A TWO-STAGE ALGORITHM FOR NOISY AND REVERBERANT SPEECH ENHANCEMENT

Yan Zhao[1] *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
{zhao.836, wang.5664, wang.77}@osu.edu

ABSTRACT

In daily listening environments, speech is commonly corrupted by room reverberation and background noise. These distortions are detrimental to speech intelligibility and quality, and also severely degrade the performance of automatic speech and speaker recognition systems. In this paper, we propose a two-stage algorithm to deal with the confounding effects of noise and reverberation separately, where denoising and dereverberation are conducted sequentially using deep neural networks. In addition, we design a new objective function that incorporates clean phase information during training. As the objective function emphasizes more important time-frequency (T-F) units, better estimated magnitude is obtained during testing. By jointly training the two-stage model to optimize the proposed objective function, our algorithm improves objective metrics of speech intelligibility and quality significantly, and substantially outperforms one-stage enhancement baselines.

Index Terms— phase, ideal ratio mask, deep neural networks, spectral mapping, speech enhancement

1. INTRODUCTION

In real-world environments, speech is distorted by both room reverberation and background noise. Such distortions together cause corrupting effects on speech, severely degrading speech intelligibility for human listeners, especially for hearing-impaired (HI) listeners [1]. Many applications, such as automatic speech recognition (ASR) and speaker identification (SID), also become much more challenging in such adverse conditions [2, 3, 4]. Therefore, better denoising and dereverberation will benefit not only human listeners but also many speech processing tasks.

In recent years, deep neural networks (DNNs) have been employed for speech enhancement or separation. Substantially better performance over conventional methods has been reported in many studies [5, 6, 7, 8]. In [9], Han *et al.* propose a spectral mapping algorithm to perform denoising and dereverberation simultaneously using a single DNN. Their key idea is to learn a mapping from the spectrum of noisy reverberant speech to that of clean anechoic speech. However, informal listening suggests that there is no improvement on speech intelligibility. Zhao *et al.* [10] point out that this is likely because of the different natures of the two distortions, which makes them difficult to address together. In their experiments, they only learn a mapping function to the spectrum of clean reverberant speech, without dereverberation. On this simpler task, they report speech intelligibility improvements for HI listeners in some noisy and reverberant conditions.

We believe that denoising and dereverberation should be addressed separately, due to their different natures. Thus, we propose a two-stage system to enhance noisy and reverberant speech. We first build two DNN-based sub-systems that are trained for denoising and dereverberation. Then, we concatenate these two DNNs to perform joint training. It is worth noting that the strategy of performing denoising and dereverberation in a step by step fashion was adopted previously [11]. Different from earlier studies, we use DNNs for denoising and dereverberation and perform joint training. Joint optimization strengthens the coupling of the two sub-systems, resulting in better performance.

Furthermore, motivated by the time-domain signal reconstruction technique [12], we propose a new objective function that incorporates clean phase information to compute the mean squared error (MSE) in the time domain. We find that this new objective function leads to consistently better performance in objective speech intelligibility and quality metrics.

The rest of this paper is organized as follows. In the next section, we describe the proposed two-stage enhancement system and objective function. Experimental setup and evaluation results are presented in Section 3 and Section 4. We conclude this paper in Section 5.

2. ALGORITHM DESCRIPTION

Fig. 1 shows the diagram of the proposed two-stage speech enhancement system. The system consists of three modules: a denoising module, a dereverberation module and a time-domain signal reconstruction (TDR) module. Note that the TDR module is only utilized at the training stage. At the test stage, the enhanced time-domain signal is resynthesized by using Griffin-Lim's iterative signal reconstruction method [13] with the noisy and reverberant phase as the initial phase.

2.1. Problem formulation

Let $s(t)$, $x(t)$, $n(t)$ and $h(t)$ denote anechoic speech, reverberant speech, background noise and room impulse response (RIR), respectively. The noisy and reverberant speech $y(t)$ is modeled by

$$y(t) = x(t) + n(t) = s(t) * h(t) + n(t) \tag{1}$$

where $*$ stands for a convolution operation. The objective of this study is to recover the anechoic signal $s(t)$ from the noisy and reverberant observation $y(t)$. This mathematical model suggests the order of denoising and dereverberation. Since $n(t)$ is uncorrelated with the desired signal $s(t)$, it is natural to remove the noise first and then to recover anechoic speech.

