Single Channel Speech Separation with Constrained Utterance Level Permutation Invariant Training Using Grid LSTM

Chenglin Xu$^{1,2}$, Wei Rao$^2$, Xiong Xiao$^3$, Eng Siong Chng$^{1,2}$ and Haizhou Li$^{2,4}$

$^1$ School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
$^2$ Temasek Laboratories@NTU, Nanyang Technological University, Singapore
$^3$ Microsoft Corporation, United States
$^4$ Department of Electrical and Computer Engineering, National University of Singapore, Singapore

17 April, 2018
Outline

1. Introduction
2. Methodology
3. Evaluation
4. Summary
Introduction

Single channel speech separation with uPIT

• The performance of single channel speech separation has been significantly improved by deep learning based techniques, such as, deep clustering (DC) [1], deep attractor network (DANet) [2], utterance-level permutation invariant training (uPIT) [3], and so on.

• However, the state-of-the-art uPIT method runs into a frame leakage problem. (Frame leakage: Frames or time-frequency bins of speaker A are wrongly aligned to the output stream of speaker B, as shown in the red box of the figure.)


Introduction

Single channel speech separation with uPIT

- The uPIT baseline framework from [1]

• **Constrain the objective using dynamic information**
  The dynamic information, e.g., the delta and acceleration, are used in the objective function to make the separation continuous across frames by using contextual information of several frames.

• **Capture temporal and spectral patterns simultaneously**
  Inspired by CASA method using heuristic rules, the grid LSTM is used to capture the heuristic patterns, e.g., common onset/offset, and learn corresponding temporal and spectral patterns from the magnitude spectrum both in time and frequency domain simultaneously.
Methodology

cuPIT-Grid LSTM system

Input mixture ($S=2$ speakers):
- constrained pairwise scores: $O(3S^2)$
- permutation error: $O(2S^2 + S!)$
- error: (summation)
The objective function in uPIT baseline:

\[
J_{c,\phi_p}(s) = \frac{1}{T} \sum_{s=1}^{S} \left( \| \hat{M}_s \odot |Y| - |X_{\phi_p(s)}| \odot \cos(\theta_y - \theta_{\phi_p(s)}) \|_F^2 \right)
\]

\[
\hat{p} = \arg \min_{p \in P} J_{c,\phi_p}(s)
\]

\[
J = J_{c,\phi_{\hat{p}}(s)}
\]

The proposed constrained objective function (cuPIT):

\[
J_{c,\phi_p}(s) = \frac{1}{T} \sum_{s=1}^{S} \left( \| \hat{M}_s \odot |Y| - |X_{\phi_p(s)}| \odot \cos(\theta_y - \theta_{\phi_p(s)}) \|_F^2 \right)
+ w_D \| f_D(\hat{M}_s \odot |Y|) - f_D(|X_{\phi_p(s)}| \odot \cos(\theta_y - \theta_{\phi_p(s)})) \|_F^2
+ w_A \| f_A(\hat{M}_s \odot |Y|) - f_A(|X_{\phi_p(s)}| \odot \cos(\theta_y - \theta_{\phi_p(s)})) \|_F^2
\]

\[
f_D(v(t)) = \frac{\sum_{l=1}^{L} l \times (v(t+l) - v(t-l))}{\sum_{l=1}^{L} 2l^2}
\]

\[
\hat{p} = \arg \min_{p \in P} J_{c,\phi_p(s)}
\]

\[
J = J_{c,\phi_{\hat{p}}(s)}
\]
• **Dataset**
  The WSJ0-2mix database* with the sampling rate at 8 kHz.
  - Training set: 20,000 utterances \( \approx 30 \)h
  - Development set: 5,000 utterances \( \approx 8 \)h
  - Test set: 3,000 utterances \( \approx 5 \)h

• **Features**
  129-dimensional spectral magnitude features computed by a STFT with a normalized square root of 32ms length hamming window and 16ms window shift.

