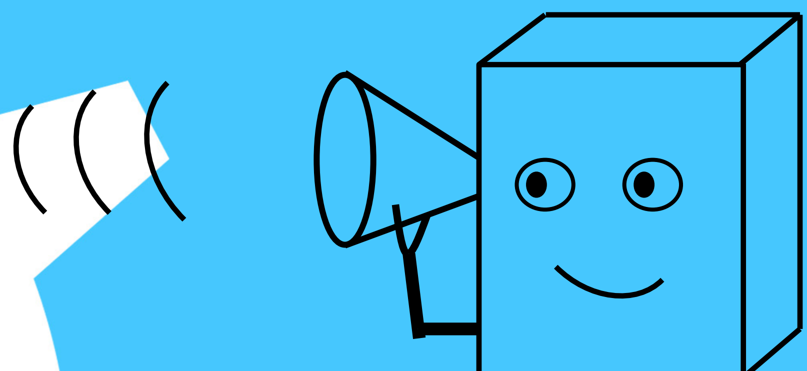
Computational Analysis of Sound Events in Realistic Multisource Environments

Tuomas Virtanen

Tampere University of Technology, Finland



Outline

- Information in everyday sounscapes
- 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

$$x(t) = \sum_n s_n(t) \star h_n(t) + n(t)$$

- 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

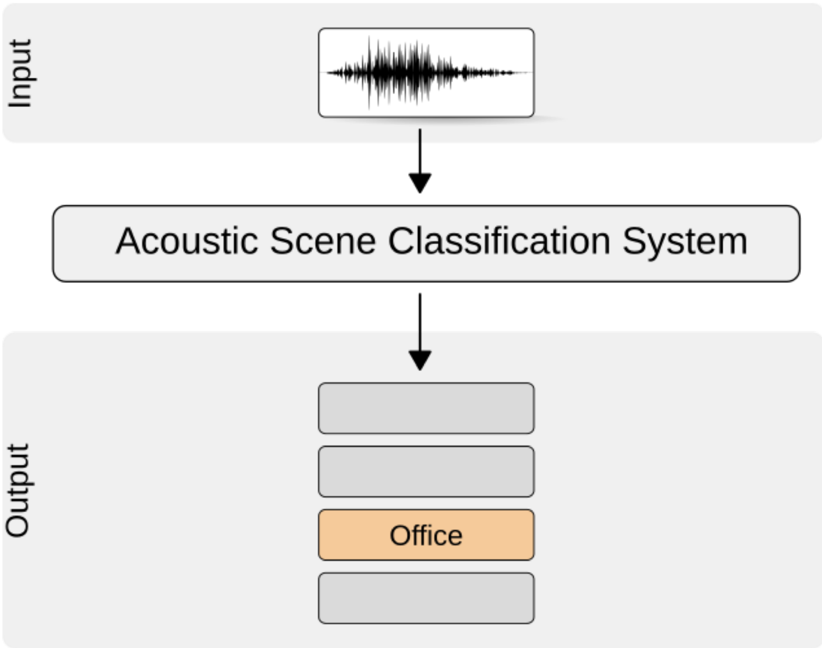
- Examples in everyday sound analysis: ”street”, ”car”, ”dog”, ”busy”, ”quiet” etc.
- Textual category labels: efficient way to present information in human-understandable way
- Estimating categorical variables allow bridging the semantic gap between signal and its semantics
- Categories can chosen to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

Pattern classification and beyond

- In simple scenarios, a signal can be characteried with one categorical variable

