INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Real-time End-to-End Monaural Multi-speaker Speech Recognition

Song Li[1], *Beibei Ouyang*[1], *Fuchuan Tong*[1], *Dexin Liao*[2], *Lin Li*[1], *Qingyang Hong*[2]

[1] School of Electronic Science and Engineering, Xiamen University, China
[2]School of Informatics, Xiamen University, China

lilin@xmu.edu.cn, qyhong@xmu.edu.cn

Abstract

The rising interest in single-channel multi-speaker speech separation has triggered the development of end-to-end multi-speaker automatic speech recognition (ASR). However, until now, most systems have adopted autoregressive mechanisms for decoding, resulting in slow decoding speed, which is not conducive to the application of multi-speaker speech recognition in real-world environments. In this paper, we first comprehensively investigate and compare the mainstream end-to-end multi-speaker speech recognition systems. Secondly, we improve the recently proposed non-autoregressive end-to-end speech recognition model Mask-CTC, and introduce it to multi-speaker speech recognition to achieve real-time decoding. Our experiments on the LibriMix data set show that under the premise of the same amount of parameters, the non-autoregressive model achieves performance close to that of the autoregressive model while having a faster decoding speed.

Index Terms: Speech separation, end-to-end multi-speaker speech recognition, non-autoregressive, real-time decoding.

1. Introduction

In recent years, end-to-end (E2E) ASR has made great progress in a relatively quiet environment with a single speaker [1–7], e.g., recordings of telephone speech or audio books. However, more realistic scenarios like spontaneous speech or meetings with multiple participants require the ASR system to recognize the speech of multiple speakers simultaneously, which is also known as the cocktail party problem. Thus, there has been a growing interest in source separation systems and multi-speaker ASR. In this paper, we focus on single-channel multi-speaker speech recognition because it is important not only in scenarios where only a single-channel is available (e.g., telephone conference recordings), but also in multi-channel recordings where conventional multi-channel processing methods, e.g., beamforming [8], cannot separate the speakers well enough in case, e.g., they are spatially too close to each other.

For single-channel source separation, the mainstream methods can be divided into two main categories: frequency-domain source separation and time-domain source separation. For frequency domain source separation, Deep Clustering (DPCL) [9] employs a neural network to map each time-frequency bin to an embedding vector by forming a cluster of embedding vectors of the same speaker in the embedding space. These clusters can be found by a clustering algorithm and be used for constructing a mask for separation in frequency domain. Another mainstream frequency-domain approach is time-frequency masking (TF-masking) [10–13], which generates the corresponding mask values for each source as training labels, such as ideal binary mask (IBM) [14], ideal ratio mask (IRM) [15] and phase-sensitive mask (PSM) [16], by estimating the weight of each source in the time-frequency bin of the mixed speech, and uses neural networks to estimate these masks. Each source can be obtained by multiplying the corresponding mask values with the amplitude spectrum of the mixed speech, and then converted to speech waveform by inverse short-time Fourier transform (iSTFT). However, TF-masking source separation uses only the amplitude spectrum as the input of the neural networks and the phase of the mixed speech to recover each source, which limits the performance of these methods. In contrast, time-domain source separation, such as time-domain audio separation network (TasNet) [17–19], directly uses the speech waveform as the input of the neural networks, utilizing both amplitude and phase information, which greatly improves the separation performance. For TasNet and TF-masking methods, the output permutation of each source is uncertain. In order to achieve speaker-independent source separation, the permutation invariant training (PIT) [20] is usually used to solve this problem by selecting the one with the lowest loss.

There are two mainstream end-to-end multi-speaker speech recognition architectures: the full E2E multi-speaker ASR [21–23], and the model that combines source separation and E2E ASR. The former uses Mel-filterbank coefficients of the mixed speech as input, and introduces speaker-dependent encoders for implicit speech separation, and then uses shared recognition encoders and decoders for multi-speaker speech recognition. This full E2E architecture is based on the PIT criterion fully optimized using the ASR loss, does not require individual sources speech as the training target of the source separation part. The latter uses the source separation methods mentioned above for speaker separation followed by speech recognition. The advantage of this joint architecture is that source separation and speech recognition can be achieved simultaneously, and the separated speech can be preserved. However, the autoregressive mechanism is usually used in both architectures for decoding, which leads to slow decoding speed and is not conducive to application in real scenarios.

