MULTICHLNNEl SPEECH SEPARATION WITH RNN FROM HOA RECORDINGS

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:
- New location-based method to estimate the parameters of a multichannel filter in overlapping speech conditions
- Ambisonics contents
1. HIGH ORDER AMBISONICS

Capture

Eigenmike

Ambeo

Rendering

Wave Field Synthesis

binaural

5.1, ATMOS...

credits: ircam
1. HIGH ORDER AMBISONICS

Order 0

Order 1

Order 2

...
2. FULL-BAND BEAMFORMING

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

Assumption: known direction of arrival of the target (at least)

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
2. MULTICOMPONENTIAL WIENER FILTERING

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

Optimization criterion:
\[
\min \mathbb{E}\{ |y(t, f) - u_1^H s(t, f)|^2 \}
\]
with \( y(t, f) = w(f)^H x(t, f) \)

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

→ Little distortion, but we need the covariance matrices!
2. MASKING-BASED COVARIANCE ESTIMATION

Omnidirectional mixture $x$

Ideal ratio mask $M_s$

$$M_s(t, f) = \frac{|s(t, f)|}{|s(t, f)| + |n(t, f)|}$$
2. MASKING-BASED COVARIANCE ESTIMATION

Estimated signal $\tilde{s}$

Ideal ratio mask $M_s$

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

$$M_s(t, f) = \frac{|s(t,f)|}{|s(t,f)| + |n(t,f)|}$$

$$\Phi_{ss}(f) = \frac{1}{T} \sum_{t=0}^{T-1} \tilde{s}(t, f)\tilde{s}^H(t, f)$$
3. PROPOSED SOLUTION

- **WER 89%**
  - directions of arrival
  - HOA beamforming

- **WER 46%**
  - masks
  - LSTM

- **WER 44%**
  - covariance estimation

- **WER 14%**
  - filter
  - Target: clean speech
  - WER 7%
3. RESULTS

**Training data:**
- 10h of mixed speech
- SIR = 0 dB
- 44 different speakers
- Room 1
  - 16 positions, RT\textsubscript{60} = 270ms

**Test data:**
- 20 min of mixed speech
- SIR = 0 dB
- 20 different speakers
- 4000 words
- Room 2
  - 42 positions, RT\textsubscript{60} = 350ms

### Table: Word Error Rate (%)

<table>
<thead>
<tr>
<th>Network inputs</th>
<th>1 spk</th>
<th>2 spk, angle diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25°</td>
</tr>
<tr>
<td>Clean speech</td>
<td>7.4</td>
<td>7.4</td>
</tr>
<tr>
<td>Mixture</td>
<td>68.5</td>
<td></td>
</tr>
<tr>
<td>Beamformer</td>
<td>24.3</td>
<td>76.0</td>
</tr>
<tr>
<td>Ideal mask</td>
<td>18.3</td>
<td>16.3</td>
</tr>
<tr>
<td>Filter from ideal mask</td>
<td>13.1</td>
<td>23.0</td>
</tr>
<tr>
<td>x, \hat{s}</td>
<td>68.6</td>
<td>91.8</td>
</tr>
<tr>
<td>mask filter</td>
<td>25.0</td>
<td>91.6</td>
</tr>
<tr>
<td>\hat{s}</td>
<td>61.2</td>
<td>90.8</td>
</tr>
<tr>
<td>mask filter</td>
<td>19.6</td>
<td>67.2</td>
</tr>
<tr>
<td>x, \hat{s}</td>
<td>55.9</td>
<td>86.4</td>
</tr>
<tr>
<td>mask filter</td>
<td><strong>17.1</strong></td>
<td><strong>80.9</strong></td>
</tr>
<tr>
<td>x, \hat{s}, \hat{n}</td>
<td>n/s</td>
<td>60.9</td>
</tr>
<tr>
<td>mask filter</td>
<td></td>
<td><strong>22.3</strong></td>
</tr>
</tbody>
</table>
CONCLUSION

order 1 ambisonics
2 speakers + noise

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

Performs as good as the filter computed from the ideal mask including with 25° apart speakers