TALKER-INDEPENDENT SPEAKER SEPARATION IN REVERBERANT CONDITIONS

Masood Delfarah[1]*, Yuzhou Liu*[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,liuyuz,dwang}@cse.ohio-state.edu

ABSTRACT

Speaker separation refers to the task of separating a mixture signal comprising two or more speakers. Impressive advances have been made recently in deep learning based talker-independent speaker separation. But such advances are achieved in anechoic conditions. We address talker-independent speaker separation in reverberant conditions by exploring a recently proposed deep CASA approach. To effectively deal with speaker separation and speech dereverberation, we propose a two-stage strategy where reverberant utterances are first separated and then dereverberated. The two-stage deep CASA method outperforms other talker-independent separation methods. In addition, the deep CASA algorithm produces substantial speech intelligibility improvements for human listeners, with a particularly large benefit for hearing-impaired listeners.

Index Terms— Talker-independent speaker separation, speech dereverberation, computational auditory scene analysis, deep CASA

1. INTRODUCTION

Room reverberation and competing speakers degrade speech perception of human listeners [1–3]. Enhancing the target speech and separating it from interfering speakers can mitigate these adverse effects. In particular hearing-impaired listeners have difficulty in reverberant and multi-talker conditions. Hearing-aids with effective speaker separation capabilities can remove this disability.

Hidden Markov models (HMMs) [4] and non-negative matrix factorization (NMF) [5] represent two traditional approaches to speaker separation. After the success of using deep neural networks (DNNs) for speech enhancement [6], the speaker separation problem has been addressed by learning a mapping from mixture features to some training target. This approach works well for talker-dependent speaker separation (i.e. separating two trained speakers) [7–10] and yields substantial improvement for human listeners in anechoic and reverberant conditions [11, 12].

Talker-independent speaker separation is the most general case where test speakers can be unseen during training. The DNN-based speech separation approach is not applicable for the talker-independent separation, because in this case the output layers cannot be uniquely assigned to individual speakers. This problem is often referred to as the permutation problem [13], and has been addressed in two approaches: deep clustering [14] and permutation invariant training (PIT) [13]. In deep clustering, embedding vectors are learned from time-frequency (T-F) units. Then using the K-means algorithm these embeddings are clustered to yield an estimated ideal binary mask (IBM) [15], which is used to reconstruct the separated signals. PIT optimizes a DNN using the minimum losses among the possible speaker-output assignments. One version of PIT, known as frame-level PIT (tPIT) [16], addresses speaker separation at the frame level. Since tPIT does not have a mechanism for assigning the separated frames to individual speakers, it does not address the speaker tracking problem. Utterance-level PIT (uPIT) [13] performs speaker tracking together with separation by calculating the loss over the entire mixture signal instead of individual frames, during training. Evaluations, however, show that tPIT with an optimal speaker tracker significantly outperforms uPIT [13].

Conv-TasNet [17] and FurcaNeXt [18] are two recent studies of talker-independent speaker separation in anechoic conditions. Both algorithms demonstrate very good separation results and even surpass the oracle results of the ideal ratio mask (IRM) [19]. Another successful approach is deep CASA [20], which first separates the speakers at the frame level, and then groups the separated frames sequentially to form the individual speaker utterances. The divide and conquer strategy in deep CASA is inspired by research in computational auditory scene analysis (CASA) [21]. CASA typically performs simultaneous grouping of temporally-overlapped speech segments, and then sequentially organizes those simultaneously separated segments into streams, each corresponding to one auditory source.

This paper addresses the monaural talker-independent speaker separation in reverberant conditions. We extend the deep CASA algorithm to a two-stage method. In the first stage reverberant speech signals are separated, and the second stage further dereverberates the separates utterances. We found two-stage processing advantageous in our talker-dependent reverberant speaker separation study [22]. The two-stage processing strategy was also pursued in other speech and speaker separation studies [23], [24], [25].

The rest of this paper is organized as follows. In Section 2, we present the background of this study including the problem formulation and the deep CASA baseline. Section 3 describes the proposed two-stage deep CASA algorithm. Evaluation results and comparisons are discussed in Section 4. We conclude the paper in Section 5.

