INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

On Synthesis for Supervised Monaural Speech Separation in Time Domain

Jingjing Chen[1], *Qirong Mao*[1,2], *Dong Liu*[1]

[1]School of Computer Science and Communication Engineering, Jiangsu University, China
[2]Jiangsu Engineering Research Center of big data ubiquitous perception and intelligent agriculture applications, Zhenjiang, China

2221808071@stmail.ujs.edu.cn, mao_qr@ujs.edu.cn, 2111908002@stmail.ujs.edu.cn

Abstract

Time-domain approaches for speech separation have achieved great success recently. However, the sources separated by these time-domain approaches usually contain some artifacts (broadband noises), especially when separating mixture with noise. In this paper, we incorporate synthesis way into the time-domain speech separation approaches to deal with above broadband noises in separated sources, which can be seamlessly used in the speech separation system by a 'plug-and-play' way. By directly learning an estimation for each source in encoded domain, synthesis way can reduce artifacts in estimated speeches and improve the speech separation performance. Extensive experiments on different state-of-the-art models reveal that the synthesis way acquires the ability to handle with noisy mixture and is more suitable for noisy speech separation. On a new benchmark noisy dataset, the synthesis way obtains 0.97 dB (10.1%) SDR relative improvement and respective gains on various metrics without extra computation cost.

Index Terms: synthesis way, speech separation, time domain, deep learning

1. Introduction

Speech separation, also known as audio source separation, is a significant task in signal processing, which is commonly known as the cocktail party problem [1, 2]. The aim of speech separation is to separate each source from mixed audio signals and this work is very important for some real-world applications. For example, it can separate clean speech from noisy speech signals to improve the accuracy of automatic speech and speaker recognition. A machine solution to this problem is critical to enable resolvability of various scenarios such as meeting transcription, hearing prosthesis, mobile telecommunication, and so on. In general, there are two types of speech separation: monaural speech separation and multi-microphone speech separation. This study focuses on monaural separation since it has low requirements for sensors. Compared to multi-microphone solutions, a monaural system is less sensitive to room reverberation and spatial source configuration [3]. On the other hand, monaural separation is a severely underdetermined figure-ground separation problem, which is typically more challenging than its multi-microphone counterpart.

Regardless of the challenge in monaural speech separation, a lot of attempts have been made in previous works to dispose of this problem. Before the rise of deep learning, many traditional methods were proposed for this task, such as non-negative matrix factorization (NMF) [4, 5] and computational auditory scene analysis (CASA) [6]. However, they only work well in the condition that there is prior information about the speakers. With the success of the deep learning techniques on various domains[7, 8], a large number of deep learning-based methods have been proposed for monaural speech separation, and they usually design data-based models to separate the mixture of unknown speakers without prior information. In general, these techniques can be divided into two categories. The first category is based on time-frequency features created by calculating the short-time Fourier transform (STFT) [9, 10, 11, 12, 13], where the time-frequency features for each source are separated and then are used to reconstruct the source waveforms by inverse STFT. The second category is end-to-end speech separation in time domain [14, 15, 16, 17, 18, 19, 20], which is a natural way to overcome the phase errors in time-frequency methods.

Because it is difficult to separate the phase, most of first-category methods only modify the magnitude, and then directly use the original phase of mixture to reconstruct the estimated source waveforms by inverse STFT. This modeling strategy is usually defective and imposes an obvious upper bound on the separation performance [14, 17]. To overcome above limits, paper [14] proposes the second category of approaches, namely end-to-end speech separation in time domain. Time-domain approaches directly model the mixture waveform using an encode-decoder framework, and perform the separation for each source on the output of the encoder. These approaches focus most of their attention on learning long-term temporal dependency and pay little attention to the way of decomposition. For the question of how to perform the separation, they empirically construct a mask for each source, which is a method that achieves best performance in time-frequency methods. However, speech estimated by the masks in time-domain approaches usually contain some artifacts (broadband noises), which could lead to degenerated separation performance, especially when it comes to the separation of mixture with noise.

