DON'T SHOOT BUTTERFLY WITH RIFLES: MULTI-CHANNEL CONTINUOUS SPEECH SEPARATION WITH EARLY EXIT TRANSFORMER

Sanyuan Chen[‡], Yu Wu[†], Zhuo Chen[†], Takuya Yoshioka[†], Shujie Liu[†], Jinyu Li[†], Xiangzhan Yu[‡]

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

ABSTRACT

With its strong modeling capacity that comes from a multi-head and multi-layer structure, Transformer is a very powerful model for learning a sequential representation and has been successfully applied to speech separation recently. However, multi-channel speech separation sometimes does not necessarily need such a heavy structure for all time frames especially when the cross-talker challenge happens only occasionally. For example, in conversation scenarios, most regions contain only a single active speaker, where the separation task downgrades to a single speaker enhancement problem. It turns out that using a very deep network structure for dealing with signals with a low overlap ratio not only negatively affects the inference efficiency but also hurts the separation performance. To deal with this problem, we propose an early exit mechanism, which enables the Transformer model to handle different cases with adaptive depth. Experimental results indicate that not only does the early exit mechanism accelerate the inference, but it also improves the accuracy.

Index Terms— speech separation, multi-channel microphone, Transformer, deep learning

1. INTRODUCTION

Speech separation plays a vital role in front-end speech processing, aiming to handle the cocktail party problem. Starting from deep clustering (DC) [1, 2] and permutation invariant training (PIT) [3, 4], a variety of separation models have been shown effective in separating overlapped speech [5, 6]. Recently, the deep learning methods have been rigorously explored for better speech separation capability, including dual-path RNN [7], Conv-tasnet [8], and deep CASA [9] employing RNN and CNN structures. With the success of Transformer model in speech community [10, 11], Transformer [12] and its variants [13] have successfully been applied to this task.

The Transformer model integrates a stack of self-attention layers to model the speech representation. Prior work shows that a deeper structure yields superior performance [14]. For example, for automatic speech recognition (ASR) tasks, a common setting is to use twelve [15] or more layers [16] in the encoder. However, continuous speech separation (CSS), which we are addressing, is a simpler task especially with a multi-channel setting. Multiple microphones combine to provide rich spatial information that allows simple models to perform the separation job with high accuracy. Applying a deep Transformer for the multi-channel CSS might be overkill for frames with only one active speaker, resulting in two problems: 1) Real-time inference is usually preferred for product deployment, especially for resource-constrained devices. The Transformer has a **heavy runtime cost** due to its deep encoder. Hence, it is necessary to speed up the execution of the Transformer-based speech separation models. 2) The Transformer model may suffer from the "**overthinking**" problem [17] as it contains too many encoder layers. We assume that a shallow Transformer encoder is sufficient to handle the non-overlapped speech well and that a deep Transformer model could potentially degrade the speech estimation.

Inspired by the depth-adaptive inference method [17], we propose to mitigate these problems with an Early Exit mechanism, which essentially makes predictions at an earlier layer for less overlapped speech while using higher layers for speech with a high overlap rate. We believe that the first few layers are sufficient to handle the less overlapped speech and thus an early exit scheme reduces the overall runtime cost. When the input contains a lot of overlaps, higher layers are automatically triggered to perform more complex analysis and generate more accurate separation results. Specifically, we introduce a mask estimator to each transformer layer and dynamically stop the inference if the predictions from two consecutive layers are sufficiently similar, based on the normalized Euclidean distance of the two prediction matrices.

We conduct experiments on the LibriCSS dataset [18]. The experimental results show that a stricter threshold (hard to exit) leads to better performance on large-overlapped utterances and worse performance on the small-overlapped utterances, which is consistent with our intuition. With threshold tuning, the proposed model improves the separation quality of the small-overlapped speech while keeping the performance on large-overlapped ones. Moreover, the early exit mechanism enables the Transformer model to achieve better separation performance while 2x the inference speed.

2. APPROACH

2.1. Problem Formulation

Given a continuously provided signal including multiple talkers, CSS aims to retrieve individual constituting utterances and route them to one of its output channels in such a way that each output signal no longer contains overlapped utterances. This is typically performed by applying a sliding window to the input signal and performing the separation at each window position. Within each window, a separation model generates a fixed number of outputs.

Let $y(t)$ denote the mixed signal and $x_s(t)$ the s-th individual target signal, where t is the time index. The mixed signal is modeled as follows:

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

We also denote their short-time Fourier transforms (STFTs) as $\mathbf{Y}(t, f)$ and $\mathbf{X}_s(t, f)$, respectively. f denotes frequency domain.

