CONTINUOUS SPEECH SEPARATION WITH CONFORMER

Sanyuan Chen[‡], *Yu Wu*[†], *Zhuo Chen*[†], *Jian Wu*[†], *Jinyu Li*[†], *Takuya Yoshioka*[†], *Chengyi Wang*[†], *Shujie Liu*[†], *Ming Zhou*[†]

[†]Microsoft Corporation [‡]Harbin Institute of Technology

ABSTRACT

Continuous speech separation was recently proposed to deal with the overlapped speech in natural conversations. While it was shown to significantly improve the speech recognition performance for multi-channel conversation transcription, its effectiveness has yet to be proven for a single-channel recording scenario. This paper examines the use of Conformer architecture in lieu of recurrent neural networks for the separation model. Conformer allows the separation model to efficiently capture both local and global context information, which is helpful for speech separation. Experimental results using the LibriCSS dataset show that the Conformer separation model achieves the state of the art results for both single-channel and multi-channel settings. Results for real meeting recordings are also presented, showing significant performance gains in both word error rate (WER) and speaker-attributed WER.

Index Terms— Multi-speaker ASR, Transformer, Conformer, Continuous speech separation

1. INTRODUCTION

The advance in deep learning has drastically improved the accuracy and robustness of modern automatic speech recognition (ASR) systems in the past decade [1, 2, 3, 4], enabling various voice-based applications. However, when applied to acoustically and linguistically complicated scenarios such as conversation transcription [5, 6], the ASR systems still suffer from the performance limitation due to overlapped speech and quick speaker turn-taking, which break the usually assumed single active speaker condition. Additionally, the overlapped speech causes the so-called permutation problem [7], further increasing the difficulty of the conversation transcription.

Speech separation is often applied as a remedy for this problem, where the mixed speech is processed by a specially trained separation network before ASR . Starting from deep clustering (DC) [7] and permutation invariant training (PIT) [8, 9], various separation models have been shown effective in handling overlapped speech [6, 10, 11, 12]. Among the network architectures proposed thus far, the Transformer [12] based approach achieved a promising result. Transformer was first introduced for machine translation [13] and later extended to speech processing [14]. A Transformer based speech separation architecture was proposed in [12], achieving the state of the art separation quality on the WSJ0-2mix dataset. It was also reported in [15] that incorporating Transformer into an end-to-end multi-speaker recognition network yielded higher recognition accuracy. However, both studies were evaluated on artificially simulated data sets that only considered overlapped speech, assuming the utterance boundaries to be provided, which significantly differs from the real conversational transcription scenario [6, 16].

In this work, inspired by the recent advances in transducer-based end-to-end ASR modeling, which has evolved from a recurrent neural network (RNN) transducer [17] to Transformer [18] and Conformer [19] transducers, we examine the use of the Conformer architecture for continuous speech separation (CSS) [20]. Unlike the prior speech separation studies, in CSS, the separation network continuously receives a mixed speech signal, performs separation, and routes each separated utterance to one of its output channels in a way that each output channel contains overlap-free signals. This allows a standard ASR system trained with single speaker utterances to be directly applied to each output channel to generate transcriptions. The proposed system is evaluated by using the LibriCSS dataset [16], which consists of real recordings of long-form multi-talker sessions that were created by concatenating and mixing LibriSpeech utterances with various overlap ratios. Our proposed network significantly outperforms the RNN-based baseline systems, achieving the new state of the art performance on this dataset. Evaluation results on real meetings are also presented along with tricks for further performance improvement.