2. BACKGROUND

In a reverberant room with two simultaneous talkers, a monaural two-speaker mixture $y(t)$ can be described as follows:

$$y(t) = s_{r_1}(t) + s_{r_2}(t) = s_1(t) * h_1(t) + s_2(t) * h_2(t) \qquad (1)$$

where s_1 and s_2 are the anechoic speech signals, h_1 and h_2 are their corresponding room impulse responses (RIRs), s_{r_1} and s_{r_2} are the reverberant speech signals, and $*$ denotes convolution. In this study, we aim at extracting s_1 and s_2 from y. This problem formulation

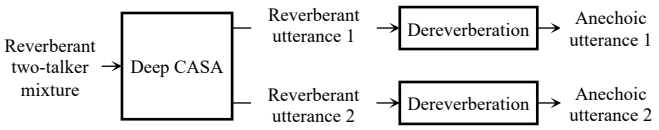

Fig. 1: Overview of the proposed two-stage deep CASA algorithm for speaker separation in reverberant conditions.

has a special case of speaker separation in anechoic conditions, i.e.:

$$h_1(t) = h_2(t) = \delta(t) \tag{2}$$

where δ is the Dirac delta function. Anechoic talker-independent speaker separation was addressed by the deep CASA approach [20] by first separating the speakers at the frame level and then assigning the separated frames to the speakers via a speaker tracker neural network.

Deep CASA can be directly extended to reverberant conditions by dereverberating and separating speech signal frames in the simultaneous grouping module and then organizing those frames into the estimates of s_1 and s_2 via the predicted speaker-frame assignments obtained from the sequential grouping module. To further improve this baseline, in the next section, we propose a two-stage deep CASA algorithm that performs the two tasks of speaker separation and speech dereverberation separately.

3. DEEP CASA FOR SPEAKER SEPARATION IN REVERBERANT CONDITIONS

An overview of the proposed two-stage algorithm is shown in Fig. 1. The first stage is deep CASA and it comprises a simultaneous grouping and a sequential grouping module. The simultaneous grouping module is fed by the reverberant mixture features to perform frame-level separation of the reverberant speech signals. Then, the sequential grouping module organizes these separated reverberant frames over time into two reverberant signals, each corresponding to one speaker. The proposed second stage is a DNN that dereverberates separated streams individually to form separated anechoic speech estimates. A description of the different modules is as follows.

3.1. Simultaneous grouping

Fig. 2a depicts the simultaneous grouping module. Given the mixture signal y, the real and imaginary short-time Fourier transform (STFT) features, Y, are extracted and fed into a DNN to produce two complex ratio masks cRM_1 and cRM_2 [26]. These ratio masks along with Y produce two STFT features $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$:

$$\hat{S}_{u_1}(m,f) = \mathrm{cRM}_1(m,f) \otimes Y(m,f) \tag{3}$$

$$\hat{S}_{u_2}(m,f) = \mathrm{cRM}_2(m,f) \otimes Y(m,f) \tag{4}$$

where m is the time frame index, f is the frequency index, and $\otimes$ denotes point-wise matrix multiplication. $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ are the estimated frame-level reverberant speaker signals, yet to be organized over time. We use the tPIT loss function [16] to organize these frames into $\hat{S}_{o_1}$ and $\hat{S}_{o_2}$ that correctly represent the speaker signals. Accordingly, two possible loss functions ℓ_1 and ℓ_2 are calculated per time frame as follows:

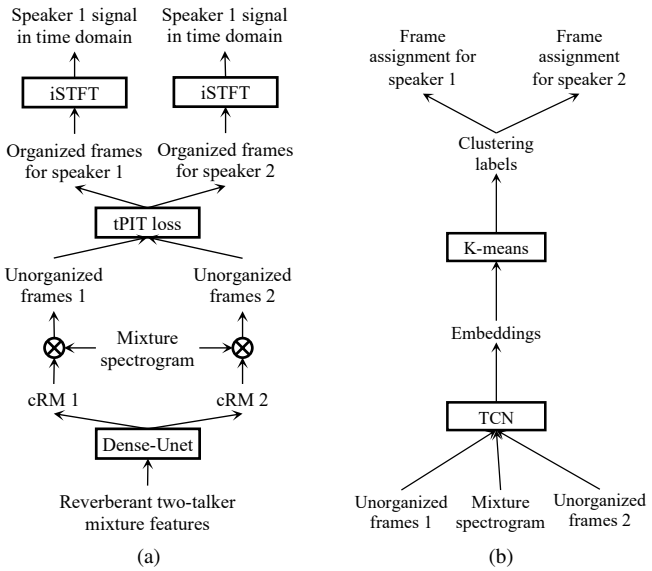

Fig. 2: Diagram of (a) the simultaneous grouping module and (b) the sequential grouping module in the proposed two-stage deep CASA algorithm.