The reason why attention-based end-to-end ASR uses the autoregressive decoding mechanism is that the decoder of it is essentially an autoregressive speech-conditional language model. Thus, if the decoder can work in a non-autoregressive mechanism, its decoding speed will be faster. Based on this idea, [24] proposed Mask-CTC, which introduces conditional masked language model (CMLM) [25] as the decoder to decode in an iterative refinement manner, greatly speeding up the decoding speed. In this paper, in order to obtain better recognition performance, we further improve the decoding strategy of the original Mask-CTC and introduce it to multi-speaker speech recognition systems to achieve real-time speech recognition decoding.

The rest of the paper is organized as follows. In Section 2, various source separation models and E2E multi-speaker ASR systems are described in detail. In Section 3, we address the experimental details. Experimental results are presented in Section 4. Finally, the paper is concluded in Section 5.

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1449

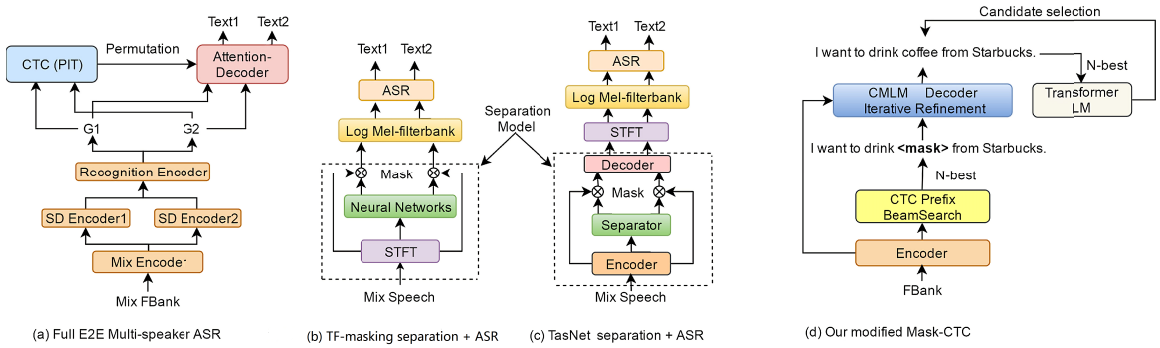

Figure 1: *Mainstream end-to-end multi-speaker speech recognition systems and modified Mask-CTC.*

2. System overview

2.1. Full E2E multi-speaker ASR

As shown in Figure 1(a), full E2E multi-speaker ASR receives the FBank coefficients of the mixed speech as input, and first preprocesses them using a shared encoder (Mix) to extract the high-level features suitable for speech separation. Then, two speaker-dependent (SD) encoders are used to separate the overlapped speech and extract speaker-dependent feature representations from the mixed speech. The recognition encoder is responsible for extracting high-level features from the separated features to facilitate speech recognition. To reduce the computational effort, permutation invariance training is performed with the connectionist temporal classification (CTC) [26] loss function, which provides the decoder with the correct permutation to compute the cross-entropy loss. The total loss function of the complete model is a weighted sum of the CTC loss function and the decoder's cross-entropy loss function:

$$L_{loss} = \alpha L_{CTC}^{\hat{\pi}} + (1-\alpha) L_{CE}^{\hat{\pi}} \tag{1}$$

where α is a hyper parameter used to balance the two losses, and $\hat{\pi}$ is the correct permutation obtained from CTC.

2.2. TF-masking separation + ASR

The most intuitive idea for multi-speaker speech recognition is to merge the source separation model with the speech recognition model. Take TF-masking frequency source separation method as an example, as shown in Figure 1(b), the mixed speech is first processed by short-time Fourier transform, and the amplitude spectrum is input to the neural networks to predict the mask of each source, which represents the weight of each source in the mixed speech. Then, the separated sources are passed through Mel filters to extract the FBank coefficients and are fed into the ASR part for recognition.

