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

Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation

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

The dominant speech separation models are based on complex recurrent or convolution neural network that model speech sequences indirectly conditioning on context, such as passing information through many intermediate states in recurrent neural network, leading to suboptimal separation performance. In this paper, we propose a dual-path transformer network (DPTNet) for end-to-end speech separation, which introduces direct context-awareness in the modeling for speech sequences. By introduces a improved transformer, elements in speech sequences can interact directly, which enables DPTNet can model for the speech sequences with direct context-awareness. The improved transformer in our approach learns the order information of the speech sequences without positional encodings by incorporating a recurrent neural network into the original transformer. In addition, the structure of dual paths makes our model efficient for extremely long speech sequence modeling. Extensive experiments on benchmark datasets show that our approach outperforms the current state-of-the-arts (20.6 dB SDR on the public WSj0-2mix data corpus).

Index Terms: direct context-aware modeling, transformer, dual-path network, speech separation, deep learning

1. Introduction

Speech separation, often referred to as the cocktail party problem [1, 2], is a fundamental task in signal processing with a wide range of real-world applications, such as separating clean speech from noisy speech signals to improve the accuracy of automatic speech recognition. The human auditory system has the remarkable ability to extract separate sources from a complex mixture, while this task seems to be difficult for automatic calculation system, especially when only a monaural recording of mixed-speech is available.

Although there are many challenges in monaural speech separation, a lot of attempts have been made in previous works to deal with this problem over the decades. Before the deep learning era, many traditional methods are introduced for this task, such as non-negative matrix factorization (NMF) [3, 4], computational auditory scene analysis (CASA) [5] and probabilistic models [6]. However, these models usually only work for closed-set speakers, which significantly restricts their practical applications. With the success of deep learning techniques on various domains [7, 8], researches start to design data-based models to separate the mixture of unknown speakers, which overcomes the obstacles of the traditional methods. In general, deep learning techniques for monaural speech separation can be divided into two categories: time-frequency (T-F) domain methods and end-to-end time-domain approaches. Based on T-F features created by calculating the short-time Fourier transform (STFT), T-F methods separate the T-F features for each source and then reconstruct the source waveforms by inverse STFT [9, 10, 11, 12, 13]. They usually use the original phase of mixture to synthesize the estimated source waveforms, which retain the phase of the noisy mixture. This strategy imposes an upper limit on the separation performance. To overcome this problem, time-domain approach is proposed in paper [14] , which directly model the mixture waveform using an encode-decoder framework and has made great progress in recent years [15, 16, 17, 18, 19, 20, 21].

However, the dominant speech separation models are usually based on recurrent neural network (RNN) or convolution neural network (CNN), which cannot model the speech sequences directly conditioning on context [22], leading to suboptimal separation performance. For example, RNN based models need to pass information through many intermediate states. And the models based CNN suffer from the problem of limited receptive fields. Fortunately, the transformer based on self-attention mechanism can resolve this problem effectively [23], in which elements of the inputs can interact directly. Nevertheless, the transformer usually only deals with sequences with length of hundreds, while end-to-end time-domain speech separation systems often model extremely long input sequences, which can sometimes be tens of thousands. Dual-path network is an effective method to deal with this problem [20].

Inspired by the above, we propose a dual-path transformer network (DPTNet) for end-to-end monaural speech separation, which introduces a improved transformer to allow direct context-aware modeling on the speech sequences, leading to superior separation performance. The major contributions of our work are summarized as follows.

1. To the best of our knowledge, this is the first work that introduces direct context-aware modeling into speech separation. This method enables the elements in speech sequences can interact directly, which is beneficial to information transmission.
2. We integrate a recurrent neural network into original transformer to make it can learn the order information of the speech sequences without positional encodings. And we embed this improved transformer into a dual-path network, which makes our approach efficient for extremely long speech sequence modeling.
3. Extensive experiments on benchmark datasets show that our approach outperforms the current state-of-the-arts (20.6 dB SDR on the public WSj0-2mix data corpus).

