Sound, Mixtures, and Learning

Dan Ellis
<dpwe@ee.columbia.edu>

Laboratory for Recognition and Organization of Speech and Audio
(LabROSA)
Electrical Engineering, Columbia University
http://labrosa.ee.columbia.edu/

Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Fragment Recognition
4. Alarm Sound Detection
5. Future Work
Auditory Scene Analysis

- **Auditory Scene Analysis**: describing a complex sound in terms of high-level sources/events
  - ... like listeners do

- **Hearing is ecologically grounded**
  - reflects ‘natural scene’ properties
  - subjective, not absolute
Sound, mixtures, and learning

- **Sound**
  - carries useful information about the world
  - complements vision

- **Mixtures**
  - are the rule, not the exception
  - medium is ‘transparent’, sources are many
  - must be handled!

- **Learning**
  - the ‘speech recognition’ lesson:
    - let the data do the work
  - like listeners
The problem with recognizing mixtures

"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?"  (after Bregman’90)

- **Received waveform is a mixture**
  - two sensors, N signals ... *underconstrained*

- **Disentangling mixtures as the primary goal?**
  - perfect solution is not possible
  - need experience-based *constraints*
Human Auditory Scene Analysis

(Bregman 1990)

- People hear sounds as separate sources

- How?
  - break mixture into small elements (in time-freq)
  - elements are grouped in to sources using cues
  - sources have aggregate attributes

- Grouping ‘rules’ (Darwin, Carlyon, ...):
  - cues: common onset/offset/modulation, harmonicity, spatial location, ...

(after Darwin, 1996)
Cues to simultaneous grouping

- **Elements + attributes**
- **Common onset**
  - simultaneous energy has common source
- **Periodicity**
  - energy in different bands with same cycle
- **Other cues**
  - spatial (ITD/IID), familiarity, ...
Context and Restoration

- Context can create an ‘expectation’: i.e. a bias towards a particular interpretation
  - e.g. auditory ‘illusions’ = hearing what’s not there

- The continuity illusion

- SWS
  - duplex perception

- How to model these effects?
Computational Auditory Scene Analysis (CASA)

- **Goal**: Automatic sound organization; Systems to ‘pick out’ sounds in a mixture
  - ... like people do

- **E.g. voice against a noisy background**
  - ... to improve speech recognition

- **Approach**:
  - psychoacoustics describes grouping ‘rules’
  - ... just implement them?
The Representational Approach  
(Brown & Cooke 1993)

- Implement psychoacoustic theory
  - ‘bottom-up’ processing
  - uses common onset & periodicity cues

- Able to extract voiced speech:

---

NCARAI Seminar - Dan Ellis  
2002-11-04 - 9/34
Adding top-down constraints

Perception is not *direct* but a *search* for *plausible hypotheses*

- Data-driven (bottom-up)... vs. Prediction-driven (top-down)
  - objects irresistibly appear
  - match observations with parameters of a world-model
  - need world-model constraints...
Approaches to sound mixture recognition

- **Recognize combined signal**
  - ‘multicondition training’
  - combinatorics..

- **Separate signals**
  - e.g. CASA, ICA
  - nice, if you can do it

- **Segregate features into fragments**
  - then missing-data recognition
Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
   - standard ASR
   - approaches to speech + noise
3. Fragment Recognition
4. Alarm Sound Detection
5. Future Work
Speech recognition & mixtures

- Speech recognizers are the most successful and sophisticated acoustic recognizers to date

• ‘State of the art’ word-error rates (WERs):
  - 2% (dictation) - 30% (phone conv’ns)
Learning acoustic models

- **Goal:** describe $p(X|M)$ with e.g. GMMs

- **Separate models for each class**
  - generalization as blurring

- **Training data labels from:**
  - manual annotation
  - ‘best path’ from earlier classifier (Viterbi)
  - EM: joint estimation of labels & pdfs

Labeled training examples \( \{ x_n, \omega_{x_n} \} \) → Sort according to class → Estimate conditional pdf for class $\omega_I$

$$p(x|\omega_I)$$
Speech + noise mixture recognition

- **Background noise**
  Biggest problem for current ASR?