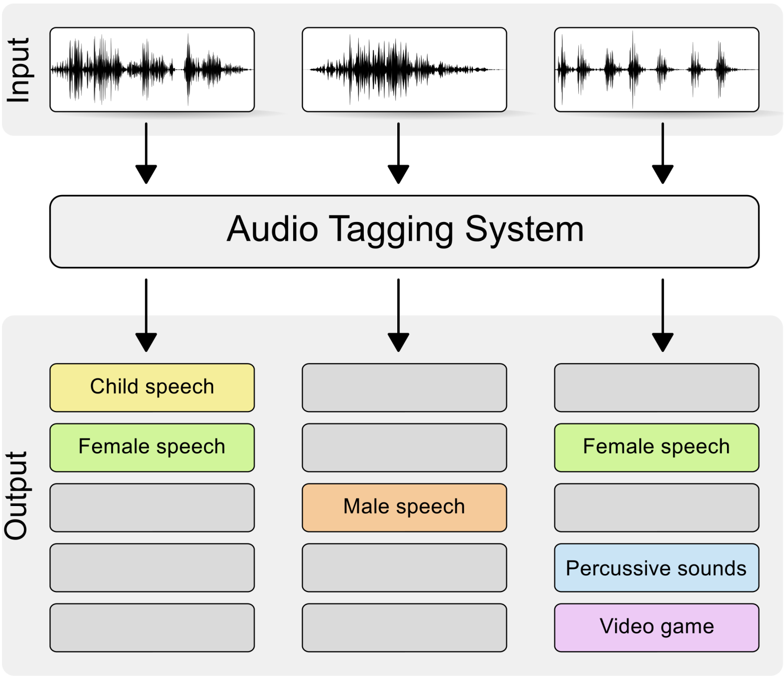
Pattern classification and beyond

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



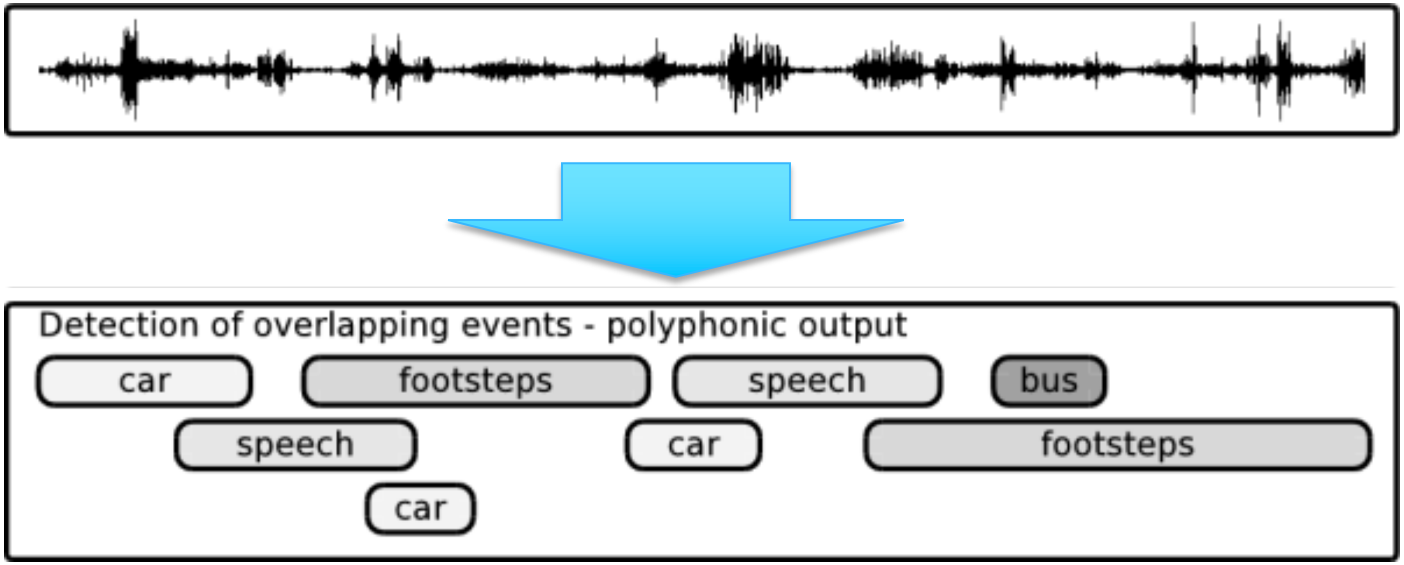
Pattern classification and beyond

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



