PROGRESSIVE MULTI-TARGET NETWORK BASED SPEECH ENHANCEMENT WITH SNR-PRESELECTION FOR ROBUST SPEAKER DIARIZATION

Lei Sun[1], *Jun Du*[1], *Xueyang Zhang*[2], *Tian Gao*[2], *Xin Fang*[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

ABSTRACT

In this paper, we design a novel front-end processing system for speaker diarization under realistic conditions with challenging background noises. To cope with diversified environments, we first extend our perviously proposed progressive learning based speech enhancement model by adding multi-task learning in each intermediate layer. The corresponding progressive multi-target (PMT) in various layers includes both progressive ratio mask (PRM) and progressively enhanced log-power spectra (PELPS) with specified signal-to-noise-ratios (SNRs). Speech distortions are commonly introduced during the front-end processing, which often deteriorate the back-end performance. However, the proposed speech enhancement model can be regarded as a bagging of models with multiple learning objectives, which provides flexibility for selecting the most appropriate output for robust speaker diarzation. In addition, a global SNR estimation is performed using the results of deep neural network (DNN) based speech activity detection (SAD) to decide whether the audio should be enhanced. We evaluate the speaker diarzation performance on the second DIHARD dataset which includes several different realistic conditions. Compared with the original data, experiments demonstrate that the enhanced data processed by our proposed method can effectively avoid the performance loss of every single domain, and achieve consistent improvements in most domains.

Index Terms— Speech enhancement, speaker diarization, speech activity detection, DIHARD data, SNR estimation

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, 2], the dialog styles, environmental scenes and so on. Good speaker diarization results are of great help for applications such as speech transcription, dominant speaker detection, speech indexing and conference summary [3]. All of these areas are extremely important for promoting and popularizing the practical application of speech technology in everyday 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 the advanced speaker diarization systems fall into two categories: the bottom-up and the top-down approaches [4]. The bottom-up method, also known as agglomerative hierarchical clustering (AHC) [5], first cuts the entire speech recording into smaller segments, each segment ideally from only one speaker. The closest segments selected by some distance metrics like Bayesian information criterion (BIC) [6], are merged iteratively until a certain stopping criterion is satisfied. Instead, the top-down approach successively divides the speech segments to new clusters until the number of speakers is reached. 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 [7, 8]. 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 function [9, 10] is learned to discriminate whether two i-vectors are from the same person.

A series of diarization challenges, namely DIHARD [11, 12], have been focused on speaker diarization for challenging recordings where there is an expectation that the current state-of-the-art will fare poorly. The data used in DIHARD were drawn from a diverse sampling of sources such as clinical interviews, speech in restaurants, extended child language acquisition recordings [13], web videos, speech in the wild and etc [14]. To overcome the noise issues in single channel speech, many great efforts have been made in the field of speech enhancement. In [15, 16], ideal ratio mask (IRM) was used to make binary classification on time-frequency (T-F) units for speech separation. Previously, we proposed a deep neural network (DNN) framework to learn the direct mapping from noisy to clean speech in log-power spectral (LPS) domain, which demonstrated its superiority to the traditional enhancement methods [17, 18], especially for tracking the non-stationary noises. For speaker diarization, we have designed a 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 in [19]. It has been shown to have stronger potentials in coping with realistic noisy environments than traditional approaches. In [20], much larger amounts of training data were adopted to guarantee better generalization ability.

However, it has been observed that speech enhancement is not always beneficial to the performance of speaker diarization. As presented in [21, 22], improvements were attained only in few domains, such as ADOS and Seedlings. Compared with the original data, the enhanced data achieved worse overall performance, especially when using oracle speech activity labels. It's also the reason why the official baseline of DIHARD-II [12] adopts the same enhancement model with [20], but not in tracks with oracle SAD. The un-

The work reported here was partly conducted at JSALT 2019 [1] hosted at École de Technologie Supérieure, and sponsored by JHU with unrestricted gifts from Amazon, Facebook, Google and Microsoft. 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 Grant Nos. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. It was also funded by Huawei Noah's Ark Lab.