2. APPROACH

2.1. Problem Formulation

The goal of speech separation is to estimate individual speaker signals from their mixture, where the source signals may be overlapped with each other wholly or partially. The mixed signal is formulated as $y(t) = \sum_{s=1}^{S} x_s(t)$, where t is the time index, $x_s(t)$ denotes the s-th source signal, and $y(t)$ is the mixed signal. Following [20], when C microphones are available, the model input to the separation model can be obtained as

$$\mathbf{Y}(t,f) = \mathbf{Y}^1(t,f) \oplus \text{IPD}(2) \ldots \oplus \text{IPD}(C), \qquad (1)$$

where $\oplus$ means a concatenation operation, $\mathbf{Y}^i(t,f)$ refers to the STFT of the i-th channel, IPD(i) is the inter-channel phase difference between the i-th channel and the first channel, i.e. IPD$(i) = \theta^i(t,f) - \theta^1(t,f)$ with $\theta^i(t,f)$ being the phase of $\mathbf{Y}^i(t,f)$. These features are normalized along the time axis. If $C = 1$, it reduces to a single channel speech separation task.

Following [21, 22], a group of masks $\{\mathbf{M}_s(t,f)\}_{1 \le s \le S}$ are estimated with a deep learning model instead of directly predicting the source STFTs. Each source STFT, $\mathbf{X}_s(t,f)$, is obtained as $\mathbf{M}_s(t,f) \odot \mathbf{Y}^1(t,f)$, where $\odot$ is an elementwise product. For the multi-channel setting, the source signals are obtained with adaptive minimum variance distortionless response (MVDR) beamforming [23]. In this paper, we employ the Conformer structure [19] to estimate the masks for (continuous) speech separation.

2.2. Model structure

Conformer [19] is a state-of-the-art ASR encoder architecture, which inserts a convolution layer into a Transformer block to increase the

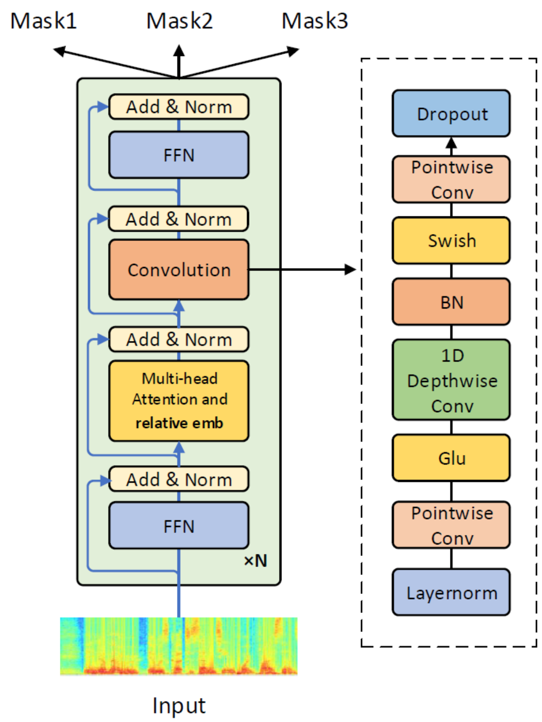

Fig. 1. Conformer architecture. There are three mask outputs, two for speakers and one for noise.

local information modeling capability of the traditional Transformer model [13]. The architecture of the Conformer is shown in Fig. 1, where each block consists of a self-attention module, a convolution module, and a macron-feedforward module. A chunk of $\mathbf{Y}(t, f)$ over time frames and frequency bins is the input of the first block. Suppose that the input to the i-th block is z, the i-th block output is calculated as

$$\hat{z} = \text{layernorm}(z + \frac{1}{2}\text{FFN}(z)) \tag{2}$$

$$z' = \text{layernorm}(\text{selfattention}(\hat{z}) + \hat{z}) \tag{3}$$

$$z'' = \text{layernorm}(\text{conv}(z') + z') \tag{4}$$

$$output = \text{layernorm}(z'' + \frac{1}{2}\text{FFN}(z'')), \tag{5}$$

where FFN(), selfattention(), conv(), and layernorm() denote the feed forward network, self-attention module, convolution module, and layer normalization, respectively. In the self-attention module, $\hat{\mathbf{z}}$ is linearly converted to $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ with three different parameter matrices. Then, we apply a multi-head self-attention mechanism

