IEEE ICASSP 2021

Don't shoot butterfly with rifles: Multi-channel Continuous Speech Separation with Early Exit Transformer

Poster #1903

Sanyuan Chen[2], Yu Wu[1], Zhuo Chen[1], Takuya Yoshioka[1], Shujie Liu[1], Jinyu Li[1], Xiangzhan Yu[2]

[1]Microsoft Corporation, [2]Harbin Institute of Technology

Multi-channel Continuous Speech Separation

The **goal** of continuous speech separation is to estimate individual speaker signals from a continuous speech input, where the source signals are fully or partially overlapped.

The mixed signal is formulated as $y(t) = \sum_{s=1}^{S} x_s(\mathrm{t})$, where $x_s(\mathrm{t})$ is the s-th source signal, t is the time index. $\boldsymbol{Y}^1(t,f)$ and $\boldsymbol{X}_s(t,f)$ refers to the STFT of the first channel of $y(t)$ and $x_s(\mathrm{t})$, respectively.

When C microphones are available, the **input** of the speech separation model is $\boldsymbol{Y}(t,f) = \boldsymbol{Y}^1(t,f) \oplus \mathrm{IPD}(2) \ldots \oplus \mathrm{IPD}(\mathrm{C})$, where $\mathrm{IPD}(\mathrm{i})$ is the inter-channel phase difference between the i-th channel and the first channel.

The speech separation model **estimates** a group of Masks $\{\boldsymbol{M}_s(t,f)\}_{1\le s\le S}$. Then each $\boldsymbol{X}_s(t,f)$ is obtained as $\boldsymbol{M}_s(t,f) \odot \boldsymbol{Y}^1(t,f)$, where $\odot$ is an elementwise product.

Transformer model

Prior work shows that the Transformer model with a deeper structure yields superior performance.

Transformer model 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.

In the **self-attention module**, we firstly convert the hidden state to Q, K, V, and then apply a multi-head self-attention mechanism with relative position embedding.

Problems when applying a deep Transformer for the multi-channel CSS:

1. **Heary run-time cost**: 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.
2. **"overthinking" problem**: We assume that a shallow Transformer encoder is sufficient to handle the nonoverlapped speech well and that a deep Transformer model could potentially degrade the speech estimation.

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.

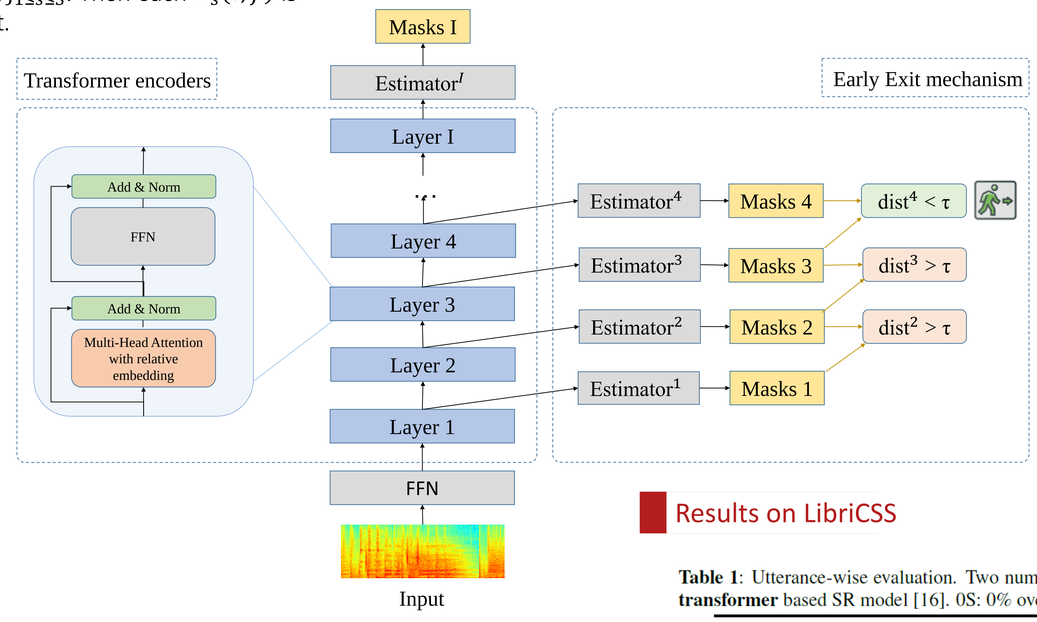
We not only **speed up inference**, but also **improves the performance** on small-overlapped testsets.

Regarding **single channel evaluation**, we obtain speed acceleration but performance degradation. We think that speech separation for single channel is much more challenging due to the absence of the microphone array signal. We leave it for future work to explore the early exit mechanism for the single channel.

Early Exit Transformer model

we propose to mitigate the Heavy run-time cost and "overthinking" 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.

Specifically, we attach 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.



Inference:

- for each i–th layer, we calculate the normalized Euclidean Distance dist^i between the estimated masks of the $(\mathrm{i}-1)$–th layer and the i–th layer.
- Given a pre-defined threshold τ, if $\mathrm{dist}^i < \tau$ for the two consecutive layers, we terminate the inference process and output the estimated masks of i–th layer as the final prediction masks.

Training:

- For each $\mathrm{Estimator}^i$, we apply PIT (permutation invariant training) to minimize Loss^i which is the Euclidean distance between the reference and the mask predicted by i–th layer.
- And the final loss is the weighted average function:

$$\mathrm{Loss} = \frac{\sum_{i=1}^{I} i \cdot \mathrm{Loss}^i}{\sum_{i=1}^{I} i}$$

Analysis

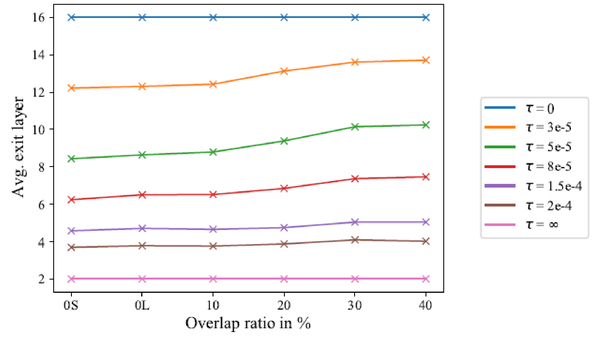


Fig. 2: The average exit layer of Early Exit Transformer across different testsets with different threshold τ for the utterance-wise evaluation.

Results on LibriCSS

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