The remains of this paper are organized as follows. We introduces monaural speech separation with DPTNet in Section

http://dx.doi.org/10.21437/Interspeech.2020-2205

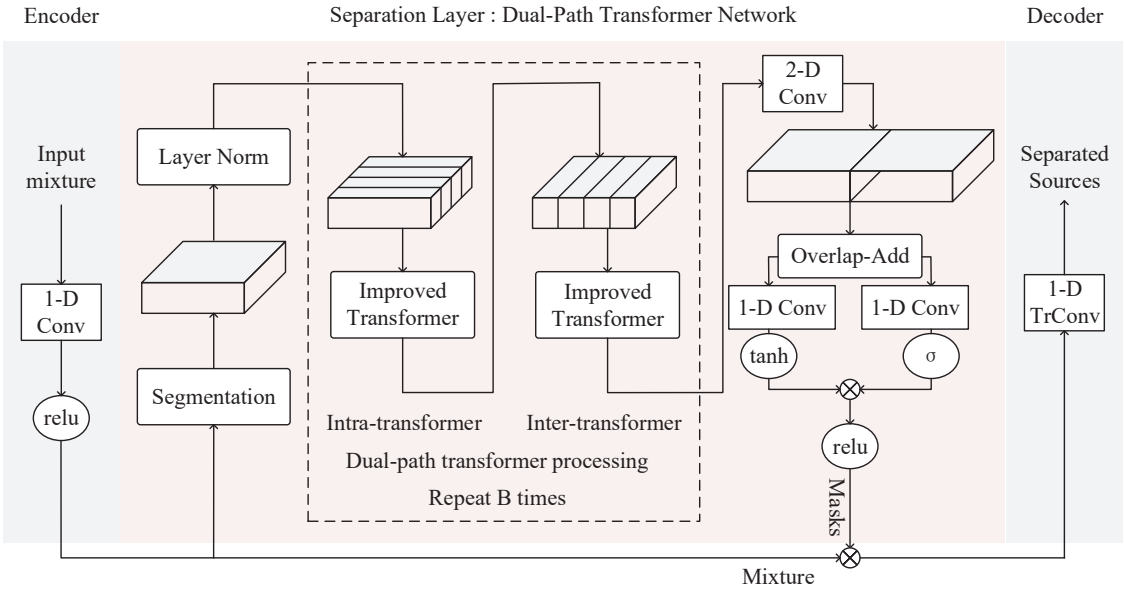


Figure 1: *Framework of speech separation with dual-path transformer network.*

2, present the experiment procedures in Section 3, analyze the experiment results in Section 4, conclude this paper and indicate future work in Section 5.

2. Speech separation with dual-path transformer network

As depicted in Figure 1, our speech separation system consists of three stages: encoder, separation layer and decoder, which is similar to that of Conv-TasNet in paper [15]. First, an encoder is used to convert segments of the mixture waveform into corresponding features in an intermediate feature space. Then the features are feed to the separation layer to construct a mask for each source. Finally, the decoder reconstructs the source waveforms by converting the masked features. In the following, we outline the encoder and decoder, and describe the separation layer, namely our dual-path transformer network, in detail.

2.1. Encoder

If we denote the speech mixture by $x \in R^{1\times T}$, then we can divide it into overlapping vectors $\boldsymbol{x} \in R^{L\times I}$ of length L samples, where I is the number of vectors. The encoder receive $\boldsymbol{x}$ and output the speech signal $X \in R^{N\times I}$ as follows:

$$X = ReLU(\boldsymbol{x} * W) \tag{1}$$

where the encoder can be characterized as a filter-bank W with N filters of length L, which is actually a 1-D convolution module.

2.2. Separation layer: dual-path transformer network

The separation layer, namely dual-path transformer network, is composed of three stages: segmentation, dual-path transformer processing and overlap-add, which is inspired by the common dual-path network [20].

2.2.1. Segmentation

Firstly, the segmentation stage splits X into overlapped chunks of length K and hop size H. Then all the chunks are concatenated to be a 3-D tensor $D \in R^{N\times K\times P}$.

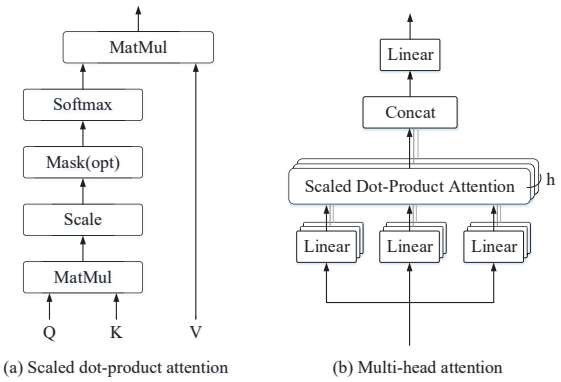
2.2.2. Dual-path transformer processing

Broadly speaking, the transformer is composed of an encoder and a decoder [23]. The encoder and decoder share the same model structure, except that the decoder is a left-context-only version for generation. To avoid confusions, the transformer in this paper refers specially to the encoder part, and it is comprised of three core modules: scaled dot-product attention, multi-head attention and position-wise feed-forward network.