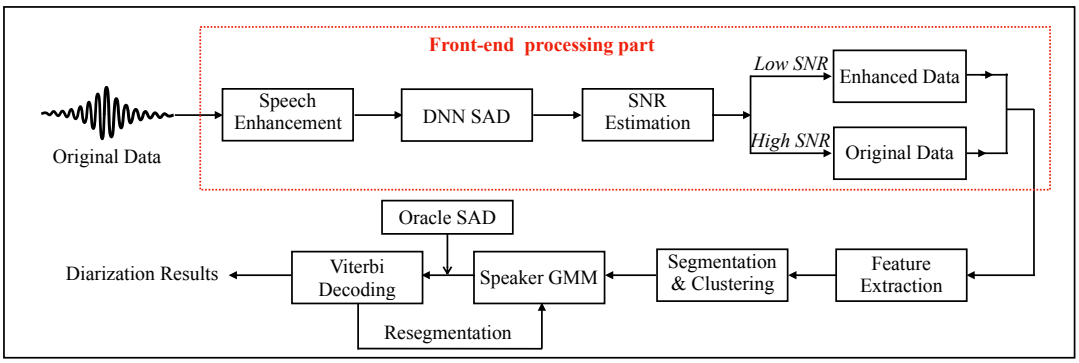

Fig. 1. The overall diagram of our proposed front-end processing system for speaker diarization.

derlying reason is often attributed to speech distortions which are inevitablely introduced during the front-end processing. It's necessary to explore a practical front-end method to control the degree of speech enhancement in realistic environments.

In this work, we first extend our previously proposed progressively learning based speech enhancement model by adding multi-task learning in each intermediate layer, namely progressive multi-target network. Comprehensive comparisons have been conducted among all intermediate targets and the outermost targets in terms of speaker diarization performance. To make it more practical in any conditions, we use a DNN SAD based SNR estimation to decide whether the recording should be enhanced or not. The proposed methods are evaluated in the development set of the second DIHARD challenge. The paper is organized as follows. In Section 2, the PMT speech enhancement network is proposed along with the SNR estimation approach. Section 3 and Section 4 present the experimental setups and results. In Section 5, we summarize several findings and conclusions.

2. FRONT-END PROCESSING

In [19, 20], the model structure adopted the advanced network design of hidden layers via densely connected progressive learning of hidden layers and multi-objective learning of the output layer. Compared with the input noisy speech, the SNR of the intermediate target increased gradually, while the output layer still corresponded to the clean speech. We used LSTM layers to fit the relationship between different targets, this stacking style network could learn multiple targets progressively and efficiently. In order to make full use of the rich set of information from the multiple learning targets, we updated the progressive learning in [23] with dense structures [24] in which the input and the estimations of intermediate targets are spliced together to learn next target. The overall method aimed to predict the clean LPS (CLPS) features and IRM given the input noisy LPS (NLPS) features with acoustic context.

Although we have evaluated the effectiveness of the deep-learning based speech enhancement method for the speaker diarization task in adverse environments, the performance analysis was not deep enough, especially for each intermediate target. Furthermore, the effect of those different learning targets on the performance of speaker diarzation remains unclear. For this purpose, we present a newly designed progressive multi-target network. To avoid possible speech distortions in quiet scenes, we also utilize a SAD based SNR estimation as an additional switch for the enhancement step.

2.1. Progressive multi-target (PMT) network

As illustrated in Fig. 2, the whole PMT network is divided into successively stacking blocks with one LSTM layer and one fully connected layer via multi-target learning per block. The fully connected layer in every block is also referred to as a target layer, which is designed to learn intermediate speech targets with a higher SNR than the targets of previous target layers. A series of progressive ratio masks (PRM) are concatenated with the progressively enhanced log-power spectra (PELPS) features together as the learning targets. The PELPS targets are set to LPS features with progressively increasing SNRs, which are the same in [19, 20].