- **Feature invariance approach:**
  Design features to reflect only speech
  - e.g. normalization, mean subtraction
  - one model for clean and noisy speech

- **Alternative:**
  More complex models of the signal
  - *separate* models for speech and ‘noise’
HMM decomposition
(e.g. Varga & Moore 1991, Roweis 2000)

• Total signal model has independent state sequences for 2+ component sources

- New combined state space $q' = \{q_1, q_2\}$
  - new observation pdfs for each combination

$$p(X | q_1, q_2)$$
Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Fragment Recognition
   - separating signals vs. separating features
   - missing data recognition
   - recognizing multiple sources
4. Alarm Sound Detection
5. Future Work
3

Fragment Recognition
(Jon Barker & Martin Cooke, Sheffield)

- **Signal separation is too hard!**
  - Instead: segregate *features* into partially-observed sources
  - then classify

- **Made possible by ‘missing data’ recognition**
  - integrate over uncertainty in observations for optimal posterior distribution

- **Goal:**
  Relate clean speech models \( P(X|M) \)
  to speech-plus-noise mixture observations
  - ... and make it tractable
Comparing different segregations

- Standard classification chooses between models $M$ to match source features $X$

$$M^* = \arg\max_M P(M \mid X) = \arg\max_M P(X \mid M) \cdot \frac{P(M)}{P(X)}$$

- Mixtures $\rightarrow$ observed features $Y$, segregation $S$, all related by $P(X \mid Y, S)$

- spectral features allow clean relationship

- Joint classification of model and segregation:

$$P(M, S \mid Y) = P(M) \int P(X \mid M) \cdot \frac{P(X \mid Y, S)}{P(X)} dX \cdot P(S \mid Y)$$

- integral collapses in several cases...
Calculating fragment matches

\[ P(M, S|Y) = P(M) \int P(X|M) \cdot \frac{P(X|Y,S)}{P(X)} \, dX \cdot P(S|Y) \]

- \( P(X|M) \) - the clean-signal feature model
- \( P(X|Y,S)/P(X) \) - is \( X \) ‘visible’ given segregation?
- Integration collapses some channels...
- \( P(S|Y) \) - segregation inferred from observation
  - just assume uniform, find \( S \) for most likely \( M \)
  - use extra information in \( Y \) to distinguish \( S \)’s
e.g. harmonicity, onset grouping

- **Result:**
  - probabilistically-correct relation between clean-source models \( P(X|M) \)
  and inferred contributory source \( P(M,S|Y) \)
Speech fragment decoder results

- Simple $P(S|Y)$ model forces contiguous regions to stay together
  - big efficiency gain when searching $S$ space

- Clean-models-based recognition rivals trained-in-noise recognition

![Graph showing WER versus SNR for AURORA 2000 Test Set A]
Multi-source decoding

- Search for more than one source
- Mutually-dependent data masks
- Use e.g. CASA features to propose masks
  - locally coherent regions
- Theoretical vs. practical limits

\[ Y(t) \rightarrow \begin{align*} S_1(t) \rightarrow & q_1(t) \\ S_2(t) \rightarrow & q_2(t) \end{align*} \]
Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Fragment Recognition
4. Alarm Sound Detection
   - sound
   - mixtures
   - learning
5. Future Work
Alarm sound detection

- Alarm sounds have particular structure
  - people ‘know them when they hear them’
  - clear even at low SNRs

- Why investigate alarm sounds?
  - they’re supposed to be easy
  - potential applications...