$$\ell_1(m) = \Sigma_f |\hat{S}_{u_1}(m,f) - S_{r_1}(m,f)| + \Sigma_f |\hat{S}_{u_2}(m,f) - S_{r_2}(m,f)| \tag{5}$$

$$\ell_2(m) = \Sigma_f |\hat{S}_{u_1}(m,f) - S_{r_2}(m,f)| + \Sigma_f |\hat{S}_{u_2}(m,f) - S_{r_1}(m,f)| \tag{6}$$

where S_{r_1} and S_{r_2} are the complex STFT features of s_{r_1} and s_{r_1}, accordingly. Then optimal speaker tracking is performed:

$$\hat{S}_{o_1}(m,f), \hat{S}_{o_2}(m,f) = \begin{cases} \hat{S}_{u_1}(m,f), \hat{S}_{u_2}(m,f) & \text{if } \ell_1(m) \leq \ell_2(m) \\ \hat{S}_{u_2}(m,f), \hat{S}_{u_1}(m,f) & \text{otherwise} \end{cases} \tag{7}$$

Next, $\hat{S}_{o_1}$ and $\hat{S}_{o_2}$ are converted into time domain signals $\hat{s}_{o_1}$ and $\hat{s}_{o_2}$ using inverse STFT (iSTFT), which are the predicted reverberant signals using optimal speaker tracking. Finally, the simultaneous grouping module is optimized by minimizing the signal-to-noise ratio loss function:

$$J^{tPIT-SNR} = -10 \sum_{i=1,2} \log \frac{\Sigma_t s_i(t)^2}{\Sigma_t [s_i(t) - \hat{s}_{o_i}(t)]^2} \tag{8}$$

The simultaneous grouping module uses a Dense-Unet architecture as used in deep CASA [20].

3.2. Sequential grouping

The simultaneous grouping module uses s_{r_1} and s_{r_2} to predict the separated reverberant signals for the speakers which are not available at the test time. For this reason, a sequential organization module is trained for speaker tracking. Fig. 2b depicts the sequential organization module. Given an M-frame mixture Y, this module is fed by $|Y|$ along with the unorganized signals $|\hat{S}_{u_1}|$ and $|\hat{S}_{u_2}|$ which are produced from Equations 3 and 4. Then, an $M \times 2$ matrix $\boldsymbol{A}$ is

predicted as:

$$\boldsymbol{A}(m) = \begin{cases} [1,0] & \text{if } \ell_1(m) \leq \ell_2(m) \\ [0,1] & \text{otherwise} \end{cases} \tag{9}$$

where $1 \leq m \leq M$. Note that $\boldsymbol{A}$ optimally organizes $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ as:

$$\begin{bmatrix} \hat{S}_{o_1} \\ \hat{S}_{o_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} \\ \mathbf{1} - \boldsymbol{A} \end{bmatrix} \begin{bmatrix} \hat{S}_{u_1} \\ \hat{S}_{u_2} \end{bmatrix} \tag{10}$$

The network predicts the embedding matrix $\boldsymbol{V} \in \mathbb{R}^{M \times D}$, where D is the embedding dimension. $\boldsymbol{V}$ encodes the information of the optimal output-speaker pairing. Accordingly, the loss function [14]:

$$J^{DC} = ||\boldsymbol{W}^{1/2}(\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{W}^{1/2}||_F^2 \tag{11}$$

is optimized, where $||.||_F$ is the Frobenius norm and:

