Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter

Jean-Marc Valin, Jean Rouat, François Michaud

Department of Electrical Engineering and Computer Engineering
Université de Sherbrooke, Québec, Canada
Jean-Marc.Valin@USherbrooke.ca
Motivations

- The context: mobile robot and cocktail party effect
- The problem: separating sound sources
- The solution: microphone array with both linear and non-linear processing
Approach

- Frequency-domain processing
- Geometric Source Separation (GSS)
  - Minimize leakage under constraints
  - Adapted for real-time processing
- Post-filter
  - Cancels remaining interferences
  - Based on Ephraim and Malah estimator
  - Handles both stationary and non-stationary noise/interference
Geometric Source Separation

- Frequency domain:
  \[ x(k) = A(k)s(k) + n(k) \]

- Constrained optimization \( y(k) = W(k)x(k) \)
  - Minimize correlation of the outputs:
    \[ J_1(W(k)) = \| R_{yy}(k) - \text{diag}[R_{yy}(k)] \|^2 \]
  - Subject to geometric constraint:
    \[ J_2(W(k)) = \| W(k)A(k) - I \|^2 \]

- Modifications to original GSS algorithm
  - Instantaneous computation of correlations
  - Stochastic-gradient descent
Post-Filter Overview

- Noise estimate as the sum of two components (stationary + transient)
Background Noise Estimation

• Minima-Controlled Recursive Average (Cohen)
  - Noise estimate is adapted during quiet periods
  - Applied for each source of interest
• Initial estimate provided directly from the microphones

\[
\lambda_{m}^{\text{stat.}}(k, \ell_0) = \frac{1}{N^2} \sum_{n=0}^{N-1} \sigma_x^2(k)
\]
Interference Estimation

- Source separation leaks
  - Incomplete adaptation
  - Inaccuracy in localization
  - Reverberation
  - Imperfect microphones

- Estimation from other separated sources

\[
\lambda^{\text{leak}}_m(k, l) = \eta \sum_{i=0, i \neq m}^{M-1} S_i(k, l)
\]

\[
S_m(k, l) = \alpha_s S_m(k, l - 1) + (1 - \alpha_s) Y_m(k, l)
\]
Suppression Rule

• Ephraim & Malah spectral estimator

\[ \hat{X}_m(k, l) = G_m(k, l)Y_m(k, l) \]

• Gain is modified to take into account probability of source being present (Cohen)

\[ G(k) = p^2(k)G_{H_1}(k) \]
Experimental Setup

- Array of 8 inexpensive microphones on a Pioneer2 robot
- Automatic localization
- Noisy conditions
- 350 ms reverberation time
Results (Signal-to-Noise Ratio)

- Three voices recorded separately so clean signal is available

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>female 1</th>
<th>female 2</th>
<th>male 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microphone input</td>
<td>-1.8</td>
<td>-3.7</td>
<td>-5.2</td>
</tr>
<tr>
<td>GSS only</td>
<td>9.0</td>
<td>6.0</td>
<td>3.7</td>
</tr>
<tr>
<td>GSS+single channel</td>
<td>9.9</td>
<td>6.9</td>
<td>4.5</td>
</tr>
<tr>
<td>GSS+proposed</td>
<td>12.1</td>
<td>9.5</td>
<td>9.4</td>
</tr>
</tbody>
</table>
Results (spectrograms)

Input

GSS

Post-filter output

Reference
Results (recognition with post-filter)

- Japanese isolated word recognition (SIG2 robot)
  - 3 simultaneous sources
  - 200 word vocabulary
  - 90 degrees separation

<table>
<thead>
<tr>
<th></th>
<th>mixed</th>
<th>GSS only</th>
<th>GSS+pf</th>
</tr>
</thead>
<tbody>
<tr>
<td>right</td>
<td>66%</td>
<td>71%</td>
<td></td>
</tr>
<tr>
<td>left</td>
<td>15%</td>
<td>21%</td>
<td></td>
</tr>
<tr>
<td>center</td>
<td>41%</td>
<td>53%</td>
<td></td>
</tr>
</tbody>
</table>

- 14% reduction in error rate
Conclusion

- Geometric Source Separation
  - Real-time minimization of leakage
- Source separation post-filter
  - Interference estimated using other sources
- Future work
  - Robustness to reverberation
  - Better integration with speech recognition
    - Using the post-filter to estimate ASR feature reliability
Questions?