Recognizing and Classifying Environmental Sounds

Dan Ellis

Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

dpwe@ee.columbia.edu

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

1. Environmental Sound Recognition
2. Foreground Events
3. Background Retrieval
4. Labels & Annotation
5. Future Directions
1. What is hearing for?

- Hearing = getting **information** from sound
  - predators/prey
  - communication

- Environmental sound recognition is **fundamental**

Environmental Sound Perception

- What do people hear?
  - sources
  - ambience

- Mixtures are the rule
**Sound Scene Evaluations**

- **Evaluations** are good for research
  - help researchers, help funders

- A decade of evaluations:

<table>
<thead>
<tr>
<th>Year</th>
<th>Meeting rooms</th>
<th>Acoustic events</th>
<th>Music transcription</th>
<th>Speech separation</th>
<th>Source separation</th>
<th>Segmentation</th>
<th>Video (soundtrack) classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>NIST MtgRm</td>
<td>ADC</td>
<td>MIREX</td>
<td>CHIL/CLEAR</td>
<td>SSC/PASCAL</td>
<td>SASSEC SiSEC</td>
<td>TRECVID</td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

  - **Metrics**: SNR, Frame Acc, Event Error Rate, mAP
- Systems submitted Mar 2013
- Results at WASPAA, Oct 2013
- 2 tasks...

Task 1: Scene classification
- 10 classes x 10 examples x 30 s
  - street, supermarket, restaurant, office, park ...
- evaluate on 100 examples (~1 hour total)
  by classification accuracy
Task 2: Event detection (“office live”)
- 16 events x 20 training examples (~ 20 min total)
  knock, laugh, drawer, keys, phone ...
- evaluate on ~15 min (?)
- metrics: frame-level AEER & event-level precision-recall
“Multimedia Event Detection”
- e.g. MED2011: 15 events × 200 example videos (~60s)
  - Making a sandwich, Birthday party, Parade, Flash mob
- evaluate by mean Average Precision
  over 10k-100k videos (200-2,000 hours)
- audio and video ...
  - participants have annotated ~1000 videos (> 10 h)

E009 Getting a Vehicle Unstuck
Consumer Video Dataset

- Columbia Consumer Video (CCV)
  - 9,317 videos / 210 hours
  - 20 concepts based on consumer user study
  - Labeled via Amazon Mechanical Turk

Mark all the categories that appear in any part of the video.

Description:
- Watch the entire video as more categories may appear over time.
- Mark all the categories that appear in any part of the video.
- Make sure the audio is on.
- If no matching category is found, mark the box in front of "None of the categories matches".
- For categories that appear to be relevant but you’re not completely sure, please still mark it.
- Please move over or click on the category name for detailed description.

Sport
- Basketball
- Baseball
- Soccer
- Ice Skate
- Ski
- Swim
- Raking

Animal
- Cat
- Dog
- Bird

Celebration
- Graduation
- Birthday
- Wedding Reception
- Wedding Ceremony
- Wedding Dance

Others
- Multi: Music Performance
- Multi: Non-music Performance
- Parade
- Beach
- Playground

Current Time: 10 sec

Submit

Original URL: http://www.youtube.com/watch?v=U3dqW6d1L0
Environmental Sound Motivations

- Audio Lifelog Diarization

- Consumer Video Classification & Search

- Real-time hearing prosthesis app

- Robot environment sensitivity

- Understanding hearing
2. Foreground Event Recognition

- “Events” are what we hear / notice
- ASR approach?

- events = words? what are subwords?
- need labeled data
- but ... mature tools are great
Transient Features

- **Transients** = foreground events?
- **Onset detector** finds energy bursts
  - best SNR
- **PCA basis** to represent each
  - 300 ms x auditory freq
- “**bag of transients**”
• **Results show a small benefit**
  • similar to MFCC baseline?

• **Examine clusters**
  • looking for semantic consistency...
  • link cluster to label
NMF Transient Features

- Decompose spectrograms into templates + activation

\[ X = W \cdot H \]

- well-behaved gradient descent algorithm
- 2D patches
- sparsity control
- computation time...

Smaragdis & Brown ’03
Abdallah & Plumbley ’04
Virtanen ’07
NMF Transient Features

- Learn 20 patches from CLEAR Meeting Room events
- Compare to MFCC-HMM detector

![Graph showing error rate in noise vs. SNR (dB)]

- NMF more noise-robust
  - combines well ...
Why Are Events Hard?

- Events are **short**
  - target sounds may occupy only a few % of time

- Events are **varied**
  - what is the vocabulary? what are the prototypes?
  - source & channel variability

- Critical information is in **fine-time structure**
  - onset transient etc.
  - poor match to classic frame-spectral-envelope features
3. Background Retrieval

- **Baseline** for soundtrack classification
  - divide sound into short frames (e.g. 30 ms)
  - calculate features (e.g. MFCC) for each frame
  - describe clip by statistics of frames (mean, covariance)
  - = “bag of features”

- Classify by e.g. Mahalanobis distance + SVM
Retrieval Evaluation

- **Rank** large test set by match to category
- **Precision-Recall**

![CCV Precision-Recall (mfcc+sbpca)](image)

- mean **Average Precision**
Retrieval Examples

- High precision for **top hits** (in-domain)
Sound Texture Features

- Characterize sounds by perceptually-sufficient statistics
  
  .. verified by matched resynthesis

- Subband distributions & env x-corrs
  - Mahalanobis distance ...

