Two-Stage Multi-Target Joint Learning for Monaural Speech Separation

Shuai Nie[1], Shan Liang[1], Wei Xue[1], Xueliang Zhang[2], Wenju Liu[1], Like Dong[3], Hong Yang[3]

[1]National Laboratory of Patten Recognition, Institute of Automation, Chinese Academy of Sciences
[2]College of Computer Science, Inner Mongolia University
[3]Electric Power Research Institute of ShanXi Electric Power Company, China State Grid Corp

{shuai.nie, sliang, wxue, lwj}@nlpr.ia.ac.cn cszxl@imu.edu.cn

Abstract

Recently, supervised speech separation has been extensively studied and shown considerable promise. Due to the temporal continuity of speech, speech auditory features and separation targets present prominent spectro-temporal structures and strong correlations over the time-frequency (T-F) domain, which can be exploited for speech separation. However, many supervised speech separation methods independently model each T-F unit with only one target and much ignore these useful information. In this paper, we propose a two-stage multi-target joint learning method to jointly model the related speech separation targets at the frame level. Systematic experiments show that the proposed approach consistently achieves better separation and generalization performances in the low signal-to-noise ratio(SNR) conditions.

Index Terms: speech separation, multi-target learning, computational auditory scene analysis (CASA)

1. Introduction

In real-world environments, the background interference substantially degrades the speech intelligibility and the performance of many applications, such as speech communication and automatic speech recognition (ASR) [1, 7, 12, 18]. To address this issue, the speech separation, which aims to extract the speech signal from the mixture, has been studied for decades. However, it is still a challenging task to achieve effective speech separations in real-world environments, especially when the signal-to-noise ratio (SNR) is low and only one microphone is available.

Speech separation can be formulated as a supervised learning problem [12, 24, 26]. Typically, a supervised speech separation learns a function that maps the noisy features extracted from the mixture to certain ideal masks or clean spectra that can be used to separate the target speech from the mixture. As a new trend, compared to the traditional speech enhancement [13], supervised speech separation has shown to be substantially promising for challenging acoustic conditions [12,24,26].

Supervised speech separation has two main types of training targets, i.e. the mask-based targets [23] and spectra-based one [26]. For the mask-based targets, the algorithm learns the best approximation of an ideal mask computed using the clean and noisy speech, such as the ideal ratio mask(IRM) [14, 25], while for the spectra-based targets, it learns the best approximation of the clean speech spectra, such as the Gammatone frequency power spectrum(GF) [9]. Both the IRM and GF can be used to generate the separated speech with the improved intelligibility and/or perceptual quality [23]. Intuitively, the IRM and the GF of clean speech present similar spectro-temporal structures as is shown by the example in Fig. 1. In fact, mathematically, the IRM can be derived from the GFs of clean speech and noise, which is computed as follows:

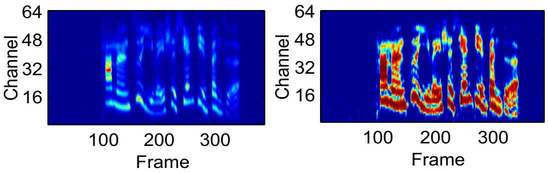

Figure 1: Left: the GF of clean speech; Right: the IRM computed with the clean speech and white noise (mixed at 0 dB)

$$IRM(t, f) = \frac{S^2(t, f)}{S^2(t, f) + N^2(t, f)} \tag{1}$$

where $S^2(t, f)$ and $N^2(t, f)$ are the GFs of clean speech and noise in the time-frequency (T-F) unit of channel f and frame t, respectively. Moreover, due to the sparsity of speech in the T-F domain, the GF keeps relatively invariant harmonic structure in various auditory environments, and the IRM is inherently bounded and less sensitive to estimation errors [15]. These correlations and complementarity can be exploited for speech separation. But they are much ignored in previous works. Therefore, jointly modeling the IRM and GF in one model will probably improve the separation performance.