When C microphones are available, the model input consists of a concatenation of the STFT features of the first channel and the inter-channel phase difference between the i-th channel and the first

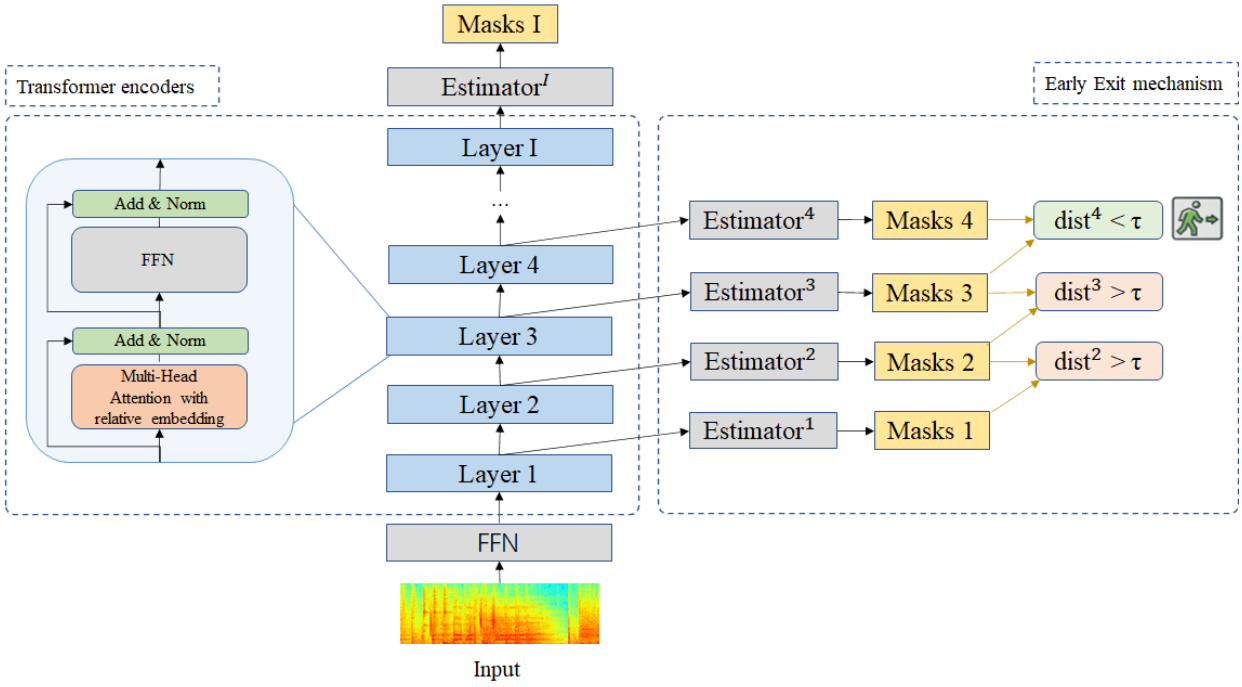

Fig. 1: The architecture of Early Exit Transformer. We attach a mask estimator to each Transformer encoder layer and dynamically stop inference if the predictions are similar between two consecutive layers.

channel. Thus, the features may be represented as

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

where $\mathbf{Y}^i(t,f)$ denotes the STFT of the i-th channel, $\text{IPD}(i) = \theta^i(t,f) - \theta^1(t,f)$, and $\theta^i(t,f)$ is the phase of $\mathbf{Y}^i(t,f)$. Each feature dimension is normalized along the time axis.

Following [19, 20], instead of directly computing the STFT of the individual signals $[\mathbf{X}_1(t,f) \ldots \mathbf{X}_S(t,f)]$, we estimate a group of masks $\mathbf{M}(t,f) = [\mathbf{M}_1(t,f) \ldots \mathbf{M}_S(t,f)]$ with a deep learning model. Then, for the s-th individual signal, $\mathbf{X}_s(t,f)$ is obtained either by beamforming or by masking, i.e., $\mathbf{M}_s(t,f) \odot \mathbf{Y}^1(t,f)$ where $\odot$ is the elementwise product.

In the following section, we will first introduce the Transformer model for speech separation and then our Early Exit Transformer network.

