Speaker Adapted Beamforming for Multi-Channel Automatic Speech Recognition

Tobias Menne, Ralf Schlüter, Hermann Ney

menne@cs.rwth-aachen.de
Content

• Introduction

• Mask based beamforming

• Integration into the acoustic model

• Speaker adaptation

• Results (on CHiME-4)
Motivation

- Integration of signal preprocessing into the acoustic model
- Joint optimization towards ASR objective
- Here: adjusting preprocessing to specific acoustic scenario (e.g. speaker of interest)
Previous Work

- Learning beamforming from the (raw) input signal
  [Sainath & Weiss\(^+\) 15, Sainath & Weiss\(^+\) 16, Li & Sainath\(^+\) 16]

- Estimating the filter matrix of filter and sum beamforming
  [Xiao & Watanabe\(^+\) 16, Meng & Watanabe\(^+\) 17]

- Incorporation of statistically optimal beamforming into the acoustic model
  [Heymann & Durde\(^+\) 17]
Mask based beamforming

Recorded noisy multi-channel signal:

\[ Y_{t,f} = X_{t,f} + N_{t,f} \] (0.1)

\( Y_{t,f}, X_{t,f}, N_{t,f} \in \mathbb{C}^M \)

- \( t \) - time frame index
- \( f \) - frequency bin index
- \( M \) - nr of microphones
Mask based beamforming

Recorded noisy multi-channel signal:

\[ Y_{t,f} = X_{t,f} + N_{t,f} \]  \hspace{1cm} (0.2)

\( Y_{t,f}, X_{t,f}, N_{t,f} \in \mathbb{C}^M \)

t - time frame index

f - frequency bin index

M - nr of microphones

Filter and sum beamforming:

\[ \hat{S}_{t,f} = w_f^H \cdot Y_{t,f} \]  \hspace{1cm} (0.3)

\( \hat{S}_{t,f} \in \mathbb{C} \) - estimate of the speech component

w_f \in \mathbb{C}^M - beamforming vector
**Mask based beamforming**

Beamforming vector by speech and noise mask:

$$w_f^H = g(\lambda^{(X)}_{t,f}, \lambda^{(N)}_{t,f}, Y_{t,f})$$  \hspace{1cm} (0.4)

$$\lambda^{(\nu)} = f^{(\nu)}_{\Theta}(Y) \quad \text{Mask estimated by neural network}$$
Mask based beamforming

Beamforming vector by speech and noise mask:

\[ w_f^H = g(\lambda_{t,f}^{(X)}, \lambda_{t,f}^{(N)}, \mathbf{Y}_{t,f}) \]  

(0.5)

\[ \lambda^{(\nu)} = f^{(\nu)}_\Theta(\mathbf{Y}) \] - Mask estimated by neural network

**GEV Beamformer**

\[ w_f^{(GEV)} = \mathcal{P}(\Phi_{\text{NN},f}^{-1} \Phi_{XX,f}) \]  

[Warsitz & Haeb-Umach 07]

**MVDR Beamformer**

\[ w_f^{(MVDR)} = \mathcal{P}(\Phi_{XX,f}) \]  

[Higuchi & Ito 16]
Mask based beamforming

Beamforming vector by speech and noise mask:

$$\mathbf{w}_f^H = g(\lambda^{(X)}_{t,f}, \lambda^{(N)}_{t,f}, \mathbf{Y}_{t,f})$$  \hspace{1cm} (0.6)

$$\lambda^{(\nu)} = f_{\Theta}^{(\nu)}(\mathbf{Y}) \quad \text{- Mask estimated by neural network}$$

**GEV Beamformer**

$$\mathbf{w}^{(GEV)}_f = \mathcal{P}(\Phi^{-1}_{NN,f} \Phi_{XX,f})$$

[Warsitz & Haeb-Umach 07]

