RECURRENT NEURAL NETWORKS FOR COCHANNEL SPEECH SEPARATION IN REVERBERANT ENVIRONMENTS

Masood Delfarah[1] *and 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

delfarah.1@osu.edu, dwang@cse.ohio-state.edu

ABSTRACT

Speech separation is a fundamental problem in speech and signal processing. A particular challenge is monaural separation of cochannel speech, or a two-talker mixture, in a reverberant environment. In this paper, we study recurrent neural networks (RNNs) with long short-term memory (LSTM) in separating and enhancing speech signals in reverberant cochannel mixtures. Our investigation shows that RNNs are effective in separating reverberant speech signals. In addition, RNNs significantly outperform deep feedforward networks based on objective speech intelligibility and quality measures. We also find that the best performance is achieved when the ideal ratio mask (IRM) is used as the training target in comparison with alternative training targets. While trained using reverberant signals generated by simulated room impulse responses (RIRs), our model generalizes well to conditions where the signals are generated by recorded RIRs.

Index Terms— Cochannel speech separation, room reverberation, deep neural network, long short-term memory

1. INTRODUCTION

A fundamental problem in speech processing is source separation. Successful separation can lead to better performance for robust automatic speech recognition (ASR), speaker identification (SID), and speech communication systems. Listeners with hearing impairment will also benefit as studies show that, in comparison to normal-hearing listeners, hearing-impaired listeners have more trouble in the presence of an interfering speaker [1, 2] and in moderate amounts of room reverberation [3, 4]. Hearing-aid devices embedded with sound separation capability should be able to help the user better understand the target speech in real acoustic environments. The focus of this study is on separating two speakers in reverberant conditions. Since reverberation degrades speech intelligibility and quality, we also aim at dereverberating the mixture signals.

Cochannel speech separation is a special case of the speech separation problem in which the goal is to recover speech of interest (i. e., target speech) distorted by background noise, room reverberation, or interfering speech. For speech separation, data-driven approaches used in supervised learning have shown better performance compared to traditional signal processing methods [5]. Supervised sound separation aims to learn a function from noisy inputs to a corresponding clean target.

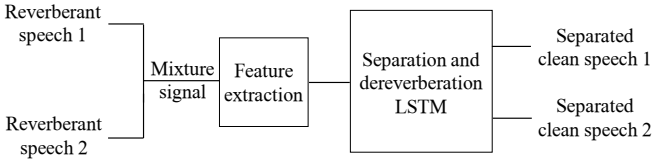

Fig. 1: Overview of the proposed separation framework.

Deep fead-forward neural networks (DFNs) have shown a strong representational capacity [6]. Wang and Wang [7] first introduced DFN for speech separation. Since then, DFNs have been increasingly used in speech separation. For example, studies in [8, 9, 10, 11] train models to separate two-talker mixtures in anechoic environments. Room reverberation is not considered in these studies, which is a major distortion in real environments. Other studies apply DFNs in reverberant conditions [12, 13, 14, 15]. These studies are on speech-noise separation and not on two-talker conditions. In our previous work [16], we showed that DFNs behave differently when the interference is human speech instead of background noise.

Recurrent neural networks (RNNs) are interesting models for speech processing due to their temporal processing mechanism. Long short-term memory (LSTM) [17] is a variant of RNN that facilitates information flow through time, via using memory cells. Erdogan *et. al.* [18] and Weninger *et. al.* [19] apply LSTMs for speech enhancement in noisy environments. In a very recent study, Chen and Wang [20] address the speech-noise separation problem and show that LSTMs have a greater capacity over DFNs in generalizing to unseen speaker and noise conditions.

Due to temporal effects of reverberation, LSTM is potentially a better model than a DFN for reverberant speech processing. In this paper we study LSTMs in separating two-talker mixtures in reverberant condition. To our knowledge LSTMs have not been applied to these conditions. In this study, we perform systematic evaluations to compare the separation performance of DFNs and LSTMs in cochannel and reverberant conditions. The evaluation also includes comparison of two different training targets.

It is important to note that our study aims at addressing the speaker-specific cochannel separation. Solutions for the problem of open speaker-set separation have been recently proposed (e.g., [21, 22, 23]). These studies are design to work in anechoic environment, and generalization to reverberant condition is not straightforward in these systems. More importantly, since these studies do not directly model the speakers, they are expected to yield worse performance in comparison with the speaker-specific models.

The rest of the paper is organized as follows. We describe our proposed cochannel speech separation method in Section 2. Section 3 presents the experimental results, and we conclude in Section 4.