2.2. Transformer Model

As shown in the left side of Figure 1, We estimate the masks from the input mixed signals with the Transformer model [21] which is composed of a stack of identical encoder layers. Each layer consists of a multi-head self-attention module and a position-wise fully connected feed-forward module.

The input of the Transformer model $\mathbf{h}_0$ is a linear conversion of the input $\mathbf{Y}(t,f)$ with a feed-forward module $\text{FFN}(\cdot)$:

$$\mathbf{h}_0 = \text{FFN}(\mathbf{Y}(t,f)). \tag{3}$$

Given the input, $\mathbf{h}_{i-1}$, of the i-th layer, the output $\mathbf{h}_i$ is calculated as

$$\mathbf{h}'_i = \text{layernorm}(\mathbf{h}_{i-1} + \text{MultiHeadAttention}(\mathbf{h}_{i-1})) \tag{4}$$

$$\mathbf{h}_i = \text{layernorm}(\mathbf{h}'_i + \text{FFN}(\mathbf{h}'_i)), \tag{5}$$

where $\text{MultiHeadAttention}(\cdot)$ and $\text{layernorm}(\cdot)$ denote the multi-head self-attention module and the layer normalization, respectively.

The multi-head self-attention module is implemented with relative position embedding as follows:

$$\text{Multihead}(\mathbf{h_{i-1}}) = [\mathbf{H}_1 \ldots \mathbf{H}_{d_{head}}]\mathbf{W}^{head} \tag{6}$$

$$\text{where } \mathbf{H}_j = \text{softmax}\left(\frac{\mathbf{Q}_j(\mathbf{K}_j + \mathbf{pos})^\intercal}{\sqrt{d_k}}\right)\mathbf{V}_j, \tag{7}$$

where d_k is the hidden layer dimensionality, and d_{head} is the number of attention heads. $\mathbf{Q_j}, \mathbf{K_j}, \mathbf{V_j}$ are linear conversions of the input $\mathbf{h}_{i-1}$ with different parameter matrices. $\mathbf{pos} = \{rel_{m,n}\} \in \mathbb{R}^{M \times M \times d_k}$ is the relative position embedding [22], where M is the maximum chunk length, and $rel_{m,n} \in \mathbb{R}^{d_k}$ represents the offset of the m-th vector in $\mathbf{Q_i}$ and the n-th vector in $\mathbf{K_i}$.

Given the output, $\mathbf{h}_I$, of the final layer, we obtain the masks $\mathbf{M}(t,f)$ with $\text{Estimator}^I(\cdot)$, an estimator consisting of a feed-forward module and a sigmoid activation function, i.e.,

$$\mathbf{M}(t,f) = \text{Estimator}^I(\mathbf{h}_I) \tag{8}$$

$$= \text{sigmoid}(\text{FFN}(\mathbf{h}_I)). \tag{9}$$

2.3. Early Exit Transformer

Despite the promising performance, the Transformer model with deep layers is prone to "**heavy runtime cost**" and "**overthinking**" in the speech separation task. To overcome this, based on the assumption that the first few layers are sufficient to handle less overlapped speech, we propose an Early Exit Transformer model (see Fig. 1) to estimate the masks by dynamically choosing the number of layers to

use. Specifically, we attach a layerwise estimator, Estimator$^i(\cdot)$, to the output of each Transformer encoder layer $\mathbf{h}_i$, based on which, we can predict the masks $\mathbf{M}^i(t,f)$ at each internal layer:

$$\mathbf{M}^i(t,f) = \text{Estimator}^i(\mathbf{h}_i) \tag{10}$$

$$= \text{sigmoid}(\text{FFN}(\mathbf{h}_i)). \tag{11}$$

During the inference, given the output of the i-th layer with $i > 1$, we calculate the normalized Euclidean Distance dist^i between the estimated masks of the $(i$-1)-th layer and the i-th layer:

$$\text{mean}_{t,f}\left(\text{EuclideanDistance}\big(\mathbf{M}^{i-1}(t,f), \mathbf{M}^i(t,f)\big)\right) \tag{12}$$

Given a pre-defined threshold τ, if dist$^i < \tau$ for the two consecutive layers, we terminate the inference process and output $\mathbf{M}^i(t,f))$ as the final prediction masks. Instead of performing the inference using all the encoder layers, the early exit mechanism makes the predictions with the first few layers for the small-overlap segments, which can accelerate the inference process and potentially reduce the "overthinking" problem.

