Especially in the single-channel case with limited spatial information, an effective speech denoising algorithm plays an important role as the front-end preprocessor. In [10], we have shown that deep learning based denoising method has stronger potentials in coping with realistic noisy environments than traditional approaches. The complicated acoustic environments also affect speech activity detection which is quite important for diarization. With a good front-end prepocessing, better speech quality and more accurate speech boundary location can ensure a higher upper bound for the performance of speaker diarization.

While state-of-the-art diarization systems perform remarkably well for some domains (e.g., conversational telephone speech such as CallHome), as was discovered at the 2017 JSALT Summer Workshop at CMU [11], this success can not transfer to more challenging corpora such as child language recordings, clinical interviews, speech in reverberant environments, web video, and speech in the wild. To explore the benchmark of current state-of-the-art systems, the first DIHARD speech diarization challenge [12] was proposed where the datasets are drawn from a diverse set of challenging domains. The challenge has two tracks, namely Track1 and Track2. Track1 uses gold speech segmentation while Track2 does diarization from scratch.

In this study, we present a novel integrated diarization system for DIHARD challenge, consisting of speech denoising, speech activity detection, the design of i-vector extraction and scoring strategy. We build the deep denoising model using the advanced LSTM architecture with the novel design of hidden layers via densely connected progressive learning and output layer via multiple-target learning [10]. Much larger amounts of training data are used to guarantee better generalization ability. A deep neural network (DNN) based speech activity detection model is trained on realistic collected data. Then we construct an i-vector extraction system for speaker representation, com-

10.21437/Interspeech.2018-1742

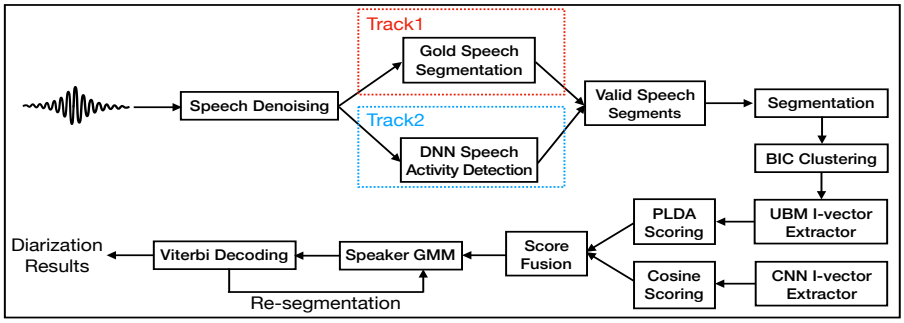

Figure 1: *Complete speaker diarization system diagram in both Track1 and Track2.*

bined with a PLDA scoring model. Furthermore, we propose a residual convolutional neural network (CNN) based i-vector extractor and make a fusion with traditional PLDA score. Finally, we evaluate the performance on both Track1 and Track2.

2. Database

A complete diarization system contains multiple sub-modules, as illustrated in Figure 1. In the section, we introduce all datasets which are used for constructing each part of those sub-modules. First for the speech enhancement, we have already explored its validity for realistic environments in [10]. Unlike only using English speech corpus WSJ0 [13] there, in this work we add a 50-hour Chinese speech corpus from 863 Program to increase the diversity of clean speech data. 115 noise types of audios are adopt to simulate noisy utterances with clean speech. To improve the stability and generalization ability, we use 400-hour simulated clean/noisy pairs of speech data instead of 36-hour in [10]. As discussed before in [10], the enhancement model often crushed when it needs to cope with any unseen speech which belongs to teenagers and babies. At that time, we ascribed the performance degradation into the vacancy of child speech data in training set. Surprisingly, in this study those problems seem partially solved with the increase of training data size, although we still do not use any child speech data. More discussion will be stated in Section 3.1.