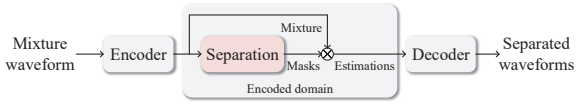
In this paper, we introduce the synthesis method to the filed of time-domain speech separation, which can synthesize the sounds lost in overlaps, thereby helping reduce the artifacts in the estimated speech. Furthermore, the introduced synthesis method has no any additional computing overheads, and it can be seamlessly used in speech separation systems by a 'plug-and-play' way. Note that the synthesis method in this paper refers to an approach to separate the features of the mixed speech, which is totally different from the Text-To-Speech based methods.

The remainder of this paper is organized as follows: In Section 2, we describe the details of the synthesis method for time-domain speech separation. Sections 3 introduces the experiment configurations. The experiment results and discussions are presented in Section 4. Section 5 concludes this paper.

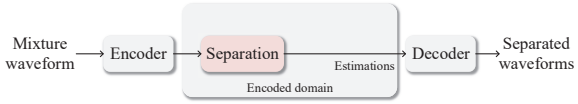
2. Synthesis method for time-domain monaural speech separation

In this section, we first describe the formal definition of the time-domain monaural speech separation. In quick succession,

http://dx.doi.org/10.21437/Interspeech.2020-1150



(a) Flowchart of the masking method



(b) Flowchart of the synthesis method

Figure 1: *Flowchart of masking and synthesis.*

we describe the synthesis method in details and then plug it into two state-of-the-art models.

2.1. Time-domain monaural speech separation

Given the discrete waveform of the mixture $y(t) \in R^{1*T}$, time-domain monaural speech separation is often formulated as directly estimating each source $x_1(t), ..., x_s(t) \in R^{1*T}$ from the mixture. Usually, we assume the sources are mixed linearly, which can be expressed by the following formula:

$$y(t) = \sum_{s=1}^{S} x_s(t) \tag{1}$$

where $s = 1, ..., S$, and S is the number of sources [14].

2.2. Synthesis method

Previous monaural speech separation approaches in time domain usually learn a mask for each source, and we present the flowchart of the masking method in Figure 1(a). The masks are multiplied with the mixed speech signals to obtain the estimation for each source in encoded domain. In the case of separating noisy mixture, the overlaps of the speech and loud noise sometimes lead to a loss of information of the clean speech, which is required to be reversible by some kind of operation. However, the masking operation cannot synthesize new sounds due to the linearity of the encoder and decoder, thus it is tough for the masking method to deal with the problem of separating noisy mixture. The underlying reason is that time-domain approaches based on masking method use an over-complete linear representation on which they apply a mask obtained from the separation network [21].

To overcome the above limitation, we propose to directly learn the estimations in encoded domain, namely the synthesis method, which is shown in Figure 1(b). Similar to the masking method, the synthesis method works on the encoded domain and generates an estimation for each source. The difference between them is that the synthesis method has no the intermediate process of learning the masks. The main advantage of the synthesis method is that it can synthesize the sounds lost in overlaps, which is quite important to reduce artifacts in separated speech and improve the performance of speech separation. In addition, we normalize the estimated waveforms to the range of [-1, 1], which does not interfere the scores of evaluation metrics but is beneficial to regulate the volume and audibility of the separated speech.

