INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Bridging the Gap Between Monaural Speech Enhancement and Recognition with Distortion-Independent Acoustic Modeling

Peidong Wang[1], *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

{wang.7642, tan.650, wang.77}@osu.edu

Abstract

Monaural speech enhancement has made dramatic advances in recent years. Although enhanced speech has been demonstrated to have better intelligibility and quality for human listeners, feeding it directly to automatic speech recognition (ASR) systems trained with noisy speech has not produced expected improvements in ASR performance. The lack of an enhancement benefit on recognition, or the gap between monaural speech enhancement and recognition, is often attributed to speech distortions introduced in the enhancement process. In this study, we analyze the distortion problem and propose a distortion-independent acoustic modeling scheme. Experimental results show that the distortion-independent acoustic model is able to overcome the distortion problem. Moreover, it can be used with various speech enhancement models. Both the distortion-independent and a noise-dependent acoustic model perform better than the previous best system on the CHiME-2 corpus. The noise-dependent acoustic model achieves a word error rate of 8.7%, outperforming the previous best result by 6.5% relatively.
Index Terms: robust automatic speech recognition, speech enhancement, speech distortion, distortion-independent acoustic modeling, CHiME-2

1. Introduction

Formulated as a supervised learning problem, speech enhancement has made major progress over the last few years with the use of data driven methods, particularly deep learning. Wang and Wang [1] first introduced deep neural networks (DNNs) to perform time-frequency (T-F) masking for speech enhancement. Lu *et al.* and Xu *et al.* proposed deep autoencoder (DAE) and DNN to map from the power spectrum of the noisy speech to that of the clean speech [2, 3], respectively. Many subsequent studies have been conducted to perform T-F masking or spectral mapping by employing a variety of deep learning models, acoustic features, and training targets [4, 5, 6, 7]. These studies have elevated the speech enhancement performance by a large margin [8]. DNN-based monaural speech enhancement has improved, for the first time, the intelligibility of noisy speech for human listeners with hearing impairment as well as with normal hearing [8, 9, 10].

Along with the progress in speech enhancement, researchers have investigated speech enhancement models as frontends for automatic speech recognition systems. Narayanan *et al.* [11, 12] proposed to combine masking-based DNN speech enhancement with speech recognition backends. In a subsequent paper using DNN as the backend, the benefit of the speech enhancement frontend is mixed, depending on training features [12]. For the acoustic model trained with cepstral features, the enhancement frontend still helps, but with log-mel features, the enhancement frontend causes performance degradation. Du *et al.* [13] applied mapping-based frontends to both GMM and DNN based recognition backends. Their observations are basically in line with those of Narayanan *et al.* The only difference is that their enhancement frontend can yield improvements on clean, noisy, and clean plus channel-mismatched conditions for the DNN acoustic model trained with noisy speech. In the fourth CHiME speech separation and recognition challenge (CHiME-4) [14], Heymann *et al.* [15] noted that the harm of processing artifacts introduced during enhancement may outweigh the benefit brought by noise reduction. Based on these studies as well as our own attempts in applying monaural speech enhancement as a frontend on the CHiME-4 task, the distortion to speech signals introduced in monaural speech enhancement is a major problem that can render enhancement useless or even harmful for robust ASR.

One way to alleviate the distortion problem is to design speech enhancement frontends that can reduce the distortion in enhanced speech. Attempts in this direction include a progressive training scheme proposed by Gao *et al.* [16], a mimic loss proposed by Bagchi *et al.* [17], and an acoustic-guided evaluation (AGE) metric proposed by Chai *et al.* [18].

In this study, we propose a distortion-independent acoustic model training scheme. It uses a large variety of enhanced speech to train the acoustic model. By using enhanced speech as training data, the distortion problem is alleviated. With the large-scale training strategy [19], the distortion-independent acoustic model is able to generalize to speech enhancement frontends not used during training. Experimental results show that the distortion-independent acoustic model is able to not only overcome the distortion problem but also generalize to various speech enhancement frontends. Both the distortion-independent acoustic model and a noise-dependent acoustic model perform better than the previous best system on the CHiME-2 corpus [20]. The noise-dependent model achieves a word error rate (WER) of 8.7%, outperforming the previous best system by 6.5% relatively.

