MULTICHLANNEL RNN-BASED SEPARATION OF OVERLAPPING SPEECH

Lauréline Perotin†‡, Romain Serizel†‡, Emmanuel Vincent†‡, Alexandre Guérin†

†Orange Labs
‡Université de Lorraine, Inria, LORIA
PROBLEM STATEMENT

Distant-microphone voice command for personal digital assistant

- Real room conditions
- Competing speakers
- Ambient babble noise

→ Enhance the target speaker
PROBLEM STATEMENT

State of the art: Neural networks to estimate time-frequency masks or multichannel filter parameters

Current challenges: Overlapping speech

Contributions:

- Ambisonics contents
- Multi-source localization
- Enhancement in overlapping speech conditions by estimating the parameters of a multichannel filter
1. HIGH ORDER AMBISONICS

Capture
- Eigenmike
- Ambeo

Rendering
- Wave Field Synthesis
- binaural
- 5.1, ATMOS...

credits: ircam

LISTEN LAURÉLINE PEROTIN

07-19-2018
1. HIGH ORDER AMBISONICS

Order 0

Order 1

Order 2

...
2. PROPOSED SOLUTION FOR DIRECTION-OF-ARRIVAL ESTIMATION

Input feature based on acoustic intensity:

\[
\begin{align*}
    \mathbf{I}_a(t, f) &= \begin{bmatrix}
        \mathcal{R}\{W(t, f) * X(t, f)\} \\
        \mathcal{R}\{W(t, f) * Y(t, f)\} \\
        \mathcal{R}\{W(t, f) * Z(t, f)\}
    \end{bmatrix} \\
    \mathbf{I}_r(t, f) &= \begin{bmatrix}
        \mathcal{I}\{W(t, f) * X(t, f)\} \\
        \mathcal{I}\{W(t, f) * Y(t, f)\} \\
        \mathcal{I}\{W(t, f) * Z(t, f)\}
    \end{bmatrix}
\end{align*}
\]
3. ENHANCEMENT: FULL-BAND BEAMFORMING

Mixture: \( x(t, f) = s(t, f) + n(t, f) \)

HOA anechoic mixing matrix:

\[
A = \begin{bmatrix}
1 & \ldots & 1 \\
\sqrt{3} \cos \theta_0 \cos \phi_0 & \ldots & \sqrt{3} \cos \theta_J \cos \phi_J \\
\sqrt{3} \sin \theta_0 \cos \phi_0 & \ldots & \sqrt{3} \sin \theta_J \cos \phi_J \\
\sqrt{3} \sin \phi_0 & \ldots & \sqrt{3} \sin \phi_J \\
\end{bmatrix}
\]

HOA beamformer:

\[
\hat{s}(t, f) = u_1^T A^\dagger x(t, f)
\]

\( \rightarrow \) not robust to reverberation and close speakers
3. ENHANCEMENT: MULTICHANNEL WIENER FILTERING

**Mixture:** \( x(t, f) = s(t, f) + n(t, f) \)

**Time-invariant multichannel Wiener filter:**
\[
 w(f) = \left[ \Phi_{ss}(f) + \Phi_{nn}(f) \right]^{-1} \Phi_{ss}(f) u_1
\]

→ Little distortion, but covariance matrices needed!

**Mask – based covariance estimation:**
\[
 M_s(t, f) = \frac{|s(t, f)|}{|s(t, f)| + |n(t, f)|}
\]

\[
 \tilde{s}(t, f) = M_s(t, f) x(t, f)
\]

\[
 \tilde{\Phi}_{ss}(f) = \frac{1}{T} \sum_{t=0}^{T-1} \tilde{s}(t, f) \tilde{s}^H(t, f)
\]
Estimation of the mask via LSTM neural network:
3. PROPOSED SOLUTION

WER 86%

DoAs

WER 49%

HOA beamforming

LSTM

.masks

WER 20%

Target: clean speech

WER 7%

X

\( \hat{s} \rightarrow \hat{n} \rightarrow x \)

\( \hat{s} \rightarrow \hat{n} \rightarrow \text{mask} \rightarrow \text{covariance estimation} \)

\( \hat{s} \rightarrow \hat{n} \rightarrow \text{filter} \rightarrow y \)

07-19-2018

LISTEN

LAURÉLINE PEROTIN

10 / 13
3. RESULTS FOR LOCALIZATION

**Test data:**
2 overlapping speakers, static, no VAD
25 to 90° angular distance
SIR = 0 dB, SNR = 20 dB

- **Simulated SRIRs** (image method)
  RT60 = 0.2 to 0.8s
- **Real SRIRs**
  RT60 ≈ 0.5s
  random source/mic orientations
- **Real recordings**
  living-room, mic on coffee table

**Training data:**
SIR = 0 to 10 dB, SNR = 20 dB, different speakers and rooms
36 h of speech made from simulated SRIRs

![Localization results graph]

<table>
<thead>
<tr>
<th></th>
<th>simulated SRIRs</th>
<th>real SRIRs</th>
<th>real recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>4.2°</td>
<td>7.7°</td>
<td>9.8°</td>
</tr>
</tbody>
</table>
### Results for Enhancement

<table>
<thead>
<tr>
<th>Acceptance Rate (%)</th>
<th>Simulated SRIRs</th>
<th>Real SRIRs</th>
<th>Real recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech</td>
<td>7.4</td>
<td>7.4</td>
<td>n/a</td>
</tr>
<tr>
<td>Mixture</td>
<td>81.9</td>
<td>86.0</td>
<td>89.5</td>
</tr>
<tr>
<td>Oracle DoA + beamformer</td>
<td>32.3</td>
<td>49.2</td>
<td>53.8</td>
</tr>
<tr>
<td>Oracle DoA + proposed filter</td>
<td>12.3</td>
<td>19.7</td>
<td>20.3</td>
</tr>
<tr>
<td>Estimated DoA + beamformer</td>
<td>33.1</td>
<td>53.1</td>
<td>57.9</td>
</tr>
<tr>
<td>Estimated DoA + proposed filter</td>
<td>13.4</td>
<td>25.3</td>
<td>26.5</td>
</tr>
</tbody>
</table>

**Training data**
- 10h of mixed speech at SIR = 0 dB
- 44 different speakers
- 16 positions in a single room at RT\_60 = 270ms
CONCLUSION

order 1 Ambisonics
2 speakers + noise

LSTM-based multichannel Wiener filter
Inputs: omnidirectional mixture
+ beamformer toward target speech
+ beamformer toward competing speech

Largely outperforms beamforming or sole masking, including with close speakers in real conditions

Directions of arrival by CRNN