This research was supported in part by an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

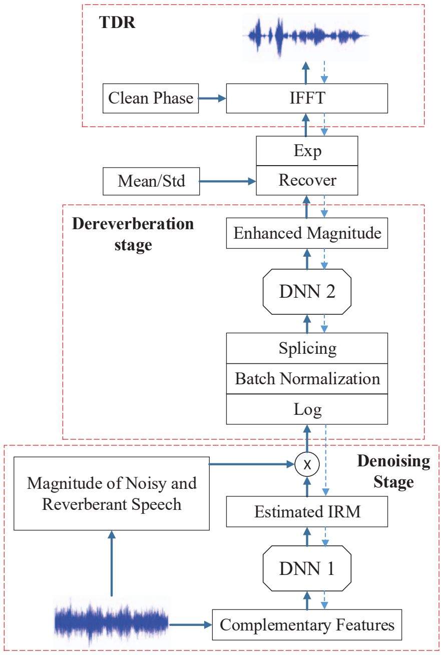


Fig. 1: System diagram of the proposed two-stage model.

2.2. Denoising stage

Given a noisy and reverberant utterance, the aim of this stage is to remove the background noise while keeping the reverberation untouched. Time-frequency (T-F) masking is a common way to suppress noise. Typically, the ideal ratio mask (IRM) is estimated by employing supervised learning approaches. The predicted mask is then applied to the T-F representation of noisy speech to perform enhancement. Recent studies [14] using DNNs to estimate the IRM for segregating speech from noise have shown substantial speech intelligibility improvements for HI listeners. In addition, within the DNN-based speech enhancement framework, an alternative method is to directly estimate the log-magnitude or log-power spectrum of clean speech [8]. However, a study on training targets [15] suggests that masking-based targets outperform mapping-based ones in both objective speech intelligibility and quality. With room reverberation, our previous work [10] also indicates that the adoption of masking-based targets can bring significant performance improvements over mapping-based targets.

Based on the above observations, we employ a DNN with 3 hidden layers to predict the IRM to remove the noise from noisy and reverberant speech. The IRM is defined as follows [15]

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

where $X^2(m,f)$ and $N^2(m,f)$ denote the energy of reverberant speech and background noise, respectively, at time frame m and frequency channel f. As shown in Fig. 1, the magnitude spectrogram of noisy and reverberant speech is multiplied by the estimated mask to form the input features for the next stage processing.

A set of complementary features is adopted as inputs for this stage [16], i.e., 15-dimensional amplitude modulation spectrogram (AMS), 31-dimensional Mel-frequency cepstral coefficients (MFCC), 13-dimensional relative spectral transform perceptual linear prediction (RASTA-PLP), 64-dimensional Gammatone filterbank power spectra (GF), and their deltas. Therefore, for each time frame, the feature dimension is 246 (2×(15+31+13+64)). It is worth noting that this set of features is originally chosen for denoising in anechoic environments.

2.3. Dereverberation stage

After the reduction of background noise, the problem is simplified as recovering the anechoic speech $s(t)$ from the reverberant speech $x(t)$. To perform dereverberation in this stage, we follow the spectral mapping method proposed by Han *et al.* [17]. Compared with the original spectral mapping algorithm, our dereverberation system has two major differences. Firstly, instead of using percent normalization, we normalize the training target, log-magnitude spectrogram of clean anechoic speech, to zero mean and unit variance as suggested in [8]. Secondly, we use the IRM-processed magnitude spectrogram of noisy and reverberant speech for feature extraction to train the dereverberation DNN. Log compression and mean-variance normalization are applied to the features before splicing adjacent frames. By using IRM-processed features, we expect closer coupling between the separately trained denoising stage and dereverberation stage, which can be beneficial for joint training. The DNN used in this stage has 3 hidden layers as well.