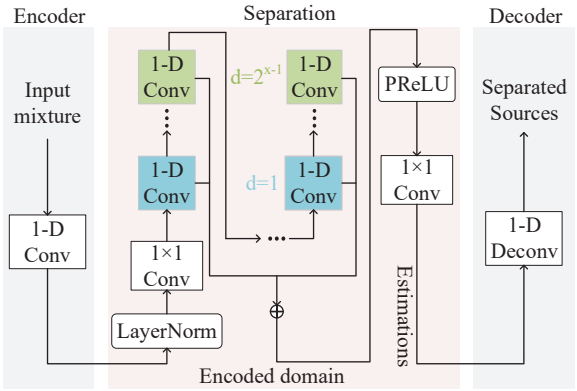


Figure 2: *Architecture of Conv-TasNet with synthesis method.*

2.3. Conv-TasNet with synthesis method for time-domain monaural speech separation

One of the utilized neural network for speech separation is the convolutional time-domain audio separation network (Conv-TasNet) as presented in [15], which is the first model that surpasses ideal time-frequency magnitude masking in this field. We show the overall framework of Conv-TasNet fused with the synthesis method in Figure 2, which consists of an encoder, a separation network and a decoder.

The mixed speech $y \in R^{1*T}$ can be divided into overlapping segments with length L, represented by $\overline{y} \in R^{L*I}$, where I is the total number of segments. The encoder can be described as a super-filter with N filters of length L, which is actually a 1-D convolution module. The output signal $Y \in R^{N*I}$ of the encoder for the mixed speech signal $\overline{y}$ is computed as:

$$Y = \overline{y} * W^{encoder} \tag{2}$$

where the rows in $W^{encoder} \in R^{N*L}$ are the encoder basis functions, each with length L. The separation network is a fully-convolution separation module that consists of stacked 1-D dilated convolutional blocks. It learns the temporal dependency in encoded domain and finally outputs an estimation $X_s \in R^{N*I}$ for the s-th source. In the decoder, the separated speech signal for the s-th source $\overline{x}_s \in R^{L*I}$ is reconstructed by a deconvolution module as below:

$$\overline{x}_s = X_s * W^{decoder} \tag{3}$$

where $W^{decoder} \in R^{N*L}$ contains N vectors (decoder basis functions) with length L each. The overlapping reconstructed segments are overlap-added to generate the final waveform $\tilde{x}_s \in R^{1*T}$.

2.4. DPRNN with synthesis method for time-domain monaural speech separation

The second utilized framework is a time-domain audio separation network with dual-path recurrent neural network (DPRNN) [20], which is a extremely effective and simple module that brings the best results for the task of monaural speech separation. The encoder and the decoder in this framework are the same as those in Conv-TasNet, while the biggest difference is that this architecture use DPRNN to replace the fully-convolution model as the separation module. The DPRNN

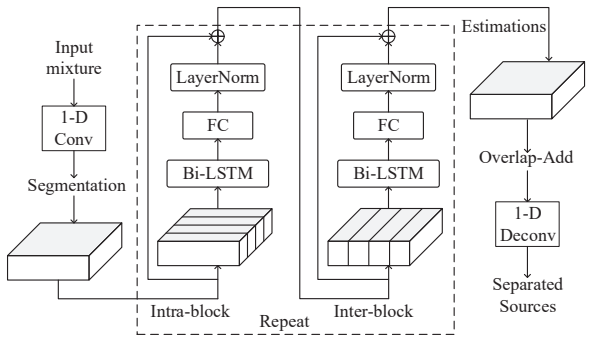


Figure 3: *Architecture of DPRNN with synthesis method. (Repeat the intra-block and inter-block B times)*

module is composed of three stages: segmentation, block processing and overlap-add, and the specific details of SDPRNN (DPRNN with synthesis method) are presented in Figure 3.

