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: \( Y^1(t, f) \) → \( X_s(t, f) \)

• Speech Separation Process:

  1. \( Y(t, f) = Y^1(t, f) \oplus \text{IPD}(2) \ldots \oplus \text{IPD}(C) \) → \( M_s(t, f) \)

  2. \( X_s(t, f) = M_s(t, f) \odot Y^1(t, f) \)
Transformer model

- Transformer block:
  \[ h_i' = \text{layernorm}(h_{i-1} + \text{MultiHeadAttention}(h_{i-1})) \]
  \[ h_i = \text{layernorm}(h_i' + \text{FFN}(h_i')) , \]

- Multi-head Self-attention
  \[ \text{Multihead}(h_{i-1}) = [H_1 \ldots H_{d_{\text{head}}}]{W_{\text{head}}} \]
  where \[ H_j = \text{softmax} \left( \frac{Q_j(K_j + \text{pos})^T}{\sqrt{d_k}} \right) V_j \]
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 $\text{dist}^i$ between the estimated masks of the $(i-1)$-th layer and the $i$-th layer.
  • Given a pre-defined threshold $\tau$, if $\text{dist}^i < \tau$ for the two consecutive layers, we terminate the inference process and output the estimated masks of $i$-th layer as the final prediction masks.

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

$$\text{Loss} = \frac{\sum_{i=1}^{I} i \cdot \text{Loss}^i}{\sum_{i=1}^{I} i}$$
Experiments on LibriCSS dataset

<table>
<thead>
<tr>
<th>System</th>
<th>Avg. exit layer</th>
<th>Speed-up</th>
<th>0S</th>
<th>0L</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>No separation [18]</td>
<td>-</td>
<td>-</td>
<td>11.8/5.5</td>
<td>11.7/5.2</td>
<td>18.8/11.4</td>
<td>27.2/18.8</td>
<td>35.6/27.7</td>
<td>43.3/36.6</td>
</tr>
<tr>
<td>BLSTM [13]</td>
<td>-</td>
<td>-</td>
<td>7.0/3.1</td>
<td>7.5/3.3</td>
<td>10.8/4.3</td>
<td>13.4/5.6</td>
<td>16.5/7.5</td>
<td>18.8/8.9</td>
</tr>
<tr>
<td>Transformer [13]</td>
<td>16.0</td>
<td>1.00×</td>
<td>8.3/3.4</td>
<td>8.4/3.4</td>
<td>11.4/4.1</td>
<td>12.5/4.8</td>
<td>14.7/6.4</td>
<td>16.9/7.2</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = 0)</td>
<td>16.0</td>
<td>0.92×</td>
<td>8.9/3.4</td>
<td>9.4/3.6</td>
<td>12.3/4.2</td>
<td>14.7/5.0</td>
<td>15.1/6.2</td>
<td>16.5/6.6</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = 8e − 5)</td>
<td>6.9</td>
<td>2.60×</td>
<td>7.6/3.2</td>
<td>7.7/3.3</td>
<td>10.1/3.8</td>
<td>12.4/4.8</td>
<td>14.4/6.2</td>
<td>16.4/6.9</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = 1.5e − 4)</td>
<td>4.8</td>
<td>4.08×</td>
<td>7.8/3.2</td>
<td>7.6/3.4</td>
<td>9.8/3.8</td>
<td>12.2/5.1</td>
<td>14.7/6.7</td>
<td>17.9/7.8</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = ∞)</td>
<td>2.0</td>
<td>6.59×</td>
<td>7.1/3.1</td>
<td>7.3/3.3</td>
<td>10.0/4.4</td>
<td>13.6/6.1</td>
<td>17.0/8.4</td>
<td>20.5/10.4</td>
</tr>
</tbody>
</table>

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.
Experiments on LibriCSS dataset

<table>
<thead>
<tr>
<th>System</th>
<th>Avg. exit layer</th>
<th>Speed-up</th>
<th>0S</th>
<th>0L</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>No separation [18]</td>
<td>-</td>
<td>-</td>
<td>15.4/12.7</td>
<td>11.5/5.7</td>
<td>21.7/17.6</td>
<td>27.0/24.4</td>
<td>34.3/30.9</td>
<td>40.5/37.5</td>
</tr>
<tr>
<td>BLSTM [13]</td>
<td>-</td>
<td>-</td>
<td>11.4/6.0</td>
<td>8.4/4.1</td>
<td>13.1/7.0</td>
<td>14.9/7.9</td>
<td>18.7/11.5</td>
<td>20.5/12.3</td>
</tr>
<tr>
<td>Transformer [13]</td>
<td>16.0</td>
<td>1.00×</td>
<td>12.0/5.6</td>
<td>9.1/4.4</td>
<td>13.4/6.2</td>
<td>14.4/6.8</td>
<td>18.5/9.7</td>
<td>19.9/10.3</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = 0)</td>
<td>16.0</td>
<td>0.76×</td>
<td>14.1/6.2</td>
<td>10.3/4.6</td>
<td>17.2/7.1</td>
<td>17.3/7.5</td>
<td>23.0/10.8</td>
<td>23.5/12.0</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = 1e − 4)</td>
<td>7.5</td>
<td>1.47×</td>
<td>11.3/5.4</td>
<td>8.9/4.4</td>
<td>12.7/6.0</td>
<td>13.8/6.7</td>
<td>17.8/9.3</td>
<td>19.7/10.5</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = 1.5e − 4)</td>
<td>5.8</td>
<td>1.88×</td>
<td>11.5/5.2</td>
<td>8.9/4.3</td>
<td>12.6/6.0</td>
<td>13.7/6.9</td>
<td>17.6/9.5</td>
<td>19.6/10.3</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = 2e − 4)</td>
<td>5.2</td>
<td>2.08×</td>
<td>11.2/5.6</td>
<td>8.8/4.5</td>
<td>12.7/6.3</td>
<td>13.9/7.2</td>
<td>18.5/9.5</td>
<td>19.6/10.9</td>
</tr>
<tr>
<td>Early Exit Transformer (τ = ∞)</td>
<td>2.0</td>
<td>4.74×</td>
<td>14.7/14.6</td>
<td>8.7/6.9</td>
<td>16.1/13.7</td>
<td>17.8/15.2</td>
<td>22.5/18.2</td>
<td>24.8/18.9</td>
</tr>
</tbody>
</table>

Table 2: Continuous speech separation evaluation
Experiments on LibriCSS

![Chart showing the average exit layer of Early Exit Transformer across different test sets with different threshold $\tau$ for the utterance-wise evaluation.](chart)

**Fig. 2:** The average exit layer of Early Exit Transformer across different test sets with different threshold $\tau$ 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.