Using Source Models in Speech Separation

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Mixtures, Separation, and Models
2. Monaural Speech Separation
3. Binaural Speech Separation
4. Conclusions
LabROSA Overview

Information Extraction

Music

Environment

Recognition

Machine Learning

Separation

Speech

Signal Processing

Retrieval
1. Mixtures, Separation, and Models

• Sounds rarely occur in isolation
  ○ so analyzing mixtures is a problem
  ○ for humans and machines
Mixture Organization Scenarios

- **Interactive voice systems**
  - human-level understanding is expected

- **Speech prostheses**
  - crowds: #1 complaint of hearing aid users

- **Archive analysis**
  - identifying and isolating sound events

- **Unmixing/remixing/enhancement...**
Separation vs. Inference

• **Ideal** separation is rarely possible
  ○ many situations where overlaps cannot be removed

• **Overlaps** → **Ambiguity**
  ○ scene analysis = find “most reasonable” explanation

• **Ambiguity** can be expressed *probabilistically*
  ○ i.e. posteriors of sources \( \{S_i\} \) given observations \( X \):

\[
P(\{S_i\} | X) \propto P(X | \{S_i\}) \cdot P(\{S_i\})
\]

  *combination physics  source models*

  ○ search over \( \{S_i\} \) ??

• **Better** source models → better inference
  ○ .. learn from examples?
Approaches to Separation

ICA
- Multi-channel
- Fixed filtering
- Perfect separation – maybe!

CASA
- Single-channel
- Time-var. filter
- Approximate separation

Model-based
- Any domain
- Param. search
- Synthetic output

or combinations ...

Source Models in Separation - Dan Ellis
EM for Model-based Separation

- **Expectation-Maximization algorithm** — for solving partially-unknown problems
  - (only local optimality guaranteed)

- **EM for model-based separation**
  - **E-step**: find distribution of unknowns \( p(u) \) given current model parameters \( \Theta \) and observations \( x \)
  - **M-step**: optimize \( \Theta \) to maximize fit to \( x \) given current \( p(u) \)

\[
E-step \\
p(u|\Theta^{(n)}) = \frac{p(x, u|\Theta^{(n)})}{p(x|\Theta^{(n)})}
\]

\[
M-step \\
\Theta^{(n+1)} = \arg\max_{\Theta} E_p(u|\Theta^{(n)})p(x, u|\Theta)
\]

- \( u \) is... GMM mixture assignment
- ... T-F cell dominance
- ... current phone of voice \( i \)
- ...

Source Models in Separation - Dan Ellis

2007-11-09 - 7/17
What is a Source Model?

- **Source Model** describes signal behavior
  - encapsulates **constraints** on form of signal
  - (any such constraint can be seen as a model...)

- A model has **parameters**
  - model + parameters
    \[ \rightarrow \text{instance} \]

- What is **not** a source model?
  - detail not provided in instance
    - e.g. using phase from **original mixture**
  - constraints on **interaction** between sources
    - e.g. independence, clustering attributes
2. Monaural Speech Separation

- Cooke & Lee's Speech Separation Challenge
  - short, grammatically-constrained utterances:
    <command:4><color:4><preposition:4><letter:25><number:10><adverb:4>
    e.g. "bin white by R 8 again"
  - task: report letter + number for "white"
  - special session at Interspeech '06

- Separation or Description?
Codebook Models

- **Given models** for sources, find “best” (most likely) states for spectra:

\[ p(x|i_1, i_2) = \mathcal{N}(x; c_{i1} + c_{i2}, \Sigma) \]

- \{i_1(t), i_2(t)\} = \arg\max_{i_1, i_2} p(x(t)|i_1, i_2)

  - can include **sequential** constraints...
  - different **domains** for combining \( c \) and defining \( \Sigma \)

- **E.g. stationary noise:**

  ![Original speech](image)
  ![In speech-shaped noise (mel magsnr = 2.41 dB)](image)
  ![VQ inferred states (mel magsnr = 3.6 dB)](image)
Speech Recognition Models

• Decode with **Factorial HMM**
  - i.e. two state sequences, one model for each voice
  - exploit sequence constraints, speaker differences?

• **IBM “superhuman” Iroquois system**
  - fewer errors than people for same speaker, level
  - exploit grammar constraints - higher-level dynamics

---

Same Gender Speakers

<table>
<thead>
<tr>
<th>6 dB</th>
<th>3 dB</th>
<th>0 dB</th>
<th>- 3 dB</th>
<th>- 6 dB</th>
<th>- 9 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDL Recognizer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No dynamics</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Acoustic dyn.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grammar dyn.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Human</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Speaker-Adapted (SA) Models

- Factorial HMM needs distinct speakers

- Use “eigenvoice” speaker space
- Iterate estimating voice & separating speech
- Performs midway between speaker-independent (SI) and speaker-dependent (SD)

Mixture: t32_swil2a_m18_sbar9n

Adaptation iteration 1

Adaptation iteration 3

Adaptation iteration 5

SD model separation

Time (sec)
3. Binaural Speech Separation

- **2 or 3 sources in reverberation**
  - assume just 2 ‘ears’

- **Tasks:**
  - identify positions of sources (and number?)
  - recover source signals
Spatial Estimation in Reverb

- Model *interaural spectrum* of each source as stationary *level* and *time* differences:

\[
\frac{L(\omega, t)}{R(\omega, t)} = a(\omega) e^{j\omega \tau} N(\omega, t)
\]

- converge via EM to \(a(), \tau\) for each source
- mask is \(Pr(X(t,\omega)\) dominated by source \(i\)
Spatial Estimation Results

- **Modeling uncertainty** improves results
  - tradeoff between constraints & noisiness

Source Models in Separation - Dan Ellis

<table>
<thead>
<tr>
<th>EM-ILD (only IPD)</th>
<th>PHAT-histogram</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.45 dB</td>
<td>0.22 dB</td>
<td>12.35 dB</td>
</tr>
<tr>
<td>8.77 dB</td>
<td>-2.72 dB</td>
<td>8.77 dB</td>
</tr>
</tbody>
</table>

2007-11-09 - 15/17
Combining Spatial + Speech Model

- Interaural parameters give
  \[ ILD_i(\omega), \ ITD_i, \ Pr(X(t, \omega) = S_i(t, \omega)) \]

- Speech source model can give
  \[ Pr(S_i(t, \omega) \text{ is speech signal}) \]

- Can combine into one big EM framework...

\[ E-step \]
\[ p(u|\Theta^{(n)}) = \frac{p(x, u|\Theta^{(n)})}{p(x|\Theta^{(n)})} \]

\[ M-step \]
\[ \Theta^{(n+1)} = \arg\max_{\Theta} E_{p(u|\Theta^{(n)})} p(x, u|\Theta) \]

\( u \) is: Pr(cell from source i)

\( \Theta \) is: interaural params

phoneme sequence

speaker params
Summary & Conclusions

• Inferring model parameters is very general
  ○ .. and very difficult, in general

• Speech models can separate single channels
  ○ .. better match to individual → better results

• Spatial cues can separate binaural signals
  ○ .. but account for uncertainty from e.g. reverb

• EM-type approach can integrate them both