Recognition & Organization of Speech & Audio

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

Outline

1 Introducing LabROSA
2 Speech recognition & processing
3 Auditory Scene Analysis
4 Projects & applications
5 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 Speech recognition & processing
   - Connectionist and tandem recognition
   - Speech and speaker detection
   - Musical information extraction

3 Auditory Scene Analysis

4 Projects & applications

5 Summary
Automatic Speech Recognition (ASR)

- Standard speech recognition structure:

  ![Diagram of speech recognition model]

  - **Acoustic model parameters**
  - **Word models**
  - **Language model**
  - **Feature calculation**
  - **Acoustic classifier**
  - **HMM decoder**
  - **Understanding/application...**

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

- Can use multiple streams...

---

ROSAtalk @ NEC - Dan Ellis

2001-08-17 - 7
The connectionist-HMM hybrid
(Morgan & Bourlard, 1995)

- Conventional recognizers use $P(X_i|S_i)$, acoustic likelihood model
  - model distribution with, e.g., Gaussian mixtures

- Can replace with posterior, $P(S_i|X_i)$:
  - neural network estimates phone given acoustics
  - discriminative

- Simpler structure for research
Tandem speech recognition
(with Hermansky, Sharma & Sivadas/OGI, Singh/CMU, ICSI)

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

Hybrid Connectionist-HMM ASR

- Feature calculation
- Neural net classifier
- Noway decoder

Conventional ASR (HTK)

- Feature calculation
- Gauss mix models
- HTK decoder

Tandem modeling

- Input sound
- Speech features
- Phone probabilities
- Words

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):**

  WER as a function of SNR for various Aurora99 systems

<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>84.6%</td>
</tr>
<tr>
<td>Tandem-mfcc</td>
<td>7.4%</td>
<td>64.5%</td>
</tr>
<tr>
<td><strong>Tandem-msg+plp</strong></td>
<td><strong>6.4%</strong></td>
<td><strong>47.2%</strong></td>
</tr>
</tbody>
</table>

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

![Graph showing WER as a function of SNR for various systems.](image-url)
Connectionist speaker recognition  
(with Dominique Genoud)

• Use neural networks to model speakers rather than phones?

• Specialize a phone classifier for a particular speaker?

• Do both at once for “Twin-output MLP”:

Feature vectors
Current vector

Output layer
Non-speech

Posterior probability vectors

MLP

53 ‘speaker’ phones

53 ‘world’ phones

9x12 features = 108 inputs

12 features

...
Speech/music discrimination
(with Gethin Williams)

- **Neural net is very sensitive to speech:**
  - characteristic jumping between phones
  - define statistics to distinguish speech regions
    e.g. entropy, ‘dynamism’ (delta-magnitude):

- **1.4% classification error on 2.5 s segments**
  - use HMM structure for segmentation

- **Good predictor of ASR success**
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’
Outline

1. Introducing LabROSA
2. Speech recognition & processing
3. Auditory Scene Analysis
   - Perception of sound mixtures
   - Illusions
   - Computational modeling
4. Projects & applications
5. Summary
Sound mixtures

- Sound ‘scene’ is almost always a mixture
  - always stuff going on
  - sound is ‘transparent’

- Need information related to our ‘world model’
  - i.e. separate objects
  - a wolf howling in a blizzard is the same as a wolf howling in a rainstorm
  - whole-signal statistics won’t do this

- ‘Separateness’ is similar to independence
  - objects/sounds that change in isolation
  - but: depends on the situation e.g. passing car vs. mechanic’s diagnosis
Human Sound Organization

- “Auditory Scene Analysis” [Bregman 1990]
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* into sources using *cues*
  - sources have aggregate *attributes*

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

(from Darwin 1996)
Cues and grouping

- Common attributes and ‘fate’
  - harmonicity, common onset
  → perceived as a single sound source/event

- But: can have conflicting cues
  - determine how $\Delta t$ and $\Delta f$ affect
    - segregation of harmonic
    - pitch of complex
The effect of context

- Context can create an ‘expectation’: i.e. a bias towards a particular interpretation
- e.g. Bregman’s “old-plus-new” principle:
  A change in a signal will be interpreted as an added source whenever possible
- a different division of the same energy depending on what preceded it
Computational ASA

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

- **Implement psychoacoustic theory?**
  - ‘bottom-up’, using common onset & periodicity

- **Able to extract voiced speech:**
Adding top-down cues

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

- **Data-driven (bottom-up)**...

  ![Diagram of data-driven](image)

  - Front end
  - Object formation
  - Grouping rules
  - Source groups

  **vs. Prediction-driven (top-down) (PDCASA)**

  ![Diagram of prediction-driven](image)

  - Front end
  - Compare & reconcile

  **Motivations**
  - detect non-tonal events (noise & click elements)
  - support ‘restoration illusions’...
    → hooks for high-level knowledge
  - ‘complete explanation’, multiple hypotheses, ...

