Technical Report OSU-CISRC-4/14-TR11
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2014**
File: **TR11.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

A Neural Network For Time-Domain Signal Reconstruction: Towards Improving The Perceptual Quality Of Supervised Speech Separation

Yuxuan Wang
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
wangyuxu@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive and Brain Sciences
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Abstract – Supervised speech separation has achieved considerable success recently. Typically, a deep neural network (DNN) is used to estimate an ideal time-frequency mask, and clean speech is produced by feeding the mask-weighted output to a resynthesizer in a subsequent step. So far, the success of DNN-based separation lies mainly in improving human speech intelligibility. In this work, we propose a new neural network that directly reconstructs the time-domain clean signal through an inverse fast Fourier transform layer. The joint training of speech resynthesis and mask estimation yields significantly improved objective quality while maintaining the objective intelligibility performance. The proposed system significantly outperforms a recent non-negative matrix factorization based separation system in both objective speech intelligibility and quality.

1 Introduction

Monaural speech separation is a long-standing problem with many important applications, such as robust automatic speech recognition and mobile speech communication. In low signal-to-noise ratio (SNR) conditions, monaural separation is particularly challenging when facing non-stationary noises. Compared to traditional speech enhancement [11], data-driven techniques have shown substantial promise in these challenging acoustic conditions [15, 17].

A successful new trend is supervised speech separation, as exemplified by its recent demonstration in improving speech intelligibility of both normal-hearing [9] and hearing-impaired listeners [4] in noisy environment. In its simplest form, supervised separation learns a mapping from noisy mixtures to an ideal time-frequency (T-F) mask. The estimated ideal mask is then used to weight the mixture in the T-F domain, and the resulting output along with the mixture phase is passed into a separate resynthesizer to produce the time-domain speech signal. Recently proposed deep neural network (DNN) based separation generalizes well to various test conditions if properly trained [17, 20]. Once trained, separation operates in a frame-by-frame fashion, making it amenable to real-time implementation.

To improve the quality of separated speech, this study proposes to directly reconstruct the time-domain clean signal, which is the ultimate target of interest. Although touched upon before [14, 16], using a standard feedforward network to learn the mapping to clean signal does not seem to work well. To tackle this, we propose a new network that incorporates the domain knowledge of speech resynthesis by adding an inverse fast Fourier transform (IFFT) layer before the output layer. The speech resynthesis and mask estimation can now be jointly trained in a single neural network. As a result, the mask is estimated (learned) in a way that directly impacts the final time-domain signal reconstruction, leading to improved quality.

This paper is organized as follows. We briefly review the supervised speech separation framework in the next section. We introduce the proposed network architecture in Section 3, and the experimental results are described in Section 4. The last section concludes this paper.

2 Supervised Speech Separation

Supervised speech separation employs data-driven, supervised learning for the separation task, unlike traditional signal processing methods. First, acoustic features are extracted from the noisy mixture. These features are fed into a learning machine, typically a deep neural network, where training targets are provided by the ideal mask of interest.

The ideal binary mask (IBM) is typically used as the training target due to its simplicity and large intelligibility improvements (e.g. [1, 2, 10]). The IBM is a binary matrix constructed from premixed signals. We set the value of a T-F unit to 1 if the local SNR is greater than a

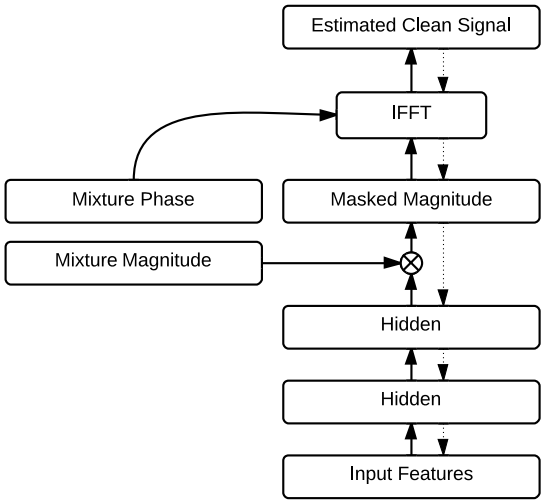

Figure 1: Schematic diagram of the proposed system. As illustrated, the time-domain signal resynthesis module is part of the network and jointly trained with mask estimation via the backpropagation algorithm.