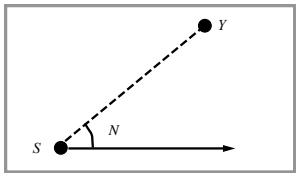
The rest of this paper is organized as follows. In Section 2, we analyze the distortion problem and describe the distortion-independent training scheme. Section 3 and 4 contain the experimental setup and results. Finally, we make a conclusion in Section 5.

2. System Description

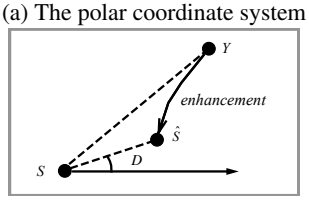
2.1. Analysis on the Distortion Problem

The distortion in this study refers to the alteration to clean speech signal introduced by speech enhancement that may cause performance degradation in an automatic speech recognition system. More specifically, this paper tackles with the dis-

http://dx.doi.org/10.21437/Interspeech.2019-1495



(a) The polar coordinate system



(b) Clean, noisy, and enhanced speech

Figure 1: *Illustration of the signal distortion problem. (a) The polar coordinate system. (b) Clean, noisy, and enhanced speech.*

tortion problem of noise-independent speech enhancement. The input to a speech enhancement systems is generated by mixing clean speech with an additive noise, as shown in equation (1):

$$Y = S + N \tag{1}$$

where Y, S, and N are the spectral representations of noisy speech, clean speech, and additive noise, respectively.

Speech enhancement for ASR typically operates on the magnitudes of frequency domain representations. Masking-based models generate a T-F mask, which is then element-wise multiplied with the magnitude of Y.

$$|\hat{S}| = |Y| \otimes M = |S + N| \otimes M \tag{2}$$

where $|\cdot|$ denotes magnitude, $\otimes$ element-wise multiplication, $\hat{S}$ enhanced speech, and M T-F mask.

Based on the definition of T-F masks, M is a real-valued matrix with element values ranging from zero to one. Equation (2) can thus be written as equation (3) below:

$$|\hat{S}| = |S + N| \otimes M = |S \otimes M + N \otimes M| \tag{3}$$

Rewriting equation (3) into the form similar to (1), we can get equation (4):

$$|\hat{S}| = |S + (S \otimes (M - A) + N \otimes M)| \tag{4}$$

where A is an all-one matrix.

The distortion D can thus be represented as equation (5):

$$D = N \otimes M - S \otimes (A - M) = N \otimes M - S \otimes \overline{M} \tag{5}$$

where $\overline{M}$ denotes the compliment of M.

Because of the element-wise multiplication with M and the subtraction with the second term in equation (5), the distortion D may deviate from noise N.

Fig. 1 shows the deviation of D from N in an intuitive way. In the two figures, the frequency representations of different signals are placed in a polar coordinate system. The center of the coordinate system corresponds to clean speech S. The distance between S and a noisy signal Y denotes the intensity of the noise, and the angle between SY and a predetermined axis denotes the noise type N. Compared with Y, enhanced speech $\hat{S}$ may typically be closer to S. In other words, the SNR of $\hat{S}$ is higher than that of Y. This may be related to the training targets of speech enhancement systems. Ideal ratio mask (IRM), a commonly adopted training target, is shown in equation (6) below:

$$IRM = \frac{|S|}{|S| + |N|} \tag{6}$$

Along with the reduction of the distance to S, $\hat{S}$ may deviate from line SY. Such a noise type change may cause the performance degradation of ASR systems designed specifically for Y. This may be the main cause of the distortion problem. In fact, for two utterances with the same kind of noise at different SNRs, experimental results show that the one with the higher SNR can yield better recognition performance. Note that because of the similarity of masking-based and mapping-based speech enhancement, the above analysis is expected to apply for mapping based systems as well.

