DNN TRAINING BASED ON CLASSIC GAIN FUNCTION FOR SINGLE-CHANNEL SPEECH ENHANCEMENT AND RECOGNITION

Yan-Hui Tu[1], Jun Du[1], Chin-Hui Lee[2] *

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China
[2]Georgia Institute of Technology, Atlanta, GA, USA
tuyanhui@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

For conventional single-channel speech enhancement based on noise power spectrum, the speech gain function, which suppresses background noise at each time-frequency bin, is calculated by prior signal-to-noise-ratio (SNR). Hence, accurate prior SNR estimation is paramount for successful noise suppression. Accordingly, we have proposed a single-channel approach to combine conventional and deep learning techniques for speech enhancement and automatic speech recognition (ASR) recently. However, the combination process is at the testing stage, which is time-consuming with a complicated procedure. In this study, the gain function of classic speech enhancement will be utilized to optimize the ideal ratio mask based deep neural network (DNN-IRM) at the training stage, denoted as GF-DNN-IRM. And at the testing stage, the estimated IRM by GF-DNN-IRM model is directly used to generate enhanced speech without involving the conventional speech enhancement process. In addition, DNNs with less parameters in the causal processing mode are also discussed. Experiments of the CHiME-4 challenge task show that our proposed algorithm can achieve a relative word error rate reduction of 6.57% on RealData test set comparing to unprocessed speech without acoustic model retraining in causal mode, while the traditional DNN-IRM method fails to improve ASR performance in this case.

***Index Terms*—** statistical speech enhancement, ideal ratio mask, deep learning, gain function, speech recognition

1. INTRODUCTION

Single channel speech enhancement is a widely researched problem in signal processing, which aims to suppress the background noise and interference from the observed noisy speech to improve the perceptual quality and the performance of automatic speech recognition (ASR) [1]. The problem of speech enhancement has been an attractive area of research in statistical signal processing for a rather long time, and short-time Fourier transform (STFT) based methods achieve relatively good performance in this field [2]. It is appropriate to further categorize this class of speech enhancement algorithms into the sub-categories of spectral subtraction [3], Wiener filtering [4], minimum mean-square error (MMSE) estimator [5], and the optimally modified log-spectral amplitude (OM-LSA) speech estimator [6]. The Wiener filtering category is restricted by linearity, and the spectral subtraction approaches are based on largely simplified mathematical expressions. For MMSE and OM-LSA, they are strictly optimal given a set of initial assumptions and optimality criteria, which can be classified into statistical approaches. These conventional methods are adaptive to the test signal, which is in general not robust enough in adverse environments, particularly when there are non-stationary noises.

Recently, a supervised learning framework has been proposed, where a deep neural network (DNN) is trained to map from input features to the output targets. In [7], a regression DNN is adopted using mapping-based method directly predicting the log-power spectra (LPS) of clean speech from LPS of the noisy speech. In [8], the new architecture with two outputs is proposed to estimate the target speech and interference simultaneously. In [9], the DNN is adopted to estimate the ideal masks including the ideal binary mask (IBM) [10] of one time-frequency (T-F) bin and ideal ratio mask (IRM) [11] of one T-F bin. And [9] also demonstrates that the IRM as the target leads to a better speech enhancement performance than IBM. The above methods are based on the DNN model using the context information. More complicated neural network architectures, such as convolutional neural network (CNN) [12] and long short-term memory (LSTM) based recurrent neural network (RNN) [13], with an expense of higher computational complexities and run-time latencies than the conventional DNN are applied to speech enhancement. In real applications, one key factor is to design a neural network architecture to simultaneously achieve good evaluation metrics and maintain a low-latency real-time capability.

In this study, the classic gain function of statistical speech enhancement is utilized to optimize the parameters of DNN based on IRM (DNN-IRM), denoted as GF-DNN-IRM. This work is comprehensively extended from our recent paper [14] with new contributions listed as follows. First, the combination process of conventional speech enhancement and deep learning methods at the testing stage has been transferred to the training stage, which simplifies the decoding process and reduces the computation complexity. Second, DNN with less parameters in the causal processing mode is also investigated. The experiments on the CHiME-4 RealData test set show that the proposed approach can achieve a relative word error rate (WER) reduction of 6.57% comparing to unprocessed speech without acoustic model retraining in the causal mode, while the traditional DNN-IRM method fails to improve ASR performance in this case.