2. PROPOSED METHOD

The proposed framework is depicted in Fig. 1. Reverberant mixtures are generated by separately convolving a target and interference utterance with a room impulse response (RIR). Reverberant target and interfering signals are mixed in the time domain, and then features are extracted from the mixture. We normalize the training features to zero mean and unit variance in each dimension. We use the same normalization factors to normalize the test data features before feeding to the DFN/LSTM, frame by frame. The estimated magnitudes are generated from the network output as described in Sec. 2.3. Lastly, using the mixture signal phase and the estimated magnitude spectrograms, the inverse short-time Fourier transform (STFT) generates an estimate of the two signals in the time domain. We briefly describe the elements of the framework in the following.

2.1. Features

In a previous study [16], we found that the combination of Gammatone Frequency Cepstral Coefficients (GFCC) [24], Power-Normalized Cepstral Coefficients (PNCC) [25], and Log-Mel Filterbank features form a complementary feature set for cochannel separation in reverberant conditions. This combination is more effective than the features used in other speech enhancement studies. We extract a 31-D GFCC, 31-D PNCC and 40-D Log-mel feature per frame of the mixture signal as described in [16]. This set can be used as a feature vector, $F(m)$, where m indicates the time frame. One can employ neighboring frames, and the feature vector, $\tilde{F}(m)$, will be:

$$\tilde{F}_{a,b}(m) = [F(m-a), \cdots, F(m+b)] \quad (1)$$

where a and b indicate the number of preceding and succeeding frames to use, respectively. Setting $b = 0$ in this formulation preserves the causality property of the system.

2.2. Learning machines

The baseline system is a DFN with 4 hidden layers, and each hidden layer has 1500 units. ReLU is used as the activation function in the hidden units. The input to this DFN is $\tilde{F}_{10,0}(.)$, and accordingly we refer to this system as $\mathrm{DFN}_{10,0}$. We also train a 4-layer LSTM, with 600 units in each of its layers. The output layer of the LSTM is a fully-connected feed-forward layer stacked on top of the recurrent layers. Due to the recurrent connections in the LSTM, it is not necessary to use a window of feature frames in the input. For that reason, we use $\tilde{F}_{0,0}(.)$ as the input feature vector for LSTM training.

Assuming that there is access to future frames (i.e., the full utterance) one can train a bidirectional LSTM (BLSTM). A BLSTM comprises of two unidirectional LSTMs one processing the signal in forward direction and the other processing it in backward. We use a BLSTM with 600 hidden units in each layer, and compare the performance with $\mathrm{DFN}_{5,5}$ which uses the feature vector $\tilde{F}_{5,5}(.)$.

Each network is trained using the Adam [26] algorithm to minimize the mean squared error loss. The algorithm is run for 100 epochs, with the learning rate of 3×10^{-4} . The LSTMs are input by 100 feature frames at a time.

2.3. Training objectives

Wang *et. al.* [5] showed that the DFN targets contribute to the separation performance. We consider two different training targets in this study. Assume $s_1(.)$, $s_2(.)$, and $m(.)$ represent the direct-sound of the first source, direct-sound of the second source, and the reverberant mixture signals in time domain, respectively. Then we apply short-time frequency transform (STFT), on each of the signals to derive $S_1(.)$, $S_2(.)$, and $M(.)$. We also define $S_1^C(.)$ and $S_2^C(.)$ as the STFT representation of $m(.) - s_1(.)$ and $m(.) - s_2(.)$, respectively. The training targets in this study are:

- **Log-magnitude spectrogram (MAG)**: This target is simply $[log|S_1(.)|, log|S_2(.)|]$. While using this type of target, we use a linear activation function in the network output layer since it ranges over $(-\infty, \infty)$. At test time the network output is decompressed by an exponential function before signal resynthesis.

- **Ideal ratio mask (IRM)**: The IRM is defined as fol-

lows [5, 9]:

$$\mathrm{IRM} = [\mathrm{IRM}_1, \mathrm{IRM}_2] \quad (2)$$

$$\mathrm{IRM}_i = \frac{|S_i|}{|S_i| + |S_i^C|}, \qquad i = 1, 2 \quad (3)$$

Since the IRM ranges over $[0, 1]$, while using the IRM as the target, we use the sigmoid function in the output layer activation. During test time, we multiply the output of the network by $[|M(.)|, |M(.)|]$ to derive estimated source magnitude responses. Note that $\mathrm{IRM}_1 + \mathrm{IRM}_2 \neq \mathbf{1}$, unlike in [9].