**MVDR Beamformer**

$$\mathbf{w}^{(MVDR)}_f = \mathcal{P}(\Phi_{XX,f})$$

[Higuchi & Ito+ 16]

$$\mathcal{P}(\cdot) \quad \text{- Principal eigenvector}$$

$$\Phi_{\nu\nu,f} = \frac{1}{\sum_{t=1}^{T} \lambda^{(\nu)}_{t,f}} \sum_{t=1}^{T} \lambda^{(\nu)}_{t,f} \mathbf{Y}_{t,f} \mathbf{Y}^H_{t,f}$$  \hspace{1cm} (0.7)
Integration

\[ \lambda_{t,f}^{(X)} \]
\[ \lambda_{t,f}^{(N)} \]

Median mask

Mask estimator

Magnitude

\[ Y_{t,f} \]
Integration

\[
\Phi_{XX,f} \quad \Phi_{NN,f} \\
\downarrow \quad \downarrow \\
\cdot \cdots \cdot \\
\downarrow \quad \downarrow \\
\lambda_{t,f}^{(X)} \quad \lambda_{t,f}^{(N)} \\
\downarrow \quad \downarrow \\
\downarrow \quad \downarrow \\
\downarrow \quad \downarrow \\
\downarrow \quad \downarrow \\
\downarrow \quad \downarrow \\
\downarrow \quad \downarrow \\
\downarrow \quad \downarrow \\
\downarrow \quad \downarrow \\
Y_{t,f} \\
\]

Spatial covariance

Median mask

Mask estimator

Magnitude
Integration

Steering vector computation

Spatial covariance

Median mask

Mask estimator

Magnitude

\[ Y_{t,f} \]

\[ \Phi_{XX,f} \]

\[ \Phi_{NN,f} \]

\[ \lambda^{(X)}_{t,f} \]

\[ \lambda^{(N)}_{t,f} \]

\[ w^{(OPT)}_{i} \]
Integration

Steering vector computation

\[ \Phi_{XX,f} \quad \Phi_{NN,f} \]

Spatial covariance

\[ \lambda_{t,f}^{(X)} \quad \lambda_{t,f}^{(N)} \]

Median mask

Mask estimator

Magnitude

Beamform filtering

\[ \hat{S}_{t,f} \]

\[ w_{i}^{(OPT)} \]

\[ Y_{t,f} \]
Integration
QR algorithm

[Francis 61]

Objective: compute \( P(A_0) \)

1. QR-decomposition of \( A_k = Q_k R_k \)
2. set \( A_{k+1} = R_k Q_k \)

For \( K \rightarrow \infty \) the diagonal of \( A_K \) contains eigenvalues of \( A_0 \)
With \( \prod_{k=0}^{K} Q_k \) containing the respective eigenvectors

QR-decomposition is differentiable as shown in [Walter & Lehmann+ 12].
Integration

Steering vector computation

Spatial covariance

Median mask

Mask estimator

Magnitude

Filterbank

Acoustic model

Beamform filtering

\[ \Phi_{XX,f}, \Phi_{NN,f} \]

\[ \lambda_t^{(X)}, \lambda_t^{(N)} \]

\[ \text{Magnitude} \]

\[ \text{Beamform filtering} \]

\[ \text{Magnitude} \]

\[ \Phi_{XX,f}, \Phi_{NN,f} \]

\[ \lambda_t^{(X)}, \lambda_t^{(N)} \]

\[ \text{Magnitude} \]

\[ \text{Beamform filtering} \]

\[ \hat{S}_{t,f} \]

\[ w_{t}^{(\text{OPT})} \]

\[ Y_{t,f} \]
Integration

\[ \Phi_{XX,f} \rightarrow \Phi_{NN,f} \rightarrow \mathbf{w}^{(GEV)}_{f} \rightarrow \text{BAN} \]