Speaking of i-vector extractor, we choose the increasingly popular VoxCeleb corpus [14] to train the i-vector extractor based on universal background model (UBM). It is a large scale speaker identification dataset derived from YouTube, containing over 100,000 utterances for 1,251 celebrities. Moreover, we use another home-made corpus in iFlytek. It is collected in daily scenario, including more than 30,000 persons. It is expected to enhance the performance of our residual CNN-based i-vector extractor.

For SAD training, 600-hour home-made realistic speech data in iFlytek was used. The speech quality is not very stable due to the complicated acoustic environments. Human annotations on each speech segment are set as the learning target.

The details of development set and evaluation set in DIHARD challenge can refer to [12, 15, 16].

3. System Description

The generic speaker diarization system often contains several main components: speech denoising, acoustic feature extraction, speech activity detection, speaker representation, speaker segmentation, speaker clustering and re-segmentation. In this section, we introduce each part in our system.

3.1. Speech denoising

Inevitably, a practical diarization system should address the environmental robustness problem in real applications. For speaker diarization, a good preprocessor should obey two rules. On one hand, it should be able to remove background noises as much as possible. On the other hand, speaker specific information should not be lost. Therefore the trade-off between noise suppression and speaker information preservation is crucial for speech enhancement in speaker diarization system. Moreover, the adverse acoustic environments require the preprocessor to have excellent stability and generalization ability. Traditional enhancement methods like Wiener filtering [3], LogMMSE [17], there are many limitations in real applications, e.g., the weakness of tracking non-stationary noises, due to the model assumptions made during the inference. Furthermore, the annoying artifact generated in denoised speech can degrade the performance of speaker diarization system.

In recent years, the emergence of deep learning techniques in speech enhancement has partly solved the problem [18, 19, 20], such as decreasing the artifact. However, the generalization ability in mismatched conditions is the main problem of deep learning based method. Inspired by our previous work [21, 22], we adopt an advanced LSTM architecture with the novel design of hidden layers via densely connected progressive learning and output layer via multiple-target learning. The overall LSTM architecture aims to predict the clean LPS features given the input noisy log-power spectra (LPS) features with acoustic context. All the target layers are designed to learn intermediate speech with higher SNRs or clean speech. For the input and multiple targets, LSTM layers are used to link between each other. This stacking style network can learn multiple targets progressively and efficiently. In order to make full use of the rich set of information from the multiple learning targets, we update the progressive learning in [22] with dense structures [23] in which the input and the estimations of intermediate target are spliced together to learn next target. Then, a weighted MMSE criterion in terms of multitask learning (MTL) is designed to optimize all network parameters randomly initialized with K target layers as follows:

$$
\begin{aligned}
E &= \sum_{k=1}^{K} \alpha_k E_k + E_{\text{IRM}} \\
E_k &= \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}_k(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k) - \mathbf{x}_n^k \|_2^2 \qquad (1) \\
E_{\text{IRM}} &= \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}_{\text{IRM}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{K-1}, \mathbf{\Lambda}_{\text{IRM}}) - \mathbf{x}_n^{\text{IRM}} \|_2^2
\end{aligned}
$$

where E_k is mean square error (MSE) corresponding to k^{th} target layer while E_{IRM} is MSE for MTL with ideal ratio masks (IRM) in the final output layer. $\hat{\mathbf{x}}_n^k$ and $\mathbf{x}_n^k$ are the n^{th} D-dimensional vectors of estimated and reference target LPS feature vectors for k^{th} target layer, respectively ($k > 0$), with N representing the mini-batch size. $\hat{\mathbf{x}}_n^0$ denotes the n^{th} vector of input noisy LPS features with acoustic context. $\mathcal{F}_k(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k)$ is the neural network function for k^{th} target with the dense structure using the previously learned intermediate targets from $\hat{\mathbf{x}}_n^0$ to $\hat{\mathbf{x}}_n^{k-1}$, and $\mathbf{\Lambda}_k$ represents the parameter set of the weight matrices and bias vectors before k^{th} target layer, which are optimized in the manner of back propagation through time (BPTT) with gradient descent. $\mathbf{x}_n^{\text{IRM}}$, $\mathcal{F}_{\text{IRM}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{K-1}, \mathbf{\Lambda}_{\text{IRM}})$, and $\mathbf{\Lambda}_{\text{IRM}}$ are corresponding versions to IRM targets. α_k is the weighting factor for k^{th} target layer. More details of the architecture can be found in [10].

