Recognition & Organization of Speech & Audio

Dan Ellis

http://labrosa.ee.columbia.edu/

Outline

1. Introducing LabROSA
2. Projects in speech, music & audio
3. Summary
Sound organization

- **Central operation:**
  - continuous sound mixture
    → distinct objects & events

- **Perceptual impression is very strong**
  - but hard to ‘see’ in signal
“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 ...

- **Disentangling mixtures as primary goal**
  - perfect solution is not possible
  - need knowledge-based constraints
The information in sound

- **A sense of hearing is evolutionarily useful**
  - gives organisms ‘relevant’ information

- **Auditory perception is ecologically grounded**
  - scene analysis is preconscious (→ illusions)
  - special-purpose processing reflects ‘natural scene’ properties
  - subjective *not* canonical (ambiguity)
Key themes for LabROSA
http://labrosa.ee.columbia.edu/

• **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. Projects in speech, music & audio
   - Tandem speech recognition
   - ‘Meeting recorder’ speech analysis
   - Musical information extraction
   - Alarm sound detection

3. Summary
Automatic Speech Recognition (ASR)

- Standard speech recognition structure:

  ![Diagram of ASR process]

  - Sound goes through feature calculation to produce feature vectors.
  - Feature vectors are processed by the acoustic classifier and acoustic model parameters to generate phone probabilities.
  - Phone probabilities are then used by the HMM decoder to produce a phone/word sequence.
  - Understanding/application uses this sequence.

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

- Can use multiple streams...

\[p(\text{“sat”|“the”, “cat”})\]
\[p(\text{“saw”|“the”, “cat”})\]
Tandem speech recognition
(with Manuel Reyes, ICSI, OGI, CMU)

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

**Hybrid Connectionist-HMM ASR**

- Input sound → Feature calculation
- Speech features → Neural net classifier
- Phone probabilities → Noway decoder
- Phone probabilities → Words

**Conventional ASR (HTK)**

- Input sound → Feature calculation
- Speech features → Gauss mix models
- Subword likelihoods → HTK decoder
- Words

**Tandem modeling**

- Input sound → Feature calculation
- Speech features → Neural net classifier
- Phone probabilities → Gauss mix models
- Subword likelihoods → HTK decoder
- Words

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

WER as a function of SNR for various Aurora99 systems

Average WER ratio to baseline:
- HTK GMM: 100%
- Hybrid: 84.6%
- Tandem: 64.5%
- Tandem + PC: 47.2%

Aurora 2 Eurospeech 2001 Evaluation

- Columbia
- Philips
- UPC Barcelona
- Bell Labs
- IBM
- Motorola 1
- Motorola 2
- Nijmegen
- ICSI/OGI/Qualcomm
- ATR/Griffith
- AT&T
- Alcatel
- Siemens
- UCLA
- Microsoft
- Slovenia
- Granada
The Meeting Recorder project
(with ICSI, UW, SRI, IBM)

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

- Data collection (ICSI, UW, ...):
  - 100 hours collected, ongoing transcription
  - headsets + tabletop + ‘PDA’
Crosstalk cancellation

- Baseline speaker activity detection is hard:

- Noisy crosstalk model: $m = C \cdot s + n$

- Estimate subband $C_{Aa}$ from A’s peak energy
  - ... including pure delay (10 ms frames)
  - ... then linear inversion
PDA-based speaker change detection

- Goal: small conference-tabletop device
- Speaker turns from PDA mock-up signals?

- SCD algo on spectral + interaural features
  - average spectral + per-channel ITD, $\Delta\phi$
Music analysis: Lyrics extraction
(with Adam Berenzweig)

- Vocal content is highly salient, useful for retrieval
- Can we find the singing?
  Use an ASR classifier:
  - Frame error rate ~20% for segmentation based on posterior-feature statistics
  - Lyric segmentation + transcribed lyrics → training data for lyrics ASR...
Music analysis: Structure recovery
(with Rob Turetsky)

• Structure recovery by similarity matrices
  (after Foote)

- similarity distance measure?
- segmentation & repetition structure
- interpretation at different scales:
  notes, phrases, movements
- incorporating musical knowledge:
  ‘theme similarity’
Alarm sound detection

- Alarm sounds have particular structure
  - people ‘know them when they hear them’

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

- Key: recognize despite background
The ‘Machine listener’

- **Goal:** An auditory system for machines
  - use same environmental information as people

- **Aspects:**
  - recognize spoken commands (but not others)
  - track ‘acoustic channel’ quality (for responses)
  - categorize environment (conversation, crowd...)

- **Scenarios**

  - personal listener → summary of your day
  - autonomous robots: need awareness
Outline

1. Introducing LabROSA
2. Projects in speech, music & audio
3. Summary
LabROSA Summary

DOMAINS

• Broadcast
• Movies
• Lectures
• Meetings
• Personal recordings
• Location monitoring

APPLICATIONS

• Structuring
• Search
• Summarization
• Awareness
• Understanding

ROSA

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