$$\boldsymbol{W} = diag(\frac{|l_1 - l_2|}{\Sigma_c |l_1 - l_2|}) \tag{12}$$

At the test time, the sequential grouping module predicts the embedding $\boldsymbol{V}$ and then a K-means algorithm clusters $\boldsymbol{V}(m)$ vectors, i.e. assigning labels $\hat{a}(m) = \{0, 1\}$. Finally, the predicted sequential organization matrix $\hat{\boldsymbol{A}}$ is obtained:

$$\hat{\boldsymbol{A}} = [\hat{\boldsymbol{a}}, \mathbf{1} - \hat{\boldsymbol{a}}] \tag{13}$$

which is used as in Eq. 9 to generate the estimated complex domain reverberant signals $\hat{S}_{r_1}$ and $\hat{S}_{r_2}$.

Following the deep CASA algorithm, a temporal convolutional network (TCN) [27, 28] is adopted as the sequence model in the sequential organization module. Using a stack of dilated convolutional layers, TCN possesses very long memory, which is desirable for the task of speaker tracking.

3.3. Speech dereverberation

In this stage, the estimated reverberant signals are individually dereverberated using a Dense-Unet. Accordingly, $\hat{S}_{r_1}$ (or $\hat{S}_{r_2}$) is fed to the network, and a complex ratio mask is generated to produce the complex domain STFT $\hat{S}_1$ (or $\hat{S}_2$). It is then converted to the time-domain signal $\hat{s}_1$ (or $\hat{s}_2$) and the network is optimized to minimize:

$$J^{SNR} = -10 \log \frac{\Sigma_t s_i(t)^2}{\Sigma_t [s_i(t) - \hat{s}_i(t)]^2} \qquad i = 1, 2 \tag{14}$$

At the inference time, the reverberant two-talker mixture is passed through the simultaneous grouping module to form the frame-level separated reverberant speaker signals. Then, the sequential grouping module uses the K-means algorithm to predict the across-frame organization sequence. This predicted sequence is used to form the two reverberant speech streams. Finally, each of the streams undergoes speech dereverberation to yield the predicted anechoic speaker signals.

4. EVALUATIONS AND COMPARISONS

4.1. Experimental setup

In this study, spectrograms are generated from signals sampled at 16 kHz and by applying a 512-point DFT on those signals divided into 32-ms frames with a frame-shift of 8 ms.

We use the WSJ0 dataset [29] to generate the two-talker mixtures[1]. To produce the training data, two different speakers from the folder `si_tr_s` are randomly picked each time to produce a pair. We use the image method to simulate reverberant room conditions. To this end, for each speaker pair a simulated room with random dimensions is picked. Next, a $T_{60} \in [0.3, 1.0]$ s is chosen for the reverberation time. A microphone is placed in this room at a random position. Then, each of the two speakers is placed in the room at different locations with a random distance of $d \in [0.5, 3]$ m from the microphone. The speakers and the microphone have the same elevation within the room. An RIR generator toolbox[2] is used to generate a room impulse response for each speaker. Then, the reverberant two-talker mixtures are generated as described in Eq. 1. The process of mixture generation is repeated to obtain 200,000 training and 1000 cross-validation mixture signals.

Test mixtures are generated using speakers from the WSJ0 corpus that were not used in the training set. For this purpose, speakers are picked from the folders `si_dt_05` and `si_et_05`, and mixture signals are generated using simulated and recorded RIRs. The test simulated room T_{60} was set to 0.3, 0.6, or 0.9 s. To generate mixtures with real room conditions, recorded RIRs from four rooms are used [30]. In each condition, 3000 test mixtures are generated.

In our experiments, each stage is trained separately and then connected to perform joint training. Performance metrics used for algorithm evaluations are ESTOI (%) [31] which is a standard speech intelligibility metric with the value range typically between 0 and 1, PESQ [32] which is a standard speech quality metric with the value range of -0.5 to 4.5, and ΔSDR (dB) [33]. In all three metrics higher scores indicate better results. The reference signals in the evaluations are the direct-sound speech signals and in each test condition the average scores of test mixtures are reported.