In this paper, we propose a multi-target deep neural network (DNN) to jointly model the IRM and GF. Its target is the combination of the IRM and the GF of clean speech. To further improve the separation performance, a two-stage method is explored. In the first stage, the multi-target DNN is trained to learn a function that maps the noisy features to the joint targets for all frequency channels in one frame. Compared to the individual T-F unit, modeling at the frame level can capture the correlations over the frequency domain in speech. Moreover, to exploit the spectro-temporal structures in speech auditory features and joint targets, we use denoising autoencoders (DAE) to model them by self-learning, respectively. Then, the learned DAEs are combined with a linear transformation matrix W_h to initialize the multi-target DNN. Finally, according to the different errors produced by output nodes, a backpropagation (BP) algorithm with bias weights is further explored to fine tune the multi-target DNN. In the second stage, the estimated IRM and GF are integrated into another DNN to obtain the final separation result with higher smoothness and perceptual quality.

This research was partly supported by the China National Nature Science Foundation (No.91120303, No.61273267, No.90820011, No.61403370 and No.61365006).

2. First Stage: Multi-Target Joint Learning

Typically, it has been recognized that by the multi-task joint learning, related tasks which share some common information, can be jointly modeled to improve the performance of each other [8]. This claim has been proven both empirically and theoretically [2]. For speech separation, as has been discussed before, the IRM and GF are highly related and have much shared information with respect to the speech separation problem. Therefore, it can be expected that the performance of speech separation can be improved by jointly modeling these two related tasks. Recently, a multi-target DNN has been extensively studied and achieved remarkable success in many applications [5]. Same with the general DNN, the multi-target DNN also consists of one input layer, multiple hidden layers and one output layer. The key difference is that the output layer of multi-target DNN is composed of multiple related targets. In this paper, in order to improve the performance of speech separation, we construct a multi-target DNN to jointly learn the IRM and the GF of clean speech. Specifically, in each frame, the IRM and GF are combined into one vector, and its each element corresponds to one output unit of the multi-target DNN.

2.1. Learning and mapping of spectro-temporal structures

As speech auditory features might be severely interfered by noise at low SNR conditions, it is very difficult to directly learn a map from noisy features to separation targets. However, due to linguistic constraints and speech production mechanisms, both speech auditory features and separation targets present prominent spectro-temporal structures that keep relatively invariant to various acoustic conditions [25]. Compared to the direct mapping from features to targets, these output structures could be used to regularize the learning and make the mapping more robust [25]. In this paper, we use two DAEs to exploit the spectro-temporal structures in speech auditory features and separation targets. Through self-learning, one DAE is trained on the auditory features and the other is trained on the joint targets of the IRM and GF. To capture temporal structures, the DAEs are trained on a window of frames instead of single time slices. Once the DAE is trained, the outputs of its coding layers can be regarded as the learned structured patterns from data [3]. Then the structured patterns learned from auditory features can be mapped into those learned from separation targets through a simple linear transformation, and the linear transformation weights $\mathbf{W}_h$ can be directly computed by $\mathbf{W}_h = (\mathbf{H}_x{}^T\mathbf{H}_x)^{-1}\mathbf{H}_x{}^T\mathbf{H}_y$, where $\mathbf{H}_x$ is the output matrix of the last coding layer in the DAE trained on the auditory features and $\mathbf{H}_y$ is the output matrix of the last coding layer in the DAE trained on the joint targets of the IRM and GF. The element of row vector in $\mathbf{H}_x$ and $\mathbf{H}_y$ corresponds to the output of one note in the last coding layer of the corresponding DAE. In fact, the above processes can be seen as the pre-training step of the multi-target DNN. The learned weights $\mathbf{W}_x$ and $\mathbf{W}_y$ in the two DAEs, as well as the mapping weights $\mathbf{W}_h$, can be used to initialize the corresponding weights in the multi-target DNN, as shown in Fig. 2.

At last, we use a supervised learning algorithm to fine tune the multi-target DNN. Interestingly, even without the fine-tuning step, the multi-target DNN pre-trained by the proposed method can also achieve relatively good separation results in our preliminary experiment, which may owe to that the initial model has the ability in learning and mapping the spectro-temporal structures in auditory features and joint targets.

