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

Enhanced Spectral Features for Distortion-Independent Acoustic Modeling

Peidong Wang[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, wang.77}@osu.edu

Abstract

It has recently been shown that a distortion-independent acoustic modeling method is able to overcome the distortion problem caused by speech enhancement. In this study, we improve the distortion-independent acoustic model by feeding it with enhanced spectral features. Using enhanced magnitude spectra, the automatic speech recognition (ASR) system achieves a word error rate of 7.8% on the CHiME-2 corpus, outperforming our previous best system by more than 10% relatively. Compared with the corresponding enhanced waveform signal based system, systems using enhanced spectral features obtain up to 24% relative improvement. These comparisons show that speech enhancement is helpful for robust ASR and that enhanced spectral features are more suitable for ASR tasks than enhanced waveform signals.

Index Terms: robust ASR, speech enhancement, enhanced spectral feature, spectral feature mapping, CHiME-2

1. Introduction

Robust automatic speech recognition (ASR) has made substantial progress in recent years. While multichannel speech enhancement techniques such as beamforming can be directly integrated to ASR systems to improve the recognition performance [1, 2, 3, 4, 5], the way to combine monaural speech enhancement frontends and ASR backends remains a challenge [6, 7, 8, 9].

The difficulty in combining monaural speech enhancement and ASR is often attributed to the speech distortion introduced in the enhancement process. To alleviate this problem, Gao *et al.* and Wang *et al.* proposed to jointly train mapping or masking based speech enhancement frontends with ASR backends [10, 11]. Gao *et al.* also proposed a progressive training scheme for speech enhancement [12, 13]. It fine-tunes enhancement models in a multitask manner. Instead of using clean speech as the only target of the output layer, they add multiple layers in a deep neural network (DNN) or a long short-term memory (LSTM) network treating speech with progressively decreased signal-to-noise ratio (SNR) as labels. This way, the enhancement model is trained to reduce noise gradually, as well as the distortion in the output layer. Bagchi *et al.* proposed a mimic loss to optimize speech enhancement frontends using the senone outputs of ASR backends [14, 15]. With a speech enhancement frontend trained with mimic loss, an off-the-shelf ASR model in Kaldi yields a WER of 9.3% on the CHiME-2 corpus. Recently, Chai *et al.* proposed an acoustic-guided evaluation (AGE) [16], which can estimate the performance of speech enhancement methods when they are used for robust ASR.

In our previous studies [17], we proposed a distortion-independent acoustic modeling method to solve the distortion problem. 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, the distortion-independent acoustic model is able to generalize to speech enhancement frontends not used during training. In the paper [17], however, the distortion-independent acoustic model does not perform as well as an acoustic model trained and tested both on the CHiME-2 noises (noise-dependent acoustic model). This study aims to improve the performance of the distortion-independent acoustic model by feeding it with enhanced spectral features directly.

There are mainly three methods to generate enhanced spectral features, masking, mapping, and signal approximation (SA) [18]. We adopt spectral mapping in this study. Note that the focus of this study is the comparison between enhanced waveform signals and enhanced spectral features. Therefore, the choice of masking, mapping, and SA may not influence the conclusion of this paper.

Spectral feature mapping is a way to extract the features of clean speech from the spectra of noisy speech. Deep learning based spectral mapping methods was introduced by Lu *et al.* and Xu *et al.* [19, 20]. They use a deep autoencoder (DAE) or DNN as the model architecture. Along with the development of deep learning, many model architectures were proposed. Weninger *et al.* and Chen *et al.* applied LSTM based neural networks for speech enhancement [21, 22]. Tan *et al.* reduced the trainable parameters of LSTM by convolutional neural networks (CNNs), forming a gated residual network (GRN) [23]. A subsequent work on convolutional recurrent network (CRN) combined LSTM and CNN [24], enabling speech enhancement to be used in real-time and with low computational complexity.

The input and output features of spectral feature mappings do not need to be in the same domain. In Huang *et al.*'s paper on the usage of masks as the reconstruction constraints of monaural source separation, they applied two spectral feature mappings for speech enhancement ("speech denoising" in the paper) [25]. Their empirical observation is that the magnitude to magnitude mapping performs better than the magnitude to log Mel mapping. Han *et al.* conducted an extensive study comparing different spectral feature mappings for robust ASR [26]. Their experimental results show that the mapping from log magnitude to log Mel works best for the DNN based ASR backend. Bagchi *et al.* conducted a subsequent work combining spectral feature mapping and multichannel source separation for robust ASR [27]. Recently, Escudero *et al.* [28] proposed to improve the dereverberation ability of Han *et al.*'s spectral mapping model [26] by combining it with weighted prediction error (WPE) [29].

