Joint Training of Speech Separation, Filterbank and Acoustic Model for Robust Automatic Speech Recognition

Zhong-Qiu 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

wangzhon@cse.ohio-state.edu, dwang@cse.ohio-state.edu

Abstract

Robustness is crucial for automatic speech recognition systems in real-world environments. Speech enhancement/separation algorithms are normally used to enhance noisy speech before recognition. However, such algorithms typically introduce distortions unseen by acoustic models. In this study, we propose a novel joint training approach to reduce this distortion problem. At the training stage, we first concatenate a speech separation DNN, a filterbank and an acoustic model DNN to form a deeper network, and then jointly train all of them. This way, the separation frontend and filterbank can provide enhanced speech desired by the acoustic model. In addition, the linguistic information contained in the acoustic model can have a positive effect on the frontend and filberbank. Besides the commonly used log mel-spectrogram feature, we also add more robust features for acoustic modeling. Our system obtains 14.1% average word error rate on the noisy and reverberant CHIME-2 corpus (track 2), which outperforms the previous best result by 8.4% relatively.

Index Terms: robust ASR, speech separation, deep neural networks, CHIME-2

1. Introduction

Deep neural networks (DNN), including convolutional neural networks (CNN) [1] and recurrent neural networks (RNN) [2], represent the state-of-the-art models for acoustic modeling in automatic speech recognition. In robust ASR, although DNN is shown to be inherently robust to slight variations of the training data because of its multi-layer architecture [3], its performance still drops significantly in the presence of rapidly changing and mismatched noises, low SNR conditions, and reverberant environments. As a result, speech enhancement or separation is still needed when using deep networks for acoustic modeling [4].

There are three common strategies when incorporating speech enhancement into robust ASR systems. The first approach is to train an acoustic model from clean speech and utilize a speech enhancement frontend to enhance noisy speech at the test stage [5]. It would be a big issue if the frontend introduces distortions not seen by the acoustic model at the training stage. The second approach avoids this problem by enhancing both training and testing data first, and then does acoustic modeling on the enhanced training set [4]. The third approach is to train an acoustic model via multi-condition training. Some studies directly feed noisy features into the acoustic model at the test stage, while other studies enhance noisy speech first. When comparing the second and third approaches, Delcroix et al. [4] can get better results using the second approach, while Seltzer et al. [6] show that the third approach is better. As suggested in [7,8], it would be better to let the acoustic model see enough variations during training. In addition, reducing the mismatch between enhanced speech and the training data for acoustic modeling is of considerable importance [5].

In our previous study [9], we proposed to jointly train a speech separation DNN with an acoustic model DNN for robust ASR. The key idea is to concatenate these two DNNs so that the error signal from the acoustic model DNN can be further back-propagated to the speech separation DNN. This way, the separation frontend can be adjusted to provide the enhanced speech desired by the acoustic model. In addition, the linguistic information from the acoustic model can influence the separation frontend. In this study, we further develop this strategy.

Here we train a speech separation DNN to enhance the noisy power spectrogram, rather than the noisy mel-spectrogram used in our previous study. We think it would probably be better to do enhancement in the power spectrogram domain since mel-spectrogram contains less information. As suggested in [10], mel-filterbank can be thought of as one layer in a neural network since mel-filtering is a linear transform of the power spectrogram. We can insert this layer between the speech separation DNN and the acoustic model DNN, and jointly train all of them so that the filterbank is adjusted accordingly.

Furthermore, in DNN-HMM hybrid approach for robust ASR, log mel-spectrogram is widely used as the only feature for acoustic modeling [5,6,11,12,13,14], partly because DNN is considered capable of automatically extracting meaningful representations through its multi-layer structure [15, 16]. However, in our experiments, we found that when using multi-condition training for acoustic modeling, adding more robust features, such as AMS [17], RASTA-PLP [18], PNCC [19], and MRCG [20], to acoustic models will significantly decrease word error rate (WER).

