DEEP CASA FOR TALKER-INDEPENDENT MONAURAL SPEECH SEPARATION

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

ABSTRACT

Monaural speech separation is the task of separating target speech from interference in single-channel recordings. Although substantial progress has been made recently in deep learning based speech separation, previous studies usually focus on a single type of interference, either background noise or competing speakers. In this study, we address both speech and nonspeech interference, i.e., monaural speaker separation in noise, in a talker-independent fashion. We extend a recently proposed deep CASA system to deal with noisy speaker mixtures. To facilitate speech enhancement, a denoising module is added to deep CASA as a front-end processor. The proposed systems achieve state-of-the-art results on a benchmark noisy two-speaker separation dataset. The denoising module leads to substantial performance gain across various noise types, and even better generalization in noise-free conditions.

Index Terms— Monaural speech separation, speech enhancement, speaker separation, deep CASA

1. INTRODUCTION

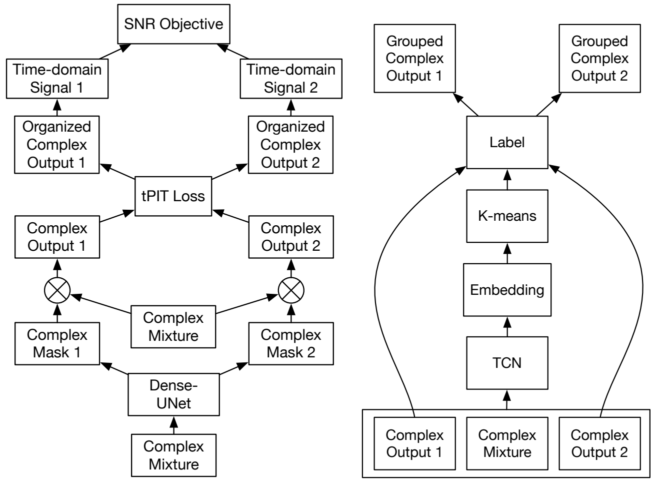
In realistic acoustical environments, the speech sound reaches our ears is usually contaminated by various types of interference, such as environmental noise and competing speakers. Although human listeners excel at attending to the target speaker while filtering out other sound sources [2], speech separation remains a difficult problem for machines over the last few decades. This study focuses on monaural (single-microphone) speech separation, which provides a flexible and cost-efficient approach to the problem.

Based on the type of interference, speech separation can be categorized into speech enhancement and speaker separation. Speech enhancement refers to the task of recovering speech from nonspeech additive noise. The difficulty of speech enhancement stems from the fact that unstructured noise can corrupt structured speech signals in unpredictable ways. With the recent development of deep learning, speech enhancement has been formulated as a supervised learning problem. Typically, a neural network is used to project noisy features to some representation of clean speech. Various learning machines have been explored for this task, including feedforwad neural networks (FNNs) [23], recurrent neural networks (RNNs) [4], and convolutional neural networks (CNNs) [18]. The models are trained to estimate spectral masks [4, 22, 26], spectral mappings [21], or time-domain signals [17]. With multi-condition large-scale training, deep learning based approaches can effectively reconstruct the contaminated patterns of speech, and generalize well to untrained conditions [1].

On the other hand, monaural speaker separation aims to separate several concurrent speakers from a single-channel recording. Unlike speech enhancement, both the target and interfering speakers are highly structured in speaker separation. A talker-dependent neural network can easily map a specific speaker to one of its output layers, leading to significant improvements in speech quality and intelligibility [3, 7, 27]. However, when it comes to talker-independent speaker separation, we must solve the permutation problem [12] due to the unknown correspondence between speakers and outputs. Many algorithms have been proposed recently to address the this problem, among which permutation invariant training (PIT) [11] and deep clustering (DC) [6] represent two major approaches. In PIT, all possible output-speaker permutations are scanned during training, and the network is optimized with respect to the permutation which leads to the minimum loss. There are two types of PIT. In frame-level PIT (tPIT), the output-speaker permutation can vary frame-by-frame, which leads to optimized frame-level separation. A fixed permutation throughout each training utterance corresponds to utterance-level PIT (uPIT), which leads to good utterance-level performance. In contrast, DC projects each time-frequency (T-F) unit of the mixture to an embedding space. Clustering the embeddings results in a set of binary masks, each of which can be used to extract one speaker. Inspired by human auditory scene analysis (ASA) mechanisms, we proposed a two-stage deep CASA system in [13], with a simultaneous grouping stage performing frame-level separation using a tPIT network, followed by a sequential grouping stage performing speaker tracking using a clustering network. Compared to PIT, DC and their variants [14, 16, 24], deep CASA substantially mitigates the mistakes in speaker tracking, and leads to better results in speaker separation.