The remainder of this paper is organized as follows. In Section 2, we present an overview of the related work. Section 3 gives detailed description of our proposed approach. Section 4 shows the ASR performance of our proposed approach on the CHiME-4 challenge. Finally, we summarize our findings in Section 5.

*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. This work was also funded by Tencent.

2. RELATED WORK

2.1. IMCRA approach

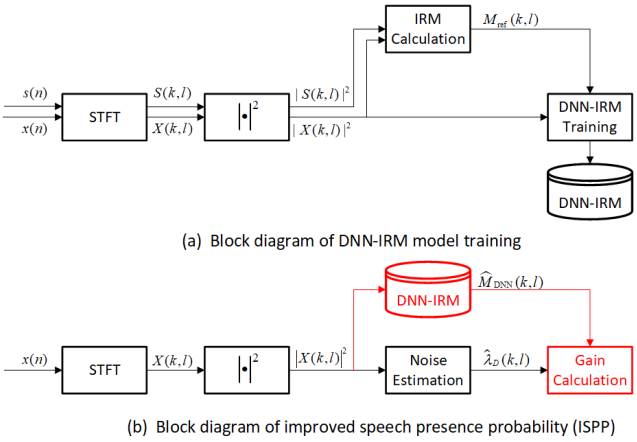

Fig. 1. Improved speech presence probability (ISPP) [14].

For the conventional speech enhancement methods, e.g, improved minima controlled recursive averaging (IMCRA) [15], the key point is the accurate estimation of the posterior signal-to-noise-ratio (SNR), denoted as $\gamma(k,l)$, and prior SNR, denoted as $\xi(k,l)$, which are defined as follows:

$$\xi(k,l) \triangleq \frac{\lambda_s(k,l)}{\lambda_d(k,l)} \tag{1}$$

$$\gamma(k,l) \triangleq \frac{|X(k,l)|^2}{\lambda_d(k,l)} \tag{2}$$

where $\lambda_s(k,l) = E[|S(k,l)|^2|H_1(k,l)]$ and $\lambda_d(k,l) = E[|D(k,l)]$ denote the variances of desired speech and noise at the T-F bin (k,l), respectively. $S(k,l)$, $D(k,l)$ and $X(k,l)$ denote the STFT of desired speech signal, noise signal and noisy speech signal, respectively. $H_1(k,l)$ indicates speech presence in the k-th frequency bin of the l-th frame. The prior SNR is estimated as follows:

$$\begin{aligned}\xi(k,l) = \ & \alpha G^2(k,l-1)\gamma(k,l-1) \\ & + (1-\alpha)\max\{\gamma(k,l-1)-1,0\}\end{aligned} \tag{3}$$

where α is a weighting factor that controls the tradeoff between noise reduction and speech distortion [5,16]. The suppression rule is based on the gain function of the prior and posterior SNRs:

$$G(k,l) = g(\gamma(k,l);\xi(k,l)) \tag{4}$$

More details can be found in [15].

2.2. Improved speech presence probability based approach

The conventional speech enhancement approach is adaptive to the test signal and with a relatively simple implementation, which is in general not robust enough in adverse environments, particularly when there are non-stationary noises. On the other hand, DNN-based regression model trained by a large amount of data can suppress non-stationary noises well, but its generalization ability is limited. When there exists mismatch between training data and test data, the robustness of DNN-based method is not good. Inspired by the above analysis, improved speech presence probability (ISPP) based approach was proposed in [14] for speech enhancement, as shown in Fig. 1.