local criterion (denoted as LC) and 0 otherwise. That is:

$$IBM(t,f) = \begin{cases} 1, & \text{if } SNR(t,f) > LC \\ 0, & \text{otherwise,} \end{cases}$$

where $SNR(t,f)$ denotes the local SNR within the T-F unit at time t and frequency f. Estimating the IBM has been shown to improve speech intelligibility [4,9], but not necessarily speech quality [19]. Following common practice, we use a 64-channel Gammatone filterbank to derive the IBM, and set LC to be 5 dB less than the input SNR to preserve adequate speech information.

Alternatively, our recent work [19] suggests to predict the ideal ratio mask (IRM), which is shown to improve both objective intelligibility and quality. The IRM is defined as:

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\beta},$$

where $S^2(t,f)$ and $N^2(t,f)$ denote the speech and noise energy in a particular T-F unit. β is a tunable parameter to scale the mask. One can see that the IRM is closely related to the frequency-domain Wiener filter [11]. In this study, we use $\beta = 1$ as it achieves the best results in terms of speech quality. The IRM is also derived using a 64-channel Gammatone filterbank.

3 Proposed Network Architecture

Figure 1 illustrates the architecture of the proposed system, which is similar to a standard DNN mask estimator with two key differences. First, the process of converting from the frequency domain to time domain is incorporated in the network. This domain knowledge enables the

reconstruction of the final time-domain signal in one pass and makes learning much easier. Second, there is no predefined ideal mask for training. Instead, the last hidden layer is treated as a masking layer and is automatically learned via the backpropagation algorithm. This can be viewed as a form of *task-dependent masking* (see also [12]). For simplicity, masking is carried out in the discrete Fourier transform (DFT) domain, so that resynthesis can be conveniently implemented as an IFFT. In the following, we describe the forward pass and backpropagation of the proposed network in detail.

3.1 Forward Pass

For the input at frame t, we denote $\mathbf{h}^t$ as the corresponding network activations from the last hidden layer, and $\mathbf{y}^t$ the corresponding DFT-domain mixture magnitude. For simplicity, we set the analysis FFT length L to the frame length. Let $d = L/2$, and $\mathbf{y}^t$ is thus a $d+1$ dimensional vector. We treat $\mathbf{h}^t$ as the mask at frame t. Therefore, the masked magnitude, or the estimated clean speech magnitude, is obtained as:

$$\mathbf{m}^t = \mathbf{h}^t \circ \mathbf{y}^t, \tag{1}$$

where $\circ$ denotes element-wise multiplication. Note that for Eq. (1) to be valid, the last hidden layer $\mathbf{h}^t$ must also be of dimension $d+1$, whereas all other hidden layers do not have this constraint. The estimated spectral magnitude along with the corresponding mixture phase are fed into an IFFT layer, generating the time-domain waveform in frame t at the output layer of the network. Specifically, the estimated clean speech $\hat{\mathbf{s}}^t$ is obtained as follows:

$$\hat{\mathbf{s}}^t = \text{IFFT}\left(\left[\mathbf{c}^t, \text{flipud}\left(\text{conj}\left(\mathbf{c}^t_{2:d}\right)\right)\right]^T\right), \tag{2}$$

where $\mathbf{c}^t$ is the complex spectrum, i.e.,

$$\mathbf{c}^t = \mathbf{m}^t \circ e^{i\mathbf{p}^t}. \tag{3}$$

Here, $\mathbf{p}^t$ is the phase angle (in radians) at frame t, and i the imaginary unit. 'flipud' denotes an operation that flips a vector upside down, and 'conj' the complex conjugation. The subscript $m:n$ denotes an operation that slices a vector from index m to n inclusively. Essentially, Eq. (2) first produces a conjugate symmetric version of $\mathbf{c}^t$, which is used as input for the subsequent IFFT to generate real time-domain signal. To isolate the impact of phase, we use mixture phase in this work. Estimated clean phase can surely be used and is expected to further improve the results.

