Audio Information Extraction

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. Audio Information Extraction
2. Speech, music, and other
3. General sound organization
4. Future work & summary
1 Audio Information Extraction (AIE)

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

- Perceptual impression is very strong
  - but hard to ‘see’ in signal
Perceptual organization: Bregman’s lake

“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)
Positioning AIE

- **Domain**
  - text ... speech ... music ... general audio

- **Operation**
  - recognize ... index/retrieve ... organize
AIE Applications

- **Multimedia access**
  - sound as complementary dimension
  - need all modalities for complete information

- **Personal audio**
  - continuous sound capture quite practical
  - different kind of indexing problem

- **Machine perception**
  - intelligence requires awareness
  - necessary for communication

- **Music retrieval**
  - area of hot activity
  - specific economic factors
Outline

1. Audio Information Extraction

2. Speech, music, and other
   - Speech recognition
   - Multi-speaker processing
   - Music classification
   - Other sounds

3. General sound organization

4. Future work & summary
Automatic Speech Recognition (ASR)

- Standard speech recognition structure:

  - Feature calculation
  - Acoustic classifier
  - HMM decoder
  - Understanding/application...

  - Acoustic model parameters
  - Word models
  - Language model

  - ‘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, ICSI)

- 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:
Aurora ‘noisy digits’
(with Manuel Reyes)

- 50% of word errors corrected over baseline
- Beat even ‘bells and whistles’ system using intensive large-vocabulary techniques
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 $x_m$

\[
p(x|q) = \int p(x_p|x_m, q)p(x_m|q)dx_m
\]

- Effective in speech recognition
  - trick is finding good/bad data mask

![Graph showing AURORA 2000 - Test Set A results with various SNR conditions and WER graphs.](image)
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
    - \( \ldots \) including pure delay (10 ms frames)
    - \( \ldots \) then linear inversion
Speaker localization
(with Wei Hee Huan)

- Tabletop mics form an array; time differences locate speakers

- Ambiguity:
  - mic positions not fixed
  - geometric symmetry

- Detect speaker activity, overlap
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’
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...
Artist similarity

- **Train network to discriminate specific artists:**
  
  w60o40 stats based on LE plp12  2001-12-28

- Focus on vocal segments for consistency

- .. then clustering for recommendation
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
Sound textures
(with Marios Athineos)

• **Textures: Large class of sounds**
  - no clear pitch, onsets, shape
  - fire, rain, paper, machines, ...
  - ‘bulk’ subjective properties

• **Abstract & synthesize by:**
  - project into low-dimensional parameter space
  - learn dynamics within space
  - generate endless versions
Outline

1. Audio Information Extraction

2. Speech, music, and other

3. General sound organization
   - Computational Auditory Scene Analysis
   - Audio Information Retrieval

4. Future work & summary
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?
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
Adding top-down cues

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

- **Data-driven (bottom-up)・・・**

  ![Diagram](image1)

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

  ![Diagram](image2)

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

-70
-60
-50
-40
dB

200
400
1000
2000
4000
f/Hz

200
400
1000
2000
4000
f/Hz

Noise1

Wefts1−4

Weft5

Wefts6,7

Weft8

Wefts9−12

Horn1 (10/10)
Horn2 (5/10)
Horn3 (5/10)
Horn4 (8/10)
Horn5 (10/10)

Crash (10/10)

Noise2, Click1

Noise1

Squeal (6/10)
Truck (7/10)

Weft5

Weft8

Wefts9−12
Audio Information Retrieval
(with Manuel Reyes)

- 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
Audio Retrieval: Results

- **Musclefish corpus**
  - most commonly reported set

- **Features**
  - mfcc, brightness, bandwidth, pitch ...
  - no temporal sequence structure

- **Results:**
  - 208 examples, 16 classes, 84% correct
  - confusions:

<table>
<thead>
<tr>
<th></th>
<th>Instr</th>
<th>Spch</th>
<th>Env</th>
<th>Anim</th>
<th>Mech</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Musical instrs.</strong></td>
<td>136 (14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Speech</strong></td>
<td></td>
<td>17 (7)</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td><strong>Envron.</strong></td>
<td></td>
<td>2</td>
<td>6 (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Animals</strong></td>
<td>2</td>
<td></td>
<td>2</td>
<td>1 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>Mechanical</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>15 (2)</td>
</tr>
</tbody>
</table>
CASA for audio retrieval

- When audio material contains mixtures, global features are insufficient

- Retrieval based on element/object analysis:

  - features are calculated over grouped subsets
Outline

1. Audio Information Extraction
2. Speech, music, and other
3. General sound organization
4. Future work & summary
Automatic audio-video analysis
(with Prof. Shih-Fu Chang, Prof. Kathy McKeown)

- **Documentary archive management**
  - huge ratio of raw-to-finished material
  - costly manual logging
  - missed opportunities for cross-fertilization

- **Problem: term <-> signal mapping**
  - training corpus of past annotations
  - interactive semi-automatic learning
  - need object-related features

---

- **Annotations**
  - **Text processing**
  - **Concept discovery**

- **A/V data**
  - **A/V segmentation and feature extraction**
  - **A/V feature unsupervised clustering**
  - **Generic detectors**

- **Multimedia Fusion**
  - **Mining Concepts & Relationships**

- **MM Concept Network**

- **Complex Spatio-Temporal Classification**
- **Question Answering**
The ‘Machine listener’

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

- **Signal understanding**
  - monitor for particular sounds
  - real-time description

- **Scenarios**

  - personal listener → summary of your day
  - future prosthetic hearing device
  - autonomous robots
LabROSA 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
Summary: Audio Info Extraction

• Sound carries information
  - useful and detailed
  - often tangled in mixtures

• Various important general classes
  - Speech: activity, recognition
  - Music: segmentation, clustering
  - Other: detection, description

• General processing framework
  - Computational Auditory Scene Analysis
  - Audio Information Retrieval

• Future applications
  - Ubiquitous intelligent indexing
  - Intelligent monitoring & description
Audio Information Extraction: panacea or punishment?