In summary, our study makes three contributions. First, we find that performing speech enhancement in the power spectrogram domain is slightly better than in the mel-spectrogram domain. Second, we jointly train the speech separation frontend, filterbank, and acoustic model to alleviate the distortion problem. Third, we find that adding more robust features to acoustic models significantly improves performance. With these observations, we achieve 14.1% WER on the challenging CHIME-2 corpus (track 2) [21], which, to our knowledge, represents the best result on this dataset.

2. System Description

In this section, we first describe the method for training a DNN-based speech separation frontend via time-frequency (T-F) masking. Then we present how we train a DNN-based acoustic model with more robust features. Note that these two DNNs are trained separately in the beginning. Finally, together with the values of the mel-filterbank, we use the parameters of the trained frontend and acoustic model to initialize the corresponding part in the joint training system. The overall joint training framework is shown in Figure 1.

2.1. Speech Separation

Recently, DNN-based time-frequency masking [22,23] has shown considerable potential for robust ASR [5,9,12,24]. These methods typically estimate the ideal ratio mask (IRM), a T-F mask that represents the ratio of speech energy to mixture energy at each T-F unit, from premixed clean speech and noise at different SNR levels. In this study, the IRM is defined in the power spectrogram domain:

$$M(t,f) = \frac{S(t,f)}{S(t,f) + N(t,f)} \quad (1)$$

where M is the ideal ratio mask, S is the power spectrogram of clean speech, and N is the power spectrogram of noise. t and f index time and frequency, respectively.

We utilize a DNN to do mask estimation. The DNN has three hidden layers, each with 1024 hidden rectified linear units (ReLU). The output layer contains 161 sigmoid units, corresponding to the number of channels in each frame of the power spectrogram. The optimization aims to minimize the cross-entropy loss function within each T-F unit. The dropout rates in the input layer and hidden layers are all set to 0.3. The maximum L_2 norm of the incoming weights of each hidden unit is set to be 1. We learn the weights starting from random initialization using stochastic gradient descent with momentum and Adagrad [25] for a maximum of 50 epochs. The mini-batch size is 256. The momentum is linearly increased from 0.1 to 0.9 in the first 12 epochs and kept fixed afterwards. The learning rate is fixed at 0.01 in the first 10 epochs, 0.005 in the following 20 epochs and 0.001 afterwards.

The window size in our study is 20 ms and the hop size is 10 ms. The features used for mask estimation are:

- 13-dimensional RASTA-PLP [18] feature;
- 15-dimensional AMS [17] feature extracted from each of the 26 channels of the mel-spectrogram;
- 31-dimensional narrowband MFCC feature with the analysis window of 20 ms;
- 31-dimensional wideband MFCC feature with the analysis window of 200 ms.

All of these features are globally mean and variance normalized before training. We splice a 7-frame window for all features except for AMS. So the total number of features for mask estimation is 915 (13*7+15*26+31*7+31*7). This feature set is shown to be complementary for mask-based speech separation in [26]. The feature set is denoted as "fIRM" for convenience.

At the test stage, given a noisy utterance, we first utilize the trained DNN to estimate the IRM of that utterance and then obtain the enhanced power spectrogram using:

$$X^* = (M^*)^\alpha \otimes X \quad (2)$$

where M^* is the estimated IRM of the noisy power spectrogram X, $\otimes$ stands for point-wise matrix multiplication, and X^* denotes the enhanced power spectrogram. Here a tunable parameter α $(0 \le \alpha \le 1)$ is used to scale the estimated masks. When α is set to 1, it means that we use the estimated masks directly. When α is set to 0, we do not perform any masking. When α is between 0 and 1, we suppress noise to some extent. Through validation, we find that $\alpha = 0.5$ is the best choice. When α is set to 0.5, Eq. (2) is similar to the square root Wiener filter which has optimal properties for power spectrogram enhancement [27].