Although monaural speech enhancement and speaker separation have been well-studied separately, little effort has been made to study the joint problem, despite its more practical usage, as real-world interference is not restricted to one type. One reason is that one has to deal with the permutation problem and unstructured noise corruption at the same time in noisy speaker separation. The two problems have adverse effects on each other, and are compounded into a more complex problem. Some early trials of talker-independent noisy speaker separation include [10] and [25]. In [10], a uPIT RNN is trained with various noise types and signal to noise ratios (SNRs). In [25], a noisy two-speaker dataset (WHAM!) is created and published, where various speech separation architectures are then benchmarked. Both studies report limited performance gain for this task.

In this study, we extend our deep CASA framework to monaural speaker separation in noise. First, we retrain a baseline deep CASA system [13] on the WHAM! dataset [25]. A light-weight denoising module is then added to deep CASA for front-end processing, resulting in what we call a denoising deep CASA system. Both systems achieve excellent results on WHAM!. Thanks to the additional denoising module, denoising deep CASA performs consistently better across various untrained noise types and SNR conditions. The



(a) (b)

Fig. 1: Diagrams of (a) the simultaneous grouping stage and (b) the sequential grouping stage in deep CASA.

denoising module also leads to better generalization to noise-free speaker mixtures.

The rest of this paper is organized as follows. The baseline deep CASA system for noisy speaker separation is described in Section 2. In Section 3, we present the denoising deep CASA system. Experimental results and comparisons are discussed in Section 4. Section 5 concludes the paper.

2. DEEP CASA FOR MONAURAL SPEECH SEPARATION

The goal of monaural speech separation is to extracting C concurrent speakers $x_c(t)$, $c = 1, ..., C$, from a single-channel recording of mixture $y(t)$:

$$y(t) = \sum_{c=1}^{C} x_c(t) + n(t) \tag{1}$$

where t indexes time and $n(t)$ denotes nonspeech noise. This study focuses on the co-channel situation where $C = 2$.

In [13], we have proposed a deep CASA system for talker-independent monaural speaker separation, which corresponds to a special case of monaural speech separation where $n(t) = 0$. There are two stages in deep CASA. The simultaneous grouping stage takes the mixture signal and separates concurrent speakers at the frame level. The frame-level separated spectra are then fed to the sequential grouping stage for speaker tracking. Deep CASA optimizes separation and speaker tracking in turn, and achieves state-of-the-art results in both objectives.

Although deep CASA is designed for the noise-free condition, it can be readily extended to speech separation in general, where competing speakers and background noise are both present. The details are presented in two stages in the following subsections.

2.1. Simultaneous Grouping

A diagram of the simultaneous grouping stage is illustrated in Fig. 1a. Given the complex short-time Fourier transform (STFT) of the mixture $Y(m, f)$, where m and f index frame and frequency, simultaneous grouping is performed to separate the two speakers at the frame level.

A Dense-UNet first takes $Y(m, f)$ as input and predicts two complex ratio masks $CRM_1(m, f)$ and $CRM_2(m, f)$. The masks are multiplied to $Y(m, f)$ to generate two complex outputs, $\hat{X}_1(m, f)$ and $\hat{X}_2(m, f)$, which estimate the complex STFT of the two speakers. The training of Dense-UNet follows the tPIT criterion, where the frame-level output-speaker pairing is chosen as the pairing that leads to the minimum loss. The outputs are then organized into two streams, $\hat{X}_{o_1}(m, f)$ and $\hat{X}_{o_2}(m, f)$, using the resulting tPIT pairings. Next, two time-domain signals, $\hat{x}_{o_1}(t)$ and $\hat{x}_{o_2}(t)$, are generated by applying inverse STFT to the organized outputs. A signal-to-noise ratio (SNR) objective $J^{tPIT-SNR}$ is calculated for backpropogation:

$$J^{tPIT-SNR} = \sum_{c=1}^{2} 10 \log \frac{\Sigma_t x_c(t)^2}{\Sigma_t [x_c(t) - \hat{x}_{o_c}(t)]^2} \tag{2}$$

The Dense-UNet in simultaneous grouping comprises a series of upsampling layers, downsampling layers, and dense convolutional blocks. The details of the Dense-UNet, including the number of layers, downsampling, upsampling, skip connections, dense convolutional blocks, and frequency mapping, follow those in [13].