2.2. Distortion-Independent Acoustic Modeling

For ASR models trained with noisy speech and evaluated on the enhanced speech, the training and evaluation audio features can be expressed as equation (7) and (8), respectively.

$$|Y_{tr}| = |S_{tr} + N_{tr}| \tag{7}$$

$$|\hat{S}_{eval}| = |S_{eval} + D_{eval}| \tag{8}$$

where $D_{eval} = N_{eval} \otimes M_{eval} - S_{eval} \otimes \overline{M_{eval}}$. Y_{tr}, S_{tr}, and N_{tr} denote the frequency representations of the noisy speech, clean speech, and additive noise in the training set, respectively. $\hat{S}_{eval}$, S_{eval}, N_{eval} are the enhanced speech, clean speech, and noise in the evaluation set, respectively. D_{eval} denotes the distortion in the enhanced evaluation speech, and M_{eval} the estimated T-F mask during evaluation.

Based on our analysis in Section 2.1, the difference between N_{tr} and D_{eval} may be the cause of the performance degradation. In order for acoustic models to generalize to D_{eval}, N_{tr} can be modified in two ways. If we view D_{eval} as a special type of noise, a straightforward way to alleviate the distortion problem is to increase the type of N_{tr}. Acoustic models trained with a large variety of noises are denoted as noise-independent acoustic models. Another way is to train the acoustic model with the enhanced speech directly, as is shown in equation (9):

$$|\hat{S}_{tr}| = |S_{tr} + D_{tr}| \tag{9}$$

where $\hat{S}_{tr}$ denotes enhanced training speech and D_{tr} refers to the distortion in it.

A concern of this method is that it may not generalize to speech enhancement frontends not used during training. We propose to use large scale training to address this problem and denote this acoustic modeling scheme as distortion-independent acoustic modeling.

Figure 2 illustrates distortion-independent acoustic modeling. The left diagram depicts the training stage and the right one testing. In the right diagram, speech enhancement blocks with dashed lines denote those not used during training. In this study, we evaluate three existing speech enhancement models: gated residual network (GRN) [21], long short-term memory (LSTM) network [22], and convolutional recurrent network (CRN) [23]. We also add the IRM as another enhancement frontend. The switch in the right diagram denotes the coupling be-

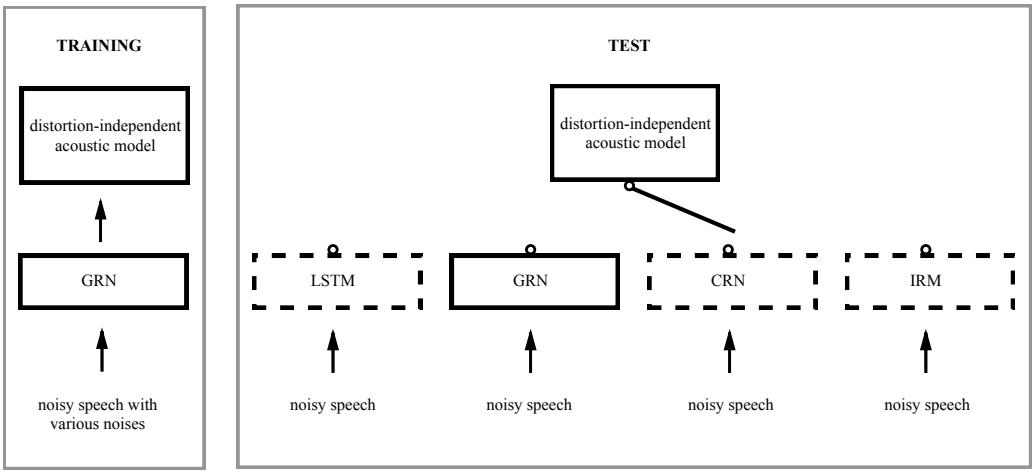


Figure 2: *Illustration of distortion-independent acoustic modeling. See text for the meaning of acronyms.*

tween a distortion-independent acoustic model and various enhancement frontends.