2.2. Acoustic Modeling

The DNN–HMM hybrid approach represents the state-of-the-art method for speech recognition. In this study, we use a DNN with 7 hidden layers for acoustic modeling. Each hidden layer contains 2048 ReLU units. We use softmax activation at the output layer and minimize the cross-entropy loss function. All the other setup and training recipes are the same as the DNN training for mask estimation.

Many previous studies only use the log mel-spectrogram as the only feature for acoustic modeling. It is believed that DNN can learn useful representations automatically from relatively raw input such as the log mel-spectrogram or power spectrogram with a large context window (normally 11 frames). For robust ASR, when the acoustic model is trained using multi-conditional data, it's sensible that adding more robust features to acoustic models would help since different features would encode different kinds of information. In this study, we use a subset of the following features for acoustic modeling:

- 26-dimensional log mel-spectrogram feature together with its delta and double delta components. We further splice the features of 11 frames together after sentence level mean normalization (denoted as "NMS" feature);
- 915-dimensional fIRM feature as described in the previous section;
- 31-dimensional PNCC feature together with its delta and double delta components. Features from 11 frames are spliced together. The PNCC feature is relatively robust to reverberation and noises as shown in [19];
- 256-dimensional multi-resolution cochleagram (MRCG) feature together with its delta and double delta components. The recently proposed MRCG feature is shown to perform well for mask estimation [20].

All of these features are globally mean and variance normalized before acoustic modeling. For comparison, we always incorporate the NMS feature as part of all the features when doing acoustic modeling.

2.3. Joint Training

The joint training framework is shown in Figure 1. After we get the estimated IRM from the speech separation frontend, we scale it exponentially and multiply it point-wisely with the power spectrogram as in Eq. (2). Then we pass the enhanced power spectrogram into the filterbank layer to get the enhanced filterbank feature. The filterbank layer gives a linear transformation similar to mel-filtering, which can be represented as one layer in the network. Afterwards, we use the log operation to compress the enhanced filterbank feature. As the sentence-level mean normalization, delta and double

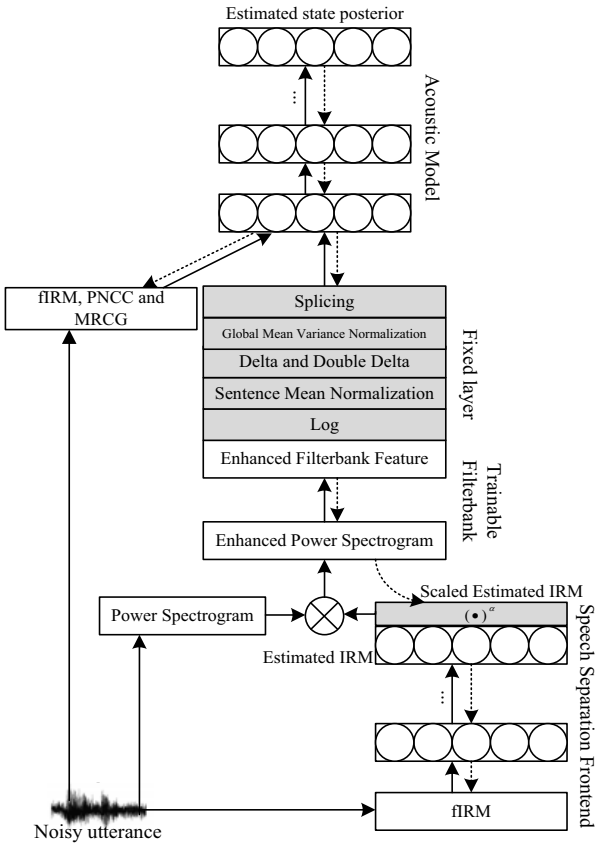

Figure 1*: Joint training framework. The layer shown in gray means that the weights or operations of that layer are fixed.*

delta, and global mean and variance normalization are all linear transformations, we can encode each of them as one layer in the network as well. Finally, after splicing several frames (11 frames in this study), together with other robust features, the output of the fixed layer is passed into the acoustic model. Interestingly, we can represent all of these steps as a big and deep neural network so that we can utilize the back-propagation algorithm to jointly train the speech separation frontend, filterbank and acoustic model.