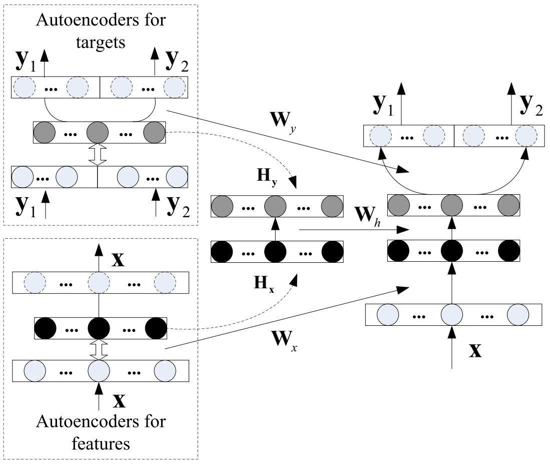

Figure 2: The initialization of the multi-target DNN through the learning and mapping of spectro-temporal structures

2.2. Fine tuning using the BP algorithm with bias weights

In the multi-target learning, the learning of different targets can be differently treated. In fact, each node in the output layer of DNN can be seen as an individual target and the corresponding learning procedure can be differently designed. Intuitively, the nodes with the larger errors probably have greater potential for decreasing the loss, and their learning should be reinforced. To this end, we use bias weights ρ to weight the errors of the output nodes. ρ is related to the errors and can be computed by Eq. (2). Accordingly, the loss function is defined by Eq. (3).

$$\rho = \frac{|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}| - \min(|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}|)}{\max(|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}|) - \min(|\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}|)} \tag{2}$$

$$J(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\sqrt{\rho} \bullet (\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y})\|^2 \tag{3}$$

where $\bullet$ and $|\cdot|$ denote the element-wise matrix multiplication and the absolute value operator, respectively. $\mathbf{W}$ and $\mathbf{b}$ stand for the connection weights and bias in the network, $\mathbf{x}$ and $\mathbf{y}$ are the inputs and the corresponding targets. $\mathbf{h}_{\mathbf{w},\mathbf{b}}(\mathbf{x})$ is the output of the network. ρ is a vector whose elements correspond to the output nodes of the network.

For each output node, the larger the error is, the greater the corresponding element in ρ is. It means that the learning of the node producing larger error will be more emphasized in the next iteration. We should note that the bias weight ρ can be determinately computed be Eq. (2) in each iteration, and it does not need to be optimized. In gradient-based optimization, the commonest method is the steepest gradient descent algorithm which is based on greedy rule [10]. However, it probably falls into local minima and may need more iterations to converge to the optimal point. In this paper, as the red arrow shows in the Fig. 3, we use the bias weights associated with errors to change the direction of the steepest gradient descent towards the direction that is farthest to the global optimal point. Compared to the direction of steepest gradient descent, this direction tends more to the direction of the global optimal point, but ensures the descent of the errors, which may bring faster convergence and avoid falling into some local minima.

To examine the effectiveness of the proposed optimization method, we design a DNN-based speech separation experiment. The input is the noisy feature and the output is the IRM. We

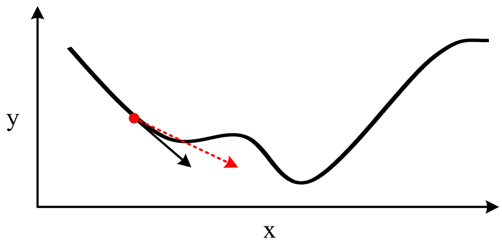

Figure 3: A example of the proposed optimization method. The black arrow is the direction of steepest gradient descent; The red arrow is the direction of the changed gradient with bias weights.