2.4. Time-domain signal reconstruction with clean phase

Most supervised learning based separation systems perform enhancement on the magnitude spectrum and use the noisy phase to synthesize the time-domain signal. In order to alleviate the mismatch between the enhanced magnitude and the noisy phase, Wang and Wang [12] employ a DNN to learn to perform TDR given the noisy phase. Improvements on objective speech quality are reported by using their method. Similarly, Erdogan *et al.* [18] propose to predict a phase-sensitive mask. However, with the noisy and reverberant phase, Wang and Wang's approach could be problematic, since the phase is corrupted more seriously. On the other hand, the magnitude and phase spectra carry complementary information [19], which implies that phase can be potentially utilized to help us obtain better magnitude enhancement. Inspired by these observations, we extend Wang and Wang's TDR method and propose a new objective function. More specifically, during training, we feed the enhanced magnitude (after denoising and dereverberation) to an inverse fast Fourier transform (IFFT) layer to reconstruct the enhanced time-domain signal with clean phase, and then optimize the loss in the time domain. During testing, the IFFT layer is removed and the enhanced signal is resynthesized by using Griffin-Lim's method. While the phase-sensitive method also utilizes clean phase information by incorporating the phase difference between clean speech and corrupted speech into an objective function, our proposed method directly employs the clean phase and isolates the influence of corrupted phase.

Mathematically, at time frame m, let $\boldsymbol{s}$, $\hat{\boldsymbol{S}}$ and $\boldsymbol{p}_c$ denote the windowed clean anechoic signal segment, corresponding enhanced magnitude after two-stage processing and clean phase, respectively. Θ denotes the parameters of learning system. Then, the objective function at the training stage is defined as follows,

$$\mathcal{L}(\boldsymbol{s},\hat{\boldsymbol{S}};\Theta) = \|\boldsymbol{s} - IFFT(\hat{\boldsymbol{S}} \circ e^{j\boldsymbol{p}_c})\|_2^2 \tag{3}$$

where $\circ$ denotes the element-wise multiplication and $\|\cdot\|_2$ denotes the L_2 norm.

From another perspective, many supervised learning based speech enhancement systems consider all the T-F units of the same importance and ignore the underlying energy of the corrupted or desired signal in each T-F unit [20]. In the proposed objective function, computing the loss in the time domain will force the learning machine to implicitly place more emphasis on the T-F units that contribute more to the time-domain signal. In other words, instead of weighting T-F units explicitly using normalized mixture energy [21] or mixture energy [6], our method weighs different units on the basis of their corresponding time-domain signal.

2.5. Joint training

As shown in Fig. 1, we concatenate the denoising DNN and the dereverberation DNN into a bigger network for joint optimization. In the denoising stage, the estimated IRM is applied to the magnitude spectrogram of noisy and reverberant speech. The enhanced magnitude is then passed through a log function to compress the dynamic range. We add a batch normalization layer [22] before the splicing operation to make sure the input to the dereverberation DNN is properly normalized. During training, this layer keeps exponential moving averages on the mean and standard deviation of each mini-batch. During testing, such running mean and standard deviation are fixed to do normalization. The normalized features of 11 frames (see Section 3) are spliced as the input features to the dereverberation DNN. After the dereverberation stage, the enhanced log-magnitude is recovered by using the standard deviation and mean of clean anechoic log-magnitude, as we have normalized the target of dereverberation DNN before training. These statistics are computed from the training data. Finally, after an exponential operation, the processed magnitude is fed to the IFFT layer to get the time-domain signal. The loss is computed by (3). Since each step above is differentiable, we can derive the error gradients to jointly train the whole system.

Before joint training, the denoising DNN and the dereverberation DNN are trained separately, and the resulting parameters are used to initialize the two-stage speech enhancement system.

3. EXPERIMENTAL SETTINGS

The proposed method is evaluated on the IEEE corpus [23] spoken by a female speaker. This corpus consists of 72 phonetically balanced lists, each containing 10 sentences. Sentences from list 1-50, list 68-72 and list 51-60 are selected to generate training data, validation data and test data, respectively. One simulated room with size $10\ m \times 7\ m \times 3\ m$ is used to generate RIRs. We generate different RIRs with the position of receiver (an omnidirectional microphone) fixed and the position of speaker randomly chosen. Moreover, we also keep the distance between the receiver and the speaker to be $1\ m$, so that the direct to reverberant ratio (DRR) does not change under each T_{60}. In our experiments, three values of T_{60} are investigated, i.e., 0.3 s, 0.6 s and 0.9 s. For each T_{60}, 10 RIRs are generated for the training and validation sets; 1 RIR is generated for the test set. We utilize an RIR generator [24] to produce the RIRs, which uses the image method [25]. In summary, we have 500×3 (T_{60}s)×10 (RIRs) = 15 k reverberant utterances in the training set, 50×3 (T_{60}s)×10 (RIRs) = 1.5 k reverberant utterances in the validation set, and 100×3 (T_{60}s)×1 (RIR) = 300 reverberant utterances in the test set.