$$\text{Multihead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{H_1} \dots \mathbf{H}_{d_{head}}]\mathbf{W}^{head} \tag{6}$$

$$\mathbf{H_i} = \text{softmax}(\frac{\mathbf{Q_i}(\mathbf{K_i} + \mathbf{pos})^{\intercal}}{\sqrt{d_k}})\mathbf{V_i}, \tag{7}$$

where d_k is the dimensionality of the feature vector, d_{head} is the number of the attention heads. $\mathbf{pos} = \{rel_{m,n}\} \in \mathbb{R}^{M \times M \times d_k}$ is the relative position embedding [24], where M is the maximum chunk length and $rel_{m,n} \in \mathbb{R}^{d_k}$ is a vector representing the offset of m and n with m and n denoting the m-th vector of $\mathbf{Q_i}$ and the n-th vector of $\mathbf{K_i}$, respectively. The Convolution starts with a pointwise convolution and a gated linear unit (GLU), followed by a 1-D

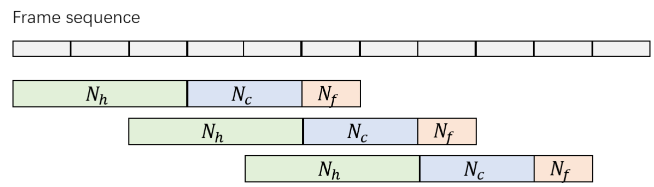

Fig. 2. Chunk-wise processing is employed to enable streaming processing for continuous speech separation.

depthwise convolution layer with a Batchnorm [25] and a Swish activation. After obtaining the Conformer output, we further convert it to a mask matrix as $\mathbf{M}_s(t, f) = \text{sigmoid}(\text{FFN}_s(output))$.

2.3. Chunk-wise processing for continuous separation

The speech overlap usually takes place in a natural conversation which may last for tens of minutes or longer. To deal with such long input signals, CSS generates a predefined number of signals where overlapped utterances are separated and then routed to different output channels.

To enable this, we employ the chunk-wise processing proposed in [26] at test time. A sliding-window is applied as illustrated in Figure 2, which contains three sub-windows, representing the history (N_h frames), the current segment (N_c frames), and the future context (N_f frames). We move the window position forward by N_c frames each time, and compute the masks for the current N_c frames using the whole N-frame-long chunk.

To further consider the history information beyond the current chunk, we also consider taking account of the previous chunks in the self-attention module. Following Transformer-XL [27], the Equation 7 is rewritten as

$$\text{softmax}(\frac{\mathbf{Q_i}(\mathbf{K_i} \oplus \mathbf{K_{cache,i}} + \mathbf{pos})^{\intercal}}{\sqrt{d_k}})(\mathbf{V_i} \oplus \mathbf{V_{cache,i}}) \tag{8}$$

where $\mathbf{Q}$ is obtained by the current chunk while $\mathbf{K}$ and $\mathbf{V}$ are the concatenations of the previous and current changes in the key and value spaces, respectively. The dimensionality of $\mathbf{K_{cache,i}}$ depends on the number of the history chunks considered.

3. EXPERIMENT

3.1. Datasets

Our training dataset consists of 219 hours of artificially reverberated and mixed utterances that sampled randomly from WSJ1 [29]. Four different mixture types described in [20] are included in the training set. To generate each training mixture, we randomly pick one or two speakers from WSJ1 and convolve each with a 7 channel room impulse response (RIR) simulated with the image method [30]. The reverberated signals are then rescaled and mixed with a source energy ratio between -5 and 5 dB. In addition, we add simulated isotropic noise [31] with a 0–10 dB signal to noise ratio. The average overlap ratio of the training set is around 50%.