ROSAtalk @ NEC - Dan Ellis 2001-08-17 - 21
PDCASA and complex scenes

[Diagram showing various sound levels and frequencies for different categories like 'City', 'Horn1 (10/10)', 'Crash (10/10)', 'Noise1', 'Squeal (6/10)', 'Truck (7/10)', etc., with corresponding frequency and level graphs.]
Outline

1 Introducing LabROSA
2 Speech recognition & processing
3 Auditory Scene Analysis
4 Projects & applications
   - Missing data recognition
   - Hearing prostheses
   - The machine listener
5 Summary
Missing data recognition
(Cooke, Green, Barker... @ Sheffield)

- Energy overlaps in time-freq. hide features
  - some observations are effectively missing

- Use missing feature theory...
  - integrate over missing data dimensions \( x_m \)
  \[
p(x|q) = \int p(x_g|x_m, q)p(x_m|q)dx_m
\]

- Effective in speech recognition
  - trick is finding good/bad data mask
Multi-source decoding
(Jon Barker @ Sheffield)

- Search of sound-fragment interpretations

"1754" + noise

Mask based on stationary noise estimate

Mask split into subbands
`Grouping' applied within bands:
Spectro-Temporal Proximity
Common Onset/Offset

CASA for masks/fragments
- larger fragments → quicker search

Use with nonspeech models?
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
CAS for audio retrieval

- When audio material contains mixtures, global features are insufficient

- Retrieval based on element/object analysis:

  - features are calculated over grouped subsets
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
Future prosthetic listening devices

- CASA to replace lost hearing ability
  - sound mixtures are difficult for hearing impaired

- Signal enhancement
  - resynthesize a single source without background
  - (need very good resynthesis)

- Signal understanding
  - monitor for particular sounds (doorbell, knocks)
    & translate into alternative mode (vibro alarm)
  - real-time textual descriptions
    i.e. “automatic subtitles for real life”
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. Tandem modeling: Neural net features
3. Meeting recorder data analysis
4. Computational Auditory Scene Analysis
5. Summary
Summary: 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

- ‘Structure discovery’ algorithms
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
Extra slides...
Positioning sound organization

- Draws on many techniques
- Abuts/overlaps various areas
Tandem recognition: 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%
- KLT over direct: +8%
- Combo-into-HTK over Combo-into-noway: +15%
- Pre-nonlinearity over posteriors: +12%
- MT over HTK: +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**

  - freq / kHz
  - time / s

  **Cepstral-smoothed mel spectrum**

  - freq / mel chan
  - time / s

  **Phone posterior estimates**

  - phone
  - time / s

  **Hidden layer linear outputs**

  - phone
  - time / s

- Neural net normalizes away noise
Acoustic Change Detection (ACD)
(with Javier Ferreiros, UPM)

- Find optimal segmentation points via Bayesian Information Criterion (BIC)
- Cluster segments to find underlying ‘sources’
- Repeat segmentation incorporating cluster assignments

![Diagram of Acoustic Change Detection (ACD)]
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
Participant motion detection

- Cross-correlation gives speaker-mic coupling:

  Changes in coupling impulse response show changes in path/orientation

- Comparison between different channels → distinguish *speaker* and *listener* motion
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, $\Delta\phi$
Computational ASA

- **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?
CASA front-end processing

- Correlogram:
  Loosely based on known/possible physiology

- linear filterbank cochlear approximation
- static nonlinearity
- zero-delay slice is like spectrogram
- periodicity from delay-and-multiply detectors
Problems with ‘bottom-up’ CASA

- **Circumscribing time-frequency elements**
  - need to have ‘regions’, but hard to find

- **Periodicity is the primary cue**
  - how to handle aperiodic energy?

- **Resynthesis via masked filtering**
  - cannot separate within a single t-f element

- **Bottom-up leaves no ambiguity or context**
  - how to model illusions?
Generic sound elements for PDCASA

- **Goal is a representational space that**
  - covers real-world perceptual sounds
  - minimal parameterization (sparseness)
  - separate attributes in separate parameters

\[
X_N[n,k] = H[k] \cdot A[n]
\]

\[
X_C[n,k] = H[k] \cdot \exp((n - n_0) / T[k])
\]

\[
X_W[n,k] = H_W[n,k] \cdot P[n,k]
\]

- **Object hierarchies built on top...**
### PDCASA for old-plus-new

- **Incremental analysis**

<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>t3</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Input signal" /></td>
<td><img src="image" alt="Time t1" /></td>
<td><img src="image" alt="Time t2" /></td>
</tr>
<tr>
<td><img src="image" alt="Time t1" />: initial element created</td>
<td><img src="image" alt="Time t2" />: Additional element required</td>
<td><img src="image" alt="Time t3" />: Second element finished</td>
</tr>
</tbody>
</table>