Interspeech 2018
2-6 September 2018, Hyderabad

A Two-Stage Approach to Noisy Cochannel Speech Separation with Gated Residual Networks

Ke Tan[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

tan.650@osu.edu, wang.77@osu.edu

Abstract

Cochannel speech separation is the task of separating two speech signals from a single mixture. The task becomes even more challenging if the speech mixture is further corrupted by background noise. In this study, we focus on a gender-dependent scenario, where target speech is from a male speaker and interfering speech from a female speaker. We propose a two-stage separation strategy to address this problem in a noise-independent way. In the proposed system, denoising and cochannel separation are performed successively by two modules, which are based on a newly-introduced convolutional neural network for speech separation. The evaluation results demonstrate that the proposed system substantially outperforms one-stage baselines in terms of objective intelligibility and perceptual quality.

Index Terms: noisy cochannel speech separation, gated residual networks, ideal ratio mask, denoising, cochannel separation

1. Introduction

Cochannel speech separation aims to separate the speech of interest (or target speech) from interfering speech [1]. This difficult problem becomes more challenging when cochannel speech is further corrupted by background noise. Applications such as hearing aids and automatic speech recognition (ASR) suffer from severe performance degradation under such real-world conditions. We refer to *noisy cochannel speech separation* as the task of separating target speech from both interfering speech and background noise.

Cochannel speech separation can be formulated as a supervised learning problem, where a mapping from acoustic features of cochannel speech to a time-frequency (T-F) mask or spectral magnitudes of target speech is learned. Huang *et al.* [2] first introduced deep neural networks (DNNs) to deal with cochannel separation. In their method, a masking layer is added to the network, which produces the spectra of the two estimated sources. Du *et al.* [3] proposed a DNN to estimate the log power spectrum of target speech from that of a cochannel mixture. Subsequently, they trained a DNN to map a cochannel mixture to the spectrum of the target speaker as well as that of an interfering speaker [4] [5]. Different from [2], they addressed a more complex situation, where the interfering speakers are different between training and test although the same target speaker is used for both training and test. More recently, Zhang *et al.* [6] developed a deep ensemble network to address speaker-dependent and target-dependent separation. In their study, multi-context networks were employed to integrate temporal information at different resolutions. Specifically, they constructed an ensemble by stacking multiple modules, each of which performs multi-context masking or mapping. Moreover, Healy *et al.* [7] utilized a DNN to deal with speaker-dependent cochannel separation. The DNN was trained to estimate the ideal ratio mask (IRM) for a male target speaker in the presence of a female interfering speaker. They found that the trained DNN provided substantial speech intelligibility improvements for hearing-impaired listeners.

In this study, we investigate supervised noisy cochannel speech separation in a gender-dependent scenario, where target speech is from a male speaker and interfering speech from a female speaker. Inspired by recent research [8] on noisy and reverberant speech enhancement, we believe that it is likely more effective to address denoising and cochannel separation in separate stages. In other words, we first separate cochannel speech and background noise, and then perform cochannel separation to reconstruct the time-domain waveforms of the two sources. We remove background noise in the first stage as the properties of speech and nonspeech noise are intrinsically different.

Additionally, motivated by our recent work on dilated convolutions and gating mechanisms [9], we propose to employ a gated residual network (GRN) with dilated convolutions to construct the two-stage system with utterance-level training. To compare with alternative modeling, we build two one-stage systems with the same GRN as the baselines, where denoising and cochannel separation are addressed simultaneously. These systems are evaluated on both trained speakers and untrained s-peakers in a noise-independent scenario. We find that the two-stage system consistently outperforms the one-stage baselines in terms of objective speech intelligibility and quality.

The rest of this paper is organized as follows. In Section 2, we describe our proposed algorithm in detail. The experimental setup and results are presented in Section 3. We conclude this paper in Section 4.

2. Algorithm description

Our proposed two-stage separation system comprises two modules, i.e. a denoising module and a cochannel separation module. A 62-layer GRN proposed in [9] is adopted to build the modules.