This study aims to improve the distortion-independent acoustic modeling method by changing the input features from enhanced waveform signals to enhanced spectral representations. Based on the ASR feature extraction process, we use

http://dx.doi.org/10.21437/Interspeech.2019-1493

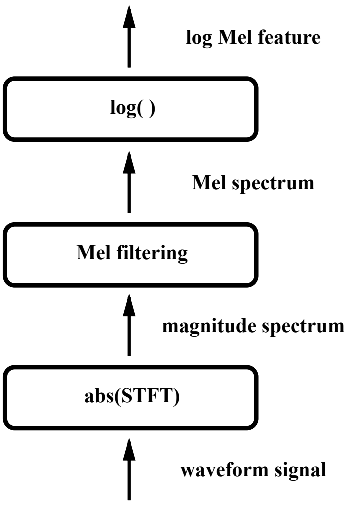

Figure 1: *Illustration of the feature extraction steps and the corresponding features before delta operations.*

four features as spectral mapping model outputs, time-domain (waveform), magnitude, Mel, and log Mel. The architectures of spectral mapping models include GRN [23], LSTM [22], and CRN [24]. With enhanced magnitude features as input, the distortion-independent acoustic model yields a WER of 7.8% on the CHiME-2 corpus, outperforming our previous best system by more than 10% relatively [17]. Compared with the corresponding waveform signal based systems, systems using enhanced magnitude features obtain up to 24% improvement. This study shows that speech enhancement is helpful for robust ASR and that enhanced spectral features are more suitable for ASR tasks than enhanced waveform signals.

The rest of the paper is organized as follows. We describe spectral features and the specifics of speech enhancement models in Section 2. Section 3 and 4 present the experiment setup and results, respectively. We give concluding remarks in Section 5.

2. System Description

2.1. Spectral Features for Distortion-Independent Acoustic Modeling

We use log Mel and its deltas as the direct inputs to the acoustic model. Figure 1 shows the feature extraction steps before delta operations. Each rectangle in the figure represents a feature extraction step and the text near arrow the corresponding features. The way to use an enhanced spectral feature is to skip the step(s) in Figure 1 below its level. Note that delta features are typically not used as the outputs of spectral feature mappings since the value ranges of the static, delta, and delta-delta features differ significantly.

2.2. Speech Enhancement for Spectral Features

This study adopts the two commonly used spectral mappings, magnitude to magnitude (mag-mag) proposed by Huang *et al.* [25] and log magnitude spectra to log Mel features (logmag-logmel) proposed by Han *et al.* [26]. We also use three other feature mappings, magnitude to waveform signals (mag-wav) [17], magnitude to Mel features (mag-mel), and magnitude to log Mel features (mag-logmel). Note that the mag-wav mapping does not involve training. It is performed by resynthesizing waveform signals from enhanced magnitude spectra. The phase of noisy speech is used to reconstruct a complex representation for enhanced speech, and an overlap-and-add technique is used to combine waveform segments.

A GRN [23] is used as the main speech enhancement model. For different output features, the output layers of GRN differ slightly. We list the specifics of the output layers in Table 1 below:

Table 1: *Output layer specifics for different spectral mapping outputs.* activation *refers to the activation function in the output layer and* dimension *is the output dimension.* plus *denotes the softplus activation and* lin *the linear activation.*

	activation	dimension
mag	plus	161
mel	plus	80
logmel	lin	80

2.3. Generalization Ability to Various Speech Enhancement Frontends

One of the main features of distortion-independent acoustic modeling is its ability to work with various speech enhancement frontends. In this study, we investigate such generalization ability on enhanced spectral features and compare the results with those on enhanced waveform signals.

