Recognition & Organization of Speech and Audio

Dan Ellis
Electrical Engineering, Columbia University
<dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/

Outline

1. Introducing LabROS
2. Tandem modeling for robust ASR
3. Other current projects
4. Future projects
5. Summary & conclusions
Organization of sound mixtures

- **Core operation:** Converting continuous, scalar signal into discrete, symbolic representation
Positioning sound organization

- Draws on many techniques
- Abuts/overlaps various areas
About auditory perception

- Received waveform is a mixture
  - two sensors, N signals ...
  - need knowledge-based constraints

- Psychoacoustics:
  the study of human sound organization
  - ‘auditory scene analysis’ (Bregman’90)

- Auditory perception is ecologically grounded
  - scene analysis is preconscious (→ illusions)
  - perceived organization:
    real-world objects + events (transient)
  - subjective not canonical (ambiguity)
Key themes for LabROSA
http://www.ee.columbia.edu/~dpwe/LabROSA/

- **Sound organization: construct hierarchy**
  - at an instant (sources)
  - along time (segmentation)

- **Scene analysis**
  - find attributes according to objects
  - use attributes to form objects
  - ... plus constraints of knowledge

- **Exploiting large data sets (the ASR lesson)**
  - supervised/labeled: pattern recognition
  - unsupervised: structure discovery, clustering

- **Special cases:**
  - speech recognition
  - other source-specific recognizers

- **... within a ‘complete explanation’**
Outline

1 Introducing LabROSA

2 Tandem modeling for robust ASR
   - ASR overview
   - Tandem modeling
   - Investigating the benefits

3 Other current projects

4 Future projects

5 Summary & conclusions
Automatic Speech Recognition (ASR)

- Standard speech recognition structure:

- ‘State of the art’ word-error rates (WERs):
  - 2% (dictation) - 30% (telephone conversations)

- Can use multiple streams...
Tandem speech recognition
(with Hermansky, Sharma & Sivadas/OGI, Singh/CMU)

- Neural net estimates phone posteriors;
  but Gaussian mixtures model finer detail

- Combine them!

- Train net, then train GMM on net output
  - GMM is ignorant of net output ‘meaning’
Tandem system results

- It works very well (‘Aurora’ noisy digits):

<table>
<thead>
<tr>
<th>System-features</th>
<th>Avg. WER 20-0 dB</th>
<th>Baseline WER ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTK-mfcc</td>
<td>13.7%</td>
<td>100%</td>
</tr>
<tr>
<td>Neural net-mfcc</td>
<td>9.3%</td>
<td>84.5%</td>
</tr>
<tr>
<td>Tandem-mfcc</td>
<td>7.4%</td>
<td>64.5%</td>
</tr>
<tr>
<td>Tandem-msg+plp</td>
<td>6.4%</td>
<td>47.2%</td>
</tr>
</tbody>
</table>
Relative contributions

- Approx relative impact on baseline WER ratio for different component:

  Combo over msg: +20%
  Combo over plp: +20%
  Combo over mfcc: +25%
  NN over HTK: +15%
  Tandem over HTK: +35%
  Tandem over hybrid: +25%

  Pre-linearity over posteriors: +12%
  KLT over direct: +8%
  Combo into HTK over Combo into noway: +15%
  Tandem over hybrid: +35%

  Tandem combo over HTK mfcc baseline: +53%
Inside Tandem systems: What’s going on?

- Visualizations of the net outputs

```
“one eight three” (MFP_183A)
```

- Spectrogram
- Cepstral-smoothed mel spectrum
- Phone posterior estimates
- Hidden layer linear outputs

```
Spectrogram
Clean 5dB SNR to ‘Hall’ noise
```

- Neural net normalizes away noise
Tandem feature space ‘magnification’

- Neural net performs a nonlinear remapping of the feature space

- small changes across critical boundaries result in large output changes
Outline

1 Introducing LabROSA

2 Tandem modeling for robust ASR

3 Other current projects
   - Alarm sound detection
   - Computational Auditory Scene Analysis
   - Multi-source and missing-data recognition
   - The Meeting Recorder project

4 Future projects

5 Summary & conclusions
Alarm sound detection

- Alarm sounds have particular structure
  - people ‘know them when they hear them’
  - build a generic detector?