The standard mean squared error between the estimated and clean signal is used as the loss function for the backpropagation training. In testing, we use the trained network to directly predict the (windowed) clean waveform snippets in each frame, which are overlap-added to produce the final time-domain reconstruction of the entire utterance.

3.2 Backpropagation

The proposed network architecture is trainable via the standard backpropagation algorithm, as the IFFT layer, i.e. Eq. (2), can be easily written in a set of matrix operations with fixed weight matrices. This is described as follows.

To begin with, we first define a permutation matrix

$$\mathbf{P}_{(d-1)\times(d+1)} = \begin{bmatrix} \mathbf{0}_{(d-1)\times 1} & \mathbf{R}_{(d-1)\times(d-1)} & \mathbf{0}_{(d-1)\times 1} \end{bmatrix},$$

where $\mathbf{0}_{(d-1)\times 1}$ is an all zero column vector of dimension $d-1$, and $\mathbf{R}_{(d-1)\times(d-1)}$ is the 90 degrees counterclockwise rotation of the identity matrix $\mathbf{I}_{(d-1)\times(d-1)}$. Then, the conjugate symmetric complex spectrum can be expressed as

$$\begin{bmatrix} \mathbf{m}^t \circ e^{i\mathbf{p}^t} \\ \mathbf{P}_{(d-1)\times(d+1)} \left(\mathbf{m}^t \circ e^{-i\mathbf{p}^t} \right) \end{bmatrix}. \tag{4}$$

By expressing the inverse DFT operation in matrix form and plugging in Eq. (1), we can rewrite Eq. (2) as

$$\hat{\mathbf{s}}^t = \mathbf{D}_{L\times L} \begin{bmatrix} \mathbf{h}^t \circ \mathbf{y}^t \circ e^{i\mathbf{p}^t} \\ \mathbf{P}_{(d-1)\times(d+1)} \left(\mathbf{h}^t \circ \mathbf{y}^t \circ e^{-i\mathbf{p}^t} \right) \end{bmatrix}, \tag{5}$$

where $\mathbf{D}_{L\times L}$ is the inverse DFT matrix of length L, i.e. $D_{nk} = e^{i\frac{2\pi}{L}(n-1)(k-1)}/L$ for $n, k = 1, 2, \cdots, L$, denoting the matrix element in row n and column k.

The inverse DFT matrix and the permutation matrix are predefined, and can be interpreted as fixed weight matrices of the network. Based on Eq. (5) and the loss function, one can easily derive the error signals with respect to the last hidden layer $\mathbf{h}^t$. Consequently, the gradients with respect to the tunable weights can be derived via the delta rule. Note that although there are no tunable weights in the IFFT layer, it affects the gradients to the downstream layers. As a result, the hidden mask $\mathbf{h}^t$ is automatically learned in light of the loss function of interest. In addition, it is known that even clean magnitude may not lead to clean speech signal due to noisy phase, which is especially true for low SNR mixtures. Therefore, another perspective is that the proposed network tries to learn an optimal masking function given the noisy (or the supplied) phase, differentiating itself from typical separation systems that are phase agnostic.

4 Experiments

4.1 Experimental Settings

All signals are sampled at the 16 kHz rate, and are framed by 20-ms windows and 10-ms frame shifts. Therefore, the length of the time-domain signal snippet in each frame is 320 samples.