- Contrast two systems:
  - standard, global features, $P(X|M)$
  - sinusoidal model, fragments, $P(M,S|Y)$
Alarms: Sound (representation)

- **Standard system: Mel Cepstra**
  - have to model alarms in noise context: each cepstral element depends on whole signal

- **Contrast system: Sinusoid groups**
  - exploit sparse, stable nature of alarm sounds
  - 2D-filter spectrogram to enhance harmonics
  - simple magnitude threshold, track growing
  - form groups based on common onset

- **Sinusoid representation is already fragmentary**
  - does not record non-peak energies
Alarms: Mixtures

- Effect of varying SNR on representations:
  - sinusoid peaks have ~ invariant properties
Alarms: Learning

- **Standard**: train MLP on noisy examples
  - **Alternate**: learn distributions of group features
    - duration, frequency deviation, amp. modulation...
    - underlying models are clean (isolated)
    - recognize in different contexts...

![Diagram showing the process of sound mixture detection](image_url)
Alarms: Results

Both systems commit many insertions at 0dB SNR, but in different circumstances:

<table>
<thead>
<tr>
<th>Noise</th>
<th>Neural net system</th>
<th>Sinusoid model system</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Del / Ins / Tot</td>
<td>Del / Ins / Tot</td>
</tr>
<tr>
<td>1 (amb)</td>
<td>7 / 25 / 36%</td>
<td>14 / 25 / 60%</td>
</tr>
<tr>
<td>2 (bab)</td>
<td>5 / 25 / 272%</td>
<td>15 / 25 / 68%</td>
</tr>
<tr>
<td>3 (spe)</td>
<td>2 / 25 / 280%</td>
<td>12 / 25 / 84%</td>
</tr>
<tr>
<td>4 (mus)</td>
<td>8 / 25 / 180%</td>
<td>9 / 25 / 576%</td>
</tr>
<tr>
<td>Overall</td>
<td>22 / 100 / 192%</td>
<td>50 / 100 / 197%</td>
</tr>
</tbody>
</table>
Alarms: Summary

• **Sinusoid domain**
  - feature components belong to 1 source
  - simple ‘segregation’ (grouping) model
  - alarm model as properties of group
  - robust to partial feature observation

• **Future improvements**
  - more complex alarm class models
  - exploit repetitive structure of alarms
Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Fragment Recognition
4. Alarm Sound Detection
5. Future Work
   - generative models & inference
   - model acquisition
   - ambulatory audio
Future work

- CASA as generative model parameterization:

\[
\begin{align*}
\theta & \rightarrow M_1 \rightarrow Y_1 \rightarrow X_1 \\
\theta & \rightarrow M_2 \rightarrow Y_2 \rightarrow X_2 \\
& \vdots
\end{align*}
\]

Source models
Source signals
Received signals
Channel parameters
Observations

Model dependence

\[
\begin{align*}
M_1 & \rightarrow Y_1 \\
M_2 & \rightarrow Y_2 \\
& \vdots
\end{align*}
\]

\[
\begin{align*}
X_1 & \rightarrow C_1 \\
X_2 & \rightarrow C_2 \\
& \vdots
\end{align*}
\]

Observations

\[
O = \{M_i\}
\]

\[
\begin{align*}
\Theta & \rightarrow \{K_i\}
\end{align*}
\]

\[
\begin{align*}
p(X|M_i,K_i)
\end{align*}
\]

Model fitting

Analysis structure

Fragment formation
Mask allocation
Likelihood evaluation
Learning source models

- The speech recognition lesson: Use the data as much as possible
  - what can we do with unlimited data feeds?

- Data sources
  - clean data corpora
  - identify near-clean segments in real sound
  - build up ‘clean’ views from partial observations?

- Model types
  - templates
  - parametric/constraint models
  - HMMs

- Hierarchic classification vs. individual characterization...
Personal Audio Applications

- Smart PDA records everything
- Only useful if we have index, summaries
  - monitor for particular sounds
  - real-time description

- Scenarios
  - personal listener → summary of your day
  - future prosthetic hearing device
  - autonomous robots

- Meeting data, ambulatory audio
Summary

• **Sound**
  - carries important information

• **Mixtures**
  - need to segregate different source properties
  - fragment-based recognition

• **Learning**
  - information extracted by classification
  - models guide segregation

• **Alarm sounds**
  - simple example of fragment recognition

• **General sounds**
  - recognize simultaneous components
  - acquire classes from training data
  - build index, summary of real-world sound