4.2. Comparison systems

We compare the performance of the proposed two-stage deep CASA algorithm with the following strong baselines:

- One-stage deep CASA [20]: This system represents the direct extension of deep CASA from anechoic to reverberant conditions and is implemented as described in Section 2.
- Dense-Unet uPIT: The Dense-Unet structure of the simultaneous grouping module in deep CASA is trained with the uPIT loss function. This system does not need a sequential organization module.
- Conv-TasNet [17]: This algorithm performs uPIT [13] in the time domain and yields very good results of speakers in anechoic conditions. Our implementation of Conv-TasNet uses a TCN structure and predicts frames of length 2 ms with 1 ms frame shift. It minimizes an utterance-level SNR loss function (Eq. 8) to predict two direct-sound speaker signals from a reverberant mixture in the time domain.
- IRM [19]: We report the results for the signals reconstructed by applying the oracle IRM. This separation is not trained and the IRM is directly calculated from premixed signals. The overlap-add method uses $\mathrm{IRM}_i \otimes |Y|$ and $\angle Y$ to resynthesize time-domain signal s_i.

[1] Mixture generation script from http://www.merl.com/demos/deep-clustering/create-speaker-mixtures.zip

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

Table 1

Objective scores for speaker separation using the WSJ0 test mixtures. Average scores for same- and different-gender mixtures are reported.

Metrics	ΔSDR (dB)		ESTOI (%)		PESQ	
Condition	Sim.	Real	Sim.	Real	Sim.	Real
Unprocessed	0.0	0.0	32.5	39.4	1.52	1.65
One-stage deep CASA	9.1	9.8	70.4	72.2	2.51	2.56
Two-stage deep CASA	8.9	10.5	72.4	75.3	2.54	2.64
uPIT Dense-UNet	7.4	8.0	63.2	67.3	2.32	2.42
Conv-TasNet [17]	7.4	7.0	62.6	62.8	2.30	2.26
IRM	9.1	10.5	76.9	82.4	2.88	3.22

Table 2

Mean speech intelligibility (in %) for normal-hearing and hearing-impaired listeners in unprocessed and processed (separated) two-talker reverberant mixtures.

TIR	T_{60} (s)	Normal-hearing		Hearing-impaired	
		Unprocessed	Processed	Unprocessed	Processed
-5 dB	0.6	39.7	84.9	-	-
	0.9	13.4	60.8		
0 dB	0.6	68.3	89.5	8.7	81.2
	0.9	51.4	84.3	5.2	67.7
5 dB	0.6	-	-	25.6	84.2
	0.9			13.7	78.8

4.3. Results and comparisons

We begin with evaluating the systems using the WSJ0 test speakers. Table 1 shows objective evaluation scores in simulated and real reverberant room conditions using such mixtures. All systems yield substantial improvement over the unprocessed mixtures. Except in the simulated room conditions with the ΔSDR metric, the two-stage deep CASA produces higher scores than one-stage deep CASA. These results suggest that addressing speech dereverberation and speaker separation in two different stages results in better generalization to unseen speakers and room conditions. The IRM has significantly better ESTOI and PESQ results than the deep CASA systems, but when measured by ΔSDR the two-stage deep CASA system has comparable performance to the IRM. Consistent with speaker separation results in anechoic conditions [20], uPIT Dense-Unet and Conv-TasNet scores are worse than the deep CASA methods across all three objective metrics. uPIT Dense-Unet and Conv-TasNet have similar performance in simulated RIR conditions, while the former has significantly better results with real room mixtures.

To examine the quality of sequential organization in the uPIT Dense-Unet and the deep CASA systems, Fig. 3 plots the rate of incorrect frame-speaker assignments for these systems in simulated and recorded RIR test conditions. Frame assignment errors are shown for frames with significant energy, i.e. those with at least -20 dB relative to the frame with the maximum energy in the mixture. Apparently, uPIT Dense-Unet has much higher assignment errors than the deep CASA systems. In general, the sequential grouping module in two-stage deep CASA has better performance than one-stage deep CASA. We think that this is because the output frames from the simultaneous grouping module in two-stage deep CASA are better separated (and thus less similar), which facilitates sequential organization.

