Computational Analysis of Sound Events in Realistic Multisource Environments

Tuomas Virtanen
Tampere University of Technology, Finland
Outline

- Information in everyday soundscapes
- Characterizing complex multisource data using categorical variables
- Supervised learning approach for polyphonic event detection
- Data acquisition
- Convolutional recurrent neural networks
- Public evaluation challenges
Information in everyday soundscapes
Information in everyday soundscapes

- Entire scene
  - Birthday party, busy street, home, etc.
- Individual sources
  - Car, beep, dog barking, etc.
- Spatial properties
  - Locations, distance, movement
- Other properties
  - Number of sources, loudness, etc.
Information in everyday soundscapes

• Entire scene
  – Birthday party, busy street, home, etc.

• Individual sources
  – Car, beep, dog barking, etc.

• Spatial properties
  – Locations, distance, movement

• Other properties
  – Number of sources, loudness, etc.
Signal characteristics of realistic soundscapes

\[ x(t) = \sum_n s_n(t) \ast h_n(t) + n(t) \]

- Large number of overlapping sources
- Non-stationary source signals \( s_n(t) \)
- Sources far away from microphones: low SNR, convolutive mixing, time-varying transfer functions \( h_n(t) \)
- Often only one microphone available
How to extract useful information from acoustic scenes?
Categorical latent variables with textual labels

- Examples in everyday sound analysis: "street", "car", "dog", "busy", "quiet" etc.
- Textual category labels: efficient way to present information in human-understandable way
- Estimating categorical variables allow bridging the semantic gap between signal and its semantics
- Categories can chosen to characterize different properties depending on target application
Pattern classification and beyond

- In simple scenarios, a signal can be characterized with one categorical variable.
Pattern classification and beyond

• In simple scenarios, a signal can be characterised with one categorical variable
• E.g. scene classification: street / home / car / park...
• Multiclass classification
• Applications: context-aware devices
Pattern classification and beyond

- In reality, classes can be overlapping
- Multilabel classification = tagging
Pattern classification and beyond

- Time-varying classes -> detection
  - Estimating start and end times of classes
- Polyphonic detection: multiple overlapping classes
Example sound event labeling
Potential applications

- **Assistive technologies**
  Tools for the hearing impaired, memory disordered, the elderly

- **Context awareness**
  E.g. robots, cars, mobile devices: reaction to events in environment

- **Surveillance**
  Detecting dangerous events and crimes

- **Urban planning**
  Analysis of human & animal activity

- **Biology**
  Recognition of species, analysis of biodiversity

- **Monitoring**
  Recognition of noise sources, machine faults

- **Multimedia information retrieval**
  Detailed description of audio in video & social networks

*Computational Sound Analysis*
Scientific challenges

Large variety of different types of sounds
Scientific challenges

Large variety of different types of sounds

Large acoustic diversity within each category
Scientific challenges

Large variety of different types of sounds

Large acoustic diversity within each category

Overlapping sounds, reverberation

\[ x(t) = \sum_{n} s_n(t) \ast h_n(t) \]
The supervised machine learning approach

• Algorithms that find mapping between training examples (audio) and labels (annotations)
• Set of possible sound classes defined in advance
  – Defines the scope of the method
• Need for annotated training material from all the classes
  – Audio recordings and its class annotations
Obtaining data

1. Real recordings
   - Relatively easy to record
   - Realistic, match with real scenarios
   - Annotations cumbersome (slow & uncertain)

2. Synthetic material
   - Mixing of sounds from sample databases
   - Easy to produce large quantities and obtain their annotations
   - Do the results / system translate to real environments?
Real audio: TUT Sound Events 2016 & 2017

- Used in DCASE 2016 & 2017 evaluations
- Environmental sound recordings from home, residential area & street context
- Binaural recordings + video
- About 4 hours of annotated audio
- Manual annotations
  - start and end times of each event
  - labels (verb+noun) based on Wordnet taxonomy
  - manually grouped to classes for supervised classification
  - in total 2000 event instances

Supervised learning for polyphonic event detection

- Sources overlapping in time
- Sound events starting and ending at different times
- How to do the supervised learning?
Segment-wise multilabel classification

- Binary encoding of class activities
- Predict the activity of each class in each frame
The multilabel deep neural network (DNN) approach

Training

Audio

Feature extraction

Input

DNN training

Multilabel encoding

Target output
Multilabel DNN approach
Acoustic features

- Signals typically represented in the spectral domain
- Mel spectrogram (log of energies in mel bands) is a commonly used perceptually motivated representation

[Diagram showing acoustic signal and Mel spectrogram]
Recurrent neural network

[Diagram of a recurrent neural network with time and various sound categories like speech, footsteps, and music]
Convolutional neural networks

- Layers of convolutions allow learning time-frequency filters to automatically find relevant representations.
**CNN**

- Pooling allows learning shift-invariant features.
- Multiple CNN layers allows learning higher-level features.
What do the CNN filters represent?

- Synthetic input maximizing the activation of selected neurons
CRNN

• Convolutional recurrent neural network
• Convolutional layers learn features
• Recurrent layers model longer temporal context
Typical CRNN parameters

- 1…4 convolutional layers
- 1…3 recurrent layers
- 96 or 256 neurons / filters per layer
- Frequency max pooling
- CNN activations: rectified linear
- Recurrent neural networks: GRU
- Dropout: 0…0.75
- Detection: binary thresholding (threshold 0.5)
- Cross entropy loss, Adam optimizer
Demonstration

• Training material: 19 hours of audio, binaural recordings
• Material from 10 contexts: basketball game, beach, inside a bus, inside a car, hallway, office, restaurant, shop, street and stadium with track and field events
• Free-label annotations, manually grouped to 61 classes
Objective evaluation with synthetic data

- Synthetic data, 16 classes: Alarms & Sirens, Baby crying, Bird singing, Cat meowing, Crowd applause...