- Isolate alarms in sound mixtures
  - representation of energy in time-frequency
  - formation of atomic elements
  - grouping by common properties (onset &c.)
  - classify by attributes...
Computational Auditory Scene Analysis (CASA)

- Implement psychoacoustic theory? (Brown’92)
  - what are the features? how are they used?
  - Additional ‘knowledge’ needed (Klassner’96)
Prediction-driven CASA

- Data-driven (bottom-up) fails for noisy, ambiguous sounds (most mixtures!)
- Need top-down constraints:
  - fit vocabulary of generic elements to sound
    ... bottom of a hierarchy?
  - account for entire scene
  - driven by prediction failures
  - pursue alternative hypotheses
Missing data recognition & CASA
(with Barker, Cooke, Green/Sheffield)

- Missing-data recognition
  - integrate across ‘don’t-know’ values
  - ‘perfect’ mask $\rightarrow$ excellent performance in noise

- Multi-source decoder
  - Viterbi search of sound-fragment interpretations

- CASA for masks/fragments
  - larger fragments $\rightarrow$ quicker search
Meeting recorder
(with ICSI, UW, SRI, IBM)

- Microphones in conventional meetings
  - for transcription/summarization/retrieval
  - informal, overlapped speech

- Data collection (ICSI and ...):
  - 10s of hours collected, ongoing
  - now being transcribed
Meeting recorder: Research issues

• Preliminary analysis
  - transcription & forced alignment
  - ground truth in turns/overlaps
  - preliminary distant-mic recordings

• Research areas
  - meeting dialog: overlaps, turns etc.
  - language modeling for meetings
  - feature design for distant acoustics

• Applications
  - information retrieval from meetings
  - ‘mapping’ meeting content
  - sociological analysis of meeting behavior
Outline

1. Introducing LabROSA
2. Tandem modeling for robust ASR
3. Other current projects
4. Future projects
   - Audio Content-Based Retrieval
   - A ‘machine listener’
   - Audio-video-text content analysis
5. Summary & conclusions
Audio Information Retrieval

- **Searching in a database of audio**
  - speech .. use ASR
  - text annotations .. search them
  - sound effects library?

- **e.g. Muscle Fish “SoundFisher” browser**
  - define multiple ‘perceptual’ feature dimensions
  - search by proximity in (weighted) feature space

- features are ‘global’ for each soundfile, no attempt to separate mixtures
CASAl for audio retrieval

- When audio material contains mixtures, global features are insufficient

- Retrieval based on element/object analysis:

  - features are calculated over grouped subsets
A ‘machine listener’

- **Goal:** Unsupervised structure discovery

- **What can you do with a large unlabeled training set (e.g. broadcast)?**
  - bootstrap learning: look for common patterns
  - have to learn generalizations in parallel: e.g. self-organizing maps, EM HMMs
  - post-filtering by humans may find ‘meaning’ in clusters
Audio-video-text content analysis
(with Shih-Fu Chang, Kathleen McKeown)

• Audio and video provide complementary info
  - correlate object features to define templates?

• Associated text annotations provide a very small amount of labeling
  - .. but for a very large number of examples
    – sufficient to obtain purchase?
  - build a ‘multimedia lexicon’ for question-answering
Outline

1. Introducing LabROSA
2. Tandem modeling for robust ASR
4. Other current projects
4. Future projects
5. Summary & conclusions
Applications for sound organization

What do people do with their ears?

• Human-computer interface
  - .. includes knowing when (& why) you’ve failed

• Robots
  - intelligence requires perceptual awareness
  - Sony’s AIBO: dog-hearing

• Archive indexing & retrieval
  - pure audio archives
  - true multimedia content analysis

• Content ‘understanding’
  - intelligent classification & summarization

• Autonomous monitoring

• Broader ‘structure discovery’ algorithms
Summary

**DOMAINS**
- Broadcast
- Movies
- Lectures
- Meetings
- Personal recordings
- Location monitoring

**ROSA**
- Object-based structure discovery & learning
- Speech recognition
- Speech characterization
- Nonspeech recognition
- Scene analysis
- Audio-visual integration
- Music analysis

**APPLICATIONS**
- Structuring
- Search
- Summarization
- Awareness
- Understanding
Conclusions

- New classification schemes for ASR
  - ... combining multiple approaches/sources

- But sound is more than just speech!
  - speech is a special case
  - need to deal with the ‘other stuff’

- Object-based analysis
  - it’s what people do
  - the world presents acoustic mixtures

- Whole-scene representation
  - it’s what people do
  - provides mutual constraints of overlap

- Broad range of approaches
  for a broad range of phenomena