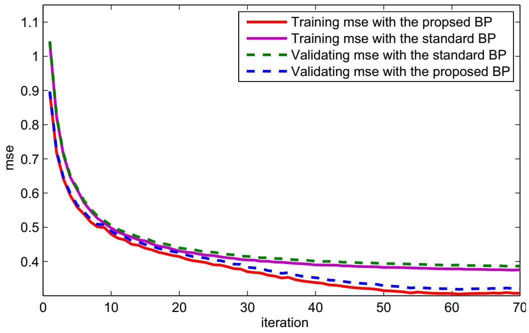

Figure 4: A example of DNN for speech separation.

train two DNNs with same configurations setting and dataset. But one is trained with the standard BP algorithm [17] and the other is trained with the proposed BP algorithm. The mean square errors(MSE) on the training set and the validating set are used to evaluate the performances of the different algorithms and depicted in Fig. 4. The results show that the DNN trained with the proposed BP algorithm can achieve smaller errors on both training set and validating set. The main reason is that the BP algorithm with bias weights may avoid falling into some local minima due to the changed gradient direction.

2.3. Auditory feature extraction

We apply a 64-channel gammatone filterbank to the input signals, and the resulting output of each channel is windowed into 20ms time frames with 50% overlap. Then we extract complementary acoustic features from the mixture speech at the frame level, and they include the amplitude modulation spectrogram(AMS), relative spectral transformed perceptual linear prediction coefficients(RASTA-PLP), mel-frequency cepstral coefficients(MFCC) and 64-channel Gammatone filterbank power spectra(GF) [22]. All features are concatenated with the corresponding delta features and smoothed by a second order auto-regressive moving average(ARMA) [4] filter along temporal trajectories.

3. Second Stage

The IRM and GF estimated by the multi-target DNN in the first stage contain much complementary information that can be exploited to further boost speech separation. In this paper, we use a DNN to integrate the estimated IRM and GF for obtaining the separated speech with higher smoothness and perception quality. As the estimated IRM and GF lose some information of speech, in addition to the estimated IRM and GF, the input features of the DNN are also concatenated with the noisy auditory features. The output of the DNN is the IRM.

4. Experiments and Results

4.1. Dataset and evaluation metrics

We systematically evaluate the proposed approach on Chinese National Hi-Tech Project 863 corpus which consists of 100,000 utterances recorded by 100 female and 100 male speakers. As pilot experiments, the evaluation is performed on a relatively small corpus. Therefore, for training, 100 utterances from 5 male and 5 female speakers, with 10 utterances for each speaker, are randomly chosen to mix with three types of non-speech noises(babble, speech shaped noise and factory) at 0 and -5 dB, respectively. For testing, we randomly choose 50 new utterances from the same 5 female and 5 male speakers to mix with six types of noises(babble, speech shaped noise, factory, traffic, machine and cocktail [6]) at -10,-7, -5, -2 and 0 dB, respectively. We should note that three types of noises in the testing set are not available to the training set, so this experimental setting can evaluate the generalization ability to noise.

We take the absolute gains of Short-Time Objective Intelligibility score (STOI) [19] and SNR compared to the unprocessed noisy speech, as well as the Perceptual Evaluation of Speech Quality (PESQ) [16] for evaluation. These metrics are relative to the speech intelligibility and/or perception quality.

4.2. Related models for comparisons and configurations

To systematically evaluate the proposed model, we choose the single-target DNN-based [23] (denoted as 'ST-DNN') and DNN-NMF-based [25] (denoted as 'DNN-NMF')speech separation models for comparison. For ST-DNN, we use a DNN with three hidden layers which have 320, 320 and 160 sigmoid units, respectively. Its inputs are a 5-frame window of complementary acoustic features described in subsection 2.3 and its outputs are the IRM computed using the clean speech and noise. We should note that the DNN simultaneously models all frequency channels rather than the individual T-F unit. The standard BP algorithm is used to optimize the DNN. We use the mini-batch gradient descent of 500 samples to train it from a initial model that is pre-trained by the stacking DAE with 100 epochs of training on the noisy auditory features [20, 21]. The training of the DNN is performed 100 epochs with a varied learning rate from 0.5 to 0.01 and couples with a dropout regularization (dropout rate 0.2) [11]. A momentum rate of 0.5 is used for the first 5 epochs, after which the rate increases to and is kept as 0.9. For simplicity, unless mentioned explicitly, the DNNs in all experiments are trained with same configurations setting, same dataset and hyper-parameters.