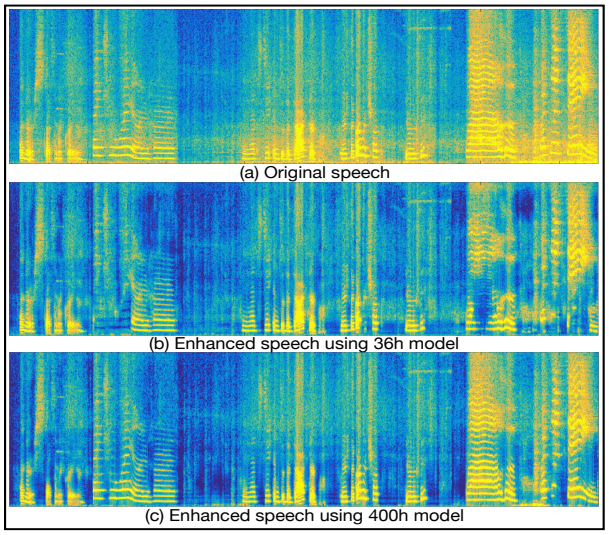

Figure 2: *A comparison of spectrograms for the proposed enhancement models with different training data setups.*

We have shown that, only replacing noisy waveforms with denoised waveforms can yield significant reductions of DERs on several challenging datasets [10]. In this study, with the same network architecture, the larger amount of training data setup brings better generalization ability not only to adult speech, but also to unseen child speech. Figure 2 is one segment example derived from DIHARD development set, where child speech exists. The result from current 400-hour model achieve better integrity comparing to former 36-hour model used in [10], even the test speech data is still unseen in training set.

3.2. Speech activity detection

Recently, several studies adopted DNNs for SAD [24, 25, 26]. Here we train a framewise binary classification DNN of speech and non-speech. The features we use are 39-dimensional perceptual linear prediction (PLP) features (13-dimensional static PLP features with Δ and $\Delta\Delta$) and include an input context of 5 neighbouring frames (± 2), yielding a final dimensionality of 195 (39×5). Considering utility efficiency, the DNN model adopts a small and compact structure using 2 hidden layers with 256 and 128 hidden units in each layer and a final dual output layer, i.e. an architecture of 195-256-128-2. All training data is from realistic collected corpus.

3.3. Speaker segmentation and clustering

To fully utilize the effective information embedded in every stage, we propose a two-pass short-long term diarization system in this section.

3.3.1. Short-term diarization

Generally, speaker changes, also known as speaker turns, may appear everywhere in conversations, within or without overlaps. Especially in scenes like meetings, debates, parties, the conversions of speakers are frequently. Given the valid speech segments from SAD, it is important to split them into speaker homogeneous segments. It is also pivotal to prevent error accumulating in the very beginning. We use the Bayesian information criterion (BIC) [5] as the hypothesis testing metric.

Then a global agglomerative hierarchical clustering (AHC) algorithm [4] is performed on all segments. At this step, every single segment is relatively short. The process is conducted iteratively, until a certain criterion is reached, upon which one separate cluster should arrive an upper limit or the number of clusters reaches a default maximum speaker number.

3.3.2. Long-term diarization

When the duration of each segment is relatively long, the i-vector can be a more powerful representative feature. We use an i-vector extraction system trained on the VoxCeleb corpus. Our UBM includes 1024 Gaussians and the total variability (TV) matrix reduces the dimension to 400. The i-vectors are denoted as UBM i-vectors and also length-normalized. In clustering, we repeatedly merge the closest two i-vectors based on a certain scoring metric. We train a PLDA scoring model to measure the similarity between the i-vectors. Moreover, we retrain the UBM i-vector/PLDA model using the denoising data.

