ENHANCING END-TO-END MULTI-CHANNEL SPEECH SEPARATION VIA SPATIAL FEATURE LEARNING

Rongzhi Gu¹, Shi-Xiong Zhang², Lianwu Chen³, Yong Xu², 
Meng Yu², Dan Su³, Yuexian Zou¹,⁴, Dong Yu²

¹ ADSPLAB, School of ECE, Peking University, Shenzhen, China
² Tencent AI Lab, Bellevue, WA, USA
³ Tencent AI Lab, Shenzhen, China
⁴ Peng Cheng Laboratory, Shenzhen, China

Author: Rongzhi Gu
Outline

- Introduction
- Proposed method
- Experiments and results
- Conclusion
Speech separation

- **Cocktail Party problem** [Cherry 1953]
  - recover the speech of each speaker from overlapped speech mixture

\[ y[n] = \sum_{s=1}^{S} x_s[n] \]

- \( y[n] = [y^1[n], ..., y^C[n]]^T \): observed multi-channel mixture signal
- \( x_s[n] = [x_{s1}[n], ..., x_{sc}[n]]^T \): reverberant image for source \( s \)
Methods for multi-channel speech separation

- **T-F masking**
  - formulate speech separation as a supervised learning task in frequency domain

- **Integration of T-F masking and beamforming**

- **End-to-end approaches**
Single-channel TasNet

- **Encoder-decoder structure**

- **Unsatisfactory performance under far-field scenario**
Our previous try – Data mismatch

- A multi-channel speech separation approach:
  - encode + IPDs
  - a multi-channel speech separation approach

- Data mismatch: encoder & STFT
  - **Encoder**: learned in purely data-driven fashion
  - **STFT**: fixed complex filters (with evenly distributed frequency responses)
**Proposed model**

- **Aim:** an end-to-end multi-channel speech separation model
  - **Encoder:** transform mixture waveform into the mixture encode
  - **Conv2d:** spatial feature learning
  - **Separator:** estimate a mask in encoder output domain for each speaker
  - **Decoder:** reconstruct the separated speech waveform
**Spatial feature learning** – multi-channel convolution sum

- **Main idea:**
  - learn time-domain filters spanning all signal channels \( \mathbf{K} = \{ \mathbf{k}^{(n)} \} \in \mathbb{R}^{C \times L \times N} \)
  to perform adaptive spatial filtering

- **Multi-channel convolution sum (MCS):**
  \[
  \text{MCS}^{(n)} = \sum_{c=1}^{C} y_c \otimes k_c^{(n)}
  \]
  - a DAS beamformer-like formation
  - Each set of filters \( k^{(n)} \) is expected to steer at a different direction

- **Implementation:**
  - Conv2d, kernel: \( N@C \times L \)
Spatial feature learning – inter-channel convolution difference

- **Inspired by IPD formulation:**
  - T-F bins that dominated by the same source share the same time delay
  - IPDs of these T-F bins naturally form a cluster within each frequency band

- **Inter-channel convolution difference (ICD)**
  \[
  ICD_{m}^{(n)} = \sum_{c=1}^{2} w_c \cdot \left( y_{m_c} \star k'^{(n)} \right)
  \]
  - \(m\): microphone pair index
  - \(w_c \in \mathbb{R}^{1 \times L}\): window function
  - Initialization: \(w_1=[1, \ldots, 1]\), \(w_2=[-1, \ldots, -1]\)

- **Implementation:**
  - A customized conv2d kernel
Visualization of learned filters

Encoder

Conv2d kernel $\mathbf{K}$
Experiments - Data

- **spatialized reverberant WSJ0 2-mix**

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td>20000 utterances / 30h / 101 speakers</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td>8000 utterances / 8h / 101 speakers</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>5000 utterances / 5h / 18 speakers (unseen)</td>
</tr>
<tr>
<td><strong>Sampling rate</strong></td>
<td>16kHz</td>
</tr>
<tr>
<td><strong>Mix SNR</strong></td>
<td>[-5, 5] dB</td>
</tr>
<tr>
<td><strong>Overlap Ratio</strong></td>
<td>100%</td>
</tr>
<tr>
<td><strong>RT60</strong></td>
<td>0.05~0.5s</td>
</tr>
<tr>
<td><strong>Microphone array</strong></td>
<td>6-microphone circular array of 7cm diameter</td>
</tr>
<tr>
<td><strong>Included angle</strong></td>
<td>0-15: 16%, 15-45: 29%, 45-90: 26%, 90-180: 29%</td>
</tr>
<tr>
<td><strong>Selected pairs for IPD/ICD</strong></td>
<td>[1, 4], [2, 5], [3, 6], [1, 2], [3, 4], [5, 6]</td>
</tr>
</tbody>
</table>
Experiments