2.1. Problem formulation

Let $s_1(t)$, $s_2(t)$ and $n(t)$ denote target speech, interfering speech, and background noise, respectively. Then the noisy speech mixture can be represented by

$$y(t) = s(t) + n(t) = s_1(t) + s_2(t) + n(t) \tag{1}$$

where $s(t) = s_1(t) + s_2(t)$ denotes cochannel speech. Given $y(t)$, the goal of noisy cochannel speech separation is to recover $s_1(t)$ and $s_2(t)$. In this study, the energy of target speech and interfering speech is equally strong. In other words, the target-

10.21437/Interspeech.2018-1406

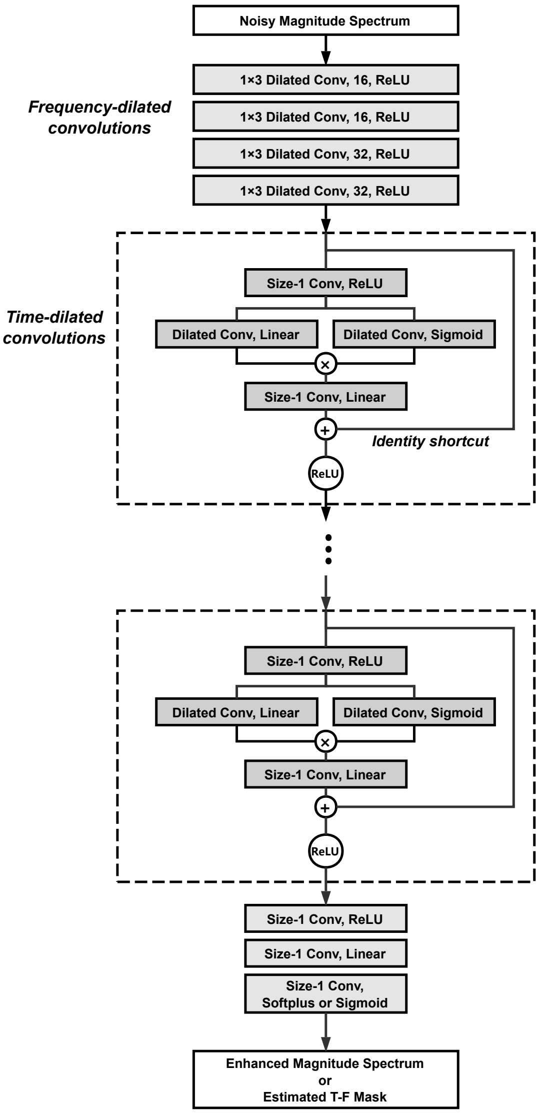

Noisy Magnitude Spectrum
Frequency-dilated convolutions
1×3 Dilated Conv, 16, ReLU
1×3 Dilated Conv, 16, ReLU
1×3 Dilated Conv, 32, ReLU
1×3 Dilated Conv, 32, ReLU
Time-dilated convolutions
Size-1 Conv, ReLU
Dilated Conv, Linear
Dilated Conv, Sigmoid
Size-1 Conv, Linear
Identity shortcut
ReLU
Size-1 Conv, ReLU
Dilated Conv, Linear
Dilated Conv, Sigmoid
Size-1 Conv, Linear
ReLU
Size-1 Conv, ReLU
Size-1 Conv, Linear
Size-1 Conv, Softplus or Sigmoid
Enhanced Magnitude Spectrum or Estimated T-F Mask

Figure 1: *Network architecture of the GRN that was developed in [9].*

to-interferer ratio (TIR) is 0 dB, where the TIR is defined by

$$TIR = 10 \log \frac{\sum_t s_1^2(t)}{\sum_t s_2^2(t)} \tag{2}$$

To avoid potential confusion, we refer to speech-to-noise ratio (SNR) as the measure that compares the level of cochannel speech to that of background noise. It is calculated as

$$SNR = 10 \log \frac{\sum_t [s_1(t) + s_2(t)]^2}{\sum_t n^2(t)} \tag{3}$$

2.2. Gated residual network

In this study, we use the gated residual network in [9] to construct the separation system. The GRN is based on dilated convolutions, which can significantly expand receptive fields. It additionally incorporates gated linear units and residual learning. Fig. 1 depicts the GRN architecture. The patterns along the frequency direction in the input magnitude spectrum are captured by frequency-dilated convolutions. Subsequently, a bunch of residual blocks are employed to perform time-dilated convolutions, which systematically aggregate temporal contexts. The high-level features learned by these residual blocks are then fed into a few convolutional layers with size-1 kernels to predict the target.