LibriCSS is used for evaluation [16]. The dataset has 10 hours of seven-channel recordings of mixed and concatenated LibriSpeech test utterances. The recordings were made by playing back the mixed audio in a meeting room. Two evaluation schemes are used: utterance-wise evaluation and continuous input evaluation. In the former evaluation, the long-form recordings are segmented into individual utterances by using ground-truth time marks to evaluate the

Table 1. Utterance-wise evaluation for seven-channel and single-channel settings. Two numbers in a cell denote %WER of the **hybrid ASR model** used in LibriCSS [16] and **E2E Transformer** based ASR model [28]. 0S and 0L are utterances with short/long inter-utterance silence.

System	**Overlap ratio in %**					
	0S	0L	10	20	30	40
No separation [16]	11.8/5.5	11.7/5.2	18.8/11.4	27.2/18.8	35.6/27.7	43.3/36.6
	Seven-channel Evaluation					
BLSTM	**7.0/3.1**	7.5/**3.3**	10.8/4.3	13.4/5.6	16.5/7.5	18.8/8.9
Transformer-base	8.3/3.4	8.4/3.4	11.4/4.1	12.5/**4.8**	14.7/6.4	16.9/7.2
Transformer-large	7.5/**3.1**	7.7/3.4	10.1/**3.7**	12.3/**4.8**	14.1/5.9	16.0/6.3
Conformer-base	7.3/**3.1**	**7.3/3.3**	**9.6**/3.9	11.9/**4.8**	13.9/6.0	15.9/6.8
Conformer-large	7.2/**3.1**	7.5/**3.3**	**9.6/3.7**	**11.3/4.8**	**13.7/5.6**	**15.1/6.2**
	Single-channel Evaluation					
BLSTM	15.8/6.4	14.2/5.8	18.9/9.6	25.4/15.3	31.6/20.5	35.5/25.2
Transformer-base	13.2/5.5	12.3/5.2	16.5/8.3	21.8/12.1	26.2/15.6	30.6/19.3
Transformer-large	13.0/**5.3**	12.4/5.1	15.5/**7.4**	**20.1**/11.1	24.6/**13.5**	27.9/**17.0**
Conformer-base	13.8/5.6	12.5/5.4	16.7/8.2	21.6/11.8	26.1/15.5	30.1/18.9
Conformer-large	**12.9**/5.4	**12.2/5.0**	**15.1**/7.5	**20.1/10.7**	**24.3**/13.8	**27.6**/17.1

pure separation performance. In the continuous input evaluation, systems have to deal with the unsegmented recordings and thus CSS is needed.

3.2. Implementation details

We use BLSTM and Transformers as our baseline speech separation models. The BLSTM model has three BLSTM layers with 1024 input dimensions and 512 hidden dimensions, resulting in 21.80M parameters. There are three masks, two for speakers and one for noise. The noise mask is used to enhance the beamforming [26]. We use three sigmoid projection layers to estimate each mask. Transformer-base and Transformer-large models with 21.90M and 58.33M parameters are our two Transformer-based baselines. The Transformer-base model consists of 16 Transformer encoder layers with 4 attention heads, 256 attention dimensions and 2048 FFN dimensions. The Transformer-large model consists of 18 Transformer encoder layers with 8 attention heads, 512 attention dimensions and 2048 FFN dimensions.

As with the Transformer baseline models, we experiment with two Conformer-based models, Conformer-base and Conformer-large. They have 22.07M and 58.72M parameters, respectively. The Conformer-base model consists of 16 Conformer encoder layers with 4 attention heads, 256 attention dimensions and 1024 FFN dimensions. The Conformer-large model consists of 18 Conformer encoder layers with 8 attention heads, 512 attention dimensions and 1024 FFN dimensions. Both Conformer and Transformer are trained with the AdamW optimizer [32], where the weight decay is set to 1e-2. We set the learning rate to 1e-4 and use a warm-up learning schedule with a linear decay, in which the warm-up step is 10,000 and the training step is 260,000.