• **Training objective:** SI-SDR

\[
\begin{align*}
\hat{x}_{\text{target}} & := \frac{\langle \hat{x}, x \rangle x}{\|x\|_2^2} \\
\epsilon_{\text{noise}} & := \hat{x} - \hat{x}_{\text{target}} \\
\text{SI-SDR} & := 10 \log_{10} \frac{\|\hat{x}_{\text{target}}\|_2^2}{\|\epsilon_{\text{noise}}\|_2^2}
\end{align*}
\]

  - \(\hat{x}\): estimated reverberant speech
  - \(x\): ground truth reverberant speech

• **Permutation invariant training**

• **Evaluation Metrics**

  - SI-SDR improvement
  - SDR improvement
## Results – different conv2d configurations

<table>
<thead>
<tr>
<th>Setup</th>
<th>window $w$</th>
<th># filters $N$</th>
<th>SI-SDR improvement (dB)</th>
<th>SDRi (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;15</td>
<td>15-45</td>
</tr>
<tr>
<td>Single-channel Conv-TasNet</td>
<td>-</td>
<td>-</td>
<td>8.5</td>
<td>9.0</td>
</tr>
<tr>
<td>+ MCS (conv2d (6×40))</td>
<td>-</td>
<td>256</td>
<td>5.7</td>
<td>10.3</td>
</tr>
<tr>
<td>+ ICD (conv2d (2×40))</td>
<td>fix -1</td>
<td>256</td>
<td>5.5</td>
<td>10.9</td>
</tr>
<tr>
<td>+ ICD (conv2d (2×40))</td>
<td>init. -1</td>
<td>256</td>
<td>6.2</td>
<td>11.2</td>
</tr>
<tr>
<td>+ ICD (conv2d (2×40))</td>
<td>init. randomly</td>
<td>33</td>
<td>8.2</td>
<td>8.1</td>
</tr>
<tr>
<td>+ ICD (conv2d (2×40))</td>
<td>fix -1</td>
<td>33</td>
<td>6.9</td>
<td>11.1</td>
</tr>
<tr>
<td>+ ICD (conv2d (2×40))</td>
<td>init. -1</td>
<td>33</td>
<td>6.7</td>
<td>11.7</td>
</tr>
</tbody>
</table>

- The performances with 33 filters are relatively superior to those with 256 filters.
- The value of $w$ contributes significantly to the separation performance
  - Init randomly: no subtraction operation
  - Fix -1: exact inter-channel convolution difference
  - Init -1: initialize $w$ as -1 and set $w$ learnable during training
## Results – IPD vs ICD

<table>
<thead>
<tr>
<th>Setup</th>
<th>SI-SDR improvement (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;15°</td>
</tr>
<tr>
<td>cosIPD, sinIPD</td>
<td>7.7</td>
</tr>
<tr>
<td>cosIPD, sinIPD (trainable kernel)</td>
<td>7.9</td>
</tr>
<tr>
<td>ICD</td>
<td>6.7</td>
</tr>
<tr>
<td>ICD, cosIPD, sinIPD</td>
<td>8.1</td>
</tr>
</tbody>
</table>

- IPD provide beneficial spatial information of sources and help the multi-channel speech separation.
- With the trainable kernel, the performance improves slightly.
- ICD based separation model obtains 0.4dB improvement over cosIPD+sinIPD based
- The incorporation of ICDs and IPDs achieves further 0.5dB improvement.
Conclusions

- This work proposes an end-to-end multi-channel speech separation model
  - learn effective spatial cues directly from the multi-channel speech waveforms
  - end-to-end optimization in a purely data-driven fashion

- The learned spatial features
  - can be computed with few parameters and computation cost
  - can be combined with well-designed IPDs and obtain better results

Thank you!
1701111335@pku.edu.cn
Thank you!
Q&A