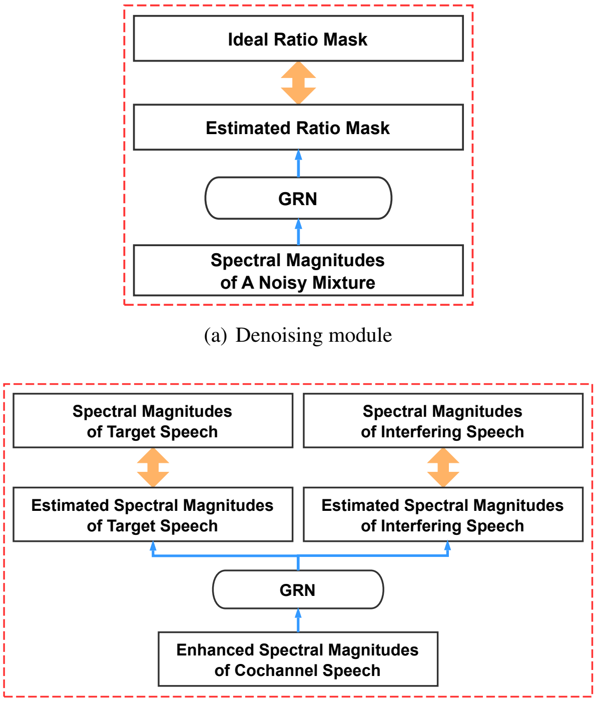
2.3. Denoising stage

Given a noisy speech mixture, the denoising module aims to separate cochannel speech from background noise. A T-F mask typically serves as the training target for noise suppression in supervised speech separation [10]. During inference, the estimated T-F mask is applied to the T-F representation of the noisy mixture to derive that of the enhanced speech. The enhanced T-F representation is subsequently used to reconstruct the time-domain waveform of the enhanced speech. The IRM [11] is a frequently used T-F mask:

$$IRM(m, f) = \sqrt{\frac{S(m, f)^2}{S(m, f)^2 + N(m, f)^2}} \tag{4}$$

where $S(m, f)^2$ and $N(m, f)^2$ represent speech energy and noise energy within a T-F unit at time frame m and frequency channel f, respectively.

As Fig. 2(a) illustrates, the denoising module in our proposed system employs a GRN to predict the IRM for noise reduction. The estimated ratio mask is subsequently applied to the spectral magnitudes of a noisy mixture to obtain the enhanced spectral magnitudes for the next stage processing.


Ideal Ratio Mask
Estimated Ratio Mask
GRN
Spectral Magnitudes of A Noisy Mixture
Spectral Magnitudes of Target Speech
Spectral Magnitudes of Interfering Speech
Estimated Spectral Magnitudes of Target Speech
Estimated Spectral Magnitudes of Interfering Speech
GRN
Enhanced Spectral Magnitudes of Cochannel Speech

(a) Denoising module

(b) Cochannel separation module

Figure 2: *Diagrams of the denoising module and the channel separation module.*

2.4. Cochannel separation stage

Once background noise is removed from a noisy mixture, one can perform cochannel speech separation to recover the two speech sources. In our proposed system, the cochannel separation module is responsible for this task. Rather than using the masking-based methods, we utilize a GRN to learn a mapping from the enhanced spectral magnitudes of cochannel speech to the spectral magnitudes of target speech and interfering speech. Different from previous works [3] on cochannel separation, the GRN predicts both the target and the interference in the output layer, following a configuration analogous to [4] (see also Fig. 2(b)). During training, we jointly minimize the mean squared error (MSE) between the dual outputs of the GRN and the corresponding references:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(\|\hat{\mathbf{S}}_n^t - \mathbf{S}_n^t\|_2^2 + \|\hat{\mathbf{S}}_n^i - \mathbf{S}_n^i\|_2^2) \tag{5}$$

where $\hat{\mathbf{S}}_n^t$, $\hat{\mathbf{S}}_n^i$, $\mathbf{S}_n^t$ and $\mathbf{S}_n^i$ denote the n-th estimated magnitude spectra of the target and the interference, and the n-th reference magnitude spectra of the target and the interference, respectively. N represents the number of the training samples. The softplus activation function [12] is applied to the output layer to fit the value range of spectral magnitudes.