• **Evaluation Metrics**
  - The global normalized signal-to-distortion ratio (GNSDR, same as “SDR improvement” in DC, DANet, uPIT baselines) using bss_eval toolbox [1].
  - Signal-to-interferences ratio (SIR).
  - Signal-to-artifacts ratio (SAR).

---

* Available at: http://www.merl.com/demos/deep-clustering
Evaluation

Experiment 1: same network architecture as baseline

- Constrained uPIT (cuPIT) vs. baseline uPIT

Evaluation results:

<table>
<thead>
<tr>
<th></th>
<th>'uPIT-BLSTM'</th>
<th>'cuPIT-BLSTM'</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNSDR (dB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Gender</td>
<td>9.53</td>
<td>11.46</td>
</tr>
<tr>
<td>Different Gender</td>
<td>9.84</td>
<td>11.71</td>
</tr>
<tr>
<td>Same Gender</td>
<td>7.31</td>
<td>7.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>'uPIT-BLSTM'</th>
<th>'cuPIT-BLSTM'</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIR (dB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Gender</td>
<td>16.52</td>
<td>19.06</td>
</tr>
<tr>
<td>Different Gender</td>
<td>17.07</td>
<td>19.50</td>
</tr>
<tr>
<td>Same Gender</td>
<td>13.61</td>
<td>14.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>'uPIT-BLSTM'</th>
<th>'cuPIT-BLSTM'</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR (dB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Gender</td>
<td>11.16</td>
<td>12.64</td>
</tr>
<tr>
<td>Different Gender</td>
<td>11.37</td>
<td>12.84</td>
</tr>
<tr>
<td>Same Gender</td>
<td>9.47</td>
<td>9.69</td>
</tr>
</tbody>
</table>

Chenglin Xu et al. ICASSP 2018
Evaluation

Experiment 1: same network architecture as baseline

- Constrained uPIT vs. baseline uPIT

Paired t-test on test set of SDR result: statistically significant
Evaluation

Experiment 2:

Grid LSTM with cuPIT objective

<table>
<thead>
<tr>
<th>GNSDR (dB)</th>
<th>'uPIT-BLSTM'</th>
<th>'cuPIT-BLSTM'</th>
<th>'cuPIT-Grid'</th>
<th>'cuPIT-Grid-RD'</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.53</td>
<td>7.31</td>
<td>9.84</td>
<td>10.11</td>
<td>11.96</td>
</tr>
<tr>
<td>11.46</td>
<td>7.70</td>
<td>11.71</td>
<td>11.85</td>
<td>10.21</td>
</tr>
<tr>
<td>12.64</td>
<td>8.12</td>
<td>12.84</td>
<td>12.96</td>
<td>8.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SIR (dB)</th>
<th>'uPIT-BLSTM'</th>
<th>'cuPIT-BLSTM'</th>
<th>'cuPIT-Grid'</th>
<th>'cuPIT-Grid-RD'</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.52</td>
<td>13.61</td>
<td>17.07</td>
<td>17.41</td>
<td>19.74</td>
</tr>
<tr>
<td>19.06</td>
<td>14.28</td>
<td>19.50</td>
<td>19.68</td>
<td>14.84</td>
</tr>
<tr>
<td>19.06</td>
<td>14.81</td>
<td>19.68</td>
<td>19.74</td>
<td>14.84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SAR (dB)</th>
<th>'uPIT-BLSTM'</th>
<th>'cuPIT-BLSTM'</th>
<th>'cuPIT-Grid'</th>
<th>'cuPIT-Grid-RD'</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.16</td>
<td>9.47</td>
<td>11.37</td>
<td>11.58</td>
<td>11.69</td>
</tr>
<tr>
<td>12.64</td>
<td>9.69</td>
<td>12.84</td>
<td>12.96</td>
<td>10.00</td>
</tr>
<tr>
<td>12.64</td>
<td>10.00</td>
<td>13.07</td>
<td>11.69</td>
<td>10.10</td>
</tr>
</tbody>
</table>

For All Gender, Different Gender, and Same Gender.
Evaluation

Experiment 2: Grid LSTM with cuPIT objective

Paired t-test on test set of SDR result: statistically significant.
Comparisons with state-of-the-art methods

**DC** \(^{[1]}\): The mixture is projected into an embedding space, where time-frequency bins belonging to the same speaker are grouped into a cluster using k-means to form a binary mask used to separate the speakers from the mixture signal.