Scaled dot-product attention is an effective self-attention mechanism that associate different positions of input sequences to calculate representations for the inputs, which is shown in Figure 2(a). The final output of this module is computed as a weighted sum of the values, where the weight for each value is computed by a attention function of the query with the corresponding keys. Multi-head attention is composed of multiple scaled dot-product attention modules, as depicted in Figure 2(b). First, it linearly maps the inputs h times with different, learnable linear projections to get parallel queries, keys and values respectively. Then the scaled dot-product attention is performed on these mapped queries, keys and values simultaneously. Position-wise feed-forward network is a fully connected feed-forward network. It is comprised of two linear transformations with a $ReLU$ activation in between. Besides the three core modules, transformer also includes several residual and normalization layers. We present the overall structure of the transformer in Figure 3(a) and it can be formulated as follows:

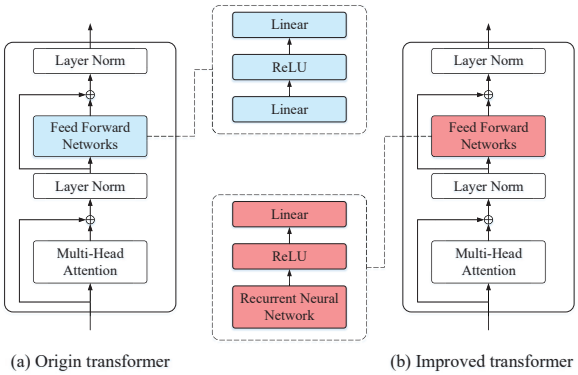
$$Q_i = ZW_i^Q, K_i = ZW_i^K, V_i = ZW_i^V \quad i \in [1, h] \tag{2}$$

$$\begin{aligned} head_i &= Attention(Q_i, K_i, V_i) \\ &= softmax(\frac{Q_i K_i^T}{\sqrt{d}})V_i \end{aligned} \tag{3}$$



(a) Scaled dot-product attention (b) Multi-head attention

Figure 2: *Attention mechanism of the transformer.*



(a) Origin transformer (b) Improved transformer

Figure 3: *Architecture of the origin and improved transformers.*

$$MultiHead = Concat(head_1,...,head_h)W^O \quad (4)$$

$$Mid = LayerNorm(Z + MultiHead) \quad (5)$$

$$FFN = ReLU(MidW_1 + b_1)W_2 + b_2 \quad (6)$$

$$Output = LayerNorm(Mid + FFN) \quad (7)$$

Here, $Z \in R^{l\times d}$ is the input with length l and dimension d, and $Q_i, K_i, V_i \in R^{l\times d/h}$ are the mapped queries, keys and values. $W_i^Q, W_i^K, W_i^V \in R^{d\times d/h}$ and $W^O \in R^{d\times d}$ are parameter matrices. FFN denotes the output of the position-wise feed-forward network, in which $W_1 \in R^{d\times d_{ff}}$, $W_2 \in R^{d_{ff}\times d}$, $b_1 \in R^{d_{ff}}$, $b_2 \in R^d$, and $d_{ff} = 4 \times d$.

The elements in speech sequences modeled by the transformer can contact directly without intermediate transmission, which introduces direct context-aware modeling into our method.

One thing missed by the above transformer is how to utilize the order information in the speech sequences. The origin transformer adds positional encodings to the input embeddings to represent order information, which is sine-and-cosine functions or learned parameters. However, we find that the positional encodings are not suitable for dual-path network and usually lead to model divergence in the training process. To learn the order information, we replace the first fully connected layer with a recurrent neural network in the feed-forward network, which is a interesting improvement from paper [22]:

$$FFN = ReLU(RNN(Mid))W_2 + b_2 \quad (8)$$

We show this improved transformer in Figure 3(b) and apply it in next dual-path transformer processing stage.

