EAD-CONFORMER: A CONFORMER-BASED ENCODER-ATTENTION-DECODER-NETWORK FOR MULTI-TASK AUDIO SOURCE SEPARATION

Chenxing Li, Yang Wang, Feng Deng, Zhuo Zhang, Xiaorui Wang, Zhongyuan Wang

Kuai Shou Technology Co., Beijing, China

ABSTRACT

In this paper, we propose a Conformer-based network to improve the performance of multi-task audio source separation. This network, named EAD-Conformer, employs Conformer blocks to capture both local and global information, and an **e**ncoder-**a**ttention-**d**ecoder manner encourages the network to perform attentive modeling based on different sources. Specifically, EAD-Conformer first parses out the feature representations from the mixture by a Conformer-based encoder. Then, an attention module extracts selective information for each track and bridges encoder and decoders. Finally, three decoders respectively process attentive features and generate output masks for different sources. In addition, the proposed discriminate loss further enlarges the distance between different sources. Experiments demonstrate the effectiveness of EAD-Conformer, which achieves 13.37 dB, 11.41 dB, 10.56 dB signal-to-distortion ratio improvement on speech, music, noise track, respectively, and shows advantages over several well-known models.

***Index Terms*—** Multi-task Audio Source Separation, EAD-Conformer, Discriminate Loss

1. INTRODUCTION

The audio signals recorded in daily life usually contain various sources, such as speech, music, and background noises. The entanglement of these sources degrades audio quality and may impair the effectiveness of subsequent technologies. For example, singing voice mixed in music can make speech recognition confused. Front speech and background noises will adversely affect technologies such as music information retrieval. Multi-task audio source separation (MTASS) [1] is proposed naturally, which aims to separate the three fixed types of sound sources into three tracks at once.

Previous works mainly focus on the single-task audio source separation, such as speech enhancement and separation, music source separation, and singing voice separation. The main purpose of speech enhancement is to extract clean speech from a noisy mixture. The classic methods [2, 3] perform enhancement in the time-frequency (T-F) domain, which estimates a learned T-F mask to remove noise interference and recover the clean speech. In order to deal with the problem of phase mismatch, [4] investigates complex spectral mapping, and SEGAN [5] directly takes raw waveform as input and outputs enhanced waveform. Besides, with the development of Deep Noise Suppression Challenge [6], DCCRN [7], TSCN [8], and SDDNet [9] suggest that the complex domain is more suitable for enhancing speech from background noises. Speech separation separates the mixture of different speakers. Single-channel-based separation has attracted the most attention. Several methods, such as frequency domain-based DPCL [10], PIT [11], CBLDNN-GAT [12] and time-domain-based TasNet [13], DPRNN [14], DPT [15], SepFormer [16], have achieved state-of-the-art (SOTA) performance. On this basis, new tasks such as multi-channel speech separation [17] and continuous speech separation [18] have been developed.

Music source separation separates different source components from a music recording, such as vocal, guitar, drum, and other musical instruments. Open-Unmix [19] applies a bidirectional LSTM to separate different music sources one by one in the frequency domain. D3Net [20] applies dilated convolution to capture long-term dependencies. Demucs [21] performs separation on the time-domain and obtains SOTA performance. Singing voice separation only needs to separate the singing voice and the musical accompaniment. Dense-UNet [22] addresses the importance of phase and obtains SOTA results by learning the complex ratio masking in self-attention.

MTASS aims to output three tracks, speech, music, and noises. The speech track outputs the speaking voice. The music track consists of full songs, vocals, or different accompaniments. Except for music and speech, any other possible background sounds are classified as noise track. When dealing with MTASS, the methods mentioned above may confront several shortcomings. (1) Due to the fixed receptive fields of the convolutional network, the convolutional-based models [1, 3, 5, 8, 9, 13] are short of capturing global dependencies, especially on long recordings. Recurrent neural network-based models [2, 7, 10, 11, 14, 15, 19] can capture long-term dependencies but with high model complexity. (2) Speech, music, and noises have different audio characteristics. We argue that just adding a classification layer to the last layer, such as [11–16, 18], is insufficient to model the internal information of each source and the differences between sources.