2.5. Joint training

When the two modules are well trained separately, we concatenate them into an integrated network for joint optimization. Specifically, the output of the denoising module, i.e. the estimated ratio mask, is applied to the spectral magnitudes of a noisy mixture. The enhanced spectral magnitudes are subsequently fed into the cochannel separation module. A batch normalization [13] layer without learnable affine parameters is inserted between the modules. Fig. 3 shows the diagram of the integrated two-stage separation system.

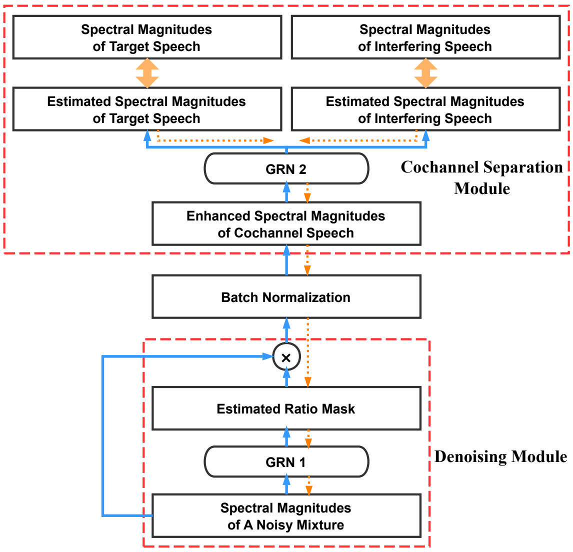

Spectral Magnitudes of Target Speech
Spectral Magnitudes of Interfering Speech
Estimated Spectral Magnitudes of Target Speech
Estimated Spectral Magnitudes of Interfering Speech
GRN 2
Cochannel Separation Module
Enhanced Spectral Magnitudes of Cochannel Speech
Batch Normalization
×
Estimated Ratio Mask
GRN 1
Denoising Module
Spectral Magnitudes of A Noisy Mixture

Figure 3: *Diagram of the two-stage system for noisy cochannel speech separation.*

2.6. One-stage baselines

We build two one-stage baseline systems for comparison, which learn mapping-based and masking-based targets [10], respectively. The mapping-based system directly predicts the spectral magnitudes of target speech and interfering speech, while the masking-based system predicts the IRMs instead:

$$IRM^t(m,f) = \sqrt{\frac{S_1(m,f)^2}{S_1(m,f)^2 + N_1(m,f)^2}} \tag{6}$$

$$IRM^i(m,f) = \sqrt{\frac{S_2(m,f)^2}{S_2(m,f)^2 + N_2(m,f)^2}} \tag{7}$$

where IRM^t and IRM^i represent the IRMs accounting for target speech and interfering speech, respectively. $S_1(m,f)^2$ and $S_2(m,f)^2$ denote the energy of target speech, $s_1(t)$, and interfering speech, $s_2(t)$, within a T-F unit at time frame m and frequency channel f, respectively. $N_1(m,f)^2$ and $N_2(m,f)^2$ denote the energy of $n_1(t) = s_2(t) + n(t)$ and $n_2(t) = s_1(t) + n(t)$ within a T-F unit at time frame m and frequency channel f, respectively. Note that $n(t)$ represents background noise. As in the cochannel separation module, both baselines are constructed with a single GRN which has dual outputs corresponding to target speech and interfering speech, respectively.

3. Experiments

3.1. Experimental setup

In our experiments, the proposed two-stage system and the one-stage baselines are evaluated on WSJ0 SI-84 dataset [14] including 7138 utterances from 83 speakers. Among these speakers, 6 speakers (3 males and 3 females) are regarded as untrained speakers. Hence, 6385 utterances from the rest of the speakers, including 39 male speakers and 38 female speakers, are utilized to create the training mixtures. To derive noise-independent models, we use 10,000 noises from a sound effect library (available at http://www.sound-ideas.com) for the training set and two challenging noises (babble and factory) from the NOISEX-92 dataset [15] for the test sets.