We use the parameters of the separately trained speech separation DNN and acoustic model DNN to initialize the corresponding parameters of the joint-training DNN. Following [10], we initialize the weights of the filterbank layer as follows:

$$W^{filterbank} = exp(W^*) \quad (3)$$

where W^* is initialized to be $\log(Mel\text{-}filterbank)$. This way, every time W^* is updated, all the values in $W^{filterbank}$ are ensured to be non-negative.

This network is further jointly trained for a maximum of 30 epochs. The learning rate is fixed at 0.001 and the momentum is fixed at 0.9. The mini-batch size is set to be 512. No dropout is performed at the filterbank layer. The network is trained to optimize the cross-entropy error of the acoustic model. All the other setup and training recipes follow those for the DNN training for mask estimation and acoustic modeling in the previous steps. The sentence level mean of each utterance and the global mean and variance are updated by running the feed-forward algorithm at the beginning of each epoch.

3. Experimental Setup

We conduct our experiments on the medium-vocabulary task of the CHIME-2 challenge (track 2) [21]. The CHIME-2 corpus is created by first convolving clean utterances in WSJ0-5k with time-varying binaural room impulse responses and then mixing with reverberant noises at six SNR levels linearly spaced from -6 dB to 9 dB. The noises contain a very rich set of sounds from a living lounge and kitchen such as background speakers, footsteps, electronic devices, laughter, distant noises outside the room etc. The multi-condition training set contains 7138 noisy and reverberant utterances (~14.5h in total). The development set contains 409 utterances for each SNR condition (~4.5h in total). The test set contains 330 utterances for each SNR condition (~4h in total).

Our system is monaural. In our experiments, we simply average the signals from the left and right ear. The training data for mask estimation (7138 mixtures in total) is created by manually mixing the reverberant training set with the given noises in the CHIME-2 corpus at the same six SNR levels. Note that this dataset is only used for mask estimation. As we mentioned before, we utilize DNNs to do acoustic modeling. All the DNN-based acoustic models are trained using the multi-condition training set. A GMM-HMM system trained with maximum likelihood using the MFCC features extracted from the corresponding clean utterances in WSJ0-5k is used to get the senone state for each frame. There are 3310 senone states in total. We use a trigram language model and the CMU pronunciation dictionary in our experiments. The HTK toolkit is used to train the GMM-HMM system. The HTK decoder is modified to do DNN-HMM hybrid system decoding.

4. Evaluation Results

Our experiments are done in an incremental way. We first compare the performance of acoustic modeling with more robust features. Then we compare the performance of T-F masking in the power spectrogram domain with T-F masking in the mel-spectrogram domain. We finally present the results of joint training and compare our results with other studies.

4.1. Expanded features for acoustic modeling

In this experiment, we directly train acoustic models with different features using multi-condition training. Note that we do not perform speech enhancement here. The results on the test set are shown in Table 1. With the commonly used NMS feature, we obtain 20.8 percent average WER on the test set. When we add the fIRM feature, the average WER drops 4.2 percent from 20.8 to 16.6. If we further add the MRCG feature, the average WER drops 0.3 more percent to 16.3. The best model we have obtained is trained with the NMS+fIRM+ MRCG+PNCC feature, and the 15.6 percent WER on the test set is absolute 5.2 percent better than the NMS baseline and is only 0.2 percent worse than the previous best result [9] on this dataset. Note that what we do is simply adding more features, and it brings us 5.2 percent WER reduction on the test set. These results suggest that when using multi-condition training, adding more features for acoustic modeling provides significant benefit, probably because manually designed features contain more useful domain knowledge. It also suggests that relying on deep networks to automatically learn optimal features from raw input may not be the best strategy. Combining the feature learning power of deep networks with domain knowledge may be a more promising way towards