2.4. Evaluation metrics

We use STOI and PESQ as the objective scores for speech intelligibility and quality, as they correlate with the human test scores. Higher STOI and PESQ scores indicate better speech intelligibility and quality. We use the direct-sound male and the direct-sound female signals as the reference in these metrics.

3. EXPERIMENTS

We use the IEEE corpus [27] to train and test the systems. This corpus consists of 1440 utterances, where half are spoken by a male speaker, and the other half by a female speaker. We randomly choose 500 sentences by each speaker for training and the remaining utterances are used for testing. Then we generate $125,000$ training signals by mixing one female and one male utterance. The reverberation time (T_{60}) is randomly chosen from the range of $[0.3, 0.9]$ seconds, and reverberant signals are generated using a RIR generator[1] based on the image method [28]. In our simulations the room size is (6.5, 8.5, 3) m and the microphone is located at (3, 4, 1.5) m. We place the male speaker at 1 m and the female speakers at 2 m distance from the microphone. Target-to-interference energy ratio (TIR) is drawn from the range of $[-12, 12]$ dB, then the female utterance is scaled and added to the male signal. Since sentences do not have the same length, a female utterance is clipped or repeated until it covers all of its corresponding male utterance in a mixture.

Test data is generated using different utterances and a slightly different simulation room, so that no RIRs in the training data is repeated in the test set.

3.1. Performance with simulated RIRs

Average STOI scores on 1000 test mixtures in different conditions are shown in Table 1. The TIR for the mixture signals is within the range of $[-12, 12]$ dB.

We observe that the DFN achieves a higher baseline performance while future frames are incorporated. Likewise, a BLSTM outperforms an LSTM. Second, LSTM outperforms DFN in all conditions, indicating that it is a better fit for speech separation in reverberant conditions. The gap between those two is as large as 7 percentage scores in high reverberation times. It is also interesting to see that the model is trained on reverberant data and generalizes well to separating anechoic mixtures. Finally, we note for the cochannel separation problem in reverberant conditions, IRM estimation is a better method than directly predicting the magnitude spectrograms of the sources.

Table 2 shows the quality of the separated signals using PESQ scores. Again, we observe that in all cases BLSTM-IRM is the best in enhancing the quality of the female and male utterances.

Spectrograms in Fig. 2 illustrate a separation example using $\mathrm{DFN}_{5,5}$-IRM and LSTM-IRM. As seen in the figures, for both systems the spectrograms of the separated signals resemble the clean spectrograms. We also observe that the LSTM was able to generate smoother spectrograms. We could also confirm this with our informal listening tests.

Table 1

Average STOI (%) scores in simulated reverberant conditions. T_{60} = 0.0 s indicates anechoic condition. Scores for female and male sentences are shown separately with the latter in parentheses. TIR for each mixture signal is in the range of $[-12, 12]$ dB.

T_{60} (s)	0.0 s	0.3 s	0.6 s	0.9 s	Average
Mixture	58.6(57.9)	54.8(47.3)	44.7(35.1)	36.9(27.1)	48.7(41.8)
$\mathrm{DFN}_{10,0}$-MAG	76.9(72.43)	73.8(67.1)	67.5(60.6)	61.7(54.9)	70.0(63.7)
$\mathrm{DFN}_{10,0}$-IRM	84.4(81.2)	78.8(70.2)	70.8(61.9)	63.5(55.1)	74.4(67.1)
LSTM-MAG	81.1(77.0)	74.2(68.9)	67.3(61.6)	72.3(56.7)	73.7(67.6)
LSTM-IRM	87.0(84.2)	81.0(71.8)	71.8(63.3)	64.8(56.9)	76.2(69.0)
$\mathrm{DFN}_{5,5}$-MAG	78.5(72.1)	76.5(70.0)	70.3(64.4)	64.17(58.9)	72.4(66.4)
$\mathrm{DFN}_{5,5}$-IRM	86.1(80.7)	80.7(71.6)	73.0(64.8)	66.0(58.9)	76.5(69.0)
BLSTM-MAG	86.2(81.9)	81.6(75.9)	74.9(70.2)	69.6(65.6)	78.1(73.4)
BLSTM-IRM	**89.9(86.4)**	**84.7(76.6)**	**77.2(70.7)**	**71.6(66.3)**	**80.9(75.0)**

Table 2

Average PESQ scores in simulated reverberant conditions.