There are three ways to train this model. The first one is to train the separation part and the ASR part separately, and combine them at inference stage, which is named as cascade model. Since the training objectives of the two models are different, cascade models usually have mismatch problems. The second way is to train the separation model and ASR jointly (ASR-driven), where the ASR loss can be directly used for optimization because the whole network is differentiable. However, using only the ASR loss is strenuous for the model training. The third method is also to train the separation part and the ASR part jointly, but the loss function is a weighted sum of the two parts:

$$L_{loss} = \beta L_{SS-PIT}^{\hat{\pi}} + (1-\beta) L_{ASR}^{\hat{\pi}} \tag{2}$$

where β is a hyper parameter used to balance the two losses, and $\hat{\pi}$ is the correct permutation obtained from the source separation part.

The loss function of the TF-masking source separation has three forms. The first one is to directly calculate the mean square error (MSE) loss between the mask predicted by the neural networks and the ground truth mask (IBM, IRM, PSM, etc.). The second form does not specify the ground truth mask, but obtains the amplitude spectrum or spectrum of each source through the mask predicted by the neural networks, and then calculates the MSE loss on either the magnitude spectrum or the complex spectrum. This form allows the neural networks to learn the most appropriate mask by itself. However, this form may distort the amplitude spectrum, so STFT consistency training [27] is usually used to solve this problem. The third form converts the separated source spectrum into the speech waveform by iSTFT and then calculates the scale invariant signal-to-distortion ratio (SI-SDR) [17] loss. In this form, the neural networks learn the most appropriate mask by themselves. In addition, all the above three forms require the use of the PIT criterion to solve the permutation problem.

2.3. TasNet separation + ASR

TasNet is an advanced time domain source separation network. As shown in Figure 1(c), TasNet takes the original speech waveform as the input directly. In addition, an encoder is employed in TasNet to replace the STFT module to learn a signal transformations suitable for source separation task automatically. The separator of TasNet serves the same role as the neural networks of the TF-masking methods, which also performs mask prediction of individual sources. The decoder of TasNet is used to replace the iSTFT, where the network reconstructs individual source speech waveforms directly. There are various architectures for TasNet, and the better performing architectures are Conv-TasNet [18] and DPRNN-TasNet [19], so these two variants of TasNet are the main architectures adopted in this paper.

The combined methods of TasNet and ASR are similar to the TF-masking methods, but since TasNet is a time-domain separation network, the separated speech waveforms need to be converted to the frequency domain through STFT before being fed into the Mel filters. Because STFT is differentiable, TasNet and ASR can be trained jointly like the TF-masking method. In addition, TasNet is optimized using the SI-SDR loss function, which is the standard loss function for time-domain source separation. Moreover, TasNet has the same permutation problem and still needs to be optimized using the PIT criterion.

2.4. Modified Mask-CTC

The decoder of the attention-based end-to-end speech recognition system is essentially a speech-conditional autoregressive language model, which requires the previous character as the input to generate the next character, resulting in slow decoding speed. In contrast, mask-based language models, such as bidirectional encoder representations from Transformers (BERT) [28], are capable of generating multiple characters simultaneously with an iterative refinement non-autoregressive mechanism. In view of this, Mask-CTC introduces a conditional mask-based language model (CMLM) [25] as the decoder for the attention-based E2E ASR, and achieves non-autoregressive fast decoding. Specifically, the speech features are input to the encoder, and the initial results are obtained through CTC greedy search. Then, the characters with low confidence in the results are replaced with the special character $<mask>$ and then are input to the CMLM decoder for iterative correction, so that the model outputs characters with higher confidence.

However, CTC greedy search is not always accurate, and the sequence length of the final recognition result is entirely determined by the CTC result, so the decoder is not able to make length correction if the length prediction is not accurate, which causes the performance degradation. In order to solve this problem, we improve the decoding strategy of Mask-CTC. As shown in Figure 1(d), we first use CTC prefix beam search to generate multiple candidate paths (N-best), making the initial recognition results as accurate as possible. Then we feed these candidates into the CMLM decoder for batch decoding, and finally we use the Transformer language model for re-scoring to select the best result. The scores of each candidate are as follows:

$$S = \sum_{i=1}^{l} (S_{yi}^{CMLM} + \gamma S_{yi}^{LM}) \tag{3}$$

where l is the length of the candidate sequence, yi is the i-th character of the candidate sequence, S_{yi}^{CMLM} and S_{yi}^{LM} are the scores of yi in the CMLM decoder and the Transformer LM outputs, respectively. The γ is a hyperparameter to adjust the weight of the Transformer LM. We select the candidate with the highest score as the final recognition result.