2.3. Types of Acoustic Models

In addition to the noise-independent and distortion-independent acoustic models in Section 2.2, we also investigate reverberant and noise-dependent acoustic models in this study. The reverberant acoustic model is trained using the reverberant-only speech in CHiME-2, whereas the noise-dependent acoustic model is trained and tested on the official noisy-reverberant utterances.

3. Experimental Setup

3.1. Dataset

Our experiments are conducted on the medium vocabulary track (track 2) of the CHiME-2 corpus. We also use noise segments from a 10000 noise database (available at https://www.soundideas.com) for speech enhancement and acoustic modeling.

The training sets for the four acoustic models are designed based on the official recipe of the CHiME-2 challenge. The reverberant acoustic model is trained using the 7138 reverberant utterances. The noise-dependent acoustic model uses the 7138 noisy-reverberant utterances in the CHiME-2 training set. For the noise-independent acoustic model, the training set is generated by mixing the reverberant utterances with noise segments from the 10k noise database at SNRs randomly chosen from {-6dB, -3dB, 0dB, 3dB, 6dB, 9dB}. The number of training utterances is 157036. For the distortion-independent acoustic model, the 157036 utterances for noise-independent training are enhanced by the gated recurrent network (GRN) [21] and the enhanced utterances are used as its training set.

For the validation sets of the acoustic models, the noise-dependent acoustic model uses $409 * 6$ noisy-reverberant utterances, whereas the other acoustic models use the 409 reverberant-only utterances in CHiME-2.

For the evaluation set, in addition to the official CHiME-2 test set, we also generate an ADT test set by mixing the reverberant-only test utterances in CHiME-2 with two ADT noises, ADTbabble and ADTcafeteria1. The average results on the two ADT noises are reported in this paper.

3.2. Implementation Details

We adopt GRN [21] as main speech enhancement model. The ASR backend is based on wide residual bidirectional LSTM (WRBN) [15] with recurrent LSTM dropout [24, 25].

During feature extraction, we skip some preprocessing operations including direct current (DC) offset removal, dithering, and pre-emphasizing. They may potentially alter the enhanced speech and influence our investigation.

For the four acoustic models, experimental setups and hyper-parameters are all kept the same. Adam optimizer [26] is used with a learning rate of 10^{-4}. The dropout rate during training is 0.2 and the acoustic model checkpoint yielding the lowest cross entropy loss on the validation set is selected for evaluation. Due to the large sizes of the noise-independent and distortion-independent training sets, we save acoustic model checkpoints every 7138 utterances for these two models.

4. Evaluation Results

4.1. Results of the Acoustic Models

Table 1 shows the results of the four acoustic models. The reverberant acoustic model clearly benefits from the speech enhancement frontend. The main reason would be that enhanced speech has higher SNR than the corresponding noisy speech.

For noise-dependent acoustic modeling, similar to previous observations [12, 13, 15], the results on the unenhanced noisy speech are better. Based on our analysis in Section 2, the reason of the performance degradation on enhanced speech is caused by the mismatch between N_{tr} and D_{eval}. Note that, although noise-dependent acoustic models perform well on N_{tr}, we find that their performance degrades significantly on untrained noises. Although trained with a large variety of N_{tr}, the noise-independent acoustic model still cannot overcome the distortion problem. This indicates that speech distortions may have different attributes from additive noises.

The distortion-independent acoustic model is able to overcome the distortion problem. Note that it is trained using noises different from those for evaluation. The good performance on the evaluation sets shows its good generalization ability to untrained noises.