Table 1. *Results (% WER) of acoustic modeling with more robust features and direct multi-condition training*

Features	-6dB	-3dB	0dB	3dB	6dB	9dB	Average
NMS	32.2	26.1	21.9	17.2	14.7	13.0	20.8
NMS+fIRM	27.0	20.3	16.8	13.9	11	10.4	16.6
NMS+fIRM+MRCG	26.5	20.7	17.1	13.1	11.0	9.8	16.3
NMS+fIRM+MRCG+PNCC	26.1	18.7	16.2	12.8	10.4	9.3	15.6

Table 2. *Results (% WER) of masking in the mel-spectrogram or power spectrogram domain with different acoustic models*

Features for acoustic modeling	Masking domain	-6dB	-3dB	0dB	3dB	6dB	9dB	Average
NMS	Mel-spectrogram	28.8	22.8	19.8	16.3	13.6	12.0	18.9
NMS	Power spectrogram	28.3	22.3	19.9	15.9	13.7	12.0	18.7
NMS+fIRM+MRCG+PNCC	Mel-spectrogram	25.3	18.5	15.3	12.0	10.4	9.0	15.1
NMS+fIRM+MRCG+PNCC	Power spectrogram	25	18.2	15.3	12.2	10.4	9.0	15.0

Table 3. *Results (% WER) of joint training and comparison with other methods*

Description	-6dB	-3dB	0dB	3dB	6dB	9dB	Average
Jointly train frontend and acoustic model	23.6	17.9	14.4	11.8	9.9	8.6	14.4
Jointly train frontend, filterbank and acoustic model	22.8	17.7	14.0	11.5	9.9	8.8	14.1
Previous best result [9]	25.1	19.2	15.1	12.8	10.5	9.5	15.4
Directly train an 11-hidden-layer DNN	25.5	20.2	16.9	13.8	10.8	9.5	16.1

improvements [28].

4.2. T-F masking in different domains

In this experiment, we compare the performance of performing T-F masking in different domains. When ideal masks are defined in the mel-spectrogram domain, the frontend is trained to get the enhanced mel-spectrogram directly. When ideal masks are defined in the power spectrogram domain, the frontend is trained to get the enhanced power spectrogram first, which is then passed into the mel-filterbank to get the enhanced mel-spectrogram. The enhanced mel-spectrogram is finally passed into a multi-conditionally trained acoustic model for decoding. The performance on the test set is shown in Table 2. When the acoustic model is trained with the NMS feature, conducting T-F masking in the power spectrogram domain can improve the average WER by around 0.2 percent. When we use the NMS+fIRM+MRCG+PNCC feature to train the acoustic model, we get about 0.1 percent improvement. We can see that defining ideal masks in the power spectrogram domain performs slightly better than in the mel-spectrogram domain. By comparing the results in Table 1 and Table 2, we can also see that performing speech separation when the acoustic model is trained with multi-conditional data can still bring us a decent amount of improvement.

4.3. Joint training

In Table 3, we present the joint training results on the test set. In this experiment, T-F masking is performed in the power spectrogram domain and the acoustic model is trained with the NMS+fIRM+MRCG+PNCC feature. To figure out whether learning parameters of the filterbank layer will help, we first fix the filterbank layer to be the mel-filterbank, and only jointly train the acoustic model and the frontend. The performance is 0.3 percent worse than joint training on all of them, which suggests that learning the filterbank helps a little. The final system achieves 14.1 percent average WER on the test set, which is absolute 1.3 percent better than the previous best result [9] on this dataset (or 8.4% relative improvement). We also point out that, by comparing the first row of Table 3 with the last row of Table 2, using joint training can improve the average WER by 0.6 percent. This is probably because of the reduction of the distortion problem and the linguistic information propagated back from the acoustic model.