For a sequential $Y \in R^{N*I}$ with feature dimension of N and time steps of I, the segmentation stage first splits Y into overlapped chunks of length K and hop size H, and then concatenates all the chunks into a 3-D tensor $U \in R^{N*K*P}$. The output U of segmentation stage is then transmitted to the stack of B SDPRNN blocks, where each block consists of two sub-modules corresponding to intra-chunk and inter-chunk processing respectively. The intra-chunk and inter-chunk processing block are applied to the second and last dimension of U alternately and iteratively as follows:

$$\begin{aligned} U_b^{intra} &= Intrablock_b[U_{b-1}^{inter}] \\ &= U_{b-1}^{inter} + LN([f_b^{intra}(U_{b-1}^{inter}[:,:,i]), i = 1, ..., P]) \end{aligned} \tag{4}$$

$$\begin{aligned} U_b^{inter} &= Interblock_b[U_b^{intra}] \\ &= U_b^{intra} + LN([f_b^{inter}(U_b^{intra}[:,j,:]), j = 1, ..., K]) \end{aligned} \tag{5}$$

where $b = 1, ..., B$, $U_0^{inter} = U$, f represents the Bi-LSTM and fully-connected layers. Finally, the output of the last SDPRNN block, namely $U_B^{inter} \in R^{N*K*P}$, is used to learn an estimation for each source. And then these estimations are transformed back into sequence $X_s \in R^{N*I}$ by overlap-adding.

3. Experiments

3.1. Dataset

We evaluate the synthesis method to the masking method on the tasks of two-speaker speech separation with or without noise. With references to the methods in [9, 22], we create a common monaural dataset (LS-2mix) and a noisy monaural dataset (LS-2mixNoise) from the Librispeech dataset [23] and a well-known noise corpus [24]. The Librispeech dataset is a new corpus of reading English speech, which is suitable for training and evaluating speech separation and recognition systems. Two speakers are randomly selected from the train-100 set of Librispeech to generate the mixtures of LS-2mix train set, at various Signal-to-Noise Ratios (SNRs) uniformly sampled between 0 dB and 5 dB. The validation and test set are similarly generated using utterances from unseen speakers in the Librispeech validation and test set. In LS-2mix, we constructed 15k and 1.5k mixtures in total for the training and validation set, respectively, and 1.5k mixtures for the test set. To create LS-2mixNoise, the two-speakers mixture is further corrupted by a random noise signal sampled from a noise corpus [24]. The SNRs between the mixture and the noise are randomly chosen between 0 dB and 5 dB, and the noise is repeated if its length is smaller than the mixture. The complete source code for creating the two datasets is available at "https://github.com/ujscjj/On-Synthesis".

3.2. Experiment setup

We use Scale-Invariant Source-to-Noise Ratio (SI-SNR) as loss function, which is similar to the standard Signal-to-Distortion Ratio (SDR) [14]. If we denote clean and estimated source by x and $\tilde{x}$ respectively, then the SI-SNR can be formulated as:

$$s_{target} = \frac{\langle \tilde{x}, x \rangle x}{\|x\|^2} \tag{6}$$

$$e_{noise} = \tilde{x} - s_{target} \tag{7}$$

$$SI - SNR := 10log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \tag{8}$$

where x and $\tilde{x}$ are both normalized to zero-mean before the calculation to ensure the scale-invariance. The utterance-level permutation invariant training (uPIT) [12] is employed in our experiments to deal with the permutation problem.

We use a training principle similar to that in [20], which means that the learning rate is initialized to 0.001 and decays by 0.98 for every two epochs. The maximum epoch is set to 100, and the criteria for early stopping is no decrease in the loss function on validation set for 10 epochs.

3.3. Evaluation metrics

For performance comparison of all tested models and methods, we use Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR) [25] and Short-Time Objective Intelligence (STOI) [26] as evaluation metrics, which are often employed in various speech separation systems.