During the training, besides the parameters for the Transformer model, the Estimators attached to the internal layers are also trained to predict the masks from the hidden states. Therefore, for each Estimator, we apply PIT [3, 4] to minimize Loss^i which is the Euclidean distance between the reference and the mask predicted by Estimator$^i(\cdot)$. The final loss is the weighted average function, following [17], as

$$\text{Loss} = \frac{\sum_{i=1}^{I} i \cdot \text{Loss}^i}{\sum_{i=1}^{I} i} \tag{13}$$

where $\frac{i}{\sum_{i=1}^{I} i}$ is used as the weight for the loss of the i-th estimator, Estimator^i. A deeper layer is assigned with a larger weight in the loss computation. The intuition behind this is that the more complex the model becomes, the more sensitive it gets to prediction errors. Moreover, while the first layer receives the gradients from all layers, only the gradients of $[i, I]$ layers back-propagate to the i-th layer. Thus, giving a larger loss weight to a deeper layer stabilizes the training process.

3. EXPERIMENT

3.1. Datasets

We train the models with 219 hours of artificially reverberated and mixed speech signals sampled randomly from WSJ1 [23]. Following [24], we include four different mixture types in the training data. Each training mixture is generated by randomly picking one or two speakers from the WSJ1 dataset and convolving each with a 7 channel room impulse response (RIR) simulated with the image method [25]. Then, we rescale and combine them with a source energy ratio between -5 and 5 dB. Simulated isotropic noise [26] is also added at a 0–10 dB signal to noise ratio. The average overlap ratio of the training set is around 50%.

We evaluate the models on the LibriCSS dataset [18], which consists of 10 hours of concatenated and mixed LibriSpeech utterances played and recorded in a meeting room. We test our model performance under a seven-channel setting. We conducted both the utterance-wise evaluation and continuous input evaluation (refer to [18] for the two evaluation schemes).

3.2. Implementation Details

Our baseline speech separation models are BLSTM and vanilla Transformer. The BLSTM model consists of three BLSTM layers with 1024 input dimensions and 512 hidden dimensions, resulting in 21.80M parameters. Three sigmoid projection layers are appended to estimate three masks, two for speakers and one for noise. We use the Adam optimizer [27] to train the BLSTM model with the learning rate initialized to 1e-3. The learning rate is decreased by half if the validation loss stops decreasing for 2 epochs. Training is performed for 100 epochs. The Transformer model consists of 16 Transformer encoder layers with 4 attention heads, 256 attention dimensions and 2048 FFN dimensions, resulting in 21.90M parameters. We use the AdamW optimizer [28] to train the Transformer model with the weight decay set to 1e-2. The learning rate is 1e-4 and the warm-up learning schedule with linear decay is used, where the warm-up step is 10,000, and the training step is 260,000.

Our Early Exit Transformer model is implemented with the same Transformer encoders as the baseline Transformer model. The model is optimized with the weighted average loss (as described in Section 2.3) with the same hyperparameters as the baseline. During inference, we vary the early exit threshold in {0, 3e-5, 5e-5, 8e-5, 1e-4, 1.5e-4, 2e-4, ∞} to control the exit layer and thus the speed-up ratio. We evaluate the speech separation accuracy with two ASR models. One is a hybrid system with a BLSTM based acoustic model and a 4-gram language model as used in the original LibriCSS paper [18]. The other is one of the best open-source end-to-end transformer [16] based ASR models [1] which achieves WERs of 2.08% and 4.95% for LibriSpeech test-clean and test-other, respectively. As with [18], by leveraging the multiple microphones, the individual target signals are generated with mask-based adaptive minimum variance distortionless response (MVDR) beamforming.

3.3. Evaluation Results

Table 1 shows the WERs of our Early Exit Transformer with different threshold τ values as well as those of the baselines for the utterance-wise evaluation. With a larger threshold, the inference process tended to exit at a lower layer, and greater speed-up was obtained. We also found the performance on low overlap ratio sets benefited from the use of fewer inference layers, which implies the "overthinking" problem of the vanilla Transformer model. Specifically, when $\tau = \infty$, the inference process always halted at the second layer. This yielded $6.59\times$ speed-up and achieved the best WERs for the two non-overlap settings. The use of a smaller threshold led to a better separation performance for high overlap ratio settings. When $\tau = 0$, its performance degraded on the small overlap ratio sets. This may have been caused by the mismatch between the training and inference, i.e., with the proposed method, the model tries to predict the mask correctly at all the layers while only the last layer's result is used at the inference time. Moreover, it is slower than the vanilla Transformer since every layer predicts the output once. With a tuned threshold, better results were obtained by mitigating the "overthinking" problem. With $\tau = 1.5e-4$, our Early Exit Transformer achieved better results in the small overlap ratio settings than the vanilla Transformer while achieving a $4.08\times$ speed-up. With a modest threshold setting ($\tau = 8e-5$), the inference time was halved while also achieving better speech separation performance with both of the two ASR models for all overlap settings.