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 CHiME-2 corpus contains reverberant and noisy utterances. The utterances in the Wall Street Journal (WSJ0) corpus are first convolved with a binaural room impulse response corresponding to a frontal position at a distance of 2m. These reverberant utterances are then mixed with binaural noise recorded in a family living room at six SNR levels: {-6dB, -3dB, 0dB, 3dB, 6dB, 9dB}. The training set contains 7138 utterances from the 83 speakers in the WSJ0 SI-84 training set. For each of the six SNR levels, the development set has 409 utterances from 10 other speakers. The test set consists of 330 utterances at each SNR level. Its speakers are different from those in the training and development set.

The speech enhancement models in this study are noise-independent [22]. Their training set is generated by mixing reverberant (more specifically "scaled") utterances in CHiME-2 with noise segments from the 10k noise database. The SNR levels of the training utterances are randomly chosen from the above six levels. We use reverberant only utterances as training targets. The distortion-independent acoustic model also uses the 10k noise during training. Its training set generation process is similar to that of the speech enhancement models. Note that the training sets for enhancement and recognition are isolated with respect to noise segments. We divide each noise segment

Table 2: *WER comparisons of spectral feature mappings using GRN.*

mapping	9dB	6dB	3dB	0dB	-3dB	-6dB	avg
mag-wav	5.51	6.54	7.10	9.70	11.04	15.45	9.2
mag-mag	**4.54**	**5.45**	6.20	7.92	**9.43**	**13.11**	**7.8**
mag-mel	5.12	5.62	6.46	8.43	10.05	15.24	8.5
mag-logmel	5.32	6.16	6.24	8.44	10.54	15.56	8.8
logmag-logmel	5.29	5.68	**6.05**	**7.57**	10.18	14.37	8.2

into two parts. The distortion-independent acoustic model can only use the first halves and the spectral mapping models the second halves. For the distortion-independent acoustic model in this study, the noisy utterances are fed to a GRN to generate enhanced waveform signals. The log Mel and delta features are extracted from the enhanced waveform signal to train the distortion-independent acoustic model.

3.2. Implementation Details

The architecture of the GRN based speech enhancement model follows the recipe in [23]. It consists of four frequency-dilated convolutional layers and three time-dilated blocks. The mean and variance of 1000 training utterances are used to normalize the features during testing. During training, the optimizer is Adam and the learning rate is 10^{-3}. The training process stops after 30 epochs and the model yielding the best performance on the development set is used for evaluation. The LSTM network is unidirectional. It has four hidden layers, each containing 512 LSTM cells. It also uses a context window of 11 for the input features. The number of maximal training epochs is 40 and the other training hyperparameters are the same as GRN. CRN is a model combining the advantages of GRN and LSTM. It has a symmetric structure consisting of five convolutional layers, two LSTM layers, and five deconvolutional layers. Each deconvolutional layer takes as input not only the output from the previous layer but also the input of the corresponding convolutional layer. The LSTM layer in CRN is bidirectional, each containing 512 units. The training hyperparameters for CRN is the same as the two models above.

As mentioned in Section 2.2, for each spectral feature output, we modify the instances of the GRN based speech enhancement model accordingly. The training hyperparameters over different spectral features are kept the same.

The distortion-independent acoustic model in this study is a wide residual bidirectional LSTM network (WRBN) [9] with utterance-wise recurrent dropout for the LSTM layers [30, 5, 17]. During feature extraction, we skip three preprocessing operations, direct current (DC) offset removal, dithering, and pre-emphasizing. They may potentially alter the enhanced speech. For the training process of the distortion-independent acoustic model, we adopt Adam optimizer with a learning rate of 10^{-4} and a dropout rate of 0.2.

Note that during the training process of the distortion-independent acoustic model, we only use enhanced waveform signals. The enhanced spectral features are only applied during evaluation.

4. Evaluation Results

4.1. Comparisons of Enhanced Spectral Features

We list the evaluation results for the five feature mappings and the four features using GRN in Table 2. Since we use GRN both in training and evaluation, this comparison avoids the potential influence of speech enhancement model mismatch.

It is clear that enhanced spectral features substantially outperform the enhanced waveform signal. More specifically, mag-mag outperforms mag-wav by more than 15% relatively. Note that the only difference between mag-wav and mag-mag is that the mag-wav mapping resynthesizes magnitude spectra to waveform signals. This performance difference may be attributed to a phase inconsistency problem caused by combining enhanced magnitude spectra and the phase information from noisy speech [31, 32, 33, 34]. Although an overlap-and-add technique is applied to alleviate its influence, the phase inconsistency problem may still degrade the quality of the resynthesized speech. Note that in this study, we use the speech enhancement model as a frontend for ASR. The phase inconsistency problem can thus be easily bypassed by using enhanced spectral features directly.