The newly designed PRM is defined as follows:

$$z^{\text{PRM}}(t,f) = \frac{S(t,f) + N_{\text{T}}(t,f)}{S(t,f) + N_{\text{I}}(t,f)} \tag{1}$$

where $S(t,f)$ represents the power spectrum of the speech signal at the time-frequency (T-F) unit (t,f), $N_{\text{T}}(t,f)$ and $N_{\text{I}}(t,f)$ represent the power spectrum of the noise in one PRM target and input signals at the T-F unit (t,f), respectively. When the numerator of Eq. (1) becomes the power spectrum of the clean speech signal, $N_{\text{T}}(t,f)$ is zero and $z^{\text{PRM}}(t,f)$ is regressed to the traditional IRM $z^{\text{IRM}}(t,f)$. Hence, in practical use, the PRM can also serve as a progressively stronger enhancing ability.

Here, the total number of target layers K is set to 3. Correspondingly, a weighted MMSE criterion is designed to optimize all network parameters randomly initialized with K target layers as follows:

$$\begin{aligned} E_{\text{MTL}}(k) &= \sum_{m=1}^{k} E_{\text{PELPS}}(m) + E_{\text{PRM}}(m) \\ E_{\text{PELPS}}(m) &= \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{F}_m(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{m-1}, \mathbf{\Lambda}_m) - \mathbf{x}_n^m\|_2^2 \\ E_{\text{PRM}}(m) &= \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{F}_{\text{PRM}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{m-1}, \mathbf{\Lambda}_{\text{PRM}}) - \mathbf{x}_n^{\text{PRM}}\|_2^2 \end{aligned} \tag{2}$$

where $E_{\text{MTL}}(k)$ corresponds to the multi-target loss in k^{th} target layer. It's the sum of two kinds of losses, namely $E_{\text{PELPS}}(m)$ and $E_{\text{PRM}}(m)$ from all lower layers. $\hat{\mathbf{x}}_n^m$ and $\mathbf{x}_n^m$ are the n^{th} D-dimensional vectors of estimated and reference target PELPS feature vectors for m^{th} target layer, respectively ($m > 0$), 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. $\mathcal{F}_m(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{m-1}, \mathbf{\Lambda}_k)$

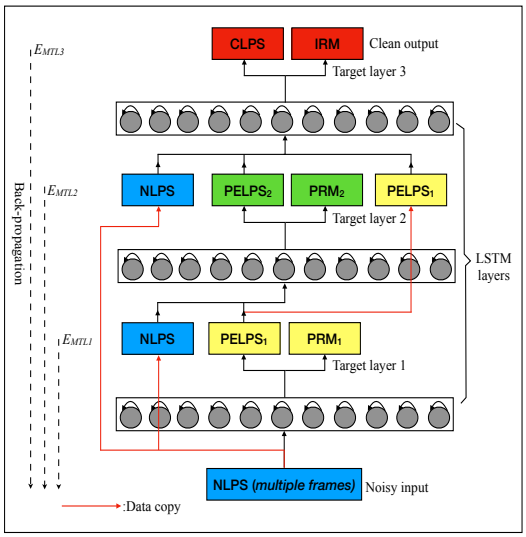

Fig. 2. An illustration of our proposed progressive multi-target network for speech enhancement, where the number of progressive stages K is set to 3.

is the neural network function for m^{th} target with the dense structure using the previously learned intermediate targets from $\hat{\mathbf{x}}_n^0$ to $\hat{\mathbf{x}}_n^{m-1}$, and $\mathbf{\Lambda}_m$ represents the parameter set of the weight matrices and bias vectors before m^{th} target layer, which are optimized in the manner of BPTT with gradient descent. $\mathbf{x}_n^{\text{PRM}}$, $\mathcal{F}_{\text{PRM}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{K-1}, \mathbf{\Lambda}_{\text{PRM}})$, and $\mathbf{\Lambda}_{\text{PRM}}$ are corresponding versions to PRM targets.

