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

Accepted by IEEE ICASSP 2021

Multi-channel Continuous Speech Separation

- To estimate individual speaker signals from a continuous speech input, where the source signals are fully or partially overlapped.
- Mixed signal: $y(t) = \sum_{s=1}^{S} x_s(t)$ → s-th source signal: $x_s(t)$
- (STFTs) short-time Fourier transforms: $\mathbf{Y}^1(t, f)$ → $\mathbf{X}_s(t, f)$
- Speech Separation Process:

1. $\mathbf{Y}(t, f) = \mathbf{Y}^1(t, f) \oplus \mathrm{IPD}(2) \ldots \oplus \mathrm{IPD}(C)$ —Separation model→ $\mathbf{M}_s(t, f)$
2. $\mathbf{X}_s(t, f) = \mathbf{M}_s(t, f) \odot \mathbf{Y}^1(t, f)$

Transformer model

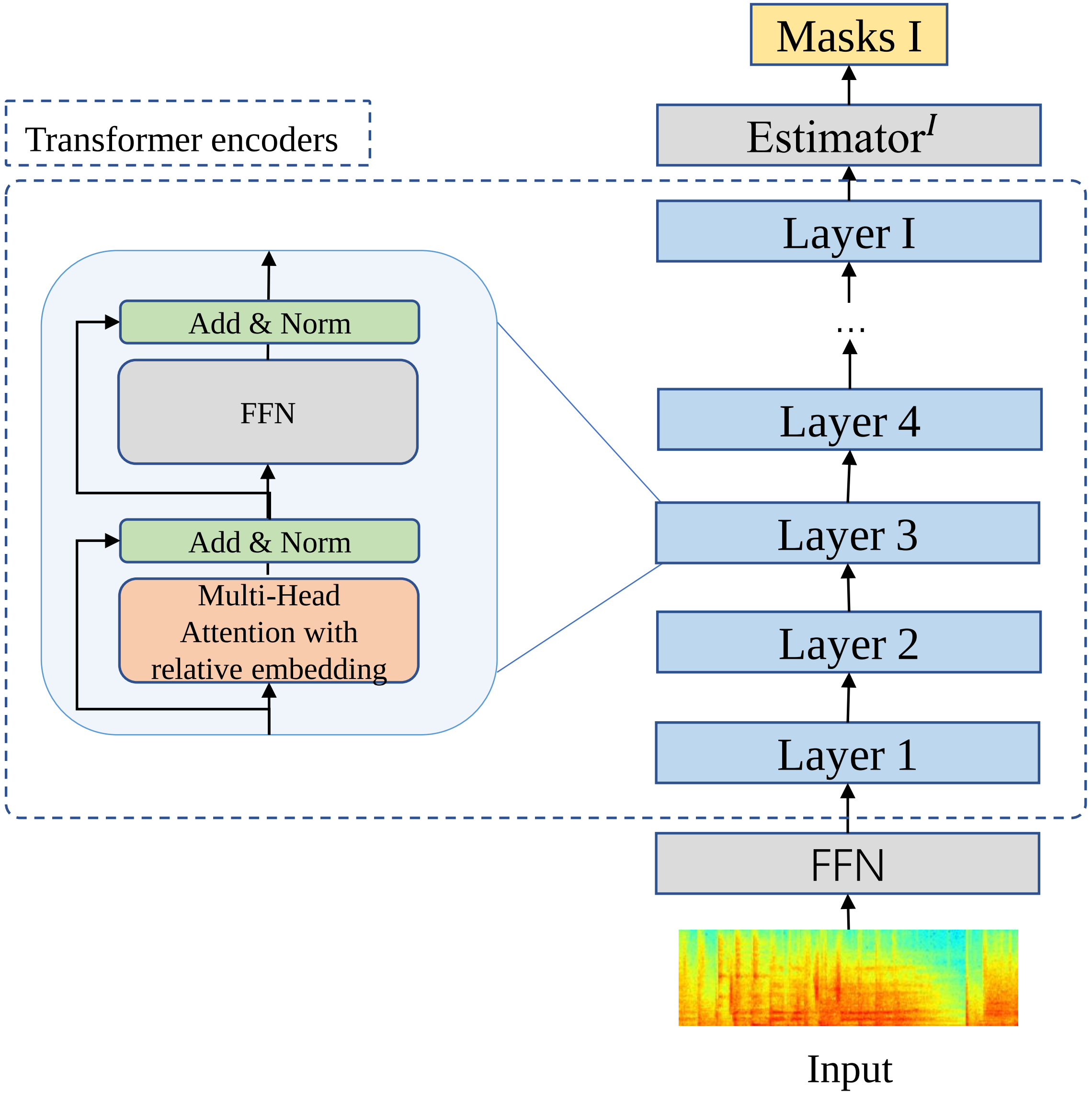
- Transformer block:

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

- Multi-head Self-attention

$$\text{Multihead}(\mathbf{h_{i-1}}) = [\mathbf{H}_1 \dots \mathbf{H}_{d_{head}}]\mathbf{W}^{head}$$
$$\text{where } \mathbf{H}_j = \text{softmax}\left(\frac{\mathbf{Q}_j(\mathbf{K}_j + \mathbf{pos})^\intercal}{\sqrt{d_k}}\right)\mathbf{V}_j$$

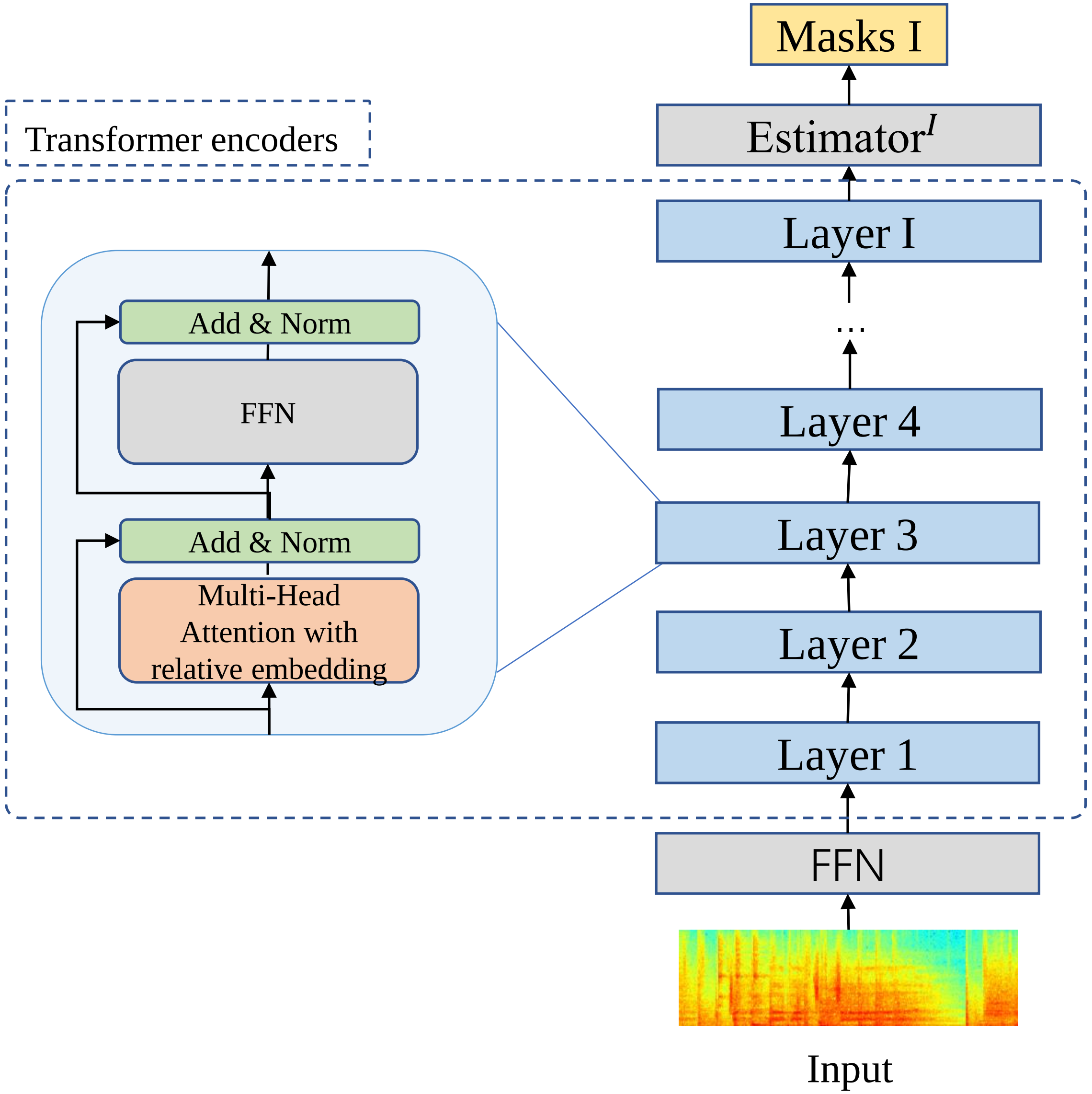
Relative position embedding



Transformer model

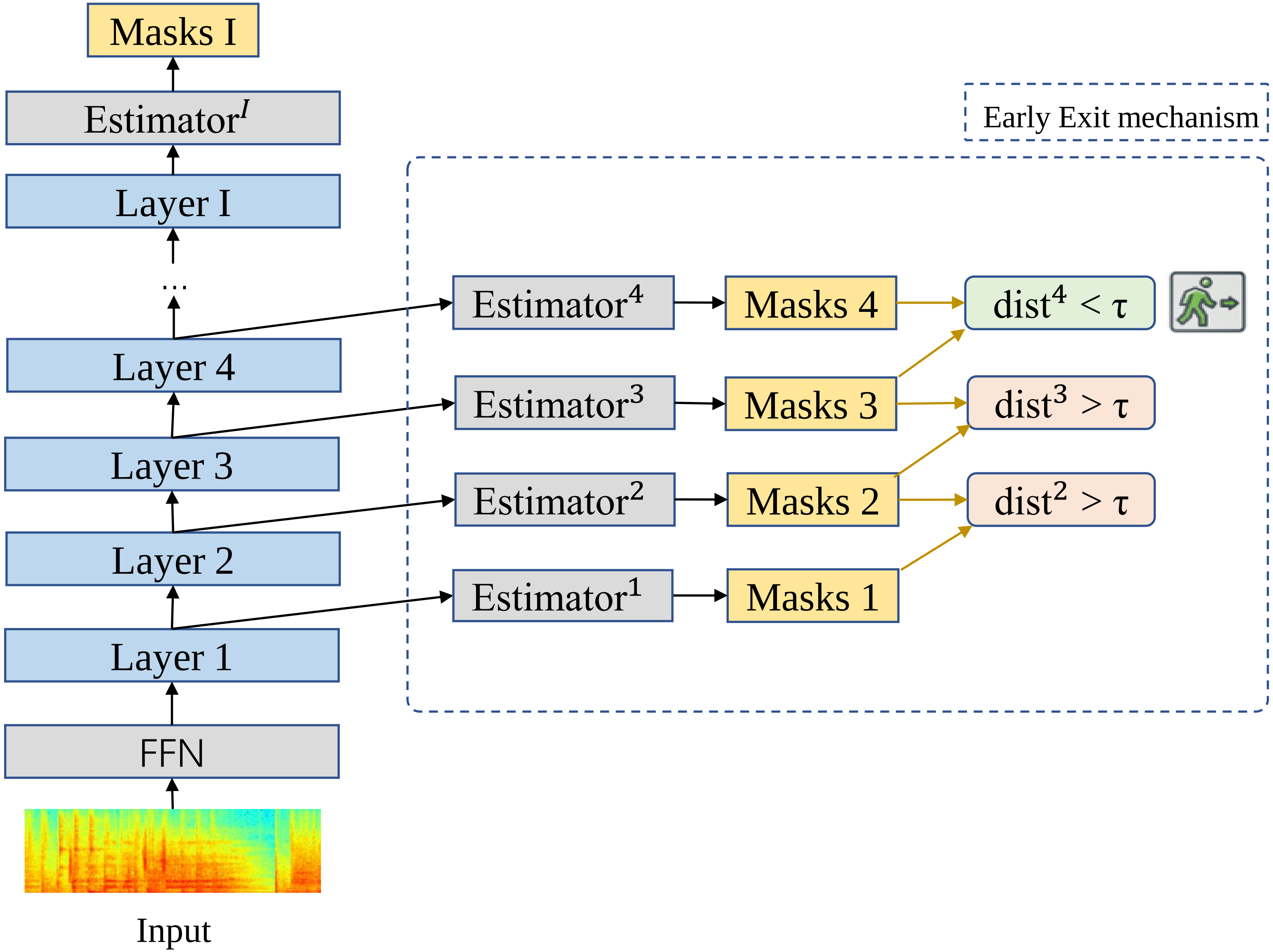
- Prior work shows that a **deeper structure** (12 or more) yields superior performance.
- **Problems**:
  - **Heavy run-time cost**
  - **“overthinking” problem:**

    a shallow Transformer is sufficient to handle the non-overlapped speech well and that a deep Transformer could potentially degrade the speech estimation.
- **Early Exit mechanism**:
  - makes predictions at an earlier layer for less overlapped speech while using higher layers for speech with a high overlap rate



Early Exit Transformer model

- **Early Exit mechanism**:
  - makes predictions at an earlier layer for less overlapped speech while using higher layers for speech with a high overlap rate
  - attach a mask estimator to each transformer layer.
  - dynamically stop the inference if the predictions from two consecutive layers are sufficiently similar.