2.2. Sequential Grouping

The sequential grouping stage tracks frame-level spectral estimates $\hat{X}_1(m, f)$ and $\hat{X}_2(m, f)$, and assigns them to the two speakers. A diagram is given in Fig. 1b.

$Y(m, f)$, $\hat{X}_1(m, f)$ and $\hat{X}_2(m, f)$ are stacked to form the input to this stage. A temporal convolutional network (TCN) projects each frame-level input to a D-dimensional embedding vector $\mathbf{V}(m) \in \mathbb{R}^{1 \times D}$. The target labels for TCN training are two-dimensional indicator vectors, denoted by $\mathbf{A}(m)$. $\mathbf{A}(m)$ gives a one-hot representation of the tPIT pairings in simultaneous grouping. $\mathbf{A}(m)$ = [1, 0] if Speaker 1 is tied to Output 1, and Speaker 2 is tied to Output 2 in Dense-UNet training, and [0, 1] otherwise. A weighted objective function between $\mathbf{V}$ and $\mathbf{A}$ (see [13] for details) forces $\mathbf{V}(m)$ corresponding to the same tPIT pairing to get closer during training, and otherwise to become farther apart.

At the inference time, clustering $\mathbf{V}(m)$ with the K-means algorithm generates a binary label for each frame, which can be used to organize the frame-level outputs from simultaneous grouping, and form the final outputs of deep CASA.

The TCN in sequential grouping consists of a sequence of dilated convolutional blocks. The long temporal context of TCN makes it suitable for speaker tracking. The details of the TCN, including feature preprocessing, the number of layers, dilated convolutional blocks, and dropDilation, follow those in [13].

3. DENOISING DEEP CASA

Although deep CASA can be readily trained for speaker separation in noise, there are several limitations to this approach. First, in the presence of noise, the simultaneous grouping module has to perform frame-level speaker separation and denoising at the same time, which is much more difficult than speaker separation alone, as unstructured noise can severely corrupt speech patterns. Second, certain types of noise, e.g., music and babble noise, contain speech-like harmonics, which may mislead the simultaneous grouping module to discard weak speakers, and retain noise instead. Third, the errors in simultaneous grouping have an adverse effect on sequential grouping, and thus may greatly degrade the performance of deep CASA.

One possible solution is to treat noise as another sound source, and employ the multi-source deep CASA system in [13] for speech separation. To do so, one additional output layer needs to be added for noise in the simultaneous grouping Dense-UNet. Such a solution may partially reduce the errors in simultaneous grouping thanks to the noise-aware training. However, since noise usually has different patterns from speech, treating them as parallel outputs may not be optimal.

In this study, we propose a denoising front-end module for simultaneous grouping, which aims to remove nonspeech noise, and recover missing speech patterns buried in noise for both speakers. The input to the denoising module is the complex STFT of the mixture $Y(m,f)$. A lite Dense-UNet is trained to predict a complex ratio mask, which is then multiplied to $Y(m,f)$ to produce an estimate of the clean speaker mixture $\hat{X}_{1+2}(m,f)$. Next, $\hat{X}_{1+2}(m,f)$ is converted back to the time domain, denoted by $\hat{x}_{1+2}(t)$. An SNR objective is calculated for denoising:

$$J^{DN-SNR} = 10 \log \frac{\Sigma_t [x_1(t) + x_2(t)]^2}{\Sigma_t [x_1(t) + x_2(t) - \hat{x}_{1+2}(t)]^2} \tag{3}$$

Components in simultaneous grouping are modified in accordance with the denoising module. The input to simultaneous grouping now becomes a stack of $\hat{X}_{1+2}(m,f)$ and $Y(m,f)$, so that reconstructed speech and raw information are both provided. The two estimated complex masks, $CRM_1(m,f)$ and $CRM_2(m,f)$, are now multiple to $\hat{X}_{1+2}(m,f)$ for speech separation. In this way, $CRM_1(m,f)$ and $CRM_2(m,f)$ are trained to only perform speaker separation. The enhancement task is left to the denoising module. During the training of simultaneous grouping, a joint SNR objective, which includes both denoising and speaker separation components, is formed.

$$J^{Joint-SNR} = J^{tPIT-SNR} + J^{DN-SNR} \tag{4}$$

Other details of simultaneous grouping and sequential grouping remain the same. We denote the deep CASA system with a denoising module by denoising deep CASA in this study.