A simplified version of the proposed deep CASA algorithm was recently evaluated on normal-hearing (NH) and hearing-impaired (HI) listeners by Healy *et al.* [34]. In this speech intelligibility test, reverberant two-talker mixtures were generated using IEEE speakers [35]. One-hundred and sixty speech mixtures were generated with T_{60} = 0.6 s or 0.9 s, and each mixture comprised one male and one female utterance. Reverberant male and female signals were mixed at target-to-interferer ratio (TIR) of -5, 0, or 5 dB, with the male speaker designated as the target. Note that the IEEE sentences and speakers were not used during training, which used the WSJ0 corpus. Ten HI and ten NH listeners participated in this experiment. Table 2 shows the average intelligibility scores for NH and HI listeners. As shown in the table, the algorithm benefit is substantial for

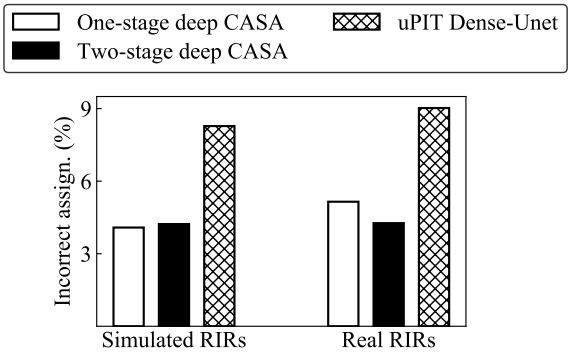

Fig. 3: Sequential organization error rates in different systems.

both groups across all test conditions. Particularly large speech intelligibility improvements are observed for HI listeners. See [34] for more details of the listening test. The listening results show that the deep CASA algorithm is remarkably effective in improving the intelligibility of human listeners in reverberant two-talker conditions.

5. CONCLUDING REMARKS

In this paper, we have proposed a two-stage deep CASA algorithm to address talker-independent speaker separation in reverberant conditions. Our evaluations with reverberant two-talker mixtures indicate that two-stage deep CASA outperforms one-stage deep CASA, which in turn outperforms Conv-TasNet and Dense-Unet systems. A major difference between deep CASA and these systems lies in their frame organization mechanisms. uPIT forces speaker frames to occur at the same output layer throughout a mixture signal which limits the flexibility of the system. Deep CASA uses tPIT and predicts the underlying speaker-frame assignments with high accuracy at the sequential grouping module. In addition, the deep CASA algorithm can substantially improve the speech intelligibility of normal-hearing and hearing-impaired listeners. In the future, we plan to extend two-stage deep CASA to speaker separation involving more than two speakers and in noisy-reverberant conditions.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2871–2876, 2003.

[2] B. C. Moore, *Cochlear hearing loss: Physiological, psychological and technical issues*. Chichester, UK: John Wiley & Sons, 2007.

[3] 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.

[4] R. J. Weiss and D. P. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Comput. Speech Lang.*, vol. 24, pp. 16–29, 2010.

[5] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1–12, 2006.

[6] 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.

[7] 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.

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

[9] ——, "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.

[10] 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.

[11] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. L. Wang, "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," *J. Acoust. Soc. Amer.*, vol. 141, pp. 4230–4239, 2017.

[12] E. W. Healy, M. Delfarah, E. M. Johnson, and D. L. Wang, "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," *J. Acoust. Soc. Amer.*, vol. 145, pp. 1378–1388, 2019.

[13] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, 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.

[14] 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.

[15] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed. Kluwer Academic, Norwell MA: Springer, 2005, pp. 181–197.

[16] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[17] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, 2019.

[18] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," *arXiv preprint arXiv:1902.04891*, 2019.

[19] 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.

[20] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 2092–2102, 2019.

[21] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[22] M. Delfarah and D. L. Wang, "Deep learning for talker-dependent reverberant speaker separation: An empirical study," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1839–1848, 2019.

[23] E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1773–1783, 2017.

[24] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1535–1546, 2017.

[25] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 53–62, 2019.

[26] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 483–492, 2016.

[27] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*, 2016, pp. 47–54.

[28] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[29] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," *Philadelphia: Linguistic Data Consortium*, 1993.

[30] 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.

[31] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 2009–2022, 2016.

[32] 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. ICASSP*, 2001, pp. 749–752.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1462–1469, 2006.

[34] E. W. Healy, E. M. Johnson, M. Delfarah, and D. L. Wang, "A talker-independent speaker separation algorithm to increase intelligibility in reverberant conditions," in preparation.

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