Recently, transformer [23] has achieved great success in speech processing [15, 16, 24]. Self-attention mechanism can draw global input-output dependencies while discarding recurrence. Based on the transformer, Conformer [25] is proposed further to capture both position-wise local features and content-based global interactions. The effectiveness of Conformer has been proven in speech enhancement [26, 27] and speech separation [18]. Conformer makes sense, but Conformer-CSS [18] may not be suitable to be directly applied to MTASS. On this basis, we propose an EAD-Conformer to further improve the performance of MTASS. Our contributions are listed as follows: (1) We extend Conformer to MTASS task. By obtaining local and global information through the Conformer, the performance can be improved. (2) Considering the different characteristics of speech, music, and noises, the proposed EAD-Conformer shares the same encoder but uses different decoders. Attention module is used to select dominant features for each track. (3) Discriminate loss increases the distance between separated outputs and alleviates the problem of misclassification. Signal-to-distortion ratio [28] improvement (SDRi) is used to evaluate the performance. Experimental results show the effectiveness of the EAD-Conformer.

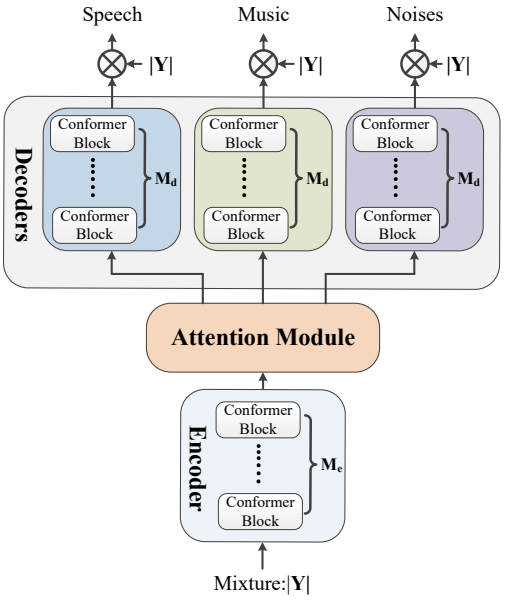

Fig. 1: The detailed framework of EAD-Conformer. The network processes the input feature in the order of encoding, attending, and decoding.

2. SYSTEM OVERVIEW

The illustration of EAD-Conformer is depicted in Fig. 1. The proposed network comprises three components: (1) The encoder processes mixture and generates audio representations. (2) The attention module extracts valuable features for each track and bridges encoder and decoders. (3) The decoders process attentive features respectively and generate the separated outputs.

2.1. Model input

In MTASS, the source signals are assumed linearly mixed, which can be represented as:

$$y(n) = \sum_{i=1}^{N} s_i(n), \tag{1}$$

where N is the number of source signals and $N = 3$ in this task. $s_i(n)$ and $y(n)$ denote the i-th source signal and the mixed signals, respectively. We choose the short-time Fourier transform (STFT) spectral magnitude as the input to the model. The following relationship is still satisfied after STFT:

$$Y(t,f) = \sum_{i=1}^{N} S_i(t,f), \tag{2}$$

where $Y(t,f)$ and $S_i(t,f)$ represent the STFT of mixture $y(n)$ and source signals $s_i(n)$, respectively. EAD-Conformer is conducted in the frequency domain. Our task is clarified as recovering each $|S_i(t,f)|$ from $|Y(t,f)|$. In this experiment, for computing STFT, we use a 1024-sample window size and a 256-sample shift.

2.2. Conformer block

Transformer [23] is famous for capturing long-term dependencies without recurrence. Both global and local correlations are essential for speech processing. Conformer [25] is proposed to combine convolutions with self-attention. The architecture of the Conformer block used in this experiment is shown in Fig. 2. The Conformer block contains two macaron-like feed-forward layers sandwiching a self-attention layer and a convolution layer. The detailed formulation of the Conformer block is:

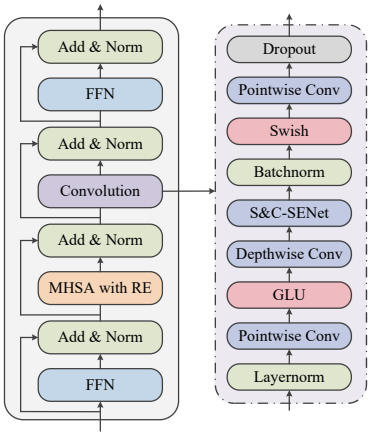

Fig. 2: The pipeline of the Conformer block and the detailed operation in the convolution module. Different from Conformer-CSS [18], S&C-SENet represents the spatial and channel squeeze-excitation layer [29]. MHSA represents the multi-head self-attention layer.

$$\begin{aligned} \hat{z} &= \textbf{norm}(z + \frac{1}{2}\textbf{FFN}(z)), \\ z' &= \textbf{norm}(\textbf{selfattention}(\hat{z}) + \hat{z}), \\ z'' &= \textbf{norm}(\textbf{conv}(z') + z'), \\ output &= \textbf{norm}(z'' + \frac{1}{2}\textbf{FFN}(z'')), \end{aligned} \tag{3}$$

where $z \in \mathbb{R}^{T\times F}$ is the input of the Conformer block. **FFN()**, **selfattention()**, **conv()**, and **norm()** denote the feed-forward layer, self-attention layer, convolution layer, and layer normalization, respectively.

In the self-attention layer, scaled dot-product attention computes a non-linear relationship between queries, keys, and values. Moreover, multi-head attention employs several scaled dot-product attention and enables the model to jointly attend to information from different representation subspaces at different positions:

$$\begin{aligned} \textbf{Multihead}(Q,K,V) &= [H_1,...,H_{d_{head}}]W^{head}, \\ H_i &= \textbf{softmax}(\frac{Q_i(K_i+RP)^T}{\sqrt{d_k}}))V_i, \end{aligned} \tag{4}$$

where Q, K, V are queries, keys, and values, respectively. d_k is the dimension of the keys, d_{head} is the number of the attention heads. RP is the relative position embedding [30].

A spatial and channel squeeze-excitation layer [29] is adopted to learn the importance of feature maps, which extracts valuable features and suppresses useless features. As Depicted in Fig. 2, the convolution layer consists of a norm layer, a pointwise convolution, a gated linear unit, a 1-D depthwise convolution layer, a norm layer, a squeeze-excitation layer [29], a Swish activation, and a pointwise convolution in order.

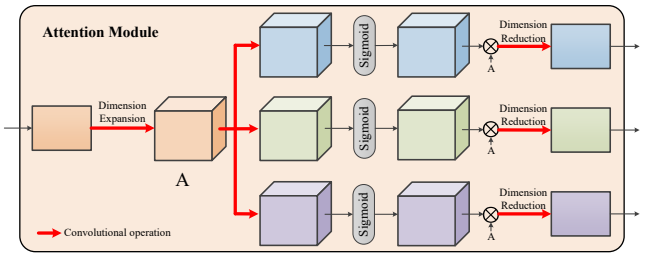

Fig. 3: The detailed framework of attention module. The red and bold arrows indicate the convolution operation.

2.3. EAD-Conformer network

As depicted in Fig. 1, the shared encoder is used to extract the fundamental features. In this experiment, the encoder is composed of M_e Conformer blocks. The detailed formula is as follows:

$$E_o = \textbf{Encoder}(|Y(t,f)|), \tag{5}$$

where the input of the encoder, $|Y(t,f)|$, is the spectral magnitude of the mixture, and E_o represents the output of the encoder.

We devise an attention module to dynamically attend to features that are only useful for different tracks, where:

$$A_1, A_2, A_3 = \textbf{Attention}(E_o). \tag{6}$$

The detailed architecture of the attention module is shown in Fig. 3. Learned from DPCL [10], the input is first expanded to get fine feature representations. In detail, the input dimension is first expanded by a (1×1) convolution, from $\mathbb{R}^{T\times F}$ to $\mathbb{R}^{T\times F\times C}$. Three convolutional layers are further adopted for non-linear transformation, which are used to learn the importance of each track. The sigmoid layer is then applied to obtaining the attention map. After obtaining the attention map, matrix multiplication is performed between the input and the attention map to obtain the attentive feature. A convolutional layer is followed for dimension reduction. The attentive feature contains rich information selected from encoders, and this information can contribute to each track.