3.3.3. Residual CNN-based i-vector extractor

Although i-vector extracted from a UBM works well in some scenes like telephone data, the modeling capability of UBM is relatively limited [27]. Inspired by the residual network in image recognition [28], an end-to-end residual CNN-based i-vector extractor was proposed in the field of speaker recognition [29]. Due to the powerful modeling ability of deep networks, we can achieve promising embedding performance as long as there is adequate training data. Thus we train a residual CNN network for i-vector which is shown in Figure 3. For the input layer, 512 frames of 64 dimensional filterbank features which belong to the same person are grouped together as a feature map. At output layer, a 512 dimensional vector is generated as the identity vector of the specific person. During the first stage in training, we pre-train the network by predicting the speaker identity using softmax loss. Then triplet loss [30] is used as the second stage training criterion. Similarities between different CNN i-vectors are measured by cosine score.

3.3.4. Realignment

At the end, a realignment over frames is performed via Viterbi decoding on the GMM of each speaker. To make it more stable, we also use some smoothing strategy to prevent erroneously detected speaker turns [31].

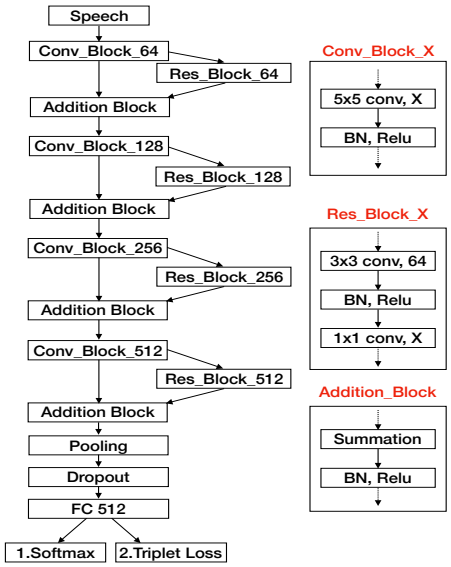

Figure 3: *The architecture of the residual CNN i-vector model.*

Table 1: *DER comparison of different speech inputs for UBM i-vector based diarization system on development set.*

DER(%)	Track1			
Speech	Miss	FA	SpkrErr	Overall
Original	8.50	0	11.76	20.26
Denoised	8.50	0	11.18	19.68
Retrained	8.50	0	11.01	19.51
DER(%)	Track2			
Speech	Miss	FA	SpkrErr	Overall
Original	18.60	6.10	8.50	33.20
Denoised	16.50	6.00	7.90	30.40
Retrained	16.50	6.00	7.60	30.10

4. Experiments

4.1. Evaluation metric

We measure the performance of the diarization system by DER, which is defined by the evaluation campaigns organized by NIST. It compares the differences between the ground-truth reference segmentation and the generated diarization output. The final DER result is the sum of three types of errors: E_{Miss}, E_{FA} and E_{Spkr}, where each represents the percent of missed speech, false alarm error speech, and speaker misclassification error speech, respectively. Lower DER indicates better diarization performance. Note that, for DIHARD challenge, non-scoring collar is not permitted which means collar is set to zero in scoring script. Moreover, multiple speakers in overlap speech segments are counted.

4.2. Results

First, we build a baseline speaker diarization system based on UBM i-vector extractor and PLDA model, which are both trained upon original VoxCeleb data. In Track1, we only use the gold speech segmentation, while Track2 uses the outputs of DNN-based SAD. As shown in Table 1, the DER on development set can benefit directly from denoised speech from 20.26% to 19.68% in Track1. Note that, our system does not tackle with overlap speech segments. That is to say, all overlap segments will be distributed to only one speaker, which generates inevitable missed error in both Tracks. Specifically, Miss is 8.5% in Track1 while FA is 0 with gold segmentation. In Track2, denoised speech can significantly reduce the percentage of Miss and FA, due to the removal of environmental interferences. Moreover, the valid speech segments can be less confusing, in terms of the reduction of SpkrErr. Furthermore, by retraining the i-vector extractor and PLDA model using denoised training data, additional improvements could be observed for both Track1 and Track2 as shown in the third row of each track.