Table 2 shows the continuous evaluation results. As with the utterance-wise evaluation, the Early Exit Transformer mod-

[1] https://github.com/MarkWuNLP/SemanticMask

Table 1: Utterance-wise evaluation. Two numbers in a cell denote %WER of the **hybrid SR model** used in LibriCSS [18] and **end-to-end transformer** based SR model [16]. 0S: 0% overlap with short inter-utterance silence. 0L: 0% overlap with a long inter-utterance silence.

System	Avg. exit layer	Speed-up	Overlap ratio in %					
			0S	0L	10	20	30	40
No separation [18]	-	-	11.8/5.5	11.7/5.2	18.8/11.4	27.2/18.8	35.6/27.7	43.3/36.6
BLSTM [13]	-	-	**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 [13]	16.0	1.00×	8.3/3.4	8.4/3.4	11.4/4.1	12.5/**4.8**	14.7/6.4	16.9/7.2
Early Exit Transformer ($\tau = 0$)	16.0	0.92×	8.9/3.4	9.4/3.6	12.3/4.2	14.7/5.0	15.1/**6.2**	**16.5/6.6**
Early Exit Transformer ($\tau = 8e-5$)	6.9	2.60×	7.6/**3.2**	7.7/**3.3**	10.1/**3.8**	12.4/**4.8**	**14.4/6.2**	**16.4**/6.9
Early Exit Transformer ($\tau = 1.5e-4$)	4.8	4.08×	7.8/**3.2**	7.6/**3.4**	**9.8/3.8**	**12.2**/5.1	14.7/6.7	17.9/7.8
Early Exit Transformer ($\tau = \infty$)	2.0	6.59×	**7.1/3.1**	**7.3/3.3**	10.0/4.4	13.6/6.1	17.0/8.4	20.5/10.4

Table 2: Continuous speech separation evaluation

System	Avg. exit layer	Speed-up	Overlap ratio in %					
			0S	0L	10	20	30	40
No separation [18]	-	-	15.4/12.7	11.5/5.7	21.7/17.6	27.0/24.4	34.3/30.9	40.5/37.5
BLSTM [13]	-	-	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 [13]	16.0	1.00×	12.0/5.6	9.1/4.4	13.4/6.2	14.4/**6.8**	18.5/9.7	19.9/**10.3**
Early Exit Transformer ($\tau = 0$)	16.0	0.76×	14.1/6.2	10.3/4.6	17.2/7.1	17.3/7.5	23.0/10.8	23.5/12.0
Early Exit Transformer ($\tau = 1e-4$)	7.5	1.47×	**11.3**/5.4	8.9/4.4	**12.7/6.0**	**13.8/6.7**	17.8/**9.3**	**19.7**/10.5
Early Exit Transformer ($\tau = 1.5e-4$)	5.8	1.88×	11.5/**5.2**	8.9/4.3	**12.6/6.0**	**13.7**/6.9	**17.6**/9.5	**19.6/10.3**
Early Exit Transformer ($\tau = 2e-4$)	5.2	2.08×	**11.2**/5.6	8.8/4.5	**12.7**/6.3	13.9/7.2	18.5/9.5	**19.6**/10.9
Early Exit Transformer ($\tau = \infty$)	2.0	4.74×	14.7/14.6	8.7/6.9	16.1/13.7	17.8/15.2	22.5/18.2	24.8/18.9

els achieved superior performance to the vanilla Transformer while improving the inference time by a factor of two. The improvements were more prominent on the small-overlapped test sets. In contrast to the utterance-wise speech separation, we observed that the inference process tended to stop at a higher layer for the same threshold and that the best results for the non-overlap settings were achieved with some internal layer rather than with the second layer. This could be because, in CSS, each model evaluation used a shorter chunk than the typical utterance length of the utterance-wise evaluation dataset, making the task harder.

3.4. Discussion and Analysis

In addition to the main experiments, there are several interesting questions that should be discussed.