The speech often has a clear harmonic structure. Music is composed of rhythmic melody. Noises are more random and irregular. The three kinds of signals have different acoustic characteristics. The features used for the separation of each track may be different. Three different signals need to be modeled separately. In this experiment, we use three decoders to model each track separately.

$$|X_i(t,f)| = \textbf{Decoder}_i(A_i) \times |Y(t,f)|, i = 1,2,3, \tag{7}$$

where $|X_i(t,f)|$ is the estimated spectral magnitude of each track. Similarly, the decoder of each track is composed of M_d Conformer blocks.

2.4. Model output

We use inverse STFT (ISTFT) to restore the separated waveforms:

$$x_i = \textbf{ISTFT}(|X_i(t,f)|, \angle Y(t,f)), \tag{8}$$

where x_i is the estimated signal. The phase of the mixture is used to restore the separated speech, music, and noises.

2.5. Loss function

Two kinds of loss are adopted: magnitude-based separation loss and discriminate separation loss. Magnitude-based separation loss is based on L_1-regularization. Discriminate separation loss measures the difference between the estimated signal and other signals, which aims to enhance the discriminative power of the model. The detailed loss function is formulated as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_1 - \lambda \times \mathcal{L}_D, \\ where\ \mathcal{L}_1 &= \sum_{i=1}^{N} |X_i - S_i|, \\ \mathcal{L}_D &= \sum_{i=1}^{N} |X_i - \sum_{j\neq i}^{N} S_j|. \end{aligned} \tag{9}$$

λ is a hyper-parameter and is used to balance each loss. $\mathcal{L}_1$ enables intra-class compactness, and $\mathcal{L}_D$ improves inter-class separability.

3. EXPERIMENTS

3.1. Experimental setup

We evaluate the proposed method using the MTASS dataset[1] [1]. MTASS dataset consists of 20,000 training data (55.6 hours), 1000 validation data (2.8 hours), and 1000 test data (2.8 hours). The speech, music, and noises are collected from aishell1 [31], DiDiSpeech [32], DSD100 [33], and DNS challenge [6]. For each speech clip, music and noise clips are added with a random signal-to-noise ratio of -5 to 5dB. All audio clips are downsampled to 16 kHz.

The encoder consists of 8 Conformer blocks. In the encoder, each Conformer block consists of 4 attention heads, 256 attention dimensions, and 1024 dimensions in the feed-forward layer. For attention module, in (inChannel, outChannel, kernel, stride)-format, the convolutions occupy $(1,32,1,1)$-, $(32,32,3,1)$-, $(32,1,1,1)$-kernels. Each decoder consists of 4 Conformer blocks. In the decoder, each Conformer block consists of 8 attention heads, 256 attention dimensions, and 512 dimensions in the feed-forward layer. The configuration of the convolutional layer in Conformer blocks is basically the same as that of Conformer-CSS [18], but with an additional squeeze-excitation layer with $reductionRatio = 2$.

EAD-Conformer is trained with the AdamW optimizer with a 1e-4 learning rate. $\lambda = 0.1$ in this experiment. The training schedule is a warm-up learning schedule with a linear decay, where the warm-up step is 6000, and the training step is 20000 (about 100 epochs).

3.2. Baselines

In this experiment, we compare the proposed EAD-Conformer with several well-known models in the field of speech enhancement, speech separation, and music source separation, which are T-F-domain-based D3Net [20], Conformer-CSS-base [18], complex-domain-based GCRN [4], Complex-MTASSNet [1], and time-domain-based Conv-TasNet [13], Demucs [21], Sepformer [16]. These baselines are trained with the same configurations mentioned in the literature. Due to the high model complexity of DPRNN [14], we adopt an 8-samples window size with a 4-samples shift in this experiment. Conformer-comp is also conducted, which is based on Conformer-CSS-base [18] and employs about the same number of parameters with the EAD-Conformer. The purpose of this is to

