Acoustic event detection
Jort F. Gemmeke
Google

November 6, 2015
A bit about me...

Jort F. Gemmeke

- **PhD 2011**  Radboud University Nijmegen, The Netherlands
  - Noise robust speech recognition (Speech enhancement, source separation)
    - Missing data techniques
    - Non-negative matrix factorization, sparse representations
- **2011-2014**  Postdoc at KU Leuven, Belgium
  - Self-taught vocal interface for dysarthric speakers
- **2014-2015**  Algorithm developer, Audience Inc.
  - Acoustic event detection
- **2015-now**  Research scientist, Google.
Acoustic event detection

- Automatically extracting information from audio: atomic units that compose it
  - E.g. footsteps, car passing by, dog barking

- Non-speech, non-music:
  - Non-speech human sounds (whistling, clapping, etc.)
  - Animal sounds
  - Environment (doors, tap running)
  - Vehicles, machinery
  - Beeps, alarm sounds
  - Natural sounds (wind, rain drops)
  - Etc.
Acoustic event detection

Demo (by Tampere University, Finland)
Goals

Motivate the research

Give an overview of the acoustic event detection field: techniques and findings

Explain the challenges, differences with other fields, future research
Why acoustic event detection?

● New research area → several emerging applications

● The audio modality has several beneficial properties:
  ○ Easy & cheap to capture and transmit
  ○ Many devices can capture audio
  ○ Sound can travel past obstacles → allows gathering information from large areas
  ○ Not affected by lighting conditions
  ○ May be complementary to visual signals
Application areas

Multimedia information retrieval

- Events may be easier to detect from sound
- Automatic tagging of videos
- Fingerprinting
- Search

Xu et al., “Creating audio keywords for event detection in soccer video,” in Proc. ICME 2003

Soccer: https://en.wikipedia.org/wiki/College_soccer
Application areas

Acoustic monitoring & surveillance

- Gunshot & scream detection
- Automatic vehicle detection
- Crash detection

Clavel et al., “Events Detection for an Audio-Based Surveillance System,” in Proc. ICME 2005
Application areas

Assistive technologies

- Sound visualization for the hearing impaired:
  - SilWatch & Sound Watcher (door chime, phone ringing, baby’s cries, fire alarm, etc)
- Acoustic fall detection
- Lifestyle monitoring

http://www.shinyu.co.jp/english/index.html

Falling sign: image by epson291
https://en.wikipedia.org/wiki/Sedentary_lifestyle
Application areas

Biodiversity monitoring

- Birds
- Marine mammals

Birds: [http://mcld.co.uk/research/](http://mcld.co.uk/research/)
Screenshot: [http://warblr.net/](http://warblr.net/)
Spectrogram: [http://sabiod.univ-tln.fr/data_samples.html](http://sabiod.univ-tln.fr/data_samples.html)
Application areas

Many more, newer possibilities

- Context aware devices
- Mechanic fault detection
- Health monitoring
- Landslide/earthquake detection
- Noise pollution monitoring (acoustic pleasantness maps)
- Life logging
  - Acoustic diary
  - Acoustic geo-sensing
Goals

Motivate the research

Give an overview of the acoustic event detection field: techniques and findings

Explain the challenges, differences with other fields, future research
Feature representations

Spectro-temporal representation

- ‘Frequency’ vs. time
- Mel-spectral, MFCC, LPC, gammatone, cochlea grams, …

Summary statistics

- Aggregate over long windows or entire recordings
- Many options:
  - Zero crossing-rate, energy, spectral bandwidth, …
  - Universal background models, l-vectors
Acoustics

Spectral characteristics:

- Harmonic events
  - Beeps, whistles, bird voices
- Non-harmonic events
  - Fans, ventilation noise
- Combinations of the above
  - E.g. wind can vary from non-harmonic broadband noise to a single tone depending on the environment
Acoustics

Temporal characteristics:

● Transients
  ○ Impact sounds, knocking, etc.