Early Exit Transformer model

- During inference:
  - we calculate the normalized Euclidean Distance dist^i between the estimated masks of the $(\mathrm{i}-1)-\mathrm{th}$ layer and the $\mathrm{i}-\mathrm{th}$ 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 $\mathrm{i}-\mathrm{th}$ layer as the final prediction masks.
- During 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 $\mathrm{i}-\mathrm{th}$ layer.
  - The final loss is the weighted average function: $$\mathrm{Loss} = \frac{\sum_{i=1}^{I} i \cdot \mathrm{Loss}^i}{\sum_{i=1}^{I} i}$$

Experiments on LibriCSS dataset

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

Experiments on LibriCSS dataset

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

Experiments on LibriCSS

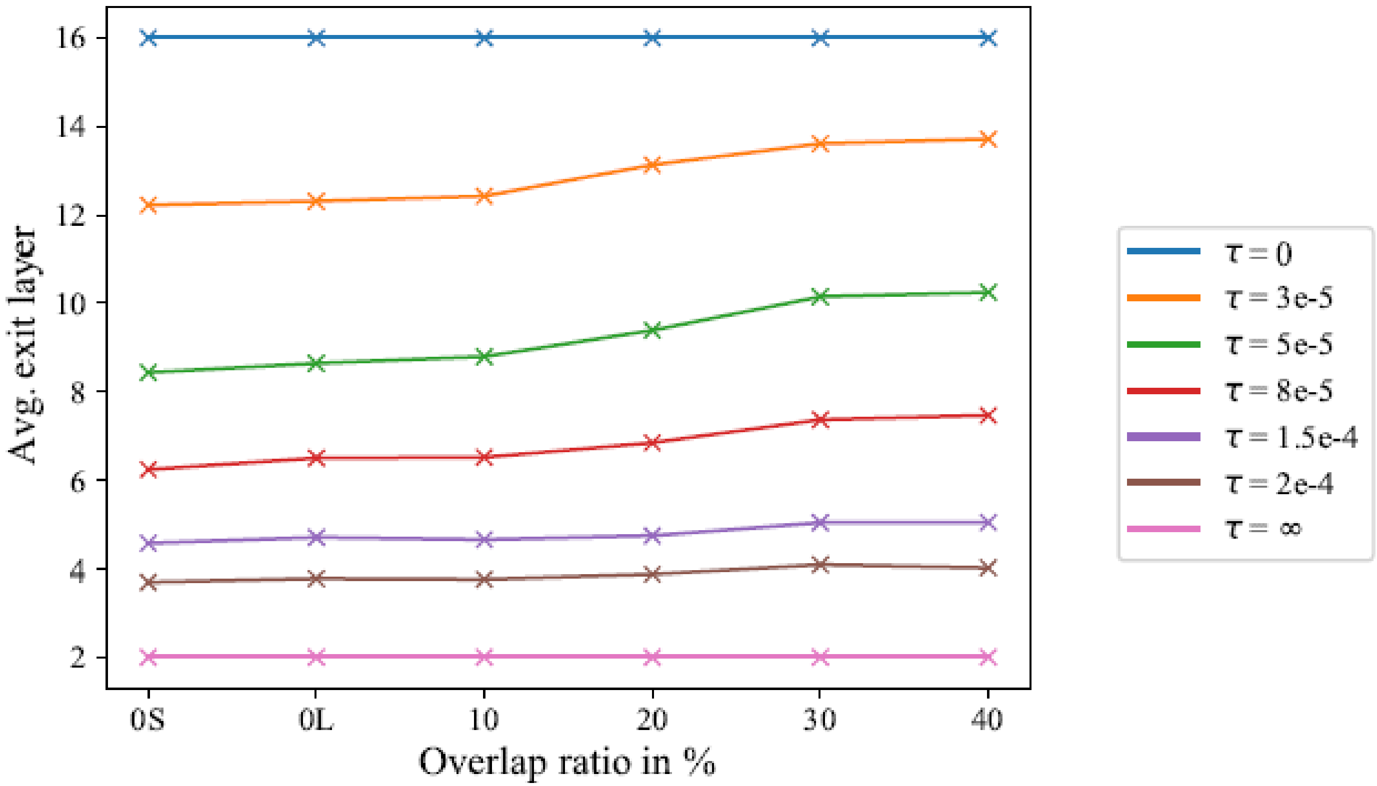


Fig. 2: The average exit layer of Early Exit Transformer across different testsets with different threshold τ for the utterance-wise evaluation.

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 observe negative results since the task is too challenging to handle.