T_{60} (s)	0.0 s	0.3 s	0.6 s	0.9 s	Average
Mixture	1.38(1.27)	1.40(1.08)	1.20(0.79)	1.08(0.65)	1.27(0.94)
$\mathrm{DFN}_{10,0}$-MAG	2.35(2.08)	2.15(1.79)	1.76(1.45)	1.51(1.21)	1.94(1.63)
$\mathrm{DFN}_{10,0}$-IRM	2.55(2.33)	2.30(1.90)	1.88(1.54)	1.63(1.29)	2.09(1.77)
LSTM-MAG	2.54(2.31)	2.28(1.84)	1.84(1.44)	1.58(1.20)	2.06(1.7)
LSTM-IRM	2.66(2.46)	2.41(1.95)	1.94(1.53)	1.67(1.30)	2.17(1.81)
$\mathrm{DFN}_{5,5}$-MAG	2.40(2.01)	2.28(1.87)	1.90(1.57)	1.63(1.32)	2.05(1.7)
$\mathrm{DFN}_{5,5}$-IRM	2.67(2.31)	2.41(1.95)	1.97(1.63)	1.70(1.37)	2.19(1.81)
BLSTM-MAG	2.71(2.48)	2.48(2.12)	2.09(1.80)	1.80(1.57)	2.27(1.99)
BLSTM-IRM	**2.85(2.58)**	**2.61(2.18)**	**2.18(1.88)**	**1.93(1.68)**	**2.39(2.08)**

[1] https://github.com/ehabets/RIR-Generator

Table 3
Average STOI (%) scores in recorded RIR conditions.

T_{60} (s)	0.32 s	0.47 s	0.68 s	0.89 s	Average
Mixture	57.0(52.7)	51.3(50.2)	57.0(52.9)	53.8(50.9)	54.5(51.7)
$\text{DFN}_{10,0}$-IRM	80.0(73.4)	72.5(69.7)	79.1(72.8)	70.1(63.5)	75.4(69.8)
LSTM-IRM	81.9(75.6)	74.8(71.4)	81.3(75)	71.2(64.6)	77.3(71.7)
$\text{DFN}_{5,5}$-IRM	81.4(74.0)	73.3(70.0)	80.0(73.0)	70.8(62.9)	76.4(70.0)
BLSTM-IRM	**84.7**(**79.8**)	**78.1**(**75.4**)	**84.1**(**78.7**)	**74.7**(**67.9**)	**80.4**(**75.4**)

Table 4
Average PESQ scores in recorded RIR conditions.

T_{60} (s)	0.32 s	0.47 s	0.68 s	0.89 s	Average
Mixture	1.43(1.26)	1.37(1.16)	1.43(1.23)	1.52(1.27)	1.44(1.23)
$\text{DFN}_{10,0}$-IRM	2.27(2.03)	1.99(1.83)	2.22(1.95)	1.90(1.62)	2.09(1.86)
LSTM-IRM	2.36(2.09)	2.06(1.83)	2.32(2.02)	1.87(1.49)	2.15(1.85)
$\text{DFN}_{5,5}$-IRM	2.36(2.05)	2.04(1.85)	2.30(1.95)	**1.95**(**1.61**)	2.16(1.87)
BLSTM-IRM	**2.50**(**2.30**)	**2.20**(**2.05**)	**2.44**(**2.19**)	1.91(**1.61**)	**2.26**(**2.04**)

3.2. Performance with recorded RIRs

In order to examine the generalizability of the methods to real room environments, we generate mixtures using recorded RIRs from [29] in 4 rooms with 37 captured RIRs in each. For each room we choose one channel of each binaural RIR and then resample it to match the sampling frequency of the mixtures. We also randomly choose two RIRs to generate reverberant mixtures. Note that no training with recorded RIRs is performed.

STOI Results in recorded RIRs are provided in Table 3. The results indicate good generalization to real acoustic environments. Finally, PESQ scores are presented in Table 4. These results also show that a BLSTM using IRM as the training targets best generalizes to recored RIR conditions.

4. CONCLUSION

In this paper we proposed using RNNs with LSTM to separate cochannel speech in reverberant conditions. Systems have been evaluated in different TIR and T_{60} conditions. We achieved substantial improvements in objective speech intelligibility and quality scores using LSTMs. Comparisons show that future frames can be very useful in separating reverberant speech signals. In future work we plan to extend this method to situations with background noise and multiple speakers.