We create 160,000 noisy mixtures for training. To create a training mixture, we mix a randomly chosen training utterance from a male speaker, a randomly chosen training utterance from a female speaker, and a random cut from the 10,000 training noises at a SNR level randomly selected from {-5, -4, -3, -2, -1, 0} dB. The two speakers are randomly drawn as well. Note that we refer to the SNR as the speech-to-noise ratio here.

For test, we use two SNR levels, i.e. -5 dB and -2 dB. To create a test mixture, we mix a pair of randomly selected utterances from a male speaker and a female speaker with a random cut from the test noise. The speakers are randomly drawn from 6 test speakers (3 males and 3 females). Specifically, we create two test sets for each noise at each SNR level:

- Test Set 1: we create 200 mixtures from utterances of 6 trained speakers (3 males and 3 females).
- Test Set 2: we create 200 mixtures from utterances of 6 untrained speakers (3 males and 3 females).

Note that the lengths of a target utterance and an interfering utterance may differ from each other. In this study, we render the length of their mixture equal to the length of the target utterance. In other words, we truncate the interfering utterance if it is longer than the target utterance; otherwise, we pad the interfering utterance by repeating it.

Table 1: *STOI and PESQ scores on trained speakers.*

metrics	STOI (in %)									PESQ								
noises	Avg.	babble				factory				Avg.	babble				factory			
SNR		-5 dB		-2 dB		-5 dB		-2 dB			-5 dB		-2 dB		-5 dB		-2 dB	
speaker		s_1	s_2	s_1	s_2	s_1	s_2	s_1	s_2		s_1	s_2	s_1	s_2	s_1	s_2	s_1	s_2
unprocessed	51.3	51.1	47.3	55.9	52.0	50.5	46.7	56.0	50.9	1.45	1.63	1.35	1.66	1.34	1.48	1.26	1.55	1.31
masking	69.4	67.0	65.1	73.7	71.6	68.4	63.7	75.2	70.7	1.84	1.79	1.60	2.00	1.87	1.85	1.65	2.08	1.91
mapping	69.8	67.0	64.6	74.5	71.1	69.4	64.9	76.2	70.9	1.91	1.83	1.65	2.06	1.92	1.94	1.78	2.14	1.99
two-stage	**72.4**	**69.7**	**68.2**	**76.7**	**74.2**	**71.6**	**67.2**	**78.3**	**73.6**	**2.01**	**1.90**	**1.79**	**2.13**	**2.05**	**2.01**	**1.88**	**2.23**	**2.10**

Table 2: *STOI and PESQ scores on untrained speakers.*

metrics	STOI (in %)									PESQ								
noises	Avg.	babble				factory				Avg.	babble				factory			
SNR		-5 dB		-2 dB		-5 dB		-2 dB			-5 dB		-2 dB		-5 dB		-2 dB	
speaker		s_1	s_2	s_1	s_2	s_1	s_2	s_1	s_2		s_1	s_2	s_1	s_2	s_1	s_2	s_1	s_2
unprocessed	51.9	50.9	48.6	55.9	52.5	50.5	47.6	56.4	52.6	1.41	1.49	1.39	1.58	1.35	1.43	1.22	1.50	1.28
masking	68.1	64.3	63.0	72.4	69.4	67.1	62.9	75.0	70.8	1.73	1.63	1.51	1.87	1.75	1.72	1.58	1.96	1.84
mapping	69.3	64.2	63.5	73.2	70.5	68.3	65.7	76.2	72.9	1.83	1.70	1.56	1.95	1.84	1.84	1.75	2.06	1.99
two-stage	**71.8**	**67.0**	**66.9**	**75.4**	**73.3**	**70.8**	**67.8**	**78.3**	**75.1**	**1.93**	**1.75**	**1.73**	**2.01**	**1.98**	**1.89**	**1.85**	**2.13**	**2.08**

In our experiments, all signals are sampled at 16 kHz. A 20-ms Hamming window is employed to segment a signal into a set of time frames. Adjacent time frames are overlapped by 50%. The feature representation and the IRMs are based on 161-dimensional short-time Fourier transform (STFT) magnitude spectra, calculated from a 320-point STFT. During training, we use Adam [16] as the optimizer. The learning rate is set to 0.001. The models are trained with the MSE objective function and a mini-batch size of 16. Within a mini-batch, all samples are padded with zeros to have the same number of time frames as the longest sample does. For all models, the network inputs are normalized to zero mean and unit variance. During inference, the enhanced STFT magnitude spectra and noisy mixture phases are passed into a resynthesizer to derive time-domain waveforms of target speech and interfering speech.