It might be argued that joint training of the separation frontend, filterbank and acoustic model is basically the same as training a deeper and bigger DNN-based acoustic model with multi-conditional data. To address this possibility, we train a DNN with 11 hidden layers and 1746 units in each layer using the NMS+fIRM+MRCG+PNCC feature for a maximum of 80 epochs as a comparison. Note that the number of parameters and other setup in this large DNN are almost the same as our jointly trained DNN. With this new DNN, as shown in Table 3, we can only obtain 16.1 percent average WER on the test set. The superiority of our approach is probably because of better network architecture and better parameter initialization.

5. Conclusions and Future Work

We have found that performing T-F masking in the power spectrogram domain is slightly better than in the mel-spectrogram domain. We have proposed a novel joint training approach that jointly adjusts the frontend, filterbank and acoustic model to alleviate the distortion problem. Furthermore, we suggest adding more features for acoustic modeling when using multi-condition training, which leads to significant improvements compared with only using the mel-spectrogram feature. Since the CHIME-2 corpus is noisy and reverberant, more experiments are needed to verify that the robust features used in this study can generalize to other datasets such as the Aurora-4 corpus which is noisy and has channel distortions. At a minimum, adding more robust features to acoustic models trained with multi-condition training appears to be a simple and effective technique towards improved robustness of ASR systems.

6. Acknowledgements

The authors would like to thank Arun Narayanan for helpful discussions. This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NSF grant (IIS-1409431), and the Ohio Supercomputer Center.

7. References

[1] T.N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. Mohamed, G. Dahl, and B. Ramabhadran, "Deep Convolutional Neural Networks for Large-scale Speech Tasks," *Neural Networks*, vol. 64, pp. 39-48, 2015.

[2] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645-6649, 2013.

[3] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[4] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?" in *Proceedings of Interspeech*, pp. 2992-2996, 2013.

[5] A. Narayanan and D.L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol 22, no. 4, pp. 826-835, 2014.

[6] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7398-7402, 2013.

[7] M. L. Seltzer, "Robustness is dead! Long live robustness!" *Reverb. Challenge Workshop*, 2014.

[8] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol 22, no. 4, pp. 745-777, 2014.

[9] A. Narayanan and D.L. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no.1, pp. 92-101, 2015.

[10] T. N. Sainath, B. Kingsbury, A. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 297-302, 2013.

[11] C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent deep neural networks for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5532-5536, 2014.

[12] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 8, pp. 1296-1305, 2014.

[13] J. T. Geiger, F. Weninger, J. F. Gemmeke, M. Wöllmer, B. Schuller, and G. Rigoll, "Memory-enhanced neural networks and NMF for robust ASR," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 6, pp. 1037-1046, 2014.

[14] S. Rennie, V. Goel, and S. Thomas, "Annealed dropout training of deep networks," *IEEE Workshop on Spoken Language Technology*, 2014.

[15] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning,* vo. 2, no. 1, pp. 1-127, 2009.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.

[17] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *The Journal of the Acoustical Society of America,* vol. 95, no. 3, pp. 1593-1602, 1994.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.

[19] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4101-4104, 2012.

[20] J. Chen, Y. Wang, and D.L. Wang, "A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1993-2002, 2014.

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

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

[23] Y. Wang, A. Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849-1858, 2014.

[24] A. Narayanan and D.L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7092-7096, 2013.

[25] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, 2011.

[26] Y. Wang, Kun Han, and D.L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 21, no. 2, pp. 270-279, 2013.

[27] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, CRC, 2007.

[28] S.-Y. Chang and N. Morgan, "Robust CNN-based Speech Recognition With Gabor Filter Kernels," in *Proceedings of Interspeech,* pp. 905-909, 2014.