We use 2000 randomly picked utterances from the TIMIT [8] training part as the training utterances. We use four types of non-stationary noises as the training and test noises: a factory noise, a babble noise, an engine noise, and an operation room noise (called "oproom"). Each noise is about 4 minutes long, and the first half is used to mix with the training utterances at -5 and 0 dB to create the training set. The TIMIT core test set, which consists of 192 utterances from unseen speakers, is used to mix with the second half of each noise to create the respective test sets at -5, 0, and 5 dB SNR, where the 5 dB SNR is unseen. Dividing the noises into two halves ensures that test noise segments are unseen. To further demonstrate the effectiveness of our method, we train and test on the IEEE corpus [7] recorded by a male speaker, where 600 IEEE utterances are used for training and 60 new ones for testing. We use two new noises, i.e. a different babble noise (called "babble2") and a cafeteria noise, to create the training and test mixtures as done for the TIMIT corpus. The new noises are about 10 minutes long each.

We call the proposed system IFFT-DNN. We first compare with two existing DNN-based supervised speech separation systems, which predict the IBM (IBM-DNN) and the IRM (IRM-DNN), respectively. We then compare with a baseline system DT-DNN that directly predicts the clean signal using the standard DNN. All systems, except for DT-DNN, use a complementary feature set [18] combined with the Gammatone filterbank energy as input features. DT-DNN is trained on the raw noisy signal, which performs better than using the complementary feature set. The DNNs in all systems have three hidden layers, each having 1024 rectified linear units (ReLUs), and are trained using adaptive stochastic gradient descent [3] with dropout regularization [5]. A special case is IFFT-DNN, where the last hidden layer has 161 linear units, as their activations are used to mask the mixture magnitude of the same length. We use linear units because clean speech magnitude can be greater than its mixture magnitude. Finally, to put the performance of IFFT-DNN in perspective, we compare with a recent non-negative matrix factorization (NMF) based system ASNA-NMF [15], which uses an active-set Newton algorithm and models a sliding window of 5 frames of DFT magnitudes. ASNA-NMF is trained on the same training set as used for the other systems.

To evaluate the objective speech quality, we use the composite measure (OVRL) proposed in [6], which shows a high correlation with subjective mean opinion scores. We evaluate the objective speech intelligibility using the short-time objective intelligibility (STOI) measure [13], which ranges from 0 to 1 and has been shown to be highly correlated with human intelligibility scores. Both OVRL and STOI are obtained by comparing separated speech with the corresponding clean speech.

4.2 Results

The separation results on the -5, 0 and 5 dB TIMIT test sets are listed in Table 1, 2, and 3, respectively. In terms of objective speech quality, most systems improve over the unprocessed

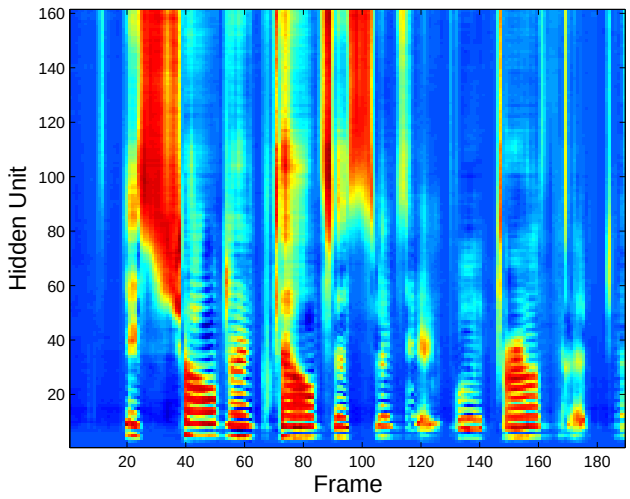

Figure 2: Visualization of the last hidden layer activations (masking layer) obtained from a TIMIT utterance mixed with the factory noise at 5 dB.