4. Results and discussions

4.1. Performance of the synthesis method on mixture with or without noise

Table 1: *The performance of the masking and synthesis method on Conv-TasNet*

(a) Performance difference on LS-2mix

Method	SDR(dB)	SIR(dB)	SAR(dB)	STOI
Masking	**12.58**	27.12	12.87	**0.90**
Synthesis	12.56	**27.36**	**12.90**	**0.90**

(b) Performance difference on LS-2mixNoise

Method	SDR(dB)	SIR(dB)	SAR(dB)	STOI
Masking	7.81	23.31	8.07	0.80
Synthesis	**8.34**	**25.32**	**8.61**	**0.81**

To prove the superiority of the synthesis method, we investigate the performance of the synthesis method as well as the masking method on the task of separating mixture with or without noise. Based on Conv-TasNet, we conduct experiments on the LS-2mix and LS-2mixNoise datasets. The results are shown in Table 1(a) and Table 1(b) respectively.

As can be seen from Table 1(a), the performance of synthesis method is almost identical to that of masking method on the common monaural dataset. However, with the occurrence of background noise in speech mixture, there is a significant difference between the two methods on all metrics, which is shown in Table 1(b). On the LS-2mixNoise dataset, the synthesis method outperforms the masking method with an improvement about 0.53 dB, 2.01 dB and 0.54 dB for SDR, SIR and SAR respectively. With respect to STOI, the synthesis method provides a 0.01 improvement, where a 1% absolute improvement of STOI is considered significant [27, 28]. Results on various metrics prove the superiority of the synthesis method on the task of speech separation with background noise.

4.2. Performance comparison on mixture with noise

Table 2: *The performance comparison of DPRNN with the masking and synthesis method on LS-2mixNoise*

Method	SDR(dB)	SIR(dB)	SAR(dB)	STOI
Masking	9.61	28.29	9.80	0.83
Synthesis	**10.58**	**34.45**	**10.72**	**0.85**

Table 3: *The performance comparison of DPRNN with the masking and synthesis method at different SNRs*

SNR	Method	SDR	SIR	SAR	STOI
10 dB	*Masking*	11.26	28.87	11.47	0.87
	Synthesis	**11.99**	**32.83**	**12.15**	**0.88**
5 dB	*Masking*	10.34	28.59	10.54	0.85
	Synthesis	**11.27**	**32.66**	**11.42**	**0.87**
0 dB	*Masking*	8.76	27.90	8.94	0.81
	Synthesis	**9.81**	**32.02**	**9.94**	**0.84**

To prove the effectiveness and generalization of the synthesis method, we conduct related experiments of separating noisy mixture on the DPRNN moudle, which is the state-of-the-art model in the field of monaural speech separation. Table 2 shows the separation performance of the two methods for learning source estimations. For the several metrics, the results of the synthesis method are 10.58 dB, 34.45 dB, 10.72 dB and 0.85, with an improvement of 0.97 dB (10.1%), 6.16 dB, 0.92 dB and 0.02 over the masking method. The success on the DPRNN module bears out the effectiveness and generalization of the synthesis method.

In addition, to study the effect of the noise intensity on the synthesis method, we investigate the separation performance of the masking and synthesis method at different SNRs. The SNRs between speech and noise are set to decrease incrementally, namely 10 dB, 5 dB and 0 dB, which means that the noise in mixture gets louder gradually. As can be seen from Table 3, the synthesis method is always superior to the masking method regardless of the intensity of the noise in the mixed speech. In particular, there is an interesting phenomenon: the increments of all metrics increase as the noise increases. For example, on the metric of STOI, the synthesis method outperforms the masking method with improvements of 0.01/0.02/0.03 at the SNRs of 10/5/0 dB. At the same SNRs, the increments for SDR, SIR and SAR show growing trends similar to that of STOI. Obviously,

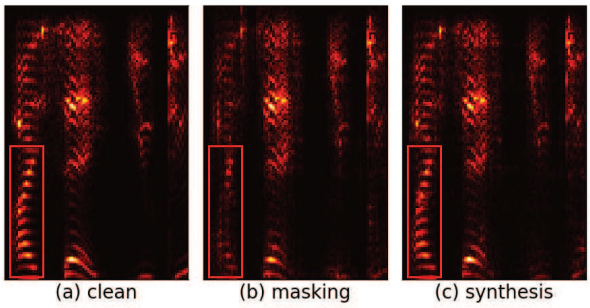


Figure 4: *Magnitude spectrogram of a source in the mixture. From left to right: (a) clean source, (b) estimation of the masking method, (c) estimation of the synthesis method*

with the increase of noise in mixed speech, the advantages of the synthesis method become more prominent. This is an interesting and unique character specific to the synthesis method. Moreover, the noisy mixture with 10 dB SNR has never been in the training set. It is clearly that the synthesis method still performs better when there is a mismatch between the SNRs of the training and test set.

