BEAMNET: End-to-End Training of a Beamformer-Supported Multi-Channel ASR System

Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, Reinhold Haeb-Umbach, University of Paderborn, Germany

{heymann, drude, haeb}@nt.uni-paderborn.de, http://www-nt.uni-paderborn.de

System overview

Highlights
- Complex valued backpropagation through statistical beamforming operation
  - Eliminates the need for parallel clean and noisy data and can be trained on real data only
  - Does not need heuristic masks as targets
  - Whole system works in the same STFT domain, no transformation between front-end and back-end necessary
- Agnostic to the array geometry
- Number of microphones can differ at test time

Backpropagation through GEV
- Most crucial step is to find the gradient for the PSD matrices given the gradient for the beamforming vector
- Generalized Eigenvalue problem:

\[
\Phi_{XX} \mathbf{w} = \Phi_{NN} \mathbf{w} \Lambda
\]

- Beamforming vector is the eigenvector corresponding to the largest Eigenvalue
- Transform Generalized Eigenvalue problem to standard Eigenvalue problem

\[
\Phi \mathbf{w} = \mathbf{w} \Lambda \quad \text{with} \quad \Phi = \Phi_{NN}^{-1} \Phi_{XX}
\]

- Normalize the eigenvector to unit norm
- Gradient can now be calculated as

\[
\frac{\partial J}{\partial \Phi} \approx \mathbf{W}^H \left( \frac{\partial J}{\partial \Lambda} + F^* \circ \left( \mathbf{W}^H \frac{\partial J}{\partial \mathbf{W}} \right) \right) \mathbf{W}^H - \mathbf{W}^H \left( F^* \circ \mathbf{W} \mathbf{W} \left( R_c \left( \mathbf{W}^H \frac{\partial J}{\partial \mathbf{W}} \right) \circ \mathbf{I} \right) \right) \mathbf{W}^H
\]

with \( F_i = (\lambda_j - \lambda_i)^{-1} \delta_{ij} \) and \( F_i = 0 \)

Acoustic model
- Hybrid approach, model estimates state posteriors
- Trained on whole utterances
- Based on Wide Residual Networks
  - Each block consists of two 2D convolutional layer and a residual connection
  - Number of channels increases with each block: 16 \( \rightarrow \) 80 \( \rightarrow \) 160 \( \rightarrow \) 320
  - Normalization across time before each non-linearity

Results
- System was evaluated on the CHiME 4 dataset with different pre-training configurations:
  - Fixed: Model pre-trained separately. The parameters are kept fixed during joint training.
  - Scratch: Model parameters initialized randomly and the trained jointly.
  - Finetune: Model pre-trained separately. The parameters can be adjusted during joint training.

<table>
<thead>
<tr>
<th>Training</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF</td>
<td>AM</td>
<td>real</td>
</tr>
<tr>
<td>BFIT+Kaldi</td>
<td>5.76</td>
<td>6.77</td>
</tr>
<tr>
<td>BFIT+WRN</td>
<td>5.53</td>
<td>6.67</td>
</tr>
<tr>
<td>fixed</td>
<td>4.26</td>
<td>4.29</td>
</tr>
<tr>
<td>scratch</td>
<td>5.51</td>
<td>5.19</td>
</tr>
<tr>
<td>scratch</td>
<td>4.14</td>
<td>4.09</td>
</tr>
<tr>
<td>fixed</td>
<td>4.09</td>
<td>3.96</td>
</tr>
<tr>
<td>finetune</td>
<td>3.77</td>
<td>3.89</td>
</tr>
</tbody>
</table>