Among different spectral outputs, the magnitude spectra generated by the mag-mag mapping performs the best. Since magnitude spectrum is a shared feature between speech enhancement and ASR, this shows that by distortion-independent acoustic modeling, the performance of the acoustic model and that of the speech enhancement frontend are correlated. The logmag-logmel mapping yields better results than mag-mel and mag-logmel. It also outperforms mag-mag at 3dB and 0dB. Note that for logmag-logmel mapping, we add a 10^{-7} to each logarithm function to prevent numerical exceptions. This may make the input and output features of the logmag-logmel mapping normalized to values greater than -7, and thus its good performances at 3dB and 0dB.

4.2. Generalization Ability to Various Speech Enhancement Methods

We list the results of various speech enhancement methods in Table 3. The ideal ratio mask (IRM) is included as an ideal speech enhancement method. Note that the unenhanced method can only use enhanced waveform signals.

Magnitude spectra outperform waveform signals for all speech enhancement methods. For LSTM and CRN, the relative improvements are 19% and 24%, respectively. Even for the IRM based speech enhancement method, enhanced magnitude spectra still obtain a performance improvement. These comparisons show that the phase inconsistency problem may cause major problems for robust ASR.

Comparing the four speech enhancement methods using magnitude spectra, we can see that with the new input feature, the distortion-independent acoustic model is still able to generalize to various speech enhancement methods. The main contribution may be from large scale training.

Table 3: *WERs of various speech enhancement methods.* unenhanced *refers to the method extracting features directly from noisy utterances, without the usage of speech enhancement frontends.*

architecture	feature	9dB	6dB	3dB	0dB	-3dB	-6dB	avg
unenhanced	wav	7.42	8.61	10.01	12.93	14.85	21.80	12.6
LSTM	wav	5.79	7.47	8.63	11.36	14.16	19.41	11.1
	mag	**5.34**	**6.07**	**6.74**	**9.58**	**10.42**	**15.93**	**9.0**
CRN	wav	6.65	7.68	9.04	11.25	13.51	18.06	11.0
	mag	**5.23**	**5.57**	**6.48**	**8.37**	**10.24**	**14.80**	**8.4**
IRM	wav	3.40	3.44	3.34	3.38	3.74	3.31	3.4
	mag	**3.27**	**3.36**	3.34	**3.33**	**3.31**	**3.14**	**3.3**

Table 4: *Comparisons with previous best systems.* distortion-independent acoustic modeling *refers to using enhanced waveform signals for distortion-independent acoustic modeling.* noise-dependent acoustic model *represents our previous best model proposed in [17]. It is trained and tested on the same (CHiME-2) noises.*

systems	9dB	6dB	3dB	0dB	-3dB	-6dB	avg
Wang and Wang [11]	6.61	6.86	8.67	10.39	13.02	18.23	10.6
Plantinga *et al.* [15]	-	-	-	-	-	-	9.3
distortion-independent acoustic modeling [17]	5.51	6.54	7.10	9.70	11.04	15.45	9.2
noise-dependent acoustic modeling [17]	5.49	6.26	6.78	8.95	9.98	14.83	8.7
magnitude features + distortion-independent acoustic modeling	**4.54**	**5.45**	**6.20**	**7.92**	**9.43**	**13.11**	**7.8**

4.3. Comparisons with Previous Best Systems

The comparisons between the systems in this study and our previous best are shown in Table 4. Using enhanced magnitude features, the distortion-independent acoustic model obtains a WER of 7.8% on the CHiME-2 evaluation set, outperforming our previous best system by 10% relatively. Moreover, multiple other systems are also better than our previous best system. The Mel features achieves a relative improvement of 2% and the log Mel feature generated by the logmag-logmel mapping obtains 6%. Note that our previous best system is a noise-dependent acoustic model trained and tested both on the noisy speech. These consistent improvements show that speech enhancement is helpful for robust ASR. Compared with [17], which uses enhanced waveform signals for distortion-independent acoustic modeling, using enhanced magnitude features for the same acoustic model obtains a 15% relative improvement. This shows that enhanced spectral features are more suitable for robust ASR than enhanced waveform signals.