Instead of directly estimating the IRM, DNN-NMF predicts the basis weights inferred by the non-negative matrix factorization (NMF), which can be used to generate the estimated mask. Compared to ST-DNN, this is the only difference. For the comparison, we use the same NMF configuration(e.g. 128 bases and a sliding window of 5 frames) with that in [25] for DNN-NMF.

4.3. Result and evaluation

In this section, we examine the effectiveness of the proposed approaches from several aspects, including the effect of the multi-target of the IRM and GF, the effect of the proposed BP algorithm with bias weights, the effect of the proposed pre-training step described in subsection 2.1, and the effect of integrating the evaluated IRM and GF with a DNN in the second stage.

First, we explore the case of single-target DNN(denoted as ‘ST-DNN’) and the case of multi-target DNN (denoted as ‘MT-DNN’). The difference between ST-DNN and MT-DNN is that ST-DNN only predicts the IRM and MT-DNN synchronously predicts the IRM and the GF of clean speech. The results are shown in the first and second rows in Table 1. We observe that MT-DNN consistently achieves significant improvement on all evaluation metrics in both matched and unmatched noise conditions. This mainly owes to that the IRM and GF contain rich complementary and correlative information that can be exploited by the multi-target DNN for speech separation.

Second, we explore the case of using the proposed BP algorithm with bias weights to optimize the multi-target DNN, denoted as ‘MT-DNN-PW’, and we compare it with MT-DNN. The results are shown in the second and third rows in Table 1. Compared to MT-DNN, MT-DNN-PW achieves further improvement. It suggests that the proposed BP algorithm with bias weights probably make the optimization reach to the better extremal point than the standard BP algorithm. This mainly owes to that the bias weights change the direction of gradient descent toward the direction of the global optimal point.

Third, we explore the case of using the proposed pre-training method to initialize the multi-target DNN, denoted as ‘MT-GM-DNN-PW’, and we compare it with MT-DNN-PW. They both are fine tuned by the proposed BP algorithm with bias weights but have different pre-training ways. MT-DNN-PW is pre-trained by the stacking DAE trained on the noisy auditory features and MT-GM-DNN-PW is pre-trained by the proposed method, and you can return to the subsection 2.1 for the detailed pre-training steps. The results are shown in the third and fourth rows in Table 1. We observe that MT-GM-DNN-PW achieves the better separation performance than MT-DNN-PW in both matched and unmatched noise conditions. On the one hand, DAE has capacity in capturing the spectro-temporal structures in speech auditory features and separation targets for speech separation, on the other hand, due to the similarity and invariance of the structure patterns in the features and targets, the mapping of structure patterns is easier and more robust than the direct mapping from features to targets.

The fifth row in Table 1 presents the results of the case using a DNN to integrate the IRM and GF evaluated by MT-GM-DNN-PW, denoted as ‘TWO-STAGE’. Its inputs consist of the noisy auditory features, the evaluated IRM and GF, and its outputs are the IRM. We observe that TWO-STAGE performs best in matched noise condition and has only a little loss of performance in unmatched noise condition. The improvements mainly owes to that the rich complementary information in the evaluated IRM and GF can be exploited by the DNN for speech separation and the integration of the IRM and GF can smooth the separation. The loss of performance in unmatched noise condition maybe caused by the over-fitting of DNN.

Finally, we compare our methods with DNN-NMF under the same setting. The results of DNN-NMF are shown in the last row in Table 1. We observe that DNN-NMF outperforms ST-DNN but perform worse than the proposed methods. It suggests that NMF captures the spectro-temporal structures in the IRM but has limited capacity in learning structure patterns compared to DAE due to its shallow and linear structure. In addition, DNN-NMF only predicts the IRM, and ignores the correlations and complementarity in the IRM and GF. Moreover, we also observe that DNN-NMF has worse generalization ability to unmatched noise. The possible reason is that the spectro-temporal bases of the IRM learned by NMF is sensitive to noise.

Table 1: The absolute gains on STOI(%) and SNR(dB), as well as the PESQ score that different methods obtains at -5dB SNR.