Table 1: Performance comparisons on -5 dB TIMIT mixtures. Boldface indicates best result

System	**Factory**		**Babble**		**Engine**		**Oproom**	
	OVRL	STOI	OVRL	STOI	OVRL	STOI	OVRL	STOI
Mixture	1.62	0.54	1.39	0.55	1.41	0.57	1.44	0.59
IBM-DNN	1.66	0.66	1.35	**0.63**	1.14	**0.78**	1.40	**0.77**
IRM-DNN	1.62	**0.67**	1.61	0.62	1.61	**0.78**	1.73	**0.77**
IFFT-DNN	**1.70**	0.65	**1.86**	0.61	**2.41**	0.77	**2.38**	0.75
DT-DNN	1.41	0.50	1.65	0.44	1.76	0.55	1.76	0.55
ASNA-NMF	**1.70**	0.60	1.71	0.57	2.15	0.71	1.99	0.68

mixtures in -5 dB. Due to the use of low *LC* values and binary gains, IBM-DNN obtains relatively worse quality results compared to the other DNN systems. IRM-DNN uses ratio masking and significantly outperforms IBM-DNN in OVRL. The proposed system, IFFT-DNN, further improves upon IRM-DNN significantly. For example, for the -5 dB engine noise, IFFT-DNN outperforms IRM-DNN by 0.8 in OVRL. A closer look (results not shown) indicate that IFFT-DNN in general has better noise suppression capability without further distorting target speech. DT-DNN uses a standard DNN to predict the time-domain clean signal. However, its performance is inferior to masking-based systems and IFFT-DNN. Without explicitly embedding the resynthesis process into the network, the parametrization of the standard DNN does not seem amenable to efficient learning on time-domain signals. Although it is also a data-driven system, ASNA-NMF does not work well in low SNR conditions; In particular, it is ineffective in noise suppression, as indicated by the poor performance in noise suppression (not shown), which in turn leads to low overall performance.

The performance trends in 0 and 5 dB SNR conditions are similar to those in -5 dB, with IRM-DNN outperforming IBM-DNN and DT-DNN performing poorly. Similarly, IFFT-DNN significantly outperforms all the other DNN systems as well as ASNA-NMF. Figure 2 visualizes the learned mask (the last hidden layer activations) of a TIMIT utterance mixed

Table 2: Performance comparisons on 0 dB TIMIT mixtures

System	Factory		Babble		Engine		Oproom	
	OVRL	STOI	OVRL	STOI	OVRL	STOI	OVRL	STOI
Mixture	2.05	0.65	1.81	0.67	1.79	0.69	1.93	0.70
IBM-DNN	2.12	**0.78**	1.59	**0.76**	1.15	**0.85**	1.34	**0.83**
IRM-DNN	2.22	**0.78**	2.14	**0.76**	2.10	**0.85**	2.23	**0.83**
IFFT-DNN	**2.30**	**0.78**	**2.40**	0.75	**2.81**	**0.85**	**2.73**	0.82
DT-DNN	1.64	0.56	1.83	0.52	1.93	0.60	1.93	0.60
ASNA-NMF	**2.30**	0.73	2.21	0.71	2.62	0.80	2.45	0.78

Table 3: Performance comparisons on 5 dB TIMIT mixtures

System	Factory		Babble		Engine		Oproom	
	OVRL	STOI	OVRL	STOI	OVRL	STOI	OVRL	STOI
Mixture	2.55	0.77	2.32	0.77	2.25	0.80	2.44	0.79
IBM-DNN	2.56	**0.86**	1.80	**0.86**	1.09	0.89	1.09	0.86
IRM-DNN	2.78	**0.86**	2.63	**0.86**	2.39	0.90	2.43	**0.88**
IFFT-DNN	**2.82**	**0.86**	**2.83**	0.85	**3.15**	**0.91**	**3.11**	0.87
DT-DNN	1.72	0.59	1.90	0.56	1.97	0.62	1.96	0.62
ASNA-NMF	2.79	0.82	2.67	0.82	3.04	0.88	2.88	0.85

with the factory noise at 5 dB SNR.