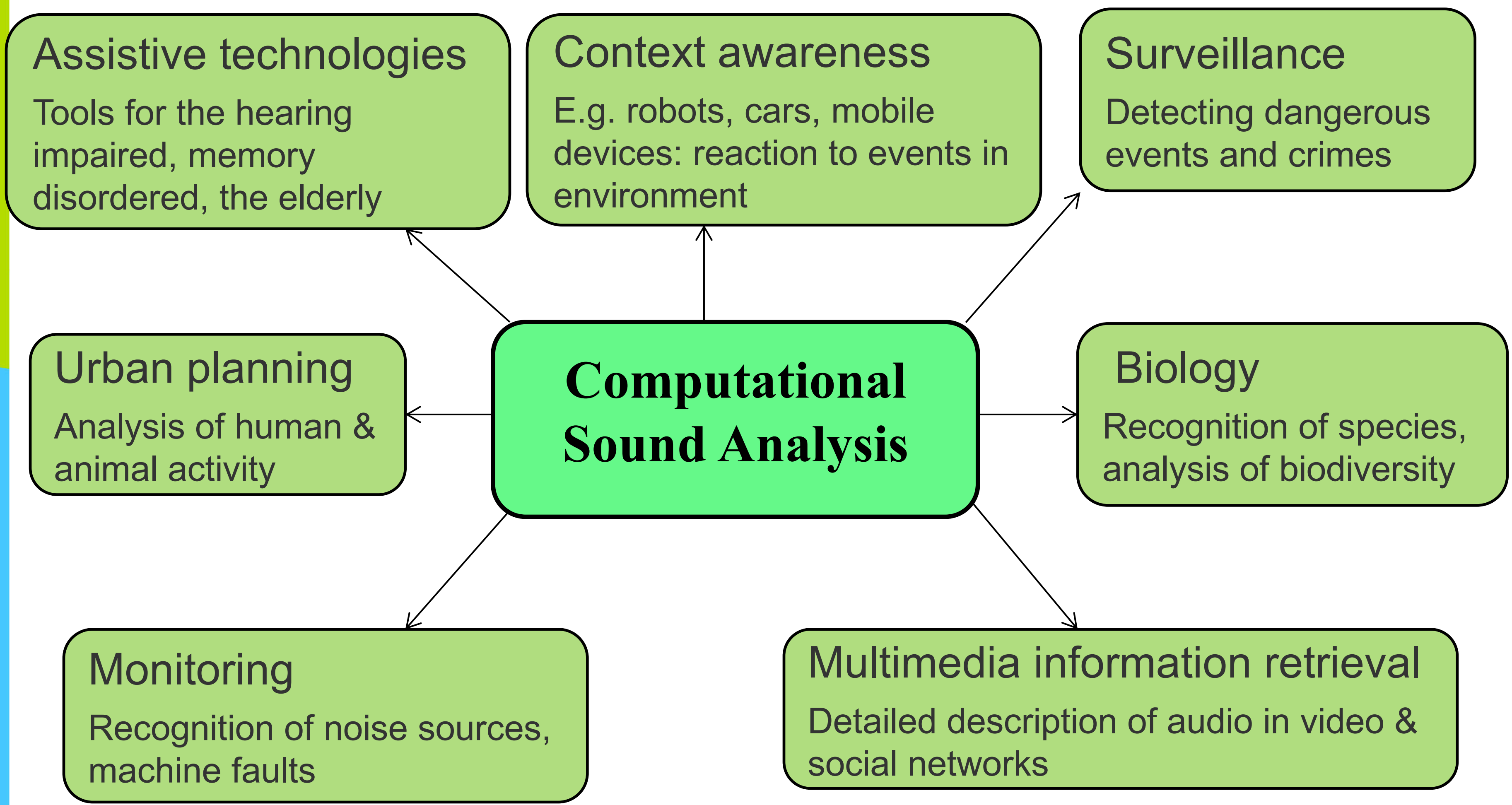
Pattern classification and beyond

- Time-varying classes -> detection
  - Estimating start and end times of classes
- Polyphonic detection: multiple overlapping classes