Exit layer across different testsets: We first explore what the exit layer distribution is with respect to different overlapped ratios. Figure 2 shows that the averaged exit layer slightly increases as the overlap ratio becomes larger. When $\tau = 8e-5$, Early Exit Transformer makes predictions one layer deeper on 40% overlap testset compared to 0S testset. The increment is more significant with $\tau = 5e-5$, demonstrating that small overlapped cases tend to exit at shallow layers which is consistent with our intuition.

Performance for the single channel scenario: We also tried the early exit mechanism on the single channel speech separation task. For this scenario, the conclusion is the more layers we use, the better performance we get. We can obtain speed acceleration but performance degradation with the early exit mechanism. Specifically, we obtain a 1.61% speed-up with 6.0% relative WER losses and a 2.07% speed-up with 10.4% relative WER losses on average. We think that speech separation for single channel is much more challenging due to the absence of the microphone array signal, and less than 16 layers are not enough to handle this task well. We leave it for future work to explore the early exit mechanism for the single channel.

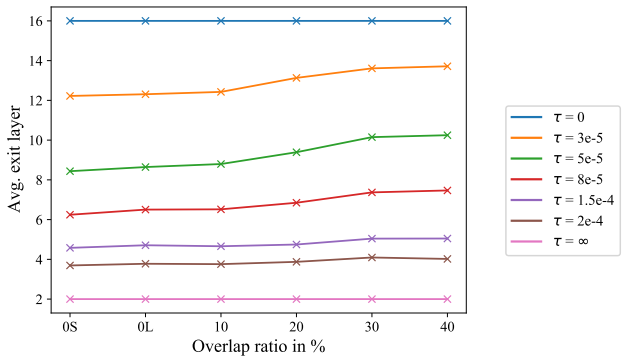

Fig. 2: The average exit layer of Early Exit Transformer across different testsets with different threshold τ for the utterance-wise evaluation.

4. CONCLUSION

We elaborate an early exit mechanism for Transformer based multi-channel speech separation, which aims to address the "overthinking" problem and accelerate the inference stage simultaneously. Each Transformer layer is equipped with a mask estimator, and the early exit is triggered if the outputs of two successive layers are similar. Experiment results show that it does not only speed up inference, but also improves the performance on small-overlapped testsets, which is consistent with our intuition. Regarding single-channel evaluation, we observe negative results since the task is too challenging to handle. In the future, we will study speeding up Transformer based separation model from other perspectives.

5. REFERENCES

[1] 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.

[2] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[3] 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.

[4] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper 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.

[5] Takuya Yoshioka, Igor Abramovski, Cem Aksoylar, Zhuo Chen, Moshe David, Dimitrios Dimitriadis, Yifan Gong, Ilya Gurvich, Xuedong Huang, Yan Huang, et al., "Advances in online audio-visual meeting transcription," in *Proc. ASRU*. IEEE, 2019, pp. 276–283.

[6] Takuya Yoshioka, Zhuo Chen, Changliang Liu, Xiong Xiao, Hakan Erdogan, and Dimitrios Dimitriadis, "Low-latency speaker-independent continuous speech separation," in *ICASSP*. IEEE, 2019, pp. 6980–6984.

[7] 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.

[8] Yi Luo and Nima 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.

[9] Ziqiang Shi, Rujie Liu, and Jiqing Han, "La furca: Iterative context-aware end-to-end monaural speech separation based on dual-path deep parallel inter-intra bi-lstm with attention," *arXiv preprint arXiv:2001.08998*, 2020.

[10] 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.

[11] 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 *Proc. Interspeech*, 2020.

[12] 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.

[13] Sanyuan Chen, Yu Wu, Zhuo Chen, Jinyu Li, Chengyi Wang, Shujie Liu, and Ming Zhou, "Continuous speech separation with conformer," *arXiv preprint arXiv:2008.05773*, 2020.

[14] Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Sebastian Stüker, and Alexander Waibel, "Very deep self-attention networks for end-to-end speech recognition," in *Proc. Interspeech*, 2019.

[15] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al., "A comparative study on transformer vs rnn in speech applications," in *Proc. ASRU*. IEEE, 2019, pp. 449–456.

[16] 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," *arXiv preprint arXiv:1912.03010*, 2019.

[17] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras, "Shallow-deep networks: Understanding and mitigating network overthinking," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3301–3310.

[18] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li, "Continuous speech separation: Dataset and analysis," in *Proc. ICASSP*. IEEE, 2020, pp. 7284–7288.

[19] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[20] 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.

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

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

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

[24] 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.

[25] Emanuel AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[26] Emanuël AP Habets and Sharon Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[27] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.