2.2. SAD based global SNR estimation

Though the oracle SAD labels are provided, we find manual annotations in different files vary a lot in scale and accuracy. Therefore, we train a deep neural network (DNN) based framewise binary classification model to detect speech frames. When testing, we choose a strict threshold to discard more silence and non-speech.

In some cases, the speaker diariziation performance suffers an obvious decline after using speech enhancement, especially for some quiet environments. The main reason can be attributed to the distortions generated in the front-end. Given the estimated speech activity information, we adopt a utterance-level SNR estimation to determine the application scope of speech enhancement. We assume that the observed signal $x(t)$ is the sum of the noise signal $n(t)$ and speech signal $s(t)$, t denotes the time index. Similarly, we also assume that noise and speech signals are independent. A global SNR can be estimated as follows [25]:

$$\begin{aligned} SNR_{\text{global}} &= 10\log_{10}\frac{P(x)-P(n)}{P(n)} \\ &= 10\log_{10}\frac{\frac{1}{A}\sum_{a=1}^{A}\|x(a)\|^2 - \frac{1}{B}\sum_{b=1}^{B}\|n(b)\|^2}{\frac{1}{B}\sum_{b=1}^{B}\|n(b)\|^2} \end{aligned} \tag{3}$$

where A denotes the total points of data which are determined as containing speech by the segmentation information from a DNN SAD or official reference SAD, while B denotes the number of those non-speech data. After estimating the SNR, only audios with low SNRs need to be processed by the speech enhancement model. The SNR threshold is set to 20 dB and the SNR preselection procedure is shown in Fig. 1.

3. EXPERIMENTAL SETUPS

3.1. Front-end processing

To improve the generalization ability of speech enhancement model, we built a 1000-hour training set. The clean speech data were collected from WSJ0[26], AIShell-1[27], THCHS-30[28] and Librispeech[29]. In addition to the 115 types of noises in [19], MUSAN [30] corpus was also adopted. The noisy mixture were made at three SNR levels (-5dB, 0dB and 5dB), and the progressive increasing SNR between two adjacent targets was set to 10 dB. The audios were sampled at 16 kHz rate and the frame length is 256 samples. Therefore, both the PELPS and the PRM were 257 dimension. As illustrated in Section 2.1, 7-frame expansion was used for the input, the number of LSTM memory cells in each layer was 1024.

Considering utility efficiency, our SAD model was based on a feed-forward neural network using 2 hidden layers with 256 and 128 hidden units in each layer. The acoustic features were 39-dimensional perceptual linear prediction (PLP) features (13-dimensional static PLP features with Δ and $\Delta\Delta$) and included an input context of 5 neighbouring frames (± 2), yielding a final dimensionality of 195 (39×5). The overall architecture was 195-256-128-2. A 600-hour home-made corpus in iFlytek was used for training with its human annotations.

3.2. Speaker diarization system

We used a standard speaker diarization system described in [20], which consisted of the following modules: Bayesian information criterion (BIC) [6] based segmentation and clustering, i-vector extraction, PLDA scoring, and resegmentation. A sliding window analysis was first conducted to detect speaker turns in extracted 39-dimensioanl PLP features of all detected segments. Then, we performed a global agglomerative hierarchical clustering (AHC) algorithm [5] on all splitted segments using the BIC. The clustering process stoped when the total cluster number reached a default maximum speaker number.

Next, we used the i-vector as a more powerful speaker representation. When training the i-vector extractor, the UBM contained 2048 Gaussians and the total variability (TV) matrix reduced the representation dimension to 400. The corpora we used here include VoxCeleb1 [31] and VoxCeleb2 [32]. The i-vectors were mean normalized, whitened, length-normalized and then used for training a PLDA model to measure the similarity. When clustering, we repeatedly merged the closest two segments based on the PLDA scoring. At the end, a resegmentation over frames was performed via Viterbi decoding on the GMM of each speaker, which should be aligned to oracle SAD boundaries.