It is important that the improvement in speech quality does not come at the expense of degraded speech intelligibility. In terms of STOI, we can see that IBM-DNN and IRM-DNN perform similarly. At -5 dB, IFFT-DNN is slightly worse than IRM-DNN (about 2%), whereas in 0 and 5 dB, IFFT-DNN achieves almost the same STOI results as IRM-DNN. DT-DNN fails to improve objective intelligibility in all conditions. ASNA-NMF also fails to compete with masking-based systems and IFFT-DNN, even in high SNR conditions.

To further evaluate our network, we train and test on a different corpus with two different non-stationary noises. The averaged results on the IEEE corpus recorded by a male speaker are shown in Table 4. Again, IFFT-DNN consistently outperforms IRM-DNN in OVRL, while still achieving comparable STOI results. In terms of STOI, ASNA-NMF is substantially worse than supervised speech separation systems in general.

5 Conclusions

We have proposed a novel supervised separation system aiming to improve the sound quality of the separated speech. The key idea is to combine speech resynthesis and mask estimation in a single neural network and have them jointly trained. The resulting system, IFFT-DNN, takes advantage of both T-F masking and direct time-domain signal reconstruction. Results in various test conditions indicate that IFFT-DNN significantly improves objective speech quality while achieving comparable intelligibility results compared to a very strong baseline IRM-DNN. IFFT-DNN also significantly outperforms a state-of-the-art NMF based separation system in terms of both quality and intelligibility.

Further work will explore improvements such as using enhanced phase, employing better

Table 4: Performance comparisons on the male IEEE corpus. Results averaged over -5, 0, and 5 dB mixtures

System	Babble2		Cafeteria	
	OVRL	STOI	OVRL	STOI
Mixture	1.97	0.67	1.68	0.66
IRM-DNN	2.06	**0.84**	1.79	**0.78**
IFFT-DNN	**2.53**	0.82	**2.38**	**0.78**
ASNA-NMF	2.24	0.71	2.21	0.69

loss functions for time-domain signals, and adapting the resynthesis to auditory frequency scales. Finally, we envision that the proposed architecture provides a means to implement an end-to-end (i.e. waveform in, waveform out) speech separation system, where raw feature extraction, T-F masking, and speech resynthesis are all trained in one pipeline.

References

[1] D. Brungart, P. Chang, B. Simpson, and D.L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[2] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Divenyi P., Ed. Kluwer Academic, Norwell MA., 2005, pp. 181–197.

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

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

[5] 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.

[6] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 229–238, 2008.

[7] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[8] J. Garofolo et al., *DARPA TIMIT acoustic-phonetic continuous speech corpus*, National Inst. of Standards and Technology, 1993.

[9] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, pp. 1486–1494, 2009.

[10] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[11] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[12] A. Narayanan and D.L. Wang, "Joint noise adaptive training for robust automatic speech recognition," *in Proc. ICASSP, to appear*, 2014.

[13] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 2125–2136, 2011.

[14] S. Tamura, "An analysis of a noise reduction neural network," in *Proc. ICASSP*, 1989, pp. 2001–2004.

[15] T. Virtanen, J. Gemmeke, and B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 2277–2289, 2013.

[16] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," *Handbook of neural networks for speech processing. Artech House, Boston, USA*, 1999.

[17] Y. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1381–1390, 2013.

[18] Y. Wang, K. Han, and D.L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 270–279, 2013.

[19] Y. Wang, A. Narayanan, and D.L. Wang, "On training targets for supervised speech separation," Ohio State University Department of Computer Science and Engineering, Tech. Rep. TR05, 2014.

[20] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, pp. 66–68, 2014.