3. THE PROPOSED APPROACH

3.1. DNN-based IRM estimation

The architecture of the DNN-based IRM estimation, which can be trained to learn the complex transformation from the noisy LPS features to the corresponding IRMs, denoted as DNN-IRM as shown in Fig. 1. Acoustic context information along both the time axis (with multiple neighboring frames) and frequency axis (with full frequency bins) can be fully exploited by the DNN to obtain a good mask estimate in adverse environments, which is strongly complementary with the conventional IMCRA-based approach to retain robustness. The estimated IRMs are restricted to be in the range between zero and one, which can be directly used to represent the speech presence probability at each T-F unit. The IRM as the learning target is defined as:

$$M_{\text{ref}}(k,l) = S_{\text{PS}}(k,l)/\left[S_{\text{PS}}(k,l) + D_{\text{PS}}(k,l)\right], \tag{5}$$

where $S_{\text{PS}}(k,l)$ and $D_{\text{PS}}(k,l)$ are clean and noise versions of power spectral features at the T-F unit (k,l). Because the training of this DNN-IRM model requires a large amount of time-synchronized stereo-data with the IRM and LPS of enhanced training data pairs, the training data are synthesized by adding different types of noise signals to the clean speech utterances with different SNR levels. Note that the specified SNR levels in the training stage are expected to address the problem of SNR variation in the testing stage with real speech data. To train the DNN-IRM model with a random initialization, supervised fine-tuning is used to minimize the mean squared error (MSE) between the DNN-IRM output $\hat{M}_{\text{DNN}}(k,l)$ and the reference IRM $M_{\text{ref}}(k,l)$, which is defined as

$$E_{\text{DNN}} = \sum_{k,l}(\hat{M}_{\text{DNN}}(k,l) - M_{\text{ref}}(k,l))^2. \tag{6}$$

This MSE is optimized using the stochastic gradient descent based back-propagation method in a mini-batch mode.

3.2. ISPP estimation using DNN-IRM

The conventional speech enhancement also retains much background noise and speech in adverse environments simultaneously. The estimated IRM based on DNN-IRM can give a realizable estimation to non-speech segmentation, but also exits mismatch between the training and testing data, which will destroy the target speech. The two approaches are complementary, and the estimated $\hat{M}_{\text{DNN}}(k,l)$ can be combined with $G(k,l)$ to yield an improved mask $\hat{M}_{\text{ISPP}}(k,l)$, i.e.,

$$\hat{M}_{\text{ISPP}}(k,l) = \delta\hat{M}_{\text{DNN}}(k,l) + (1-\delta)G(k,l) \tag{7}$$

where δ was set to 0.5.

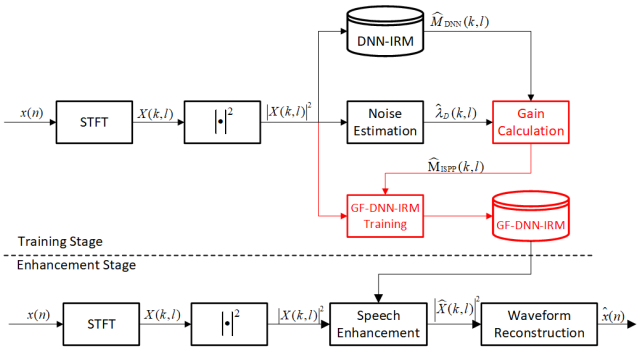

Fig. 2. The proposed GF-DNN-IRM approach.

3.3. DNN training based on classic gain function

In Section 3.2, we give a simple description of the proposed ISPP approach [14] at the testing stage. And in this section, the combination of two types of approaches is directly utilized to optimize the IRM estimation based on DNN at the training stage. In this study, DNN-IRM based on classic gain function training is illustrated in Fig. 2. [17] also demonstrated that the directly mapping from the noisy features to clean features for DNN regression model is not the best strategy. For the conventional DNN-IRM training, the reference IRM is obtained according to Eq. (5), which is based on the clean speech. In this study, the estimated ISPP features by gain function are directly adopted as the learning target to improve the generalization of DNN model.