Table 2: *DER comparison of different scoring strategies on development set.*

DER(%)	Track1			
Scoring	Miss	FA	SpkrErr	Overall
PLDA	8.50	0	11.01	19.51
PLDA+Cosine	8.50	0	8.90	17.40
DER(%)	Track2			
Scoring	Miss	FA	SpkrErr	Overall
PLDA	16.50	6.00	7.60	30.10
PLDA+Cosine	16.50	6.00	6.90	29.40

Table 3: *DER comparison of overall top-3 teams on evaluation set.*

DER (in %)	USTC-iFlytek (Ours)	Team1	Team2
Track1	24.56	23.73	25.07
Track2	36.05	37.19	35.51

System fusion [32, 33] is an effective strategy to improve the performance of speaker diarization system, including feature-level fusion [34], system output-level fusion [35], and multi-model fusion like audio-visual fusion [36]. To fully utilize the complementarity between UBM i-vector and CNN i-vector, in our fusion system we directly conduct a scoring fusion between PLDA score of UBM i-vector and cosine score of CNN i-vector. Comparing to single PLDA scoring, the fusion method obtains relative SpkrErr reductions of 19.2% in Track1 and 9.2% in Track2, respectively. Using this fusion system, we achieve both the second place on the evaluation set of DIHARD challenge among 14 teams of Track1 and 11 teams of Track2, as illustrated in Table 3.

5. Summary and future work

First, a well-designed speech enhancement algorithm can help both detection and diarization of valid speech segments. Second, different designs of i-vector extractor could be strongly complementary. In the future, we aim to improve the diarization performance by investigating the overlap detection and separation in realistic scenes.

6. Acknowledgment

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. This work was also funded by Huawei Noah's Ark Lab.

7. References

[1] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *Multimodal Technologies for Perception of Humans*, pp. 509–519, 2008.

[2] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[4] K. Han and S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," in *Interspeech 2010, September 26-30, Makuhari, Japan*, 2010, pp. Interspeech–2010.

[5] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, 1978.

[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] L. Sun, J. Du *et al.*, "A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *ICASSP*, 2018.

[11] K. Church, A. Cristiab *et al.*, "Enhancement and Analysis of Conversational Speech: JSALT 2017," in *ICASSP*, 2018.

[12] 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.

[13] 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.

[14] A. Nagrani, J. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[15] E. Bergelson, "Bergelson Seedlings HomeBank Corpus," doi:10.21415/T5PK6D.

[16] N. Ryant, "DIHARD Corpus," Linguistic Data Consortium.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[18] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[19] ——, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 826–835, 2014.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[21] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*. IEEE, 2017, pp. 136–140.

[22] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement," in *INTERSPEECH*, 2016, pp. 3713–3717.

[23] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[24] X. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[25] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *INTERSPEECH*, 2013, pp. 728–731.

[26] Q. Wang, J. Du, and et al, "A universal VAD based on jointly trained deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[27] 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.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Interspeech 2010, September 26-30, Makuhari, Japan*, 2010, pp. Interspeech–2010.

[29] 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.

[30] 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.

[31] C. Fredouille, S. Bozonnet, and N. Evans, "The LIA-EURECOM RT ?9 Speaker Diarization System," in *RT?9, NIST Rich Transcription Workshop*, vol. 15, 2009, pp. 17–23.

[32] S. Tranter, "Two-way cluster voting to improve speaker diarisation performance," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–753.

[33] 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.

[34] 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.

[35] 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.

[36] 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.