Two kinds of noises including babble noise and speech shaped noise (SSN) are studied. Both of them are about 10 min long. Random cuts from the first 8 min and the remaining 2 min of each noise are mixed with the reverberant speech at a specified signal-to-noise ratio (SNR) to generate the noisy and reverberant speech for the training/validation set and test set, respectively. Three levels of SNRs are considered, namely, -5 dB, 0 dB and 5 dB. Note that the reverberant speech is taken as the signal when calculating the SNR. Consequently, for each type of noise, there are 15 k×3 (SNRs) = 45 k utterances for training, 1.5 k×3 (SNRs) = 4.5 k utterances for validation, and 300×3 (SNRs) = 900 utterances for testing. Neither the noises nor the RIRs of test data are seen during training.

In our experiments, signals are sampled at 16 kHz. A 20-ms Hamming window is applied to divide the signal into frames, with a 50% overlap between adjacent frames. For the time-domain optimization, the clean anechoic signal segment of each frame is also windowed by a Hamming window. We use a 320-point fast Fourier transform (FFT) analysis, resulting in 161 frequency bins. 5 frames on each side of the current frame and itself (11 frames in total) are combined as a context window to incorporate the temporal information. The 11-frame context window is suggested by [9]. We employ overlap-add (OLA) method with Griffin-Lim's phase enhancement algorithm to resynthesize the time-domain signal. The number of iterations is set to 20.

For DNN training, the input features are normalized to zero mean and unit variance by using the statistics of the training data. All DNNs are trained with exponential linear units (ELUs) [26], which lead to faster convergence and better performance over rectified linear units (ReLUs) [27], especially when the networks become deeper. In each hidden layer, there are 1024 hidden units. We utilize Adam [28] as the optimizer to train the networks. Dropout regularization [29] is adopted to prevent overfitting. The dropout rates for the input layer and all the hidden layers are set to 0.2. The hyperparameters are chosen according to the performance on the validation data. For the ratio mask target which is bounded by [0,1], we employ sigmoid activation units in the output layer; for the others, linear activation functions are used.

4. EVALUATION RESULTS

In this study, two objective metrics, short-time objective intelligibly (STOI) [30] and perceptual evaluation of speech quality (PESQ) [31], are employed to evaluate speech intelligibility and quality, respectively. The value range for STOI is between 0 and 1, and for PESQ, it is between -0.5 and 4.5. For both metrics, the higher is the better. The clean anechoic speech is used as the reference signal.

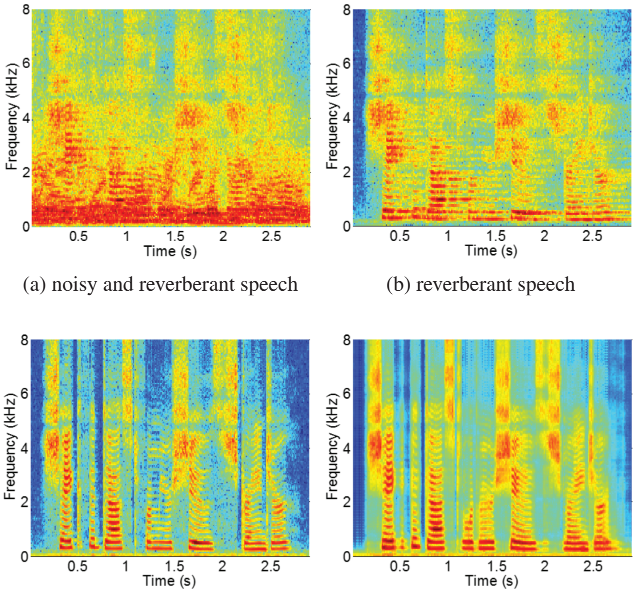
We have two baseline systems to compare. One is using the spectral mapping method, denoted as "**mapping**" for convenience. Same as the dereverberation stage, we normalize the target log-magnitude of clean anechoic speech to zero mean and unit variance. The other baseline system is denoted as "**masking**". Simply speaking, we utilize a DNN with the complementary features to predict the IRMs which are constructed by taking the clean anechoic speech as desired signal and the rest as interference. In order to maintain the same network depth with our proposed two-stage system, for the baseline systems, we employ DNNs with 6 hidden layers. To investigate the proposed objective function, we also add the TDR module to the masking baseline. This method is denoted as "**masking+TDR**" (the network structure is similar to that proposed in [12]). Note that the network is initialized by using the parameters of the masking baseline. We denote the proposed two-stage system as "**two-stage+TDR**". In order to investigate how much performance gain the two-stage strategy alone can bring, another two-stage system without TDR module is also included in the experiments. This method is denoted as "**two-stage**".