To train the GF-DNN-IRM model with a random initialization, supervised fine-tuning is used to minimize MSE between the DNN-IRM output $\hat{M}_{\text{GF-DNN}}(k,l)$ and the ISPP $\hat{M}_{\text{ISPP}}(k,l)$, which is defined as

$$E_{\text{GF-DNN}} = \sum_{k,l} (\hat{M}_{\text{GF-DNN}}(k,l) - \hat{M}_{\text{ISPP}}(k,l))^2. \quad (8)$$

This MSE is also optimized using the stochastic gradient descent based back-propagation method in a mini-batch mode. Please note that $\hat{M}_{\text{ISPP}}(k,l)$ is calculated according to Eq. (7) as illustrated in Fig. 2. One advantage of GF-DNN-IRM model training is the stereo-data pairs is not necessary with the well-trained DNN-IRM model. Our proposed training procedure of GF-DNN-IRM is summarized as in Algorithm 1:

Algorithm 1 DNN training based on gain function (GF-DNN-IRM)

Input: Noisy training data and well-trained DNN-IRM model
Output: The parameter set of GF-DNN-IRM model

```
for each mini-batch do
    for each T-F bin (k,l) in one mini-batch do
        Compute a posterior SNR γ(k,l) using Eq. (2), a prior
        SNR ξ(k,l) using Eq. (1), and obtain G(k,l) via Eq. (4).
        Compute M̂_DNN(k,l) using DNN-IRM model.
        Compute the learning target of GF-DNN-IRM model
        M̂_ISPP(k,l) with M̂_DNN(k,l) and G(k,l) using Eq. (7).
        Accumulate the corresponding gradients based on Eq. (8).
    end for
    Update GF-DNN-IRM using stochastic gradient descent.
end for
```

4. EXPERIMENTAL EVALUATION

4.1. Data corpus

We present the experimental evaluation of our framework in the CHiME-4 task [18], which was designed to study real-world ASR scenarios where a person is talking to a mobile tablet device equipped with 6 microphones in a variety of adverse environments. Four conditions were selected: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). For each case, two types of noisy speech data were provided: RealData and SimData. RealData was collected from talkers reading the same sentences from the WSJ0 corpus [19] in the four conditions. SimData, on the other hand, was constructed by mixing clean utterances with environmental noise recordings using the techniques described in [20]. CHiME-4 offers three tasks (1-channel, 2-channel, and 6-channel) with different testing scenarios. In this paper, we focus only on the 1-channel case to make the paper concise. The readers can refer to [18] for more detailed information regarding to CHiME-4.

4.2. Implementation details

For front-end configurations, waveform was sampled at 16 kHz, and the corresponding frame length was set to 512 samples (or 32 msec) with a frame shift of 128 samples. A short-time Fourier transform (STFT) analysis was used to compute the DFT of each overlapping windowed frame. To train the DNN-IRM model, the 257-dimensional feature vector was used for IRM target. The PyTorch was used for neural network training [21]. The learning rate for the first 15 epochs was initialized as 0.01, then reduced to 0.001 for the last 15 epochs. To build the training data, clean speech was derived from the WSJ0 corpus [19], and the 4 noise types provided by CHiME-4 in [22] were selected as our noise database. 7138 utterances (about 12 hours of reading style speech) from 83 speakers, were corrupted with the above mentioned 4 noise types at three SNR levels (-5dB, 0dB and 5dB) to build a 36-hour training set, consisting of pairs of clean speech and noisy speech utterances.

For the back-end configurations, the baseline ASR recognition system was trained on the speech recognition toolkit Kaldi [23]. For time delay neural network (TDNN) based acoustic model training, backstitch optimization method was used. The decoding was based on 3-gram language models with explicit pronunciation and silence probability modeling. The model was re-scored by a 5-gram language model first. Then the Kaldi-RNNLM was used for training the RNN language model, and n-best re-scoring was used to improve performance. The model was trained according to the scripts downloaded from the official GitHub website[1].

4.3. Experimental results

Table 1 shows the comparison of GF-DNN-IRM using different settings of (τ, N_L, N_U) on the test sets of RealData. τ, N_L, N_U denoted the number of frames in the input layer, the number of hidden layers and the number of hidden units, respectively, and there are three blocks in Table 1. The results in Table 1 were obtained by the Kaldi tools without acoustic model retraining.

For the first block of Table 1, "Noisy" denoted the original noisy speech randomly selected from channel 1-6 (except channel 2), namely 1-channel case. "Baseline" denoted the method proved by Kaldi tools and the enhanced speech was obtained by BLSTM-IRM model [24]. "IMCRA" denoted the enhanced speech was

[1]https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4

Table 1. WER (%) comparison of GF-DNN-IRM models using different settings of (τ, N_L, N_U) on the test sets of RealData. τ, N_L, N_U denoted the number of frames in the input layer, the number of hidden layers and the number of hidden units, respectively.