Example sound event labeling

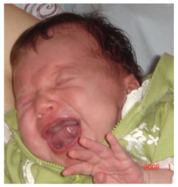
Potential applications



Scientific challenges

Large variety of different types of sounds




Scientific challenges

Large variety of different types of sounds



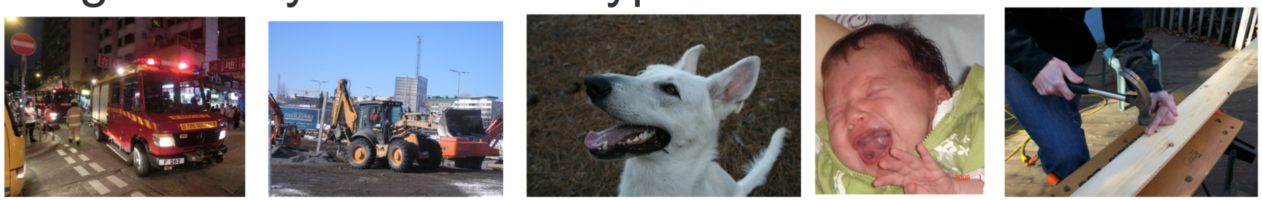


Large acoustic diversity within each category




Scientific challenges

Large variety of different types of sounds



Large acoustic diversity within each category

Overlapping sounds, reverberation

$$x(t) = \sum_n s_n(t) \star h_n(t)$$

Images by edwin.11, Robert Southworth, willc2, sannse, Baminnick, Palatkwapi, Oxyman, Brazzouk, Wistula

The supervised machine learning approach

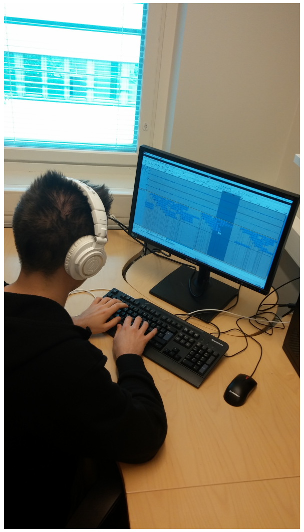
- Algoritms that find mapping between training examples (audio) and labels (annotations)
- Set of possible sound classes defined in advance
  – Defines the scope of the method
- Need for annotated training material from all the classes
  – Audio recordings and its class annotations

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?