	STOI (in %)										**PESQ**									
T_{60} (s)	**0.3**			**0.6**			**0.9**			**Avg.**	**0.3**			**0.6**			**0.9**			**Avg.**
SNR (dB)	**-5**	**0**	**5**	**-5**	**0**	**5**	**-5**	**0**	**5**		**-5**	**0**	**5**	**-5**	**0**	**5**	**-5**	**0**	**5**	
unprocessed	59.0	67.0	74.6	57.1	64.3	70.9	53.8	59.8	65.0	63.5	0.817	1.202	1.611	0.778	1.103	1.462	0.749	1.006	1.301	1.114
mapping	75.7	81.8	85.5	73.2	80.2	83.8	69.2	76.8	80.4	78.5	1.836	2.192	2.453	1.722	2.100	2.310	1.573	1.906	2.112	2.023
masking	78.6	83.7	86.8	76.3	81.9	85.0	73.3	79.1	82.0	80.7	1.967	2.293	2.568	1.873	2.172	2.399	1.717	2.013	2.202	2.134
masking+TDR	80.0	85.2	88.2	77.8	83.5	86.7	74.8	80.8	83.8	82.3	2.054	2.374	2.648	1.950	2.248	2.477	1.788	2.091	2.274	2.212
two-stage	81.3	86.4	89.0	79.0	84.9	87.8	75.8	82.1	84.7	83.4	2.182	2.549	**2.779**	2.061	2.412	**2.612**	1.895	2.227	**2.414**	2.348
two-stage+TDR	**82.9**	**87.6**	**90.0**	**80.8**	**86.3**	**88.9**	**77.9**	**83.5**	**85.9**	**84.9**	**2.221**	**2.562**	2.759	**2.119**	**2.443**	**2.612**	**1.960**	**2.244**	2.408	**2.370**

Table 1: STOI and PESQ scores at each condition for SSN.

	STOI (in %)										**PESQ**									
T_{60} (s)	**0.3**			**0.6**			**0.9**			**Avg.**	**0.3**			**0.6**			**0.9**			**Avg.**
SNR (dB)	**-5**	**0**	**5**	**-5**	**0**	**5**	**-5**	**0**	**5**		**-5**	**0**	**5**	**-5**	**0**	**5**	**-5**	**0**	**5**	
unprocessed	55.4	64.4	72.9	53.6	61.6	69.3	50.6	57.5	63.7	61.0	1.015	1.297	1.654	0.943	1.200	1.506	0.833	1.089	1.339	1.208
mapping	72.3	79.9	84.2	70.8	78.7	82.6	66.5	74.7	78.7	76.5	1.723	2.117	2.389	1.684	2.048	2.282	1.475	1.833	2.063	1.957
masking	75.5	82.4	86.3	73.0	80.5	84.4	69.3	76.9	80.8	78.8	1.841	2.200	2.503	1.729	2.087	2.353	1.564	1.913	2.159	2.039
masking+TDR	78.0	84.3	87.9	75.7	82.8	86.4	72.0	79.3	83.1	81.1	1.962	2.334	2.647	1.851	2.211	2.473	1.674	2.010	2.256	2.158
two-stage	81.4	86.3	88.8	79.3	84.9	87.6	75.8	81.5	84.5	83.3	2.219	**2.576**	**2.775**	2.101	**2.438**	**2.620**	1.919	**2.226**	**2.414**	2.365
two-stage+TDR	**83.6**	**87.6**	**90.0**	**81.6**	**86.3**	**88.9**	**78.3**	**83.1**	**85.8**	**85.0**	**2.272**	2.573	2.761	**2.139**	2.433	2.608	**1.951**	2.223	2.395	**2.373**

Table 2: STOI and PESQ scores at each condition for babble noise.