We introduce the modified Mask-CTC into the end-to-end multi-speaker speech recognition systems mentioned above to achieve real-time end-to-end monaural multi-speaker speech recognition. Specifically, for the full E2E multi-speaker ASR, we use the CMLM decoder of the modified Mask-CTC to replace the original attention-decoder module and keep the rest of the modules same as before. For the combined source separation and ASR architecture, we use the modified Mask-CTC to replace the ASR part entirely.

3. Experimental setup

3.1. LibriMix dataset

LibriMix [29] is a new open source speech separation benchmark dataset, which contains artificial mixtures of utterances from LibriSpeech, including four conditions (2/3 speakers, clean or noisy). We choose the subset (Libri2Mix) from LibriMix for our experiments, which contains 292 hours of speech data, 1252 speakers, and the sampling rate is 8 kHz. The training set, development set, and test set contain 270 hours, 11 hours, and 11 hours of two speaker's mixed speech, respectively. For training efficiency, we removed the speech larger than 15 seconds from the training set, so the final training set is 137 hours. Another popular speech separation dataset is wsj0-2mix [12, 30], which consists of 30 hours of fully overlapped speech data. However, real-world overlap ratios are typically in the order of 20% or less in natural meetings and casual dinner parties, which shows that wsj0-2mix is not applicable to real scenarios. In contrast, Libri2Mix contains a variety of mixed speech data with different overlap rates: 0 %, 20 %, 40 %, 60 %, 80 %, and 100 %, which can be adapted to different application scenarios. In addition, Hanning windows with length of 256, 128 hop length, and FFT length of 256 are used for STFT computation, and 80-dimensional FBank coefficients are used for speech recognition. For the modeling unit, we use English characters due to the training data size limitation.

3.2. Implementation details

Since the training period of Mask-CTC model is relatively high (300 epochs), we first experiment on the autoregressive Transformer ASR (80 epochs), and then choose the optimal architecture to replace with Mask-CTC. The parameters of the Transformer ASR and the Mask-CTC are entirely the same, only the decoding strategy is different. As suggested in [5], we use the configuration named $big\ model$ for Transformer, in which d^{att}=256, d^{conv1d}=2048, d^{head}=4 , e=12, and d=6. In addition, a convolutional front-end was used to subsample the acoustic features by a factor of 4, and CTC loss (weight=0.3) is used to assist in training. For full E2E multi-speaker ASR, the convolutional front-end just mentioned is used as the mix encoder. The SD encoder and the recognition encoder are the 4-layer and 8-layer Transformer encoder respectively, and the decoder is the 6-layer Transformer decoder. Their parameters are the same as the layers of Transformer ASR mentioned above. Moreover, a 6-layer Transformer LM is used as an option to further performance improvement.

We have investigated a variety of source separation models, and due to the space limitations, we provide the specific details of each source separation model via the github web page: `https://github.com/Alex-Songs/Multi-speaker-ASR`. During training, for the architecture combining source separation and ASR, we default to train jointly and adjust the hyperparameter β so that the loss ratio between the two is 1:1. Moreover, we also compare the cascade system and the ASR-driven training method to ensure the integrity of our experiments.

In all experiments, we set the number of parameters of all the models to be close for fair comparison. The Adam [31] optimizer is used to optimize all models. The initial learning rate is set to 0.001, and it will decay 0.5 when the validation loss goes up. The final models are selected by early stopping mechanism [32].

4. Results

4.1. Comprehensive results analysis

We have conducted sufficient experiments on the model architectures mentioned in the paper, including full E2E multi-speaker ASR, the combination of frequency domain source separation and ASR, and the combination of time domain source separation and ASR. As shown in Table 1, the best results were achieved with the combined architecture of time-domain source separation and ASR, especially DPRNN-TasNet. This is because the time-domain source separation uses the speech waveform as the input, which can make full use of the amplitude and phase information, and in turn facilitates the speech separa-