Enhancement	(τ, N_L, N_U)	BUS	CAF	PED	STR	AVG
Noisy		21.03	13.90	11.29	9.21	13.85
Baseline [24]		32.34	24.08	18.91	14.57	22.47
IMCRA [15]		26.68	18.96	13.77	10.55	17.49
DNN-IRM	(1,3,2048)	23.67	18.58	14.44	10.61	16.82
	(5,3,2048)	23.41	18.71	14.37	10.29	16.69
	(7,3,2048)	23.54	17.48	13.32	10.16	16.12
GF-DNN-IRM	(1,3,2048)	19.26	13.26	10.76	8.48	12.94
	(5,3,2048)	19.36	13.12	10.65	8.35	12.87
	(7,3,2048)	19.50	13.00	10.46	8.24	12.80

obtained by IMCRA-based speech enhancement [15]. We could observe that the IRM estimated by "Baseline" significantly degraded the ASR performance, comparing to "Noisy". For example, the average WER of "Noisy" was 13.85%, while the average WER of "Baseline" was 22.47%, respectively. "IMCRA", namely classic speech enhancement approach, also failed to improve the ASR performance.

For the second block of Table 1, "DNN-IRM" denoted the enhanced speech was obtained by the estimated IRM using different DNN settings of (τ, N_L, N_U). The setting of input frame number τ as the acoustic context determined the hard latency of deep models. τ=1 denoted the causal mode where only the central frame was adopted with no hard latency. τ=5 and 7 employed 2 and 3 history/future frames respectively. For DNN-IRM model, the acoustic context was quite important for recognition performance. For example, "DNN-IRM(7,3,2048)" outperformed "DNN-IRM(1,3,2048)", a relative WER reduction of 4.16%. But all DNN-IRM results were still worse than those of "Noisy".

For the last block, "GF-DNN-IRM" denoted the enhanced speech was obtained by the estimated IRM using GF-DNN-IRM models. We could observe that the IRM estimated by GF-DNN-IRM model could directly improve the ASR performance without acoustic model retraining. For example, "GF-DNN-IRM(7,3,2048)" improved the ASR performance with a relative WER reduction of 7.58%, comparing to "Noisy". In the causal mode, our proposed algorithm, "GF-DNN-IRM(1,3,2048)", achieved a relative WER reduction of 6.57%, comparing to "Noisy". Finally, the performance gaps among GF-DNN-IRM models with different architectures were smaller than those among DNN-IRM models. For example, a relative WER reduction of 4.16% was yielded from "DNN-IRM(1,3,2048)" to "DNN-IRM(7,3,2048)" while only a relative WER reduction of 1.08% was generated from "GF-DNN-IRM(1,3,2048)" to "GF-DNN-IRM(7,3,2048)".

Fig. 3 gives an utterance example from the RealData test set of CHiME-4 to illustrate the motivation of using classic single-channel speech enhancement algorithm to optimize the DNN-based IRM estimation at the training stage. Fig. 3 a) and b) plot the spectrograms from channel 0 (the close-talking microphone to record the reference "clean" speech) and one corresponding channel with noisy speech. Fig. 3 c) and d) plot the IRMs estimated by IMCRA and DNN-IRM methods. Comparing these two plots, we observed that the estimated IRM by the DNN model might misclassify the T-F regions dominated by speech to non-speech/noise, while the estimated IRM by the IMCRA method could alleviate this problem, where the values of IRM estimated by the IMCRA method were much higher than those of DNN model at the circled region marked in the black rectangle. But we also could find that the values of IRM estimated by the IMCRA method were noncontinuous among neighbouring speech frames. Finally, in Fig. 3 e), the IRMs estimated by GF-DNN-IRM model could fully utilize the complementarity of IMCRA and DNN-IRM based approaches, namely preserving better speech regions than DNN-IRM and yielding better continuity than IMCRA.