(a) noisy and reverberant speech

(b) reverberant speech

(c) anechoic speech

(d) enhanced speech

Fig. 2: (Color online) Example spectrograms of noisy and reverberant speech (babble noise, SNR = -5 dB, T_{60} = 0.9 s), reverberant speech (T_{60} = 0.9 s), anechoic speech and enhanced speech (two-stage+TDR).

Table 1 and Table 2 list the STOI and PESQ values of unprocessed and processed signals under different noisy and reverberant conditions. Boldface number highlights the best result of each condition. Similar performance trends can be observed for SSN and babble noise. Clearly, each approach improves STOI and PESQ substantially. By switching to the masking-based target and using the complementary features, we find a performance boost in both STOI and PESQ, which is consistent with the denoising results reported in [15, 10].

Taking the masking method as the stronger baseline, we study whether our proposed methods can further improve the performance. We use the average performance of each approach for comparison. Firstly, combining the masking method with the proposed TDR module can bring us 1.6% and 2.3% STOI improvements for SSN and babble noise, respectively. Some improvements on PESQ are also observed. These results suggest that the new objective function provides an effective way to improve the current supervised speech enhancement system. Secondly, more performance gains are obtained by employing the two-stage strategy. Specifically speaking, for SSN, additional 2.7% STOI and 0.214 PESQ scores are gained over the masking method; for babble noise, we get 4.5% STOI and 0.326 PESQ improvements. Finally, the two-stage system with the TDR module (two-stage+TDR) performs best in terms of STOI. Compared with the masking baseline, 4.2% and 6.2% STOI improvements are obtained for SSN and babble noise, respectively. Interestingly, two-stage+TDR method only brings some slight PESQ improvements over two-stage method in low SNR conditions.

Fig. 2 gives an enhancement example of the sentence "Shake the dust from your shoes, stranger". Fig. 2(a) presents the spectrogram of noisy and reverberant speech with babble noise at SNR = -5 dB and reverberation time at 0.9 s. Figs. 2(b), (c) and (d) show the corresponding spectrograms of reverberant speech, anechoic speech and speech enhanced by the proposed algorithm (two-stage+TDR), respectively. As shown in Fig. 2(d), additive noise and smearing effects caused by reverberation have been largely removed, and the spectrotemporal patterns are much restored, demonstrating that the proposed algorithm can effectively enhance noisy and reverberant speech.

5. CONCLUSION

In this paper, we have proposed a two-stage system aiming to enhance speech in noisy and reverberant environments. Two DNN sub-systems are utilized to perform denoising and dereverberation separately, and then form a coherent system by joint optimization. In addition, we have developed a new objective function for supervised speech separation, which incorporates clean phase. Systematic evaluation using objective metrics indicates that the proposed system should improve speech intelligibility and quality in a wide range of noisy and reverberant conditions.

6. REFERENCES

[1] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *Journal of Speech, Language, and Hearing Research*, vol. 53, pp. 1429–1439, 2010.

[2] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition.," in *INTERSPEECH*, 2002.

[3] 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, pp. 745–777, 2014.

[4] K. A. Al-Karawi, A. H. Al-Noori, F. F. Li, and T. Ritchings, "Automatic speaker recognition system in adverse conditions - implication of noise and reverberation on system performance," *International Journal of Information and Electronics Engineering*, vol. 5, pp. 423–427, 2015.

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

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

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

[8] 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*, vol. 23, pp. 7–19, 2015.

[9] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 982–992, 2015.

[10] Y. Zhao, D. L. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6525–6529.

[11] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech dereverberation in noisy reverberant environment.," in *INTERSPEECH*, 2007, pp. 854–857.

[12] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4390–4394.

[13] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–243, 1984.

[14] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, pp. 2604–2612, 2016.

[15] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[16] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 270–279, 2013.

[17] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4628–4632.

[18] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[19] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[20] Z.-Q. Wang, Y. Zhao, and D. L. Wang, "Phoneme-specific speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 146–150.

[21] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 625–638, 2009.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[23] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, pp. 225–246, 1969.

[24] E. Habets, "Room impulse response generator," Available at `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[26] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.

[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

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

[29] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[30] 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, pp. 2125–2136, 2011.

[31] 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*, 2001, vol. 2, pp. 749–752.