Table 1: *Performance comparison of different models.*

Model	Min_dev_WER(%)	Min_test_WER(%)	Min_dev_CER(%)	Min_test_CER(%)	SDRi_dev	SDRi_test	RTF (CPU, nj=1)
Full End-to-End Multi-speaker ASR							
Multi-speaker LAS [22] (100.63M)	48.11	48.39	26.69	26.89			3.939
+ Transformer LM (shallow fusion, weight=0.6, 7.6M)	50.28	50.37	39.33	39.39			6.097
Multi-speaker Transformer (32.38M)	26.58	26.55	12.68	12.52			6.794
+ Transformer LM (shallow fusion, weight=0.6, 7.6M)	**24.28**	**24.34**	11.85	11.76			8.953
Multi-speaker Mask-CTC (max_iteration=10) (32.38M)	28.92	28.77	12.06	12.01			**0.091**
+ CTC Prefix BeamSearch (max_iteration=10, beam=2)	28.45	28.32	11.67	11.65			0.201
+ CTC Prefix BeamSearch (max_iteration=10, beam=4)	28.11	28.03	11.65	11.62			0.411
+ CTC Prefix BeamSearch (max_iteration=10, beam=8)	27.01	27.02	11.53	11.52			0.903
+ Transformer LM (rescore, weight=0.6, 7.6M)	**26.15**	**26.11**	**11.41**	**11.43**			0.939
Frequency domain speech separation + ASR							
TF-masking (BLSTM) + ASR (ASR-driven) (30.38M)	69.10	69.70	51.40	52.30	-0.291	-0.373	4.740
TF-masking (BLSTM)_log_spectrum_mse + ASR (30.38M)	25.50	25.80	12.40	12.60	10.640	10.923	4.740
TF-masking (BLSTM)_IBM_mse + ASR (30.38M)	23.30	23.50	9.80	10.00	1.504	1.551	4.740
TF-masking (BLSTM)_IRM_mse + ASR (30.38M)	23.20	23.70	9.70	10.10	3.493	3.566	4.740
TF-masking (BLSTM)_PSM_mse + ASR (30.38M)	28.20	28.70	13.40	13.70	5.973	5.770	4.740
TF-masking (BLSTM)_SI-SDR + ASR (30.38M)	22.70	23.00	10.70	10.70	**11.337**	**11.050**	4.740
TF-masking (BLSTM)_magnitude_mse + ASR (30.38M)	22.10	22.40	10.40	10.40	7.411	6.988	4.740
+ STFT Consistency [27]	21.00	21.90	9.90	10.20	7.521	7.042	4.740
+ Transformer LM (shallow fusion, weight=0.6, 7.6M)	**18.90**	**19.50**	**9.70**	**10.00**	7.521	7.042	5.895
+STFT Consistency (cascade)	67.00	67.10	41.20	41.20	8.701	8.374	4.740
TF-masking (TCN)_magnitude_mse + ASR (30.38M)	24.20	24.40	12.40	12.30	1.369	1.310	5.146
TF-masking (Transformer)_magnitude_mse + ASR (31.16M)	24.90	25.20	12.80	12.90	0.535	0.488	4.604
TF-masking (Conformer)_magnitude_mse + ASR (31.96M)	23.90	23.50	12.20	11.70	1.047	1.082	4.875
Time domain speech separation + ASR							
Conv-TasNet + ASR (31.58M)	19.80	19.50	8.10	8.70	10.916	10.838	5.792
DPRNN-TasNet + ASR (ASR-driven) (29.73M)	60.70	61.10	45.80	46.10	-0.110	-0.130	5.125
DPRNN-TasNet + ASR (cascade) (29.73M)	30.70	31.32	15.57	15.76	**13.790**	**13.681**	5.125
DPRNN-TasNet + ASR (29.73M)	15.30	14.50	6.70	6.80	10.232	10.081	5.125
+ Transformer LM (shallow fusion, weight=0.6, 7.6M)	**14.50**	**13.60**	**6.30**	**6.50**	10.232	10.081	5.875
DPRNN-TasNet+ Mask-CTC (max_iteration=10) (29.73M)	18.37	17.56	7.90	8.20	10.537	10.457	**0.078**
+ CTC Prefix BeamSearch (max_iteration=10, beam=2)	18.01	17.21	7.85	8.10	10.537	10.457	0.184
+ CTC Prefix BeamSearch (max_iteration=10, beam=4)	17.85	17.01	7.75	7.95	10.537	10.457	0.388
+ CTC Prefix BeamSearch (max_iteration=10, beam=8)	16.95	16.27	7.52	7.62	10.537	10.457	0.837
+ Transformer LM (rescore, weight=0.6, 7.6M)	**15.98**	**15.19**	7.28	7.58	10.537	10.457	0.897