System	Matched noise			Unmatched noise		
	gSTOI	gSNR	PESQ	gSTOI	gSNR	PESQ
ST-DNN	20.08	17.17	2.14	11.54	10.60	1.32
MT-DNN	21.14	17.67	2.22	12.35	10.75	1.36
MT-DNN-PW	21.60	18.13	2.24	12.38	10.62	1.35
MT-GM-DNN-PW	22.09	18.69	2.21	**12.86**	**10.89**	**1.36**
TWO-STAGE	**23.06**	**19.41**	**2.39**	12.51	10.68	1.30
DNN-NMF	20.48	16.94	2.15	10.18	9.69	1.20

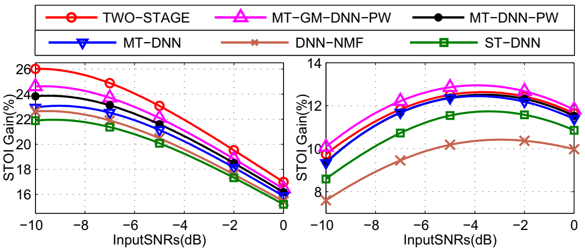

Figure 5: The STOI gains of different systems at different SNRs. Left Fig: the matched-noise case; Right Fig: the unmatched-noise case.

Moreover, we also compare the generalization ability to unmatched SNRs with other methods. As the results show in Fig. 5, the proposed methods perform best in both matched and unmatched noise conditions, which mainly owes to incorporating the spectro-temporal structure in speech and the correlation between the IRM and GF into the supervised speech separation, and these useful information keeps relatively invariant to various SNR conditions due to the sparsity of speech.

5. Conclusion and Related Works

The works presented here mainly focus on the supervised speech separation. Along this research line, many methods have been proposed and achieved considerable success [12, 23, 24, 26]. Among these methods, many methods formulate speech separation as a binary classification problem, such as [12], [24]and [22]. They independently model the individual T-F unit with only one target such as the IBM, and ignore the spectro-temporal structure in auditory features and the correlation of various separation targets. Recently, several supervised speech separation methods are proposed to simultaneously model all T-F units [23, 25] in one frame, and use the IRM as the target. Although they take the correlation over the T-F domain into account, only one separation target is used and the correlation between different separation targets is ignored. Compared to the previous works, the main contributions of our work are: 1) proposing a multi-target DNN to jointly model the IRM and GF for exploiting their correlation. 2) exploring a pre-training method for the multi-target DNN, which can capture the spectro-temporal structures in auditory features and separation targets; 3) proposing a BP algorithm with bias weights to change the direction of steepest gradient descent toward that of the global optimal point, which may avoid to fall into some local minima; 4) integrating the estimated IRM and GF with a DNN to further improve the separation performance.

6. References

[1] J. B. Allen, "Articulation and intelligibility," *Synthesis Lectures on Speech and Audio Processing*, vol. 1, no. 1, pp. 1–124, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[3] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," in *Proc. Advances in Neural Information Processing Systems (NIPS'2013')*, 2013, pp. 899–907.

[4] C.-P. Chen and J. A. Bilmes, "Mva processing of speech features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[5] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. the 25th International Conference on Machine Learning (ICML'2008')*, 2008, pp. 160–167.

[6] M. Cooke, *Modelling auditory processing and organisation*. Cambridge University Press, 2005, vol. 7.

[7] H. Dillon, *Hearing aids*. Thieme, 2001.

[8] A. Evgeniou and M. Pontil, "Multi-task feature learning," *Proc. Advances in Neural Information Processing Systems (NIPS'2008')*, vol. 19, p. 41, 2007.

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

[10] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[11] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[12] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[13] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[14] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2013)*, 2013, pp. 7092–7096.

[15] ——, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 92–101, 2015.

[16] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2001)*, vol. 2, 2001, pp. 749–752.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, 1988.

[18] M. L. Seltzer, B. Raj, and R. M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

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

[20] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. the 25th International Conference on Machine Learning (ICML'2008')*, 2008, pp. 1096–1103.

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

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

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

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

[25] ——, "A structure-preserving training target for supervised speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2014)*, 2014, pp. 6107–6111.

[26] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.