Spatial covariance

Median mask

Mask estimator

Magnitude

Filterbank

Acoustic model

Beamform filtering

\[ \mathbf{Y}_{t,f} \]

\[ \mathbf{w}^{(OPT)}_{f} \rightarrow \hat{s}_{t,f} \]
Speaker adaptation

- QR-alg.
- Spatial covariance
- Median mask
- Mask estimator
- Magnitude
- Beamform filtering
- First pass transcriptions
- Acoustic model
- Filterbank
- Magnitude
- Beamform filtering

\[ w_f^{(GEV)} \]
\[ \Phi_{XX,f} \]
\[ \Phi_{NN,f} \]
\[ \lambda_{t,f}^{(X)} \]
\[ \lambda_{t,f}^{(N)} \]
\[ Y_{t,f} \]
CHiME-4 dataset

- 18 h of simulated and real 6-channel data (87 speakers)
- recorded with handheld device
- noise scenarios: caffee, bus, pedestrian area, street junction

<table>
<thead>
<tr>
<th>Set</th>
<th>female</th>
<th>male</th>
<th>time (h) per speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>2</td>
<td>2</td>
<td>0.7</td>
</tr>
<tr>
<td>Eval</td>
<td>2</td>
<td>2</td>
<td>0.5</td>
</tr>
</tbody>
</table>
System

**Mask estimator**
- 1 BLSTM (256 units)
- 2 fully connected ReLU (512 units)
- 5 iterations of QR-algorithm
- hyper parameters of speaker adaptation tuned on dev set

**Acoustic model**
- unnormalized 80 dimensional log mel filterbank features
- linear layer (80 units) with batch normalization
- 5 BLSTM layers (600 units each)
System performance at different training stages

Average WER (%) for the described systems for different stages of the integrated training.

<table>
<thead>
<tr>
<th>System id</th>
<th>Front-end</th>
<th>Joint training</th>
<th>Speaker adapted</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BFIT</td>
<td>-</td>
<td>-</td>
<td>4.36</td>
<td>7.17</td>
</tr>
<tr>
<td>1</td>
<td>GEV</td>
<td>-</td>
<td>-</td>
<td>3.46</td>
<td>5.18</td>
</tr>
<tr>
<td>2</td>
<td>GEV</td>
<td>+</td>
<td>-</td>
<td>3.32</td>
<td>4.84</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>+</td>
<td>3.09</td>
<td>4.58</td>
</tr>
</tbody>
</table>
Speaker specific performance

WER (%) of separate speakers for the jointly trained system and the speaker adapted system

<table>
<thead>
<tr>
<th>System id</th>
<th>Dev</th>
<th></th>
<th></th>
<th>Eval</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F01</td>
<td>F04</td>
<td>M03</td>
<td>M04</td>
<td>F05</td>
<td>F06</td>
</tr>
<tr>
<td>2</td>
<td>4.19</td>
<td>3.23</td>
<td>2.77</td>
<td>3.07</td>
<td>6.88</td>
<td>4.09</td>
</tr>
<tr>
<td>3</td>
<td>3.55</td>
<td>3.20</td>
<td>2.48</td>
<td>3.14</td>
<td>6.35</td>
<td>4.09</td>
</tr>
</tbody>
</table>
Mask estimations
Conclusion

• Joint optimization of preprocessing and acoustic modeling is beneficial

• Speaker adaptation of mask estimator for beamforming can be beneficial depending on target speaker

• Adaptation does not decrease performance otherwise

• Masks after adaptation show stronger emphasis of speakers fundamental frequency and harmonics
Thank you for your attention

Any questions?
J. G. Francis.
The QR transformation a unitary analogue to the LR transformation - part 1.

J. Heymann, L. Durde, C. Boeddeker, P. Hanebrink, R. Haeb-Umbach.
Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system.

T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani.
Robust MVDR beamforming using time frequency masks for online offline ASR in noise.

B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, M. Bacchiani.
Neural network adaptive beamforming for robust multichannel speech recognition.

Z. Meng, S. Watanabe, J. R. Hershey, H. Erdogan.

T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, A. Senior.
Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms.

T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani.
Factored spatial and spectral multichannel raw waveform cldnns.