tion and speech recognition performance. For the architecture of combining frequency domain source separation and ASR, we first performed a comparison of various loss function forms (as mentioned in Section 2.2), where the combination of amplitude spectrum MSE loss and STFT consistency has the best performance, which indicates that speech recognition is more sensitive to the amplitude spectrum, and using the amplitude spectrum as the training target of the source separation part improves the performance of speech recognition. In addition, we have tried various networks for mask prediction, such as temporal convolutional network (TCN) [33], Transformer [34], and Conformer [7], but the performance is not as good as BLSTM. This is because TCN has limited perceptual field, while Transformer and Conformer usually need deeper network structure to achieve comparable performance. For full End-to-end multi-speaker ASR, although in general the performance is not as good as the other two structures, it is more suitable for scenarios where there is no source speech and only the corresponding text is available.

4.2. Comparison of different training methods

As shown in Table 1, we trained the combined source separation and ASR architecture with different training methods (as described in Section 2.2), where the best results were obtained by training source separation and ASR jointly with the weighted summation of loss functions. We also found that the cascaded system has better speech separation performance (SDRi) and poorer speech recognition performance, which indicates that the cascade system does have a mismatch problem that makes performance degrade. In addition, the pure ASR-driven approach has poor performances due to the lack of source speech as the training targets for the source separation part. In this scenario, the full E2E multi-speaker ASR is a better choice.

4.3. The effect of the modified Mask-CTC

We can see from Table 1 that the real time factor (RTF) metrics of the models using the autoregressive mechanism for decoding, tested on a server configured with a 4-core Intel(R) Xeon(R) CPU with 125G memory, are all much larger than 1. This indicates that the decoding speed of these models is slow, which limits the deployment of end-to-end multi-speaker speech recognition systems in real application scenarios. In view of this, we introduced the modified Mask-CTC into the best model of the structure of full E2E multi-speaker ASR and the combined source separation and ASR architectures (as mentioned in Section 2.4). From the experiments, it is clear that the modified Mask-CTC achieves a performance close to that of the autoregressive ASR model in both architectures with the RTF metric less than 1, which achieves our desired goal of real-time end-to-end monaural multi-speaker speech recognition.

5. Conclusions

In this paper, we conducted a sufficient investigation of various mainstream source separation models and end-to-end multi-speaker models. In addition, we improved the non-autoregressive ASR model Mask-CTC and introduced it into end-to-end multi-speaker ASR, which greatly accelerates the decoding speed and facilitates the application of end-to-end multi-speaker ASR in real scenarios.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.61876160 and No.62001405).

7. References

[1] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[2] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[4] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[5] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[6] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for ASR," *arXiv preprint arXiv:1904.11660*, 2019.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*, 2020, pp. 5036–5040.

[8] D. Ba, D. Florencio, and C. Zhang, "Enhanced mvdr beamforming for arrays of directional microphones," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 1307–1310.

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

[10] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[11] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[12] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[13] J. Li, H. Zhang, X. Zhang, and C. Li, "Single channel speech enhancement using temporal convolutional recurrent neural networks," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 896–900.

[14] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, pp. 181–197, 2005.

[15] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[17] 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)*, 2018, pp. 696–700.

[18] 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. PP, no. 99, pp. 1–1, 2019.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.

[20] D. Yu, M. Kolbaek, 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)*, 2017, pp. 241–245.

[21] H. Seki, T. Hori, S. Watanabe, J. L. Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2620–2630.

[22] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6256–6260.

[23] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6134–6138.

[24] Y. Higuchi, S. Watanabe, N. Chen, T. Ogawa, and T. Kobayashi, "Mask ctc: Non-autoregressive end-to-end asr with ctc and mask predict." in *Interspeech 2020*, 2020, pp. 3655–3659.

[25] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models." in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6111–6120.

[26] A. Graves, S. Fernndez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[27] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 900–904.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[30] 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)*, vol. 2017, 2017, pp. 246–250.

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

[32] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the Trade, this book is an outgrowth of a 1996 NIPS workshop*, 1998, pp. 55–69.

[33] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[34] 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.