We use two ASR models to evaluate the speech separation accuracy. One is the ASR model used in the original LibriCSS publication [16], which is a hybrid system using a BLSTM acoustic model and a 4-gram language model. The other one is one of the best open-source end-to-end Transformer based ASR models [28], which achieves 2.08% and 4.95% word error rates (WERs) for LibriSpeech test-clean and test-other, respectively. Following [16], we generate the separated speech signals with spectral masking and mask-based adaptive minimum variance distortionless response (MVDR) beamforming for the single-channel and seven-channel cases, respectively. For a fair comparison, we follow the LibriCSS setting for chunk-wise CSS processing, where N_h, N_c, N_f are set to 1.2s, 0.8s, 0.4s respectively.

3.3. Results for utterance wise evaluation

Table 1 shows the WER of the utterance wise evaluation for the seven-channel and single-channel settings. Our Conformer models achieved state-of-the-art results. Compared with BLSTM, Conformer-base yielded substantial WER gains for the 7-channel setting. The fact that the Conformer-base model outperformed Transformer-base for almost all the settings indicates Conformer's superior local modeling capability. Also, the larger models achieved better performance in the highly overlapped settings. As regards the single-channel case, while the overall WERs were higher, the trend was consistent between the single- and multi-channel cases, except for the non-overlap scenario. With the seven channel input, all models showed similar performance for 0S and 0L. On the other hand, when only one channel was used, the self-attention models were markedly better. This could indicate that the seven-channel features contain sufficiently rich information for simpler networks to do the beamforming well. Meanwhile, the information in the single-channel signal is quite limited, requiring a more advanced structure.

3.4. Results for continuous input evaluation

Table 2 shows the continuous input evaluation results. The Conformer and Transformer models performed consistently better than BLSTM, but their performance gap became smaller in the large overlap test-set. The relative WER gains obtained with Conformer-base over BLSTM were 4% and 15% for the hybrid and transducer ASR systems, respectively, which were smaller than those obtained for the utterance-wise evaluation. A possible explanation is that the self-attention based methods are good at using global information while the chunk-wise processing limits the use of the context information.

It is noteworthy that 0S results were much worse than those of 0L only in the continuous evaluation, which is consistent with the previous report [16]. The 0S dataset contains much more quick speaker turn changes, imposing a challenge for both speech separation and ASR. The self-attention-based models showed a clear improvement over BLSTM, indicating that they are also helpful for dealing with turn-takings in natural conversations.

Table 2. Continuous speech separation evaluation for seven-channel and single-channel settings.

System	**Overlap ratio in %**					
	0S	0L	10	20	30	40
No separation [16]	15.4/12.7	11.5/5.7	21.7/17.6	27.0/24.4	34.3/30.9	40.5/37.5
	Seven-channel Evaluation					
BLSTM	11.4/6.0	**8.4**/4.1	13.1/7.0	14.9/7.9	18.7/11.5	20.5/12.3
Transformer-base	12.0/5.6	9.1/4.4	13.4/6.2	14.4/6.8	18.5/9.7	19.9/10.3
Transformer-large	**10.9**/5.4	8.8/**4.0**	**12.6**/6.0	13.6/**6.7**	**17.2**/9.3	**18.9**/10.2
Conformer-base	11.1/5.6	8.7/**4.0**	12.8/6.1	13.8/**6.7**	17.6/9.4	19.6/10.4
Conformer-large	11.0/**5.2**	8.7/**4.0**	**12.6/5.8**	**13.5**/6.8	17.6/**9.0**	19.6/**10.0**
Conformer$_{xl}$-base	11.4/5.4	8.7/4.1	13.2/6.2	13.6/**6.7**	17.8/9.5	20.0/10.8
Conformer$_{xl}$-large	11.0/**5.2**	8.8/4.1	12.9/**5.8**	13.7/**6.7**	17.5/9.4	19.8/10.6
	Single-channel Evaluation					
BLSTM	19.1/11.7	16.1/9.7	22.1/14.5	27.4/19.1	33.0/25.9	37.6/30.1
Transformer-base	13.8/7.1	**11.5**/6.6	16.7/9.6	20.8/13.3	26.7/18.6	31.0/21.6
Transformer-large	**13.0**/7.2	12.3/6.9	**15.8**/9.5	**19.8/12.2**	**25.3**/16.9	**28.6/19.3**
Conformer-base	14.1/7.7	13.0/7.1	17.4/10.6	21.9/13.7	27.4/18.7	32.0/22.4
Conformer-large	13.3/**6.9**	11.7/**6.1**	16.3/**9.1**	20.7/12.5	25.6/**16.7**	29.3/**19.3**