5. ACKNOWLEDGEMENTS

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

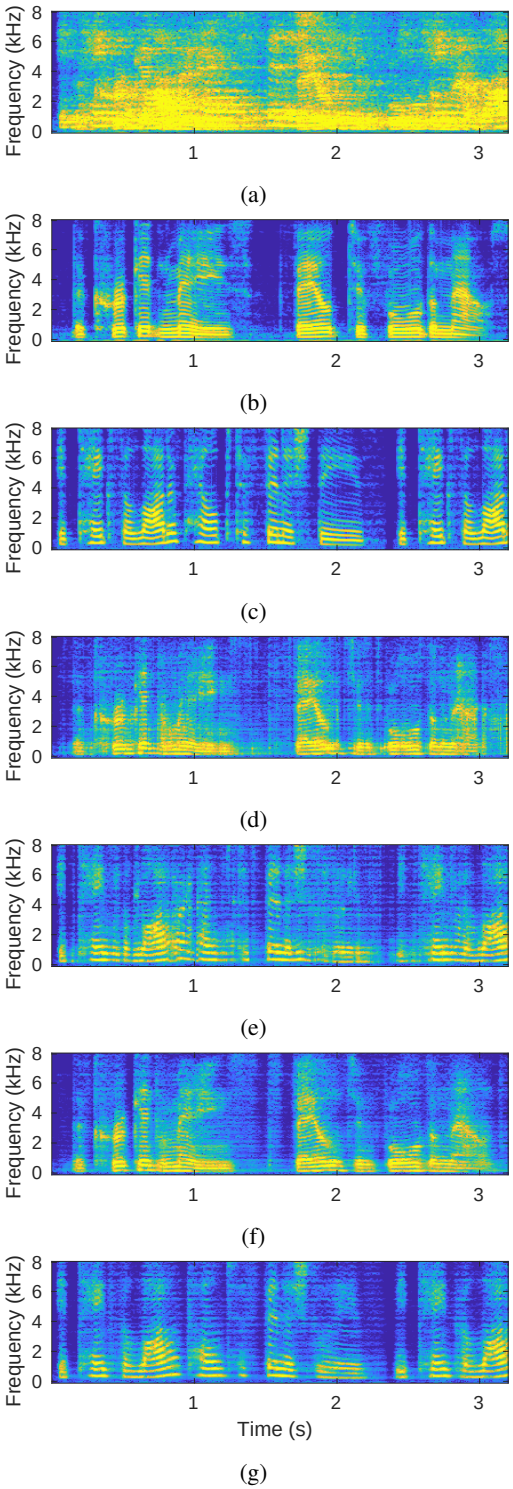

Fig. 2: (Color online) Separation illustration for an IEEE male sentence mixed with a female sentence at TIR of 0 dB and T_{60} of 0.9 s. Spectrogram for (a) reverberant mixture, (b) clean male speech (c) clean female speech, (d) estimated male speech from $\text{DFN}_{5,5}$-IRM, (e) estimated female speech from $\text{DFN}_{5,5}$-IRM (f) estimated male speech from BLSTM-IRM, and (g) estimated female speech from BLSTM-IRM.

6. REFERENCES

[1] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Amer.*, vol. 88, pp. 1725–1736, 1990.

[2] R. Carhart and T. W. Tillman, "Interaction of competing speech signals with hearing losses," *Arch. Otolaryngol.*, vol. 91, pp. 273–279, 1970.

[3] O. Hazrati and P. C. Loizou, "Tackling the combined effects of reverberation and masking noise using ideal channel selection," *J. Speech Lang. Hear. Res.*, vol. 55, pp. 500–510, 2012.

[4] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1581–1592, 1994.

[5] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 1849–1858, 2014.

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[7] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1381–1390, 2013.

[8] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. ICSP*, 2014, pp. 473–477.

[9] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. ICASSP*, 2014, pp. 1562–1566.

[10] ——, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 2136–2147, 2015.

[11] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 967–977, 2016.

[12] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. ICASSP*, 2014, pp. 4628–4632.

[13] 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 Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 982–992, 2015.

[14] Y. Zhao, D. L. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *Proc. ICASSP*, 2016, pp. 6525–6529.

[15] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *Proc. ICASSP*, 2017, pp. 5580–5584.

[16] M. Delfarah and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1085–1094, 2017.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[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 *Proc. ICASSP*, 2015, pp. 708–712.

[19] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. LVA/ICA*, 2015, pp. 91–99.

[20] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, pp. 4705–4714, 2017.

[21] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[22] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[23] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017, pp. 246–250.

[24] Y. Shao, S. Srinivasan, and D. L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*, 2007, pp. IV–277.

[25] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 1315–1329, 2016.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICML*, 2015.

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

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[29] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1867–1871, 2010.