**DC+** \(^{[2]}\): The cluster stage is connected with the embedding learning network to do end-to-end mask estimation.

**DANet** \(^{[3]}\): Attractor points, which attract the time-frequency bins corresponding to each target speaker, are created in the embedding space. The network is trained in end-to-end to estimate the masks, which are used to separate the mixture signal.

**PIT-DNN** \(^{[4]}\): The magnitude approximation masks are estimated in end-to-end by using a permutation invariant training with context expansion in inputs and calculating the cost using DNN.

**PIT-CNN** \(^{[4]}\): The magnitude approximation masks are estimated in end-to-end by using a permutation invariant training using CNN.

**uPIT-BLSTM** \(^{[5]}\): The magnitude approximation masks are estimated in end-to-end by using an utterance level permutation invariant training to solve the label ambiguity problem in training and inference stage.

---


## Evaluation

### Comparative Study

- **Comparisons with state-of-the-art methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Opt Assign (GNSDR, dB)</th>
<th>Def Assign (GNSDR, dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev Set</td>
<td>Test Set</td>
</tr>
<tr>
<td>DC [1]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DC+ [2]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DANet [3]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PIT-DNN [4]</td>
<td>7.3</td>
<td>7.2</td>
</tr>
<tr>
<td>PIT-CNN [4]</td>
<td>8.4</td>
<td>8.6</td>
</tr>
<tr>
<td>uPIT-BLSTM [5]</td>
<td>10.9</td>
<td>10.8</td>
</tr>
<tr>
<td>uPIT-BLSTM*</td>
<td>10.8</td>
<td>10.7</td>
</tr>
<tr>
<td>cuPIT-BLSTM</td>
<td>11.1</td>
<td>11.0</td>
</tr>
<tr>
<td>cuPIT-Grid</td>
<td>11.2</td>
<td>11.2</td>
</tr>
<tr>
<td>cuPIT-Grid-RD</td>
<td>11.3</td>
<td>11.3</td>
</tr>
<tr>
<td>IRM</td>
<td>12.4</td>
<td>12.7</td>
</tr>
<tr>
<td>IPSM</td>
<td>14.9</td>
<td>15.1</td>
</tr>
</tbody>
</table>

**Opt Assign**: realign the output streams by using target speaker’s speech to show the upper bound without frame leakage.

**Def Assign**: default output streams from the system without realignment.

**uPIT-BLSTM***: Our reimplementation of uPIT-BLSTM baseline.

---

Example: two female speakers’ mixture (‘050a050i_2.1935_421c020b_-2.1935’)

(a) Mixed Speech
(b) Target Speaker 1
(c) Target Speaker 2

(d) Separation 1 by uPIT
(e) Separation 2 by uPIT
(f) Separation 1 by cuPIT-G-RD
(g) Separation 2 by cuPIT-G-RD

SDR: 17.2  SIR: 22.1  SAR: 18.9
SDR: 8.8  SIR: 13.6  SAR: 10.7
SDR: 21.2  SIR: 30.2  SAR: 21.7
SDR: 13.9  SIR: 24.9  SAR: 14.3
Example: male-female speakers’ mixture (‘441c020m_2.4506_447o030z_-2.4506’)

SDR: 17.2 SIR: 22.1 SAR: 18.9
SDR: 8.8 SIR: 13.6 SAR: 10.7
SDR: 21.2 SIR: 30.2 SAR: 21.7
SDR: 13.9 SIR: 24.9 SAR: 14.3
Summary

• We propose a constrained cost function in uPIT by using dynamic information to solve the frame leakage problem.

• We further propose to use a Grid LSTM to learn temporal and spectral patterns from the time and frequency domain of the mixture signal simultaneously.

• The proposed method achieves better results than the current state-of-the-art uPIT method.
Thank you!