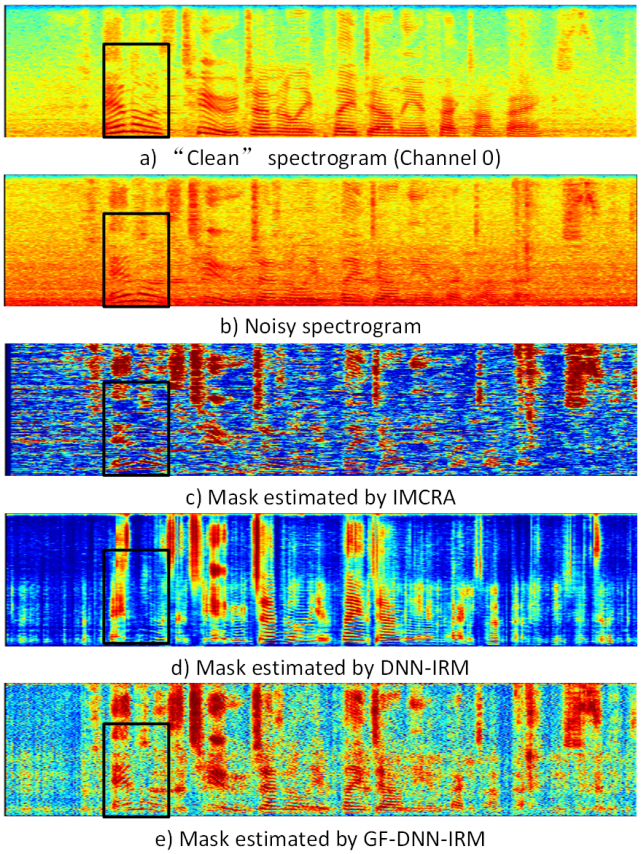

Fig. 3. The comparison of estimated masks from different approaches for an utterance of CHiME-4 RealData test set.

5. CONCLUSION

In this work we proposed a novel architecture for single-channel speech enhancement utilizing the gain function of classic speech enhancement to guide the DNN-IRM training, denoted as GF-DNN-IRM. As evaluation parameters we used speech recognizer performance, under the assumption that the speech recognizer is a black box without retraining of acoustic model. Experiments demonstrate that the IRM estimated by the DNN-IRM model is not effective for ASR performance, while the IRM estimated by GF-DNN-IRM model can directly improve the ASR performance without acoustic model retraining. For example, "GF-DNN-IRM(7,3,2048)" (WER of 12.80%) can improve the ASR performance with a relative WER reduction of 7.58%, compared to "Noisy" (WER of 13.85%), while "DNN-IRM(7,3,2048)" are 16.82%. In the future, we will explore how to utilize more powerful neural network to optimize the training of neural networks with less parameters.

6. REFERENCES

[1] Li Deng and Xiao Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[2] Philipos C. Loizou, *Speech Enhancement : Theory and Practice*, CRC press, second edition, 2013.

[3] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] Jae S. Lim and Alan V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[5] Yariv Ephraim and David 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.

[6] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[7] Yong Xu, Jun Du, Lirong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[8] Yanhui Tu, Jun Du, Yong Xu, Lirong Dai, and Chin-Hui Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *International Symposium on Chinese Spoken Language Processing.(ISCSLP)*, 2014.

[9] Yuxuan Wang and DeLiang Wang, "Towards scaling up classification-based speech separation," *Trans. Audio, Speech and Lang. Proc.*, vol. 21, no. 7, pp. 1381–1390, July 2013.

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[11] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[12] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement.," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2016.

[13] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Bjorn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[14] Yan-Hui Tu, Ivan Tashev, Shuayb Zarar, and Chin-Hui Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *international conference on acoustics, speech, and signal processing.(ICASSP)*, 2018, pp. 2531–2535.

[15] Israel Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[16] Olivier Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[17] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Snr-based progressive learning of deep neural network for speech enhancement.," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2016.

[18] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, 2016.

[19] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csri (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[20] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933—1950, 2007.

[21] Nikhil Ketkar, *Deep Learning with Python: A Hands-on Introduction*, Apress, Berkeley, CA, 2017.

[22] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[24] Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, and Shinji Watanabe, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2018.