4. EVALUATION AND COMPARISON

4.1. Experimental Setup

We conduct experiments on the noisy two-speaker dataset WHAM! [25], which consists of two-speaker mixtures from the WSJ0-2mix dataset [6] combined with real ambient noise samples. The WSJ0-2mix dataset has a 20,000-mixture training set and a 5,000-mixture validation set generated by selecting random speaker pairs in the Wall Street Journal (WSJ0) training set si_tr_s, and mixing them at SNRs between 0 dB and 5 dB. The test set in WSJ0-2mix has 3,000 mixtures, which are similarly generated using 16 untrained speakers from the WSJ0 development set si_dt_05 and si_et_05. WHAM! pairs each two-speaker mixture in WSJ0-2mix with a nonspeech ambient noise sample, recorded in real environments such as coffee shops, restaurants, and bars. To generate a mixture in WHAM!, a random noise signal and a random SNR between -3 dB and 6 dB are first sampled. The first speaker in the WSJ0-2mix mixture is then scaled so that the SNR between the first speaker and the noise is equal to the randomly sampled value. The same scale is applied to the second speaker. A min version of WHAM! is adopted, where the longer of the two speech signals is truncated. All mixtures are sampled at 8 kHz.

Table 1: Number of parameters, average ΔSDR (dB), ΔSI-SNR (dB), PESQ and ESTOI (%) for various systems evaluated on WHAM! (separate-noisy, 8k-min [25]).

	# of param.	ΔSDR	ΔSI-SNR	PESQ	ESTOI
Mixture	-	0.0	0.0	1.68	34.4
BLSTM-TasNet [25]	23.0M	-	9.8	-	-
One-stage chimera [25]	29.6M	-	9.9	-	-
Two-stage chimera [25]	59.2M	-	10.3	-	-
uPIT Dense-UNet	4.8M	12.2	11.8	2.42	65.3
Deep CASA	12.8M	13.8	13.4	2.58	70.4
Denoising deep CASA	14.0M	**14.7**	**14.4**	**2.63**	**73.0**
IBM	-	14.0	13.4	2.72	77.0

To facilitate comparison, we create another test set on the basis of WHAM!. The first 1,000 mixtures in WHAM!'s test set are remixed, where the two speakers in a mixture are mixed exhaustively with one of the four noises in the CHiME-4 challenge [20], namely bus (BUS), cafe (CAF), pedestrian area (PED) and street (STR) noise, at a speech-to-noise ratio (SNR) of -2, 5, 10 dB, resulting a total of 12,000 noisy two-speaker test mixtures. This test set is denoted by WHAM!-CHiME4. To make a test mixture in WHAM!-CHiME4, the noise needs to be scaled to reach the desired SNR, while the speakers in WHAM! keep their original scales. Here the SNR is calculated using loudness units full-scale (LUFS) [5] to remove silent regions from computation. To reveal the generalization in noise-free conditions, evaluation is also conducted on the first 1,000 two-speaker mixtures in WHAM! test, without mixing them with any noise. STFT with a frame length of 32 ms, a frame shift of 8 ms, and a square root Hanning window is calculated for all systems.

We evaluate the algorithms using signal-to-distortion ratio improvement (ΔSDR) [19], perceptual evaluation of speech quality (PESQ) [8], and extended short-time objective intelligibility (ESTOI) [9]. To make a systematical comparison with other systems, results are also reported in terms of scale-invariant signal-to-noise ratio improvement (ΔSI-SNR) [16].

4.2. Comparison Systems

All of the following comparison systems are trained on WHAM!. The results of the first three systems are reported in [25].