Table 2 also shows that the Conformer$_{xl}$ models using longer context information did not result in lower WERs especially in the large overlap ratio settings. Two factors may have contributed to the performance degradation. 1) The unexpected noise may have been introduced from the use of the longer history, which may contain more speakers' voices. 2) Also, we did not consider the overlap regions of the adjacent windows during training, possibly making the training/testing gap greater and resulting in sub-optimal performance. We leave the training with overlap regions for future work.

3.5. Results on large scale real meetings

To further verify the effectiveness of our method, we further conduct an experiment on an internal real conversation corpus which consists of 15.8 hours of single channel recordings of daily group discussions, noted as the Real Conversation dataset. In this dataset, the per-meeting speaker number ranges from 3 to 22. We applied a modified version of the conversation transcription system of [6], where a large scale trained speech recognizer and speaker embedding extractor were included, to obtain speaker attributed transcriptions.

Compared with LibriCSS, those real meetings are significantly more complex with respect to the acoustics, linguistics, and interspeaker dynamics. To deal with the real data challenges, three improvements were made. Firstly, we increased the training data amount to 1500 hours. Additional clean speech samples were taken from a Microsoft internal corpus and they were mixed with the simulation setup as Section 3.1. Secondly, the separation network sometimes generated a low volume residual signal from the redundant output channel for single speaker regions, which increased the word insertion errors. To mitigate this, we introduced a merging scheme, where the two channel outputs were merged when a single active speaker was judged to be present. The merger was triggered when only one masked channel had significantly large energy. Lastly, to reduce the distortion introduced by the masking operation, we used single speaker signals corrupted by background noise as a training target. This noisy label (nlabel) scheme allowed the separation network to focus only on the separation task and leave the noise to the ASR model. The WER and speaker attributed WER (SA-WER) were used for evaluation, where the latter assesses the combined quality of speech transcription and speaker diarization [6].

Table 3. Continuous evaluation on a real meeting dataset.

system	Data	WERR	SA-WERR
Original	N/A	0	0
BLSTM	219hr	-6.4%	-18.8%
Conformer-base	219hr	-7.2%	-6.3%
Conformer-large	219hr	-2.5 %	1.9 %
Conformer-base	1500hr	9.5%	8.8%
Conformer-base-merge	1500hr	8.4%	10.13%
Conformer-base-merge-nlabel	1500hr	11.8%	13.7%
Conformer-large-merge-nlabel	1500hr	8.08%	18.4%

Table 3 shows the WER and SA-WER reduction rates. With the three improvements described above, the proposed model reduced the WER and SA-WER by 11.8% and 18.4% relative, respectively, compared with a system without the separation front-end. Although the BLSTM based network improved the recognition result for the LibriCSS dataset especially for the high overlap ratio settings, it largely degraded the speech recognition and speaker diarization performance on the Real Conversation dataset. Because the speech overlap happens only sporadically in real conversations, it is important for the separation model not to hurt the performance for less overlap cases. Thanks to the better modeling capacity, the Conformer based models significantly mitigate the performance degradation. In addition, it can be seen that each introduced step brought about consistent improvement for both performance metrics.