● Sustained
  ○ Wind, engine noises
  ○ Often slowly varying in time, e.g. wind, car passing by

● Combinations of the above
  ○ Many natural sounds consist of elementary units
  ○ E.g. an individual raindrop vs. a texture of rain drops
Tasks

Acoustic event:
- Localized in time (can be given a start and end time)
- Can be given a label from a set of possible event labels
- E.g. dog barking, car passing by, alarm beep

Acoustic scene / context:
- Medley of multiple sound events
- E.g. busy street, birthday party
Event classification

Class 1

Class 2

Class 3
Scene classification

Class 1

Class 2
Event detection
Recognition

<table>
<thead>
<tr>
<th>alert</th>
<th>drawer</th>
<th>keys</th>
<th>knock</th>
<th>pageturn</th>
<th>pendrop</th>
<th>phone</th>
<th>printer</th>
<th>speech</th>
<th>switch</th>
</tr>
</thead>
</table>

Time [s]
Classification

Linear classifier, SVM

K-NN

GMM

DNN
Temporal structure

- Within-event structure
- Between-event structure

Hidden Markov Model

- Speech
- Music
- Beep
- Car
General findings

- Optimal features depend on data
- Optimal approach depends on type and amount of data
- Technology seems to follow developments in speech recognition:
  - Dynamic Time Warping
  - Gaussian Mixture Models
  - Gaussian Mixture Models + Hidden Markov Models
  - Deep Neural Networks (+ Hidden Markov Models)
  - Recurrent Neural Networks (?)
Goals

Motivate the research

Give an overview of the acoustic event detection field: techniques and findings

**Explain the challenges, differences with other fields, future research**
Challenges

- Environmental mismatch
- Overlapping events
- Relations between events
- Taxonomy
- Evaluation
- Lack of accurate annotation
Environmental mismatch

● Often events originate a distance from the microphone
  ○ Results in additional background events
  ○ But also different ‘room’ impulse response

● Recording conditions can be very different
  ○ Bandwidth limited microphones (smartphone capture, …)
  ○ Microphone location (microphone inside pocket, jacket, …)
  ○ Encoding (raw, mp3, …)

● Sound of events can differ substantially
  ○ Different geographic locations (culture, device brands, …)
Overlapping events

Approaches to deal with overlapping events:

- Front-end separation prior to single-stream decoding
- Explicitly model compositionality of events in feature space
- Multi-condition (“robust”) per-label modelling
Relations between events

- In speech processing
  - Strong relation between words, through grammar and semantics
  - Lot’s of training data to model these relations

- In acoustic event detection
  - Less training data: no cross-domain model
  - Some relation between events, but varies widely
    - Car: take keys, open door, close door, start engine
  - Relations at various levels: sequential, but also via context
    - Not so likely to hear a car passing by when inside an airplane

Key: https://en.wikipedia.org/wiki/Remote_keyless_system
Car: https://en.wikipedia.org/wiki/Presidential_state_car_(United_States)
Taxonomy

- Example: “My dog Charlie is barking happily”
- Possible class labels:
  - Dog
  - Dog barking
  - Animal
  - Barking
  - My dog
  - Charlie
  - Dog barking happily
  - Golden retriever (dog breed) barking
  - ...
- No established taxonomy or way to define the class labels
Evaluation

No agreed-upon metrics

- Events may be overlapping - should we count “substitutions”?
- Ground truth annotation may be temporally inaccurate
- Application specific - onset vs duration
- Events are typically sparse
- Relations between events
Data

- “Wash Hand Basin, Running Tap, Plu - FX Sound Effect”
- “Coonhound barking at dog park”
- “Door Bell Sounds”
And then you’ve got this...
Conclusions

Field with many opportunities

Commonalities with speech recognition and image/video recognition

- Multiple sources of interest, overlapping sounds
- No language model for sound
- Taxonomy: sound classes, atomic units

Data is a major challenge

- Annotation is costly
- Annotation is inaccurate
- Much acoustic variation
- Taxonomy is application-driven – difficult to reuse data
Future directions

- **Applications**
  - More advanced pattern recognition, more classes
  - More diverse usage scenarios

- **Research**
  - Datasets from more diverse environments
  - Standardized evaluations with more diverse datasets
  - Robustness and accuracy in realistic scenarios
  - Learning from datasets with sparse or loose annotation
  - Human-in-the-loop active learning

- **Information fusion with other sensors**
  - Video, movement, location, etc.

- **Multichannel audio capture**
  - Source tracking, sensor networks