- BLSTM-TasNet [15]: This system performs uPIT [11] in the time domain using a bi-directional long short-term memory (BLSTM) RNN.
- One-stage chimera [25]: This system uses a chimera BLSTM network which simultaneously estimates DC embeddings and uPIT outputs. uPIT outputs are used during inference.
- Two-stage chimera [25]: A denoising BLSTM is added as the first stage before the chimera network [25]. The two stages are then jointly optimized.
- uPIT Dense-UNet: The Dense-UNet structure in simultaneous grouping of deep CASA is trained with a uPIT SNR objective. The detailed settings follow those in [13].
- Deep CASA: This system is described in Section 2. It represents the direct extension of deep CASA to noisy conditions. The detailed training recipes follow those in [13]. The two stages are jointly optimized.
- Denoising deep CASA: A lite version of Dense-UNet with a channel size of 32 is used for front-end denoising. Other

Table 2: ΔSDR (dB), PESQ and ESTOI (%) for various systems evaluated on WHAM!-CHiME4. The results are averaged across -2, 5 and 10 dB, and reported with respect to different noise types.

	BUS			CAF			PED			STR			Average		
	ΔSDR	PESQ	ESTOI	ΔSDR	PESQ	ESTOI	ΔSDR	PESQ	ESTOI	ΔSDR	PESQ	ESTOI	ΔSDR	PESQ	ESTOI
Mixture	0.0	1.80	40.2	0.0	1.68	33.4	0.0	1.66	33.6	0.0	1.70	36.7	0.0	1.71	36.0
uPIT Dense-UNet	13.7	2.62	71.5	11.1	2.34	62.6	10.7	2.31	61.0	12.3	2.44	65.5	12.0	2.43	65.2
Deep CASA	15.7	2.85	77.5	12.8	2.50	68.0	12.1	2.44	65.7	13.9	2.60	70.8	13.7	2.60	70.5
Denoising deep CASA	**16.7**	**2.91**	**79.7**	**13.6**	**2.54**	**70.4**	**12.9**	**2.48**	**68.3**	**14.7**	**2.64**	**73.2**	**14.5**	**2.64**	**72.9**
IBM	15.4	2.98	79.1	13.2	2.68	75.6	12.7	2.58	76.5	14.1	2.73	77.2	13.8	2.74	77.1

Table 3: ΔSDR (dB), PESQ and ESTOI (%) for various systems evaluated on WHAM!-CHiME4. The results are averaged across BUS, CAF, PED and STR noise, and reported with respect to different SNR values.

	SNR = -2 dB			SNR = 5 dB			SNR = 10 dB			Noise free		
	ΔSDR	PESQ	ESTOI	ΔSDR	PESQ	ESTOI	ΔSDR	PESQ	ESTOI	ΔSDR	PESQ	ESTOI
Mixture	0.0	1.56	26.1	0.0	1.73	37.4	0.0	1.84	44.4	0.0	2.03	56.2
uPIT Dense-UNet	12.4	2.10	51.5	11.7	2.50	68.6	11.8	2.68	75.4	12.6	2.88	82.4
Deep CASA	13.6	2.19	56.3	13.5	2.69	74.1	13.8	2.91	81.0	15.4	3.19	88.7
Denoising deep CASA	**14.6**	**2.24**	**59.8**	**14.3**	**2.73**	**76.3**	**14.5**	**2.96**	**82.5**	**16.4**	**3.32**	**90.2**
IBM	15.4	2.46	69.0	13.3	2.80	78.9	12.9	2.97	83.4	13.8	3.28	89.1

details follow Section 3 and [13]. The two stages are jointly optimized.

- Ideal binary mask (IBM): The IBM is defined as:

$$\mathrm{IBM}_c(m,f) = \begin{cases} 1, & \text{if } |X_c(m,f)| > |Y(m,f) - X_c(m,f)| \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

4.3. Results and Comparisons

Table 1 compares deep CASA with other talker-independent speech separation methods on the noisy two-speaker dataset WHAM! [25]. For all methods, we list the best reported results, and leave unreported fields blank. The numbers of parameters in BLSTM-TasNet, one-stage and two-stage chimera are estimated according to [25]. As shown in the table, all proposed methods significantly outperform the benchmark results presented in [25]. The baseline deep CASA system produces better results than uPIT Dense-UNet, thanks to its two-stage processing scheme. The denoising module for deep CASA introduces a substantial gain in terms of all metrics, with only a small portion of additional parameters. Denoising deep CASA surpasses the IBM in terms of ΔSDR and ΔSI-SNR, and generates very close PESQ and ESTOI.