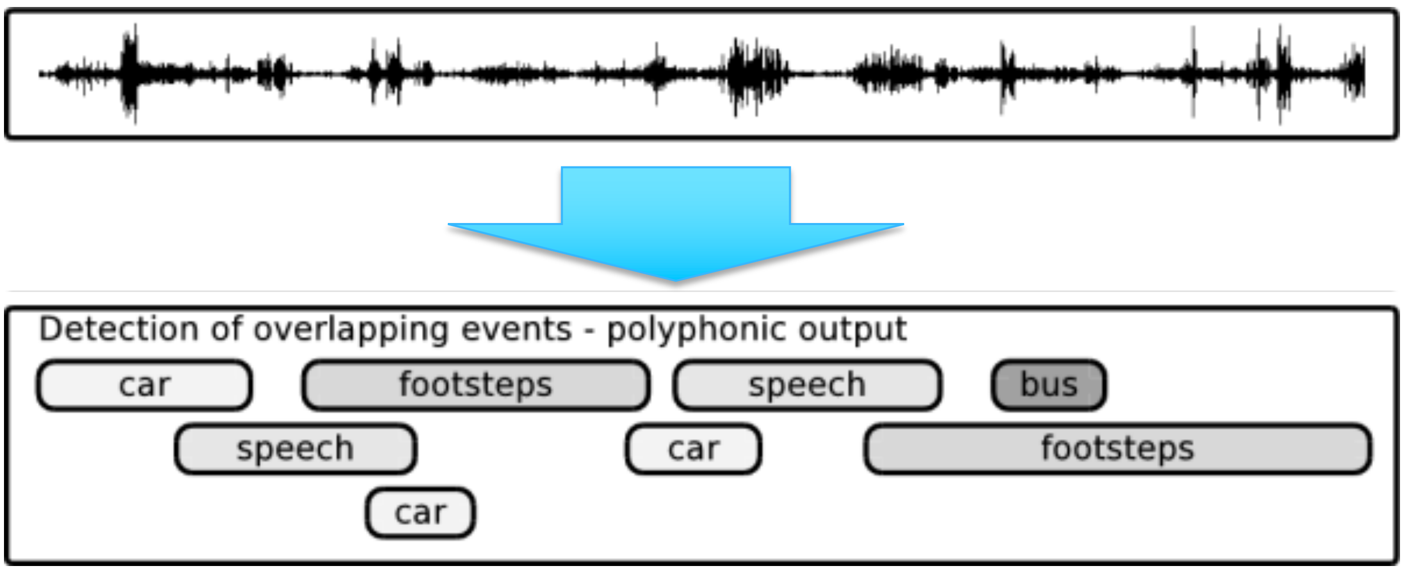
Real audio: TUT Sound Events 2016 & 2017

- 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
- Available online: http://www.cs.tut.fi/sgn/arg/dcase2016/



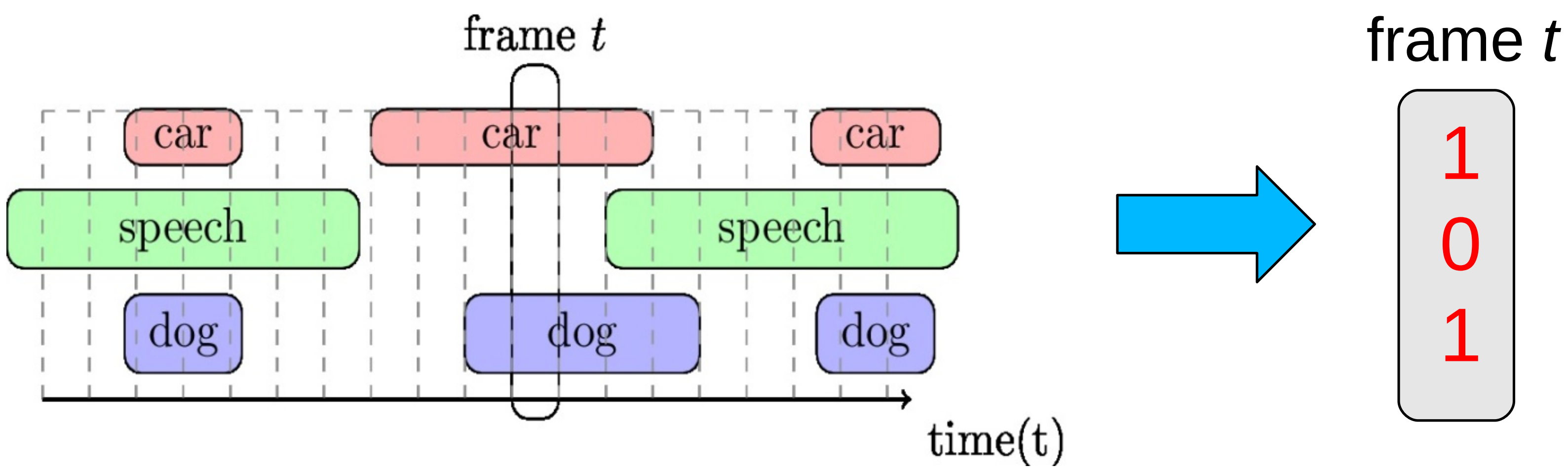
Supervised learning for polyphonic event detection

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

