Tight integration of spatial and spectral features for BSS with Deep Clustering embeddings

Lukas Drude, Reinhold Haeb-Umbach
Table of contents

1. Problem Statement
2. Known Solutions
   - Spectral: Deep Clustering
   - Spatial: Time-Variant cGMM
3. Integrated Model (proposed)
4. Evaluation
5. Conclusion
Problem Statement

Goal:
- Transcribe speech from an observed mixture

Application:
- Meeting transcription
- Home assistants/ automation
- (Surveillance)

Focus:
- Robust source separation front-end exploiting spectral and spatial features
Motivation

Spectral / deep learning:

- Leverage training data
- Model speech characteristics
- Overfit, possibly poor generalization

Spatial clustering (multi-channel):

- Training data agnostic
- No concept of human speech
- Exploit spatial selectivity
Motivation

Spectral/ deep learning:

+ Leverage training data
+ Model speech characteristics
- Overfit, possibly poor generalization

Spatial clustering (multi-channel):

+ Training data agnostic
- No concept of human speech
+ Exploit spatial selectivity

Joint formulation desired!
Exploit speaker specific **spectral** characteristics for separation.

Concept:

- BLSTM yields an embedding vector $e_{tf}$ for each tf-bin.
- Encourage tendency to form clusters in embedding space.
- Cluster using $k$-means on $e_{tf}$.

---

Exploit speaker specific **spectral** characteristics for separation.

Concept:

- BLSTM yields an embedding vector $e_{tf}$ for each tf-bin.
- Encourage tendency to form clusters in embedding space.
- Cluster using $k$-means on $e_{tf}$.

Relevance:

- Speaker independent
- Number of speakers not fixed at training time

Exploit speaker specific **spectral** characteristics for separation.

**Concept:**
- BLSTM yields an embedding vector $e_{tf}$ for each tf-bin.
- Encourage tendency to form clusters in embedding space.
- Cluster using **k-means** on $e_{tf}$. 

**Relevance:**
- Speaker independent
- Number of speakers not fixed at training time

---

Exploit **spatial** diversity to separate speakers.

**Concept:**
- Complex random vectors $y_{tf}$
- Acoustic transfer function captured by spatial correlation matrix.

$$p(y_{tf}) = \sum_k \pi_{fk} \cdot \mathcal{N}_\mathbb{C}(y_{tf}; 0, \sigma_{tfk} \cdot R_{fk})$$

---

[2] Ito et al., Relaxed disjointness based clustering […], IWAENC 2014
Exploit spatial diversity to separate speakers.

Concept:

- Complex random vectors $y_{tf}$
- Acoustic transfer function captured by spatial correlation matrix.

$$p(y_{tf}) = \sum_k \pi_{fk} \cdot \mathcal{N}_C(y_{tf}; 0, \sigma_{tfk} \cdot R_{fk})$$

Relevance:

- Competitive: Winning system of CHiME 3 and CHiME 4 challenges

[2] Ito et al., Relaxed disjointness based clustering [...], IWAENC 2014
Integrated Model

\[ p(e_{tf}, y_{tf}) = \sum_k \pi_{fk} \cdot vMF(e_{tf}; \mu_k, \kappa_k) \cdot N(y_{tf}; \text{zero.pnum}, \sigma_{tfk} \cdot R_{fk}) \]

• Observations independent, given class label
• Estimate of all parameters better, when estimated jointly

Risk/ opportunity: Weighting between both models

L. Drude, R. Haeb-Umbach: Integration of spatial and spectral features
Integrated Model

$$p(e_{tf}, y_{tf}) = \sum_k \pi_{fk} \cdot \text{vMF}(e_{tf}; \mu_k, \kappa_k) \cdot N_{C}(y_{tf}; \text{zero.pnum}, \sigma_{tfk} \cdot R_{fk})$$

- Observations independent, given class label
- Estimate of all parameters better, when estimated jointly
- Risk/ opportunity: Weighting between both models

Spatial observation model
Integrated Model

- Statistical model:
  \[ p(e_{tf}, y_{tf}) = \sum_k \pi_{fk} \cdot vMF(e_{tf}; \mu_k, \kappa_k) \cdot \mathcal{N}_{\mathbb{C}}(y_{tf}; 0, \sigma_{tfk} \cdot R_{fk}) \]

- Observations independent, given class label
- Estimate of all parameters better, when estimated jointly
- Risk/ opportunity: Weighting between both models
Integrated Model

