Neural Network based Spectral Mask Estimation for Acoustic Beamforming

Jahn Heymann
Lukas Drude
Reinhold Haeb-Umbach
Multi channel processing with neural networks

MOTIVATION
Motivation

• Single-channel:
  • Neural networks rendered many feature enhancement techniques superfluous

• Multi-channel:
  • Stack channels (features)
  • Work on raw waveforms

• Our approach: Combine neural network with a traditional beamformer
GEV & MVDR

ACOUSTIC BEAMFORMING
Acoustic beamforming

- **MVDR**
  - Minimize noise
  - Source distortionless

- **GEV**
  - Maximize SNR
  - Introduces distortions

\[
\text{argmin}_{F} \mathbf{F}^H \Phi_{NN} \mathbf{F} \quad \text{s.t.} \quad \mathbf{F}^H \mathbf{d} = 1.
\]

\[
\mathbf{d} = \mathcal{P} \{ \Phi_{xx} \}
\]

\[
\mathbf{F}_{MVDR} = \frac{\Phi_{NN}^{-1} \mathcal{P} \{ \Phi_{xx} \} \mathcal{P} \{ \Phi_{xx} \}^H}{\mathcal{P} \{ \Phi_{xx} \}^H \Phi_{NN}^{-1} \Phi_{xx} \mathcal{P} \{ \Phi_{xx} \}}
\]

\[
\mathbf{F}^H \Phi_{xx} \mathbf{F} = \lambda \Phi_{NN} \mathbf{F}
\]

Jahn Heymann - Neural Network based Spectral Mask Estimation for Acoustic Beamforming
Acoustic beamforming

- Both beamformers depend only on signal statistics
  - Cross-Power Spectral Density of speech and noise
  - Independent of microphone array
  - No assumption on acoustic transfer function
- We estimate PSD matrices using masks

\[
\Phi_{\nu\nu} = \frac{1}{T} \sum_{t=1}^{T} M_\nu(t) \mathbf{Y}(t) \mathbf{Y}(t)^H \quad \text{where} \quad \nu \in \{X, N\}
\]

- This allows us to incorporate a neural network
Neural mask estimation

SYSTEM OVERVIEW
System overview

noise-aware

Neural network

Median

PSD

Beamformer

\( Y \)

\( \hat{X} \)

\( M_X \)

\( M_N \)

\( \Phi_{XX} \)

\( \Phi_{NN} \)
System overview

Neural network

Median

PSD

Beam-former

$Y$

$\cdot$

$\hat{X}$

$\hat{X}$

$M_X$

$\Phi_{XX}$

$M_N$

$\Phi_{NN}$

$\hat{X}$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

Jahn Heymann - Neural Network based Spectral Mask Estimation for Acoustic Beamforming
Network configurations and experimental setup

SETUP
## Network configurations

### BLSTM

<table>
<thead>
<tr>
<th>Layer</th>
<th>Units</th>
<th>Type</th>
<th>Non-linearity</th>
<th>dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>256</td>
<td>BLSTM</td>
<td>Tanh</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>513</td>
<td>FF</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>513</td>
<td>FF</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>513/1026</td>
<td>FF</td>
<td>Sigmoid</td>
<td>0.0</td>
</tr>
</tbody>
</table>

### FF

<table>
<thead>
<tr>
<th>Layer</th>
<th>Units</th>
<th>Type</th>
<th>Non-linearity</th>
<th>dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>513</td>
<td>FF</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>513/1026</td>
<td>FF</td>
<td>Sigmoid</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Experimental setup

- CHiME III challenge
  - 6 channels
  - 4 different real-world background noise types
- Metrics
  - PESQ / WER
- Compared to
  - Parametric source separation approaches [Tran10] & [Ito13]
  - BeamformIt! (only ASR)


MVDR vs. GEV, Speech Enhancement, Speech Recognition

RESULTS
Results

- GEV works better with our masks as it avoids the matrix inversion
Results

Jahn Heymann - Neural Network based Spectral Mask Estimation for Acoustic Beamforming
Results

Jahn Heymann - Neural Network based Spectral Mask Estimation for Acoustic Beamforming
Results

Jahn Heymann - Neural Network based Spectral Mask Estimation for Acoustic Beamforming
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>clean</th>
<th>noise-aware</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>40.17</td>
<td></td>
</tr>
<tr>
<td>BLSTM</td>
<td>22.28</td>
<td>15.42</td>
</tr>
<tr>
<td>FF</td>
<td>21.93</td>
<td>17.85</td>
</tr>
<tr>
<td>BeamformIt!</td>
<td>22.65</td>
<td></td>
</tr>
<tr>
<td>Ito13</td>
<td>27.32</td>
<td></td>
</tr>
<tr>
<td>Tran10</td>
<td>22.70</td>
<td></td>
</tr>
<tr>
<td>BeamformIt!*</td>
<td>12.79</td>
<td></td>
</tr>
<tr>
<td>BLSTM*</td>
<td>-</td>
<td>7.45</td>
</tr>
</tbody>
</table>

*new Baseline with DNN AM

---

Jahn Heymann - Neural Network based Spectral Mask Estimation for Acoustic Beamforming
CONCLUSIONS
Conclusion

• Beamformer supported by Neural Network
• Significant performance gains
• Independent of microphone array configuration
• Small & simple network possible
• Robust against mismatch conditions

Code available:
https://github.com/fgnt/nn-gev