4. CONCLUSION

In this work, we investigated the use of Conformer for continuous speech separation. The experimental results showed that it outperformed RNN-based models for both utterance-wise evaluation and continuous input evaluation. The superiority of Conformer to Transformer was also observed. This work is also the first to report substantial WER and SA-WER gains from the speech separation in a single-channel real meeting transcription task. The results indicate the usefulness of appropriately utilizing context information in the speech separation.

5. REFERENCES

[1] Jinyu Li, Rui Zhao, Zhuo Chen, Changliang Liu, Xiong Xiao, Guoli Ye, and Yifan Gong, "Developing far-field speaker system via teacher-student learning," in *Proc. ICASSP*. IEEE, 2018, pp. 5699–5703.

[2] Ladislav Mošner, Minhua Wu, et al., "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," in *Proc. ICASSP*. IEEE, 2019, pp. 6475–6479.

[3] Lei Sun, Jun Du, et al., "A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the chime-5 challenge," *IEEE JSTSP*, vol. 13, no. 4, pp. 827–840, 2019.

[4] Jinyu Li, Yu Wu, Yashesh Gaur, Chengyi Wang, Rui Zhao, and Shujie Liu, "On the comparison of popular end-to-end models for large scale speech recognition," in *Interspeech*, 2020.

[5] Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[6] Takuya Yoshioka, Igor Abramovski, et al., "Advances in online audio-visual meeting transcription," in *Proc. ASRU*. IEEE, 2019, pp. 276–283.

[7] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*. IEEE, 2016, pp. 31–35.

[8] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*. IEEE, 2017, pp. 241–245.

[9] Morten Kolbæk, Dong Yu, and Others, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM TASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

[10] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[11] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*. IEEE, 2020, pp. 46–50.

[12] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[14] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5884–5888.

[15] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP*. IEEE, 2020, pp. 6134–6138.

[16] Zhuo Chen, Takuya Yoshioka, et al., "Continuous speech separation: Dataset and analysis," in *Proc. ICASSP*. IEEE, 2020, pp. 7284–7288.

[17] Yanzhang He, Tara N Sainath, et al., "Streaming end-to-end speech recognition for mobile devices," in *Proc. ICASSP*. IEEE, 2019, pp. 6381–6385.

[18] Qian Zhang, Han Lu, Hasim Sak, et al., "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *Proc. ICASSP*, 2020.

[19] Anmol Gulati, James Qin, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[20] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, and Fil Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *ICASSP*. IEEE, 2018, pp. 5739–5743.

[21] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM TASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.

[22] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio," in *New Era for Robust Speech Recognition*, pp. 165–186. Springer, 2017.

[23] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel mmse-based framework for speech source separation and noise reduction," *IEEE/ACM TASLP*, vol. 21, no. 9, pp. 1913–1928, 2013.

[24] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations," in *NAACL*, 2018, pp. 464–468.

[25] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[26] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Interspeech*, 2018, pp. 3038–3042.

[27] Zihang Dai, Zhilin Yang, and Others, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[28] Chengyi Wang, Yu Wu, Yujiao Du, Jinyu Li, Shujie Liu, Liang Lu, Shuo Ren, Guoli Ye, Sheng Zhao, and Ming Zhou, "Semantic mask for transformer based end-to-end speech recognition," *Interspeech*, 2020.

[29] Linguistic Data Consortium Philadelphia, "CSR-II (WSJ1) Complete," 1994, `http://catalog.ldc.upenn.edu/LDC94S13A`.

[30] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, pp. 943–950, 1979.

[31] Emanuël AP Habets and Sharon Gannot, "Generating sensor signals in isotropic noise fields," *JASA*, vol. 122, no. 6, pp. 3464–3470, 2007.

[32] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *ICLR*, 2018.