- Statistical model:
  \[ p(e_{tf}, y_{tf}) = \sum_k \pi_{fk} \cdot \text{vMF}(e_{tf}; \mu_k, \kappa_k) \cdot \mathcal{N}_c(y_{tf}; 0, \sigma_{tfk} \cdot R_{fk}) \]

- Observations independent, given class label
- Estimate of all parameters better, when estimated jointly
- Risk/ opportunity: Weighting between both models
Update equations: E-step

- Update class affiliation posterior:

\[
\gamma'_{tfk} = \frac{\pi_{fk}}{\sum_k \gamma'_{tk}} \cdot \text{Prior} \cdot \text{Spectral Model} \cdot \text{Spatial Model}
\]

\[
\gamma_{tfk} = \frac{\gamma'_{tfk}}{\sum_k \gamma'_{tk}}
\]
Evaluation: Setup

- Train/CV deep clustering on single channel WSJ utterances [1, 3]
- GMM-HMM recognizer trained on clean, 3-Gram LM [3]
- 3000 test mixtures, 2 speakers per mixture:
  - Random source and array positions (6 sensors)
  - Image method to generate room impulse responses

Example: Posterior Masks (200 ms < $T_{60}$ < 300 ms)

- **DC + vMFMM**
- **vMF-TV-cGMM**
- **TV-cGMM**

Integration of spatial and spectral features

L. Drude, R. Haeb-Umbach
Evaluation: Signal to Distortion Ratio Gain

Low reverb ($T_{60} = 50 \ldots 100$ ms)

<table>
<thead>
<tr>
<th>Method</th>
<th>SDR gain / dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC + Masking</td>
<td>10</td>
</tr>
<tr>
<td>DC + BF</td>
<td>20</td>
</tr>
<tr>
<td>VMF-TV-cGMM</td>
<td>10.9 dB</td>
</tr>
<tr>
<td>Spatial only</td>
<td>19.8 dB</td>
</tr>
</tbody>
</table>

Medium reverb ($T_{60} = 200 \ldots 300$ ms)

<table>
<thead>
<tr>
<th>Method</th>
<th>SDR gain / dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC + Masking</td>
<td>10</td>
</tr>
<tr>
<td>DC + BF</td>
<td>20</td>
</tr>
<tr>
<td>VMF-TV-cGMM</td>
<td>10.9 dB</td>
</tr>
<tr>
<td>Spatial only</td>
<td>19.8 dB</td>
</tr>
</tbody>
</table>

Integration of spatial and spectral features
Evaluation: Signal to Distortion Ratio Gain

Low reverb ($T_{60} = 50 \ldots 100$ ms)

- DC + Masking
- DC + BF
- VMF-TV-cGMM spatial only

SDR gain / dB

Medium reverb ($T_{60} = 200 \ldots 300$ ms)

- DC + Masking
- DC + BF
- VMF-TV-cGMM spatial only

SDR gain / dB
Word Error Rates \((T_{60} = 50 \ldots 100 \text{ ms})\)

<table>
<thead>
<tr>
<th>Features</th>
<th>Model</th>
<th>Extraction</th>
<th>WER / %</th>
</tr>
</thead>
<tbody>
<tr>
<td>spectral</td>
<td>DC [3]</td>
<td>Masking</td>
<td>65.8</td>
</tr>
<tr>
<td>spectral</td>
<td>DC</td>
<td>Beamforming</td>
<td>42.4</td>
</tr>
<tr>
<td>integrated</td>
<td>vMF-TV-cGMM</td>
<td>Beamforming</td>
<td>29.7</td>
</tr>
<tr>
<td>spatial</td>
<td>TV-cGMM [2]</td>
<td>Beamforming</td>
<td>33.6</td>
</tr>
</tbody>
</table>

[2] Ito et al., Relaxed disjointness based clustering [...], IWAENC 2014

Audio Example

(Switch to external audio player.)

• Observed mixture: y.wav
• Estimates: z_1.wav, z_2.wav
• Clean speech: x_1.wav, x_2.wav
Conclusion and Future Work

• EM algorithm for joint model

Spectral/ deep learning:
  ▶ Leverages training data
  ▶ Models speech characteristics

Spatial clustering:
  ▶ Generalizes to unseen conditions
  ▶ Exploits spatial selectivity

• Outperforms both very different baselines
• Bridges model-based research and deep learning research

Future Work:
• Joint training
• Real recordings