4.3. Visualization of estimated speech's magnitude

To gain an intuitive understanding of the superiority of the synthesis method, we visualize and compare the magnitude spectrogram of speech separated by the masking and synthesis method. From the red box in Figure 4, we can observe that the masking method performs really inferior on one note. The reason is that the mixture in the note contains loud noise, which leads to losing some information of the clean speech. Because the encoder and decoder are both linear in time-domain separation systems, it is difficult for the masking operation to generate new sounds, which brings a pack of troubles for recovering the lost information from the mixture. Therefore, artifacts come into being. Owing to the synthetic nature, the synthesis method can deal with this problem by generating the sounds lost in the mixture, ultimately leading to superior performance in the case of separating noisy mixture.

5. Conclusion

In this paper we investigate the effectiveness of the synthesis method for multi-talker monaural speech separation in the presence of a large amount of noise. Experiments on two state-of-the-art models prove the significance of suggested synthesis method. Benefiting from the character of generating new sounds, the synthesis method brings about 0.97 dB (10.1%) relative SDR improvement over the state-of-the-art approach on a new benchmark noisy dataset without any additional computing cost. Clearly, the synthesis method is a 'plug-and-play' way, which can be used seamlessly in speech separation systems. We hope that it can shed light for other speech processing problems.

6. Acknowledgement

This work is supported in part by the Key Projects of the National Natural Science Foundation of China under Grant U1836220, the National Nature Science Foundation of China of 61672267, the Project funded by China Postdoctoral Science Foundation of 2020M671376, and Qinlan Talent Program of Jiangsu Province.

7. References

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] J. F. Woodruff, "Integrating monaural and binaural cues for sound localization and segregation in reverberant environments," Ph.D. dissertation, The Ohio State University, 2012.

[4] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," 2006.

[5] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 66–70.

[6] E. B. D. Wang, G. J. Brown, and C. Darwin, "Computational auditory scene analysis: Principles, algorithms and applications," *Acoustical Society of America Journal*, vol. 124, p. 13, 2008.

[7] J. Gou, Z. Yi, D. Zhang, Y. Zhan, X. Shen, and L. Du, "Sparsity and geometry preserving graph embedding for dimensionality reduction," *IEEE Access*, vol. 6, pp. 75 748–75 766, 2018.

[8] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu, and Y. Xue, "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition," *IEEE Access*, vol. 7, pp. 93 847–93 857, 2019.

[9] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[10] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[11] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[12] 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 Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[13] G.-P. Yang, C.-I. Tuan, H.-Y. Lee, and L.-s. Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," in *Proc. Interspeech*, 2019, pp. 1363–1367.

[14] Y. Luo and N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[15] Y. Luo and Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation," in *Proc. Interspeech*, 2019, pp. 3183–3187.

[17] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, S. Harada, and J. Han, "End-to-end monaural speech separation with multi-scale dynamic weighted gated dilated convolutional pyramid network," in *Proc. Interspeech*, 2019, pp. 4614–4618.

[18] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. Interspeech*, 2019, pp. 1348–1352.

[19] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via TasNet," *arXiv preprint arXiv:1910.11615*, 2019.

[20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," *arXiv preprint arXiv:1910.06379*, 2019.

[21] A. Defossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[22] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6394–6398.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] G. Hu, "100 nonspeech sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[27] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

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