4. RESULTS AND ANALYSIS

4.1. Evaluation metric

We evaluate the diarization error rate [33] on the DIHARD development set where oracle SAD labels are provided. Collar is set to zero and multiple speakers in overlap speech segments are counted. Here

we use oracle SAD and directly compare the speaker error part which can fairly represent the trend of the overall DER. Rather than targeting the best number, we would like to present the detailed analysis on the effects of speech enhancement to robust speaker diarization.

4.2. Comparison between baseline model and PMT model

Since the proposed model structure is an improved version of the original method [20] which is also adopted in the official DIHARD-II baseline[1] [12], here we also take it as our baseline model and compare the performance of the two methods. The details of them are presented in Table 1, we use the mask outputs in the outermost layer of both models to make the comparison fair which keeps the same usage with [12, 20]. Given that the speaker error of the original data is 12.1%, both methods yield improvements. The main difference between these two model architectures is the adoption of PRM targets. However, we use a larger training data set here, which is also the reason for the better results.

Table 1. Details comparison between baseline model and the proposed PMT model.

SE models	Training data	PL stages	PRM targets	SpkErr
Baseline[12, 20]	400h	3	No	11.62%
Proposed model	1000h	3	Yes	11.39%

4.3. Detailed analysis of different outputs of PMT model

Fig. 3 lists the diarization performance of all individual outputs from PMT speech enhancement network in terms of speaker error. Several observations can be made. First, it's obvious to see that PELPS in the outermost target layer 3 yields the worst performance, which is a common method that directly learns the clean speech during the training process [18, 21]. An important reason is that in a mismatched real test scenario, excessive noise reduction can lead to great nonlinear distortions which can severely hurt the back-end diarization performance. However, performance of the PELPS greatly improves as the hierarchy is reduced, given that PELPS outputs of target layer 1 and 2 correspond to 11.0% and 11.96%, respectively. It indicates that the introduced distortions of PELPS targets can be well controlled when using the shallow outputs of PMT network.

Second, as a contrast, the performance of mask-based targets PRM is much more stable which is at least better than the original data. Still, the results from target layer 1 are superior than those higher layers. As described in Section 2.1, each target layer is given a specified SNR gain. It's shown that the target of PELPS1 and PRM1 with +10dB SNR gain is optimal in speaker diarzation. As suggested in [34], a simple ensemble method by averaging the PRM and PELPS outputs in feature level could improve the enhancement metrics. However, similar improvement doesn't migrate to speaker diarization. Compared with original data, the best performance by PRM1 reduces the average speaker error by absolute 1.15%, from 12.1% to 10.95%. It indicates that in order to reduce mismatches in realistic situations, it is better to consider using relatively higher SNR speech as the training target instead of pure clean speech.

4.4. Performance on different data domains

To explore the difference between the original data and the enhanced data, we list the domain-wise results in Table 2. As we can see,

[1]https://github.com/staplesinLA/denoising_DIHARD18

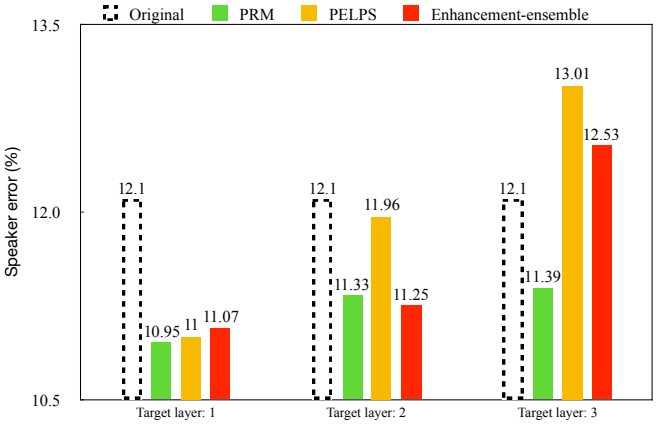

Fig. 3. Histogram of the speaker error of different enhanced outputs in the PMT network.