[1] https://github.com/Windstudent/Complex-MTASSNet/

Table 1: *Experimental results of ablation study.*

Methods	SDRi (dB)			
	Speech	Music	Noise	Ave
Conformer-encoder	9.27	7.90	8.72	8.63
+squeeze-excitation	10.18	8.82	8.81	9.27
+decoders	12.28	10.58	9.89	10.91
+attention module	12.55	10.73	10.21	11.16
+discriminate loss	**13.37**	**11.41**	**10.56**	**11.78**
Conformer-comp	11.31	9.80	9.35	10.15

Table 2: *SDRi for the comparison models obtained on the speech, music, and noise signal tracks.*

Methods	SDRi (dB)			
	Speech	Music	Noise	Ave
GCRN-RI [4]	9.11	5.76	5.51	6.79
GCRN-cRM [4]	8.73	6.25	6.50	7.16
Demucs [21]	9.93	6.38	6.29	7.53
D3Net [20]	10.55	7.64	7.79	8.66
Conv-TasNet [13]	11.80	8.35	8.07	9.41
DCCRN [7]	11.24	9.15	8.80	9.73
Sepformer [16]	11.33	8.52	9.42	9.76
DPRNN [14]	11.34	9.41	8.63	9.79
Conformer-CSS-base [18]	11.39	9.64	9.18	10.07
Complex-MTASSNet [1]	12.57	9.86	8.42	10.28
Conformer-comp	11.31	9.80	9.35	10.15
EAD-Conformer	**13.37**	**11.41**	**10.56**	**11.78**

compare EAD-Conformer with the Conformer-CSS-base [18] under similar size. Specifically, Conformer-comp has 19 Conformer blocks. These models are trained with the same datasets as our models.

3.3. Ablation study

We perform an ablation study to investigate the contribution of each module in depth. Before applying discriminate loss, the models are trained under the guidance of $\mathcal{L}_1$. At first, Conformer-encoder is conducted as the base, which has 8 Conformer blocks. The squeeze-excitation layer helps the network to capture interdependence between the feature channels. The experimental results show that this part of the information contributes to the final performance improvement. The performance has been improved with the help of three decoders, which achieves 1.64 dB SDRi improvement. This can be accounted for two reasons. The first reason is the increase in the number of parameters. The second can be attributed to the proposed pipeline, in which the network models audio signals via three different decoders. Conformer-comp has the similar number of parameters as the proposed model but achieves poorer performance. This indicates the effectiveness of the proposed mechanism. The attention module connects the encoder and the decoders and extracts useful information for each track. The experimental results also prove the effectiveness of attention selection, which brings 0.25 dB SDRi. Finally, discriminate loss further contributes to performance improvement, which shows that the discriminate loss allows the network to enlarge the difference between the three tracks.

3.4. Model analysis

Conformer-based models have obtained more SDRi, which shows the effectiveness of the Conformer structure. Complex-MTASSNet achieves the best among baselines. Complex-MTASSNet can capture spectrum details between different sources by applying stacked convolutions. However, the ability to capture dependencies is limited because of the fixed receptive field. Compared with Complex-MTASSNet, EAD-Conformer achieves performance improvement. That means local features and global dependencies are both critical for MTASS. Like Complex-MTASSNet and EAD-Conformer, these methods both model the source signals in three tracks. These methods also have achieved performance improvements compared with other methods. For example, EAD-Conformer occupies a similar size as Conformer-comp, but EAD-Conformer performs much better than Conformer-comp. It is shown that for tasks with multiple outputs such as MTASS, and each output has obviously different acoustic characteristics, it is meaningful to model each track separately. This also indicates the effectiveness of the proposed pipeline. In general, EAD-Conformer achieves the best SDRi on all three tracks, which are 13.37 dB, 11.41 dB, and 10.56 dB on speech, music, and noise track, respectively.