Note that Wang and Wang's, Plantinga *et al.*'s, and the noise-dependent acoustic model all follow the official CHiME-2 recipe, i.e. they use the same (CHiME-2) noises during training and evaluation. For the distortion-independent acoustic model in this study, the training noises are from the 10k noise database and the test noises are CHiME-2. This training noise difference cannot be avoided since it relates directly to the design of the training schemes. Distortion-independent acoustic modeling requires the usage of a large variety of noises for its generalization ability to unseen noises and speech enhancement frontends. One important thing to note is that the training set of the distortion-independent acoustic model does not include the CHiME-2 noises. Therefore, the distortion-independent acoustic model is tested on unseen noises, whereas the other models are trained and tested on matched noises. This may disadvantage distortion-independent acoustic modeling based methods, but it does not influence the conclusion of this study.

5. Concluding Remarks

In this study, we have investigated enhanced spectral features for distortion-independent acoustic modeling. For each of the four features in the ASR feature extraction process, we design five spectral mappings. The magnitude to magnitude mapping appears to perform the best. Using enhanced magnitude features, the distortion-independent acoustic model yields a WER of 7.8% on the CHiME-2 corpus, outperforming our previous best system by more than 10% relatively. In addition, multiple other spectral feature based systems also perform better than our previous best. The observation that the new distortion-independent acoustic modeling based system outperforms the noise-dependent acoustic model convincingly shows that speech enhancement is helpful for speech recognition. Compared with waveform signal based systems, systems using spectral features achieve up to 24% relative improvement. This suggests that enhanced spectral features are more suitable for speech recognition tasks than enhanced waveform signals. In addition to the two comparisons above, we have shown that using enhanced magnitude features in this study, the distortion-independent acoustic model can still generalize to various speech enhancement methods. Future work includes applying a self-attention mechanism to the distortion-independent acoustic model, exploring time-domain speech enhancement for robust ASR, investigating distortion-independent training for end-to-end speech recognition, and using distortion-independent training for post-filtering.

6. Acknowledgements

We would like to thank K. Tan for sharing the GRN and CRN code. This work was supported in part by two NSF grants (IIS-1409431 and ECCS-1808932) and the Ohio Supercomputer Center.

7. References

[1] D. L. 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.

[2] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, pp. 1529–1539, 2007.

[3] X. Zhang, Z. Q. Wang, and D. L. Wang, "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust asr," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 276–280.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 196–200.

[5] P. Wang and D. L. Wang, "Filter-and-convolve: A CNN based multichannel complex concatenation acoustic model," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 5564–5568.

[6] A. Narayanan and D. L. 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.

[7] ——, "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.

[8] J. Du, Q. Wang, T. Gao, Y. Xu, L. R. Dai, and C. H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 616–620.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proceedings of the 4th International Workshop on Speech Processing in Everyday Environments (CHiME16)*, 2016, pp. 12–17.

[10] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4375–4379.

[11] Z. Q. Wang and D. L. 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.

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

[13] ——, "Densely connected progressive learning for lstm-based speech enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 5054–5058.

[14] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 5609–5613.

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

[16] 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.

[17] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *arXiv preprint arXiv:1903.04567*, 2019.

[18] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. of 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 577–581.

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

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

[21] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[22] J. Chen and D. L. 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, J. Chen, and D. L. 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.

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

[25] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 2136–2147, 2015.

[26] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. L. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Proc. of INTERSPEECH*, 2015, pp. 2484–2488.

[27] D. Bagchi, M. I. Mandel, Z. Q. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *Proc. of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 496–503.

[28] J. P. Escudero, J. Novoa, R. Mahu, J. Wuth, R. Stern, and N. B. Yoma, "An improved dnn-based spectral feature mapping that removes noise and reverberation for robust automatic speech recognition," *arXiv preprint arXiv:1803.09016*, 2018.

[29] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1717–1731, 2010.

[30] P. Wang and D. L. Wang, "Utterance-wise recurrent dropout and iterative speaker adaptation for robust monaural speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 4814–4818.

[31] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction." in *Proc. of SAPA INTERSPEECH*, 2008, pp. 23–28.

[32] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: A state of the art," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, 2011, pp. 375–386.

[33] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, pp. 55–66, 2015.

[34] Z. Q. Wang, J. L. Roux, D. L. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," pp. 2708–2712, 2018.