the baseline speech enhancement model obtains improvements in many domains, and the PRM1 yields more stable and better overall performance. In most domains, the enhanced speech data reduce the speaker error except for domains like LIBRIVOX, DCIEM, which were recorded under quiet conditions [12]. A threshold of 20 dB was adopted as SNR-preselection to divide the audios into two groups: high SNR and low SNR, and we only performed speech enhancement on those low SNR data. As seen in Table 2, performance losses in these two areas have been recovered substantially. At the same time, the average speaker error rate also dropped to 10.7%, indicating that the SNR-based preselection of speech enhancement is reasonable in practical application. The combination of the above technologies has led to improvements in most domains.

Table 2. The speaker error of each domain in the second DIHARD development set.

Data domains	Original	Baseline	PRM1	SNR preselection
LIBRIVOX	0.63	**1.09**	**0.82**	0.63
YOUTHPOINT	1.70	1.61	1.45	1.16
SEEDLINGS	30.09	28.83	27.00	26.90
ADOS	21.23	14.02	13.99	13.99
SCOTUS	5.24	3.67	3.66	3.78
DCIEM	4.04	**4.82**	**7.66**	4.04
RT04	12.80	10.37	11.28	11.28
CIR	27.93	**28.52**	27.86	27.86
SLX	7.55	**9.92**	5.29	5.51
MIXER6	5.74	**5.93**	3.28	3.28
VAST	20.56	19.58	16.38	17.32
Ave.	12.10	11.62	10.95	10.70

5. CONCLUSIONS AND FUTURE WORK

In this work, we propose a progressive multi-target network for single channel speech enhancement which jointly learns the progressive multiple targets PELPS and PRM. Through comprehensive experiments, PRM1 obtained from the shallowest target layer has the best performance in speaker diarization. A DNN SAD based SNR estimation is adopted to select recordings which need to be enhanced. Compared with original speech data, the proposed method effectively avoids performance loss and achieves consistent improvements under most conditions in the DIHARD corpora.

6. REFERENCES

[1] P. Garcia-Perera and et al., "Speaker detection in the wild: lessons learned from jsalt 2019," submitted to ICASSP 2020.

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

[3] 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.

[4] 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.

[5] K. Han and S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system." Interspeech, 2010.

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

[7] 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.

[8] 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.

[9] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, ICCV 2007. IEEE 11th International Conference on*. IEEE, pp. 1–8.

[10] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," *CRIM, Montreal, Technical Report*, 2008.

[11] 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.

[12] ——, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

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

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

[15] 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.

[16] ——, "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.

[17] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[18] 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.

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

[20] L. Sun, J. Du, C. Jiang, X. Zhang, S. He, B. Yin, and C.-H. Lee, "Speaker diarization with enhancing speech for the first dihard challenge," *Proc. Interspeech 2018*, pp. 2793–2797.

[21] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. molkov, O. Novotn, K. Vesel, O. Glembek, O. Plchot, L. Moner, and P. Matjka, "But system for dihard speech diarization challenge 2018," in *Proc. Interspeech 2018*, pp. 2798–2802.

[22] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge," in *Proc. Interspeech 2018*, pp. 2808–2812.

[23] 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.

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

[25] P. Papadopoulos, A. Tsiartas, J. Gibson, and S. Narayanan, "A supervised signal-to-noise ratio estimation of speech signals," in *IEEE International Conference on Acoustics*, 2014.

[26] 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.

[27] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[28] W. Dong and Z. Xuewei, "Thchs-30 : A free chinese speech corpus," 2015. [Online]. Available: http://arxiv.org/abs/1512.01882

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

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

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

[32] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," 2018.

[33] "Nist rich transcription evaluations." [Online]. Available: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

[34] 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)*. IEEE, 2017, pp. 136–140.