Table 3: *Model parameters, MAC/S and RTF of the comparison models.*

Models	Parameters	MAC/S	RTF (GPU)
GCRN [4]	9.88 M	2.5 G	0.031
Demucs [21]	243.32 M	5.6 G	0.006
D3Net [20]	7.93 M	3.5 G	0.002
Conv-TasNet [13]	5.14 M	5.2 G	0.017
DCCRN [7]	3.70 M	14.5 G	0.018
Sepformer [16]	25.75 M	77.71 G	0.052
DPRNN [14]	2.65 M	21.9 G	0.012
Conformer-CSS-base [18]	21.87 M	1.4 G	0.004
Complex-MTASSNet [1]	28.18 M	1.8 G	0.019
Conformer-comp	26.17 M	1.7 G	0.004
EAD-Conformer	26.09 M	2.12 G	0.005

Model parameters and computational complexity are also concerned in this experiment. We have counted the model size (parameters), multiply-accumulate operations per second (MAC/S), and the processing time consumption per second (real-time factor, RTF) on GPU (Nvidia 2080Ti) of these models, and the detailed results are listed in Table 3. EAD-Conformer occupies 26.09 M parameters but shows advantages on MAC/S and RTF. Compared with Conformer-comp, EAD-Conformer achieves performance improvement with roughly equivalent size. EAD-Conformer performs relatively balanced on complexity tests while ensuring the best separation performance.

4. CONCLUSIONS

We propose an EAD-Conformer to further improve the performance of MTASS, which performs in an encoder-attention-decoder manner. In detail, the encoder is used to extract fundamental features. An attention module follows to select attentive features for each track. Finally, three different decoders process these features and generate outputs. Experimental results show that EAD-Conformer effectively separates the mixture. EAD-Conformer achieves 13.37 dB, 11.41 dB, and 10.56 dB SDRi on speech, music, and noise track and outperforms several popular methods. In MTASS, the model not only needs to capture the long-term dependencies of signals but also needs to model the characteristics of each signal and the difference among signals.

5. REFERENCES

[1] L. Zhang, C. Li, F. Deng, and X. Wang, "Multi-task audio source separation," in *IEEE ASRU*, 2021.

[2] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE ICASSP*, 2015, pp. 708–712.

[3] F. Deng, T. Jiang, X. Wang, C. Zhang, and Y. Li, "Naagn: Noise-aware attention-gated network for speech enhancement." in *Interspeech*, 2020, pp. 2457–2461.

[4] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Interspeech*, 2017, pp. 3642–3646.

[6] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," *arXiv preprint arXiv:2101.01902*, 2021.

[7] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2020.

[8] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *IEEE ICASSP*, 2021, pp. 6628–6632.

[9] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A simultaneous denoising and dereverberation framework with target decoupling," in *Interspeech*, 2021.

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

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

[12] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *IEEE ICASSP*, 2018, pp. 711–715.

[13] Y. Luo and N. 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.

[14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE ICASSP*, 2020, pp. 46–50.

[15] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Interspeech*, 2020.

[16] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE ICASSP*, 2021, pp. 21–25.

[17] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

[18] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *IEEE ICASSP*, 2021, pp. 5749–5753.

[19] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[20] N. Takahashi and Y. Mitsufuji, "D3net: Densely connected multidilated densenet for music source separation," *arXiv preprint arXiv:2010.01733*, 2020.

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

[22] Y. Zhang, Y. Liu, and D. Wang, "Complex ratio masking for singing voice separation," in *IEEE ICASSP*, 2021, pp. 41–45.

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

[24] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *IEEE ICASSP*, 2018, pp. 5884–5888.

[25] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020.

[26] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "Df-conformer: Integrated architecture of conv-tasnet and conformer using linear complexity self-attention for speech enhancement," *arXiv preprint arXiv:2106.15813*, 2021.

[27] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *IEEE ICASSP*, 2021, pp. 7118–7122.

[28] 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.

[29] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *MICCAI*. Springer, 2018, pp. 421–429.

[30] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *NAACL*, 2018, pp. 464–468.

[31] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *O-COCOSDA*, 2017, pp. 1–5.

[32] T. Guo, C. Wen, D. Jiang, N. Luo, R. Zhang, S. Zhao, W. Li, C. Gong, W. Zou, K. Han *et al.*, "Didispeech: A large scale mandarin speech corpus," in *IEEE ICASSP*, 2021, pp. 6968–6972.

[33] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 186–190.