Objective evaluation with synthetic data

Objective evaluation with synthetic data

<table>
<thead>
<tr>
<th>Classifier</th>
<th>F-score (framewise)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary GMM</td>
<td>40.5%</td>
</tr>
<tr>
<td>FNN</td>
<td>49.2%</td>
</tr>
<tr>
<td>CNN</td>
<td>52.8%</td>
</tr>
<tr>
<td>RNN</td>
<td>59.8%</td>
</tr>
<tr>
<td>CRNN</td>
<td>66.4%</td>
</tr>
</tbody>
</table>

Case study: acoustic monitoring

- Detection of a target sound over a long period of time (e.g. months)
Existing applications

• Automatic captioning of acoustic events in Youtube videos:
Existing applications

- Prominent event detection, several suppliers:
  - Baby cry monitoring, window breakage, bog barking monitoring, etc.
DCASE evaluation campaigns

- Previously, each group focused on specific application, with different data
- DCASE = Detection & Classification of Acoustic Scenes and Events
- Public evaluation data challenge:
  1. Provide open data that researchers can use
  2. Encourage reproducible research
  3. Attract new researchers into the field
  4. Create reference points for performance comparisons
DCASE over the years

- **DCASE 2013**
  - 3 Tasks: Acoustic Scenes; Office Live; Office Synthetic
  - 25 challenge submissions, presented at WASPAA 2013
- **DCASE 2016**
  - 4 Tasks: Acoustic Scenes, Office Synthetic, Real Events, Domestic Tagging
  - 82 challenge submissions, one-day workshop in Budapest
- **DCASE 2017**
  - 4 Tasks: Acoustic Scenes, Rare Events, Real Events, Large-scale Weak Labels
  - 200 challenge submissions, two-day workshop in Munich
- **DCASE 2018**
  - 5 Tasks: Acoustic Scenes, Audio Tagging, Bird Detection, Weak Labels, Multichannel Activity Classification
DCASE 2017

Task 1: Scene Classification

15 classes:

- Bus
- Cafe/restaurant
- Car
- City center
- Forest path
- Grocery store
- Home
- Lakeside beach
- Library
- Metro station
- Office
- Residential area
- Train
- Tram
- Urban park
Task 1: Results

- 97 Systems / 39 Teams / 129 Authors
- Top system performance 83.3 %, baseline system 61%
- Convolutional neural networks most popular, good performance in general
- Top system used GAN to generate more training examples
Task 2: Detection of rare sound events

- Detecting target sound event within 30-second synthetic mixture
- Target sound events: baby crying, glass breaking, gunshot
- Motivation: Surveillance and smart home applications
- Examples: Alarm the user based on detected hazardous activity
Task 2: Results

• 33 Systems (13 Teams / 38 Authors)
• Metrics:
  – event-based Error Rate (ER)
  – F1-score (secondary metric)
  – both calculated with 500ms onset collar
Task 3: Sound Event Detection in Real-life Audio
Task 3: Results

- 36 Systems (13 Teams / 32 Authors)
- Evaluated using **segment-based Error Rate** (ER) and F1-score (secondary metric), both calculated in one second segments
- Top system ER 0.79, F-score 41.7%
Task 4: Large-Scale Weakly Supervised Sound Event Detection for Smart Cars
Task 4: overall results

- 34 submissions / 9 Teams / 25 Authors (for both subtasks)
- Significant improvement over MLP-based baseline
DCASE 2017: General trends

- Convolutional neural networks were widely used and obtained good results
- Recurrent layers help in detection tasks
- Powerful classifiers are sensitive to training-test mismatch
- Spectral features dominating
DCASE 2018

• 5 tasks:
  1. Acoustic scene classification
  2. General-purpose audio tagging of Freesound content with AudioSet labels
  3. Bird audio detection
  4. Large-scale weakly labeled semi-supervised sound event detection in domestic environments
  5. Monitoring of domestic activities based on multi-channel acoustics

http://dcase.community/challenge2018/
DCASE 2018 Schedule

• 30 March: Challenge open, data and baseline methods released
• 30 June: Release of evaluation datasets
• 31 July: Submission deadlines
• 15 September: Challenge results
• 19-20 November: Workshop in Woking, Surrey, UK

http://dcase.community/challenge2018/
Future research directions

• Weakly labeled data
• Opportunistic data collection (online sources)
• Robust classification
• Spatial audio (localization, tracking, separation of sources)
• Audio + video + other modalities
Contributors

- Toni Heittola, Annamaria Mesaros, Emre Cakir, Heikki Huttunen, Giambattista Parascandolo, Konstantinos Drossos, Sharath Adavanne, Eemi Fagerlund, Aku Hiltunen, Archontis Politis
References


www.cs.tut.fi/~tuomasv/publications.html
Summary

• Estimating categorical variables represented by textual labels allows characterizing complex data
• Sound event detection: research field with several potential applications
• Scientific challenges: robust classification, dealing with overlapping sounds, reverberation
• Practical challenges: acquisition of annotated data
• Convolutional recurrent networks enable learning suitable representations and give state of the art performance
• Public evaluation campaigns allow comparison of different methods and reproducible research