In dual-path transformer processing stage, the output D of the segmentation stage is passed to a heap of B dual-path transformers (DPTs), as presented in Figure 1. Each DPT consists of intra-transformer and inter-transformer, which are committed to modeling local and global information respectively. The intra-transformer processing block first model the local chunk independently, which acts on the second dimension of D:

$$\begin{aligned} D_b^{intra} &= IntraTransformer_b[D_{b-1}^{inter}] \\ &= [transformer(D_{b-1}^{inter}[:,:,i]), i = 1,...,P] \end{aligned} \quad (9)$$

Then the inter-transformer is used to summarize the information from all chunks to learn global dependency with performing on the last dimension of D:

$$\begin{aligned} D_b^{inter} &= InterTransformer_b[D_b^{intra}] \\ &= [transformer(D_b^{intra}[:,j,:]), j = 1,...,K] \end{aligned} \quad (10)$$

where $b = 1,...,B$ and $D_0^{inter} = D$. Note that the layer normalization in each sub-transformer is applied to all dimensions.

Indeed, this structure makes each element in speech sequences interact directly with only some elements and interact with the rest elements through an intermediate element. This fact imposes a slight negative impact on the direct context-aware modeling. However, the structure of dual paths allows our approach to model for extremely long speech sequences efficiently. In general, the small shortcoming caused by the dual-path structure is far less than the benefits it brings.

2.2.3. Overlap-Add

The output of the last inter-transformer D_B^{inter} is used to learn a mask for each source by a 2-D convolution layer. The masks are transformed back into sequences $M_s \in R^{N\times I}$ by overlap-add, and masked encoder features for s-th source are obtained by the element-wise multiplication between X and M_s:

$$Y_s = X \cdot M_s \quad (11)$$

2.3. Decoder

In decoder, a transposed convolution module is used to reconstruct separated speech signals $\boldsymbol{y}_s \in R^{L\times I}$ for s-th source:

$$\boldsymbol{y}_s = Y_s * V \quad (12)$$

where values in $V \in R^{N\times L}$ are the parameters of the transposed convolution module. Then the overlap-add method is applied to obtain the final waveforms $y_s \in R^{1\times T}$. The structure and function of decoder are both symmetrical with those of the encoder.

3. Experiment

3.1. Dataset

We evaluate our proposed model on two-speaker speech separation using the WSJ0-2mix [9] and LS-2mix dataset [24].

The WSJ0-2mix dataset is derived from the WSJ0 data corpus [25]. The 30 hours of training data and 10 hours of validation data contain two-speaker mixtures generated by randomly selecting utterances from different speakers in the WSJ0 training set si_tr_s, and mixing them at random signal-to-noise ratios (SNR) between -5 dB and 5 dB. 5-hours test set is similarly generated using utterances from unseen speakers in WSJ0 validation set si_dt_05 and evaluation set si_et_05.

LS-2mix is created based on the Librispeech dataset [24], which is a new corpus of reading English speech. Two speakers are randomly selected from the train-100 set to generate training mixtures, at various 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. Generated LS-2mix dataset contains 20000, 5000 and 3000 utterances in the train/validation/test set.

3.2. Experiment setup

In encoder and decoder, the window size is 2 samples and a 50% stride size is used. The number of filters is set to 64. As for the separation layer, the number of dual-path transformers, namely B, is set to 6, and $h = 4$ parallel attention layers are employed.

In the training stage, we train proposed model for 100 epochs on 4-second long segments, and the criteria for early stopping is no decrease in the loss function on validation set for 10 epochs. Adam [26] is used as the optimizer and gradient clipping with maximum L2-norm of 5 is applied during training. We increase the learning rate linearly for the first $warmup\_n$ training steps, and then decay it by 0.98 for every two epochs:

$$\begin{aligned} lrate &= k_1 \cdot d_{model}^{-0.5} \cdot n \cdot warmup\_n^{-1.5} \\ &\qquad when\ n \leq warmup\_n \\ &= k_2 \cdot 0.98^{epoch//2} \qquad when\ n > warmup\_n \end{aligned} \tag{13}$$

where n is the step number, k_1, k_2 are tunable scalars, and $k_1 = 0.2$, $k_2 = 4e^{-4}$, $warmup\_n = 4000$ in this paper.

These hyper-parameters are selected empirically according to the setups in the dual-path network [20] and transformer [23]. A Pytorch implementation of our DPTNet model can be found at "https://github.com/ujscjj/DPTNet".