To evaluate the generalization of deep CASA to untrained noise types, we test the proposed systems on WHAM!-CHiME4, and present the results in Table 2. Similar to Table 1, denoising deep CASA substantially outperforms uPIT Dense-UNet and the baseline deep CASA across different noise types. For CAF and PED noise, which contain noticeable background speech, there is still a small gap between denoising deep CASA and the IBM, due to the interference of background speakers. For BUS and STR, where noise mostly comes from traffic, the gap has been greatly reduced. The average results across all noise types are very similar to those in Table 1, demonstrating the noise generalization ability of the proposed methods.

Table 3 presents the results on WHAM!-CHiME4 with respect to different speech-to-noise ratios (SNRs). Again, the performance of denoising deep CASA leads consistently across all reported SNRs. The last three columns correspond to the results on noise-free two-speaker mixtures, where uPIT Dense-UNet and deep CASA deliver much worse performance than their WSJ0-2mix results in [13], mainly attributed to two factors. First, the mixtures in WHAM! have a much larger range of energy levels than WSJ0-2mix, making the task slightly harder for speech separation systems. Second, the uPIT Dense-UNet and deep CASA systems here are trained on noisy data, and tested on mismatched noise-free data. Better results are expected if noise-free samples are also included in training. With the help of the denoising module, denoising deep CASA makes the tasks of speech enhancement and speaker separation relatively independent from each other, and keeps the speaker separation module robust to different noise levels. It also leads to much better generalization in the noise-free condition, as shown by the underlined results.

5. CONCLUSION

We have extended the deep CASA framework [13] for general monaural speech separation, namely, speaker separation in noise. A denoising preprocessing module is added to deep CASA to improve the noise robustness of the model. The proposed denoising deep CASA system substantially outperforms all published results, and surpasses the IBM in terms of ΔSDR. Further examination reveals that the denoising module leads to good generalization to different noise types and levels, as well as the noise-free condition. In the future, we plan to extend deep CASA for more realistic acoustical environments, which include conversational speech and room reverberation.

6. ACKNOWLEDGEMENTS

This research was supported in part by an NIDCD grant (R01DC012048) and the Ohio Supercomputer Center.

7. REFERENCES

[1] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, “Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises,” *J. Acoust. Soc. Amer.*, vol. 139, pp. 2604–2612, 2016.

[2] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, 1953.

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

[4] H. Erdogan, J. R. Hershey, and S. Watanabe, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *Proc. ICASSP*, 2015, pp. 708–712.

[5] E. Grimm, R. Van Everdingen, and M. Schöpping, “Toward a recommendation for a European standard of peak and LKFS loudness levels,” *SMPTE Motion Imaging Journal*, vol. 119, pp. 28–34, 2010.

[6] 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.

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

[8] ITU-R, “Perceptual evaluation of speech quality (PESQ) An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” *Recommendation P.862*, 2001.

[9] 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.

[10] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training,” in *Proc. IEEE MLSP*, 2017.

[11] 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.

[12] Y. Liu and D. L. Wang, “A CASA approach to deep learning based speaker-independent co-channel speech separation,” in *Proc. ICASSP*, 2018, pp. 5399–5403.

[13] ——, “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.

[14] Y. Luo, Z. Chen, and N. Mesgarani, “Speaker-independent speech separation with deep attractor network,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 787–796, 2018.

[15] Y. Luo and N. Mesgarani, “Tasnet: Time-domain audio separation network for real-time, single-channel speech separation,” in *Proc. ICASSP*, 2018, pp. 696–700.

[16] ——, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, 2019.

[17] A. Pandey and D. L. Wang, “A new framework for CNN-based speech enhancement in the time domain,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1179–1188, 2019.

[18] S. R. Park and J. W. Lee, “A fully convolutional neural network for speech enhancement,” in *Proc. Interspeech*, 2017, pp. 1993–1997.

[19] 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.

[20] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” in *Computer Speech and Language*, 2016.

[21] D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[22] 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.

[23] 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.

[24] Z.-Q. Wang, K. Tan, and D. L. Wang, “Deep learning based phase reconstruction for speaker separation: A trigonometric perspective,” in *Proc. ICASSP*, 2019, pp. 71–75.

[25] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, “WHAM!: Extending speech separation to noisy environments,” in *Proc. Interspeech*, 2019, pp. 1368–1372.

[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] 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.