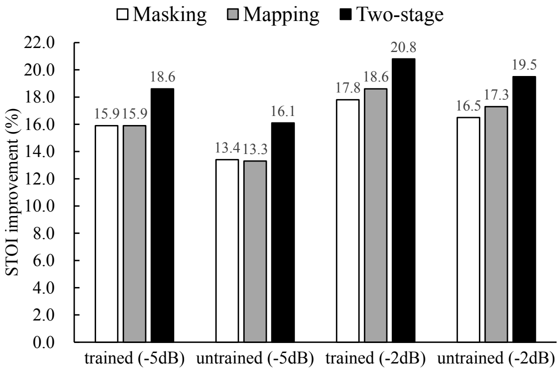


Figure 4: *Comparison of noisy cochannel speech separation systems in terms of STOI improvements for target speech on the untrained babble noise. The STOI improvements are calculated as* $\Delta STOI(\%) = 100 \times (STOI_{processed} - STOI_{unprocessed})$.

3.2. Experimental results

In this study, we use two objective metrics, i.e. short-time objective intelligibility (STOI) [17] and perceptual evaluation of speech quality (PESQ) [18], to evaluate objective speech intelligibility and quality, respectively.

The STOI and PESQ scores on trained speakers and untrained speakers are presented in Table 1 and Table 2, respectively. The best results in each case are highlighted by boldface numbers. In both tables, we use "masking" to indicate the masking-based baseline and "mapping" to indicate the mapping-based baseline. Additionally, target speech (male) and interfering speech (female) are denoted as $\mathbf{s_1}$ and $\mathbf{s_2}$, respectively.

Generally, regardless of the system of choice, the GRN proposed by [9] provides substantial improvements in terms of both STOI and PESQ scores over the unprocessed mixtures in a noise-independent scenario. As shown in Table 1 and Table 2, the STOI scores achieved by the two baselines are close on trained speakers, while the mapping-based baseline yields more than 1% STOI improvements on untrained speakers compared to the masking-based baseline. This indicates that the mapping-based baseline generalizes better to untrained speakers than the masking-based baseline. With the proposed two-stage framework, the STOI scores further improve by more than 2.5% on both trained speakers and untrained speakers. With regard to speech quality, a 0.1 PESQ improvement over the mapping-based baseline is achieved by the two-stage system.

The STOI improvements for target speech on the babble noise are shown in Fig. 4. Overall, the proposed two-stage system consistently outperforms the two one-stage baselines in terms of STOI improvements. In the most challenging case, where cochannel speech from untrained speakers is mixed with the babble noise at -5 dB, the one-stage baselines leads to an around 13.3% STOI improvement over the unprocessed mixtures. The proposed two-stage system, however, yields a 16.1% STOI improvement, which is substantially better than the one-stage baselines.

4. Conclusions

In this study, we have proposed a two-stage system to deal with noisy cochannel speech separation. In the proposed system, we use two successive modules to perform denoising and cochannel separation separately, and subsequently integrate them for joint optimization. A newly-introduced network for speech separation, named GRN, is employed to construct the modules. Our experimental results indicate that the proposed system consistently outperforms the one-stage baselines for both trained speakers and untrained speakers. In future research, we will extend the present work to the gender-independent scenario.

5. Acknowledgements

The authors would like to thank Masood Delfarah for providing his DNN code. This research was supported in part by an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

6. References

[1] K. Hu and D. L. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 122–131, 2013.

[2] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[3] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *12th International Conference on Signal Processing (ICSP)*, 2014, pp. 473–477.

[4] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 250–254.

[5] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2016.

[6] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 5, pp. 967–977, 2016.

[7] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. L. Wang, "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4230–4239, 2017.

[8] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5580–5584.

[9] K. Tan, J. Chen, and D. L. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, to appear.

[10] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *arXiv preprint arXiv:1708.07524*, 2017.

[11] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[12] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, "Improving deep neural networks using softplus units," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–4.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[14] D. B. Paul and J. M. 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.

[15] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.