3.3. Training objective

We train proposed model with utterance-level permutation invariant training (uPIT) [12] to maximize scale-invariant source-to-noise ratio (SI-SNR) [14]. SI-SNR is defined as:

$$s_{target} = \frac{\langle \tilde{x}, x \rangle x}{\|x\|^2} \tag{14}$$

$$e_{noise} = \tilde{x} - s_{target} \tag{15}$$

$$SI - SNR := 10log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \tag{16}$$

where x, $\tilde{x}$ are clean and estimated source respectively, both of which are normalized to zero-mean before the calculation.

4. Performance evaluation

In all experiments, we report the scale-invariant signal-to-noise (SI-SNR) and signal-to-distortion ratio (SDR) to measure the separation performance of our DPTNet, both of which are often employed in various speech separation systems.

We first report the SI-SNR and SDR scores on the WSJ0-2mix dataset obtained by our model and the well-known separation methods in recent years. As shown in Table 1, our DPTNet achieves 20.2 dB and 20.6 dB on the metrics of SI-SNR and SDR respectively, where a new state-of-the-art performance is achieved. Benefiting from the direct context-aware modeling, the elements in the speech sequences modeled by our DPTNet can interact directly, which results in the optimal monaural speech separation performance. In addition, our model maintains a small model size.

Table 1: *Comparison with other methods on WSJ0-2mix in SI-SNR (dB), SDR (dB) and Model Size*

Method	SI-SNR	SDR	Model Size
DPCL++ [27]	10.8	-	13.6M
uPIT-BLSTM-ST [12]	-	10.0	92.7M
Deep Attractor [10]	10.5	-	-
ADANet [28]	10.4	10.8	9.1M
Grid LSTM PIT [29]	-	10.2	-
ConvLSTM-GAT [30]	-	11.0	-
Chimera++ [31]	11.5	12.0	-
WA-MISI-5 [32]	12.6	13.1	32.9M
BLSTM-TasNet [14]	13.2	13.6	-
Conv-TasNet-gLN [15]	15.3	15.6	5.1M
Conv-TasNet+MBT [33]	15.5	15.9	-
Deep CASA [34]	17.7	18.0	12.8M
FurcaNeXt [35]	-	18.4	51.4M
DPRNN [20]	18.8	19.0	**2.6M**
DPTNet	**20.2**	**20.6**	**2.69M**

Table 2: *Comparison with baselines on the LS-2mix dataset*

Method	SI-SNR	SDR	Model Size
Conv-TasNet-gLN [15]	12.9	13.5	5.1M
DPRNN [20]	15.0	15.6	**2.6M**
DPTNet	**16.2**	**16.8**	**2.69M**

To prove the generalization of our approach, we conduct related experiments on the LS-2mix dataset. Compared to those in the WSJ0-2mix data corpus, the mixtures in LS-2mix is difficult to separate, but this does not interfere with the comparison between our method and the baselines. We reproduce two classical methods, namely Conv-TasNet [15] and DPRNN [20], as baselines. Table 2 lists the average SI-SNR and SDR obtained by our DPTNet and the two baselines, where our direct context-aware modeling is still significantly superior to the state-of-the-art approach DPRNN. This presents the generalization of our method and further demonstrates the effectiveness of it.

5. Conclusion and future work

In this paper, we propose the dual-path transformer network for end-to-end multi-speaker monaural speech separation, which models the speech sequences directly conditioning on context. Our model can learn the order information in speech sequences without positional encodings and model effectively for extremely long sequences of speech signals. Experiments on two benchmark datasets demonstrate the effectiveness of proposed model, and we achieve a new state-of-the-art performance on the public WSJ0-2mix data corpus. In the future, we would like to extend this work by directly modeling long speech feature sequences without the dual-path structure. It is promising to further improve the separation performance.

6. Acknowledgements

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 and 61906077, 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] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," 2006.

[4] 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.

[5] 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.

[6] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Ninth International Conference on Spoken Language Processing*, 2006.

[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] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[22] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," *Proc. Interspeech 2018*, pp. 3723–3727, 2018.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] 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.

[25] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *Interspeech 2016*, pp. 545–549, 2016.

[28] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[29] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6–10.

[30] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 711–715.

[31] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 686–690.

[32] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *arXiv preprint arXiv:1804.10204*, 2018.

[33] M. W. Lam, J. Wang, D. Su, and D. Yu, "Mixup-breakdown: a consistency training method for improving generalization of speech separation models," *arXiv preprint arXiv:1910.13253*, 2019.

[34] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2092–2102, 2019.

[35] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 653–665.