McDermott et al. '09
Ellis, Zheng, McDermott '11
Sound Texture Features

- Test on MED 2010 development data
  - 10 audio-oriented manual labels

• Per-class stats
  - relate dimensions to classes?

- Perform ~ same as MFCCs
  - covariance ~ texture?
Auditory Model Features

- **Subband Autocorrelation PCA (SBPCA)**
  - Simplified version of Lyon et al. system
  - 10x faster \((RT \times 5 \rightarrow RT/2)\)
- Captures **fine time structure** in multiple bands
  - .. missing in MFCC features

Lyon et al. 2010
Cotton & Ellis 2013
Subband Autocorrelation

- Autocorrelation stabilizes fine time structure

- 25 ms window, lags up to 25 ms
- calculated every 10 ms
- normalized to max (zero lag)
Auditory Model Feature Results

- **SAI** and **SBPCA** close to **MFCC** baseline

- **Fusing** MFCC and SBPCA improves mAP by 15% rel
  - mAP: 0.35 → 0.40

- **Calculation time**
  - MFCC: 6 hours
  - SAI: 1087 hours
  - SBPCA: 110 hours
What is Being Recognized?

- **Soundtracks represented by global features**
  - MFCC covariance, codebook histograms
  - What are the critical parts of the sound?
Semantic Audio Features

- Train classifiers on related labeled data

- defines a new “semantic” feature space

- Use for target classifier

- or combo
4. Labels & Annotation

- “Semantic Features” are a promising approach
  - but we need good coverage...
  - how to learn more categories?

- Annotation is expensive
  - fine time annotation
    > 10x real-time
  - a few hours are available

- What to label?
  - generic vs. task-specific

<table>
<thead>
<tr>
<th>animal</th>
<th>singing</th>
<th>clatter</th>
</tr>
</thead>
<tbody>
<tr>
<td>anim_bird</td>
<td>music_sing</td>
<td>rustle</td>
</tr>
<tr>
<td>anim_cat</td>
<td>music</td>
<td>scratch</td>
</tr>
<tr>
<td>anim_ghoat</td>
<td>knock</td>
<td>hammer</td>
</tr>
<tr>
<td>anim_horse</td>
<td>thud</td>
<td>washtub</td>
</tr>
<tr>
<td>human_noise</td>
<td>clap</td>
<td>applause</td>
</tr>
<tr>
<td>laugh</td>
<td>click</td>
<td>whistle</td>
</tr>
<tr>
<td>scream</td>
<td>bang</td>
<td>squeak</td>
</tr>
<tr>
<td>child</td>
<td>beep</td>
<td>tone</td>
</tr>
<tr>
<td>mumble</td>
<td>engine_quiet</td>
<td>sirene</td>
</tr>
<tr>
<td>speech</td>
<td>engine_light</td>
<td>water</td>
</tr>
<tr>
<td>speech_ne</td>
<td>power_tool</td>
<td>micro_blow</td>
</tr>
<tr>
<td>radio</td>
<td>engine_heavy</td>
<td></td>
</tr>
<tr>
<td>white_noise</td>
<td>cheer</td>
<td></td>
</tr>
<tr>
<td>other_creak</td>
<td>crowd</td>
<td></td>
</tr>
</tbody>
</table>
BBC Audio Semantic Classes

- **BBC Sound Effects Library**
  - 2238 tracks (60 h)
  - short descriptions

- **Use top 45 keywords**

- **Added as “semantic units”**
  - some redundancy visible in mutual APs
BBC Audio Semantic Classes

- Limited semantic correspondence
Label Temporal Refinement

- **Audio Ground Truth at coarse time resolution**
  - better-focused labels give better classifiers?
  - but little information in very short time frames

- **Train classifiers on shorter (2 sec) segments?**
  - Initial labels apply to whole clip
  - Relabel based on most likely segments in clip
  - Retrain classifier
Label Temporal Refinement

- Refining labels is "Multiple Instance Learning"
  - "Positive" clips have at least one +ve frame
  - "Negative" clips are all –ve

- Refine based on previous classifier’s scores

- threshold from CDFs of +ve and –ve frames

- mAP improves ~10% after a few iterations
5. Future: Tasks & Metrics

• Environmental sound recognition: What is it **good for**?
  - media content description ("Recounting")
  - environmental **awareness**

• What are the right ways to **evaluate**?
  - task-specific metrics: AEER, F-measure
  - downstream tasks: WER, mAP
  - real **applications**: archive search, aware devices
Labels & Annotations

- **Training data:** quality vs. quantity
  - quality costs:
    - DCASE ~ 0.3 h
    - TRECVID MED (Aladdin) ~ 10 h
  - quantity always wins

- **Opportunistic labeling**
  - e.g. Sound Effects library, subtitles ...
  - need refinement strategies

- **Existing annotations indicate interest**
Source Separation

• Separated sources makes event detection easy
  • “separate then recognize” paradigm

• integrated solution more powerful...

• Environmental Source Separation is ill-defined
  • relevant “sources” are listener-defined
  • environment description addresses this
  • Environment recognition for source separation
Summary

• (Machine) Listening:
  Getting useful information from sound

• Foreground event recognition
  ... by focusing on peak energy patches

• Background sound retrieval
  ... from long-time statistics

• Data, Labels, and Task
  ... what are the sources of interest?
References 1/2


• Keansub Lee, Dan Ellis, Alex Loui, “Detecting local semantic concepts in environmental sounds using Markov model based clustering,” *IEEE ICASSP*, 2278-2281, Dallas, Apr 2010.


Acknowledgment

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.