Table 1: *WERs of different acoustic models.* chime-2 *denotes the official CHiME-2 evaluation set.* ADT *refers to the average WER of ADTbabble and ADTcafeteria1.* w/o *refers to noisy evaluation data without speech enhancement (i.e. unenhanced speech), and* w/ *evaluation data with enhancement. The distortion-independent acoustic model is trained using GRN enhanced speech.*

model	eval	enhance	9dB	6dB	3dB	0dB	-3dB	-6dB	avg
reverberant	chime-2	w/o	31.27	38.69	46.85	57.33	62.94	72.31	51.6
		w/	**10.50**	**13.67**	**17.26**	**23.73**	**29.91**	**39.87**	**22.5**
	ADT	w/o	31.03	47.53	67.50	85.96	93.49	95.58	70.2
		w/	**11.40**	**19.00**	**31.68**	**50.89**	**71.89**	**88.79**	**45.6**
noise-dependent	chime-2	w/o	**5.49**	**6.26**	**6.78**	**8.95**	**9.98**	**14.83**	**8.7**
		w/	5.81	7.98	8.33	11.26	14.48	19.05	11.2
noise-independent	chime-2	w/o	6.63	**7.72**	8.82	**10.69**	**13.06**	**17.45**	**10.7**
		w/	**6.37**	7.92	**8.78**	11.62	13.30	19.80	11.3
	ADT	w/o	**6.59**	**8.66**	**14.00**	**23.72**	**39.04**	**60.73**	**25.5**
		w/	7.98	11.04	19.35	30.99	51.60	76.08	32.8
distortion-independent	chime-2	w/o	7.42	8.61	10.01	12.93	14.85	21.80	12.6
		w/	**5.51**	**6.54**	**7.10**	**9.70**	**11.04**	**15.45**	**9.2**
	ADT	w/o	10.20	13.27	20.99	32.30	51.02	75.54	33.9
		w/	**6.60**	**8.64**	**14.73**	**22.76**	**37.74**	**58.24**	**24.8**

4.2. Generalization Ability to Various Speech Enhancement Models

The results of the distortion-independent acoustic model evaluated with different speech enhancement frontends are shown in Table 2. Trained on the speech enhanced by GRN, the distortion-independent model is able to work with LSTM [22], CRN [23], and IRM. Note that GRN, LSTM, and IRM are masking based, whereas CRN is mapping based. The results in Table 2 show that there may be a pattern in the distortions caused by speech enhancement models. In real-world applications, this suggests that a distortion-independent acoustic model may not need to be retrained when a new speech enhancement frontend is applied.

Table 2: *WERs of the distortion-independent acoustic model with other frontends.* unenh *denotes unenhanced speech. See Table 1 caption for other notations.*

model	eval	9dB	6dB	3dB	0dB	-3dB	-6dB	avg
unenh	chime-2	7.42	8.61	10.01	12.93	14.85	21.80	12.6
	ADT	10.20	13.27	20.99	32.30	51.02	75.54	33.9
LSTM	chime-2	5.79	7.47	8.63	11.36	14.16	19.41	11.1
	ADT	7.61	10.21	17.57	28.47	44.83	67.08	29.3
CRN	chime-2	6.65	7.68	9.04	11.25	13.51	18.06	11.0
	ADT	7.50	10.09	15.57	25.73	41.12	62.33	27.1
IRM	chime-2	3.40	3.44	3.34	3.38	3.74	3.31	3.4
	ADT	3.66	3.64	3.62	3.73	3.95	4.10	3.8

4.3. Comparisons with Previous Best Systems

Table 3 shows the comparisons of the ASR systems in this study with previous best systems. The distortion-independent acoustic model achieves a WER of 9.2%, outperforming the previous best systems on the CHiME-2 corpus [27, 28]. For the noise-dependent acoustic model, we achieve an average WER of 8.7%, outperforming the previous best system by 6.5% relatively. Note that the distortion-independent acoustic model is trained without using the CHiME-2 noises, whereas the noise-dependent acoustic model is trained and tested on the same CHiME-2 noises. This may be the reason why the distortion-independent acoustic model does not perform better than the noise-dependent acoustic model. The excellent results of our proposed models suggest that the observations in this study are likely valid for real world systems.

Table 3: *WER comparisons between the proposed models and prior work.*

model	9dB	6dB	3dB	0dB	-3dB	-6dB	avg
Wang and Wang [28]	6.61	6.86	8.67	10.39	13.02	18.23	10.6
Plantinga *et al.* [27]	-	-	-	-	-	-	9.3
distortion-independent	5.51	6.54	7.10	9.70	11.04	15.45	9.2
noise-dependent	**5.49**	**6.26**	**6.78**	**8.95**	**9.98**	**14.83**	**8.7**

5. Concluding Remarks

In this study, we have analyzed the distortion problem in monaural speech enhancement for speech recognition. Viewing the distortion problem as a noise type mismatch between training and evaluation, we have proposed a distortion-independent acoustic modeling scheme. Experimental results show that the distortion-independent acoustic model can not only overcome the distortion problem but also work with various speech enhancement frontends. Both the distortion-independent and a noise-dependent acoustic model perform better than the previous best system on the CHiME-2 corpus. The noise-dependent acoustic model achieves a WER of 8.7%, outperforming the previous best result by 6.5% relatively. Future work includes investigating time domain speech enhancement for distortion-independent acoustic modeling and applying distortion-independent training to the post filtering process in multichannel speech recognition.

6. Acknowledgments

This work was supported in part by two NSF grants (IIS-1409431 and ECCS-1808932) and the Ohio Supercomputer Center.

7. References

[1] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381–1390, 2013.

[2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of INTERSPEECH*, 2013, pp. 436–440.

[3] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19, 2015.

[4] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. of LVA/ICA*, 2015, pp. 91–99.

[5] Y. Xu, J. Du, Z. Huang, L. Dai, and C. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *arXiv preprint arXiv:1703.07172*, 2017.

[6] Y. Xu, J. Du, L. Dai, and C. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. of INTERSPEECH*, 2014, pp. 2670–2674.

[7] T. Gao, J. Du, Y. Xu, L. Liu, C.and Dai, and C. Lee, "Improving deep neural network based speech enhancement in low SNR environments," in *Proc. of LVA/ICA*, 2015, pp. 75–82.

[8] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[9] E. Healy, S. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, pp. 3029–3038, 2013.

[10] M. Kolbak, Z. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 153–167, 2017.

[11] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of ICASSP*, 2013, pp. 7092–7096.

[12] ——, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 826–835, 2014.

[13] J. Du, Q. Wang, T. Gao, Y. Xu, L. Dai, and C. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Proc. of INTERSPEECH*, 2014, pp. 616–620.

[14] 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 & Language*, vol. 46, pp. 535–557, 2017.

[15] J. Heymann, L. Drude, and H. Reinhold, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. of CHiME16*, 2016, pp. 12–17.

[16] T. Gao, J. Du, L. Dai, and C. Lee, "SNR-based progressive learning of deep neural network for speech enhancement." in *Proc. of INTERSPEECH*, 2016, pp. 3713–3717.

[17] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," *arXiv preprint arXiv:1803.09816*, 2018.

[18] L. Chai, J. Du, and C. H. Lee, "Acoustics-guided evaluation (age): a new measure for estimating performance of speech enhancement algorithms for robust asr," *arXiv preprint arXiv:1811.11517*, 2018.

[19] A. Narayanan, A. Misra, K. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 441–447.

[20] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 126–130.

[21] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 189–198, 2019.

[22] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, pp. 4705–4714, 2017.

[23] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. of INTERSPEECH*, 2018, pp. 3229–3233.

[24] P. Wang and D. Wang, "Utterance-wise recurrent dropout and iterative speaker adaptation for robust monaural speech recognition," in *Proc. of ICASSP*, 2018, pp. 4814–4818.

[25] ——, "Filter-and-convolve: A CNN based multichannel complex concatenation acoustic model," in *Proc. of ICASSP*, 2018, pp. 5564–5568.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic gradient descent," in *Proc. of International Conference on Learning Representations (ICLR)*, 2015, pp. 1–15.

[27] P. Plantinga, D. Bagchi, and E. Fosler-Lussier, "An exploration of mimic architectures for residual network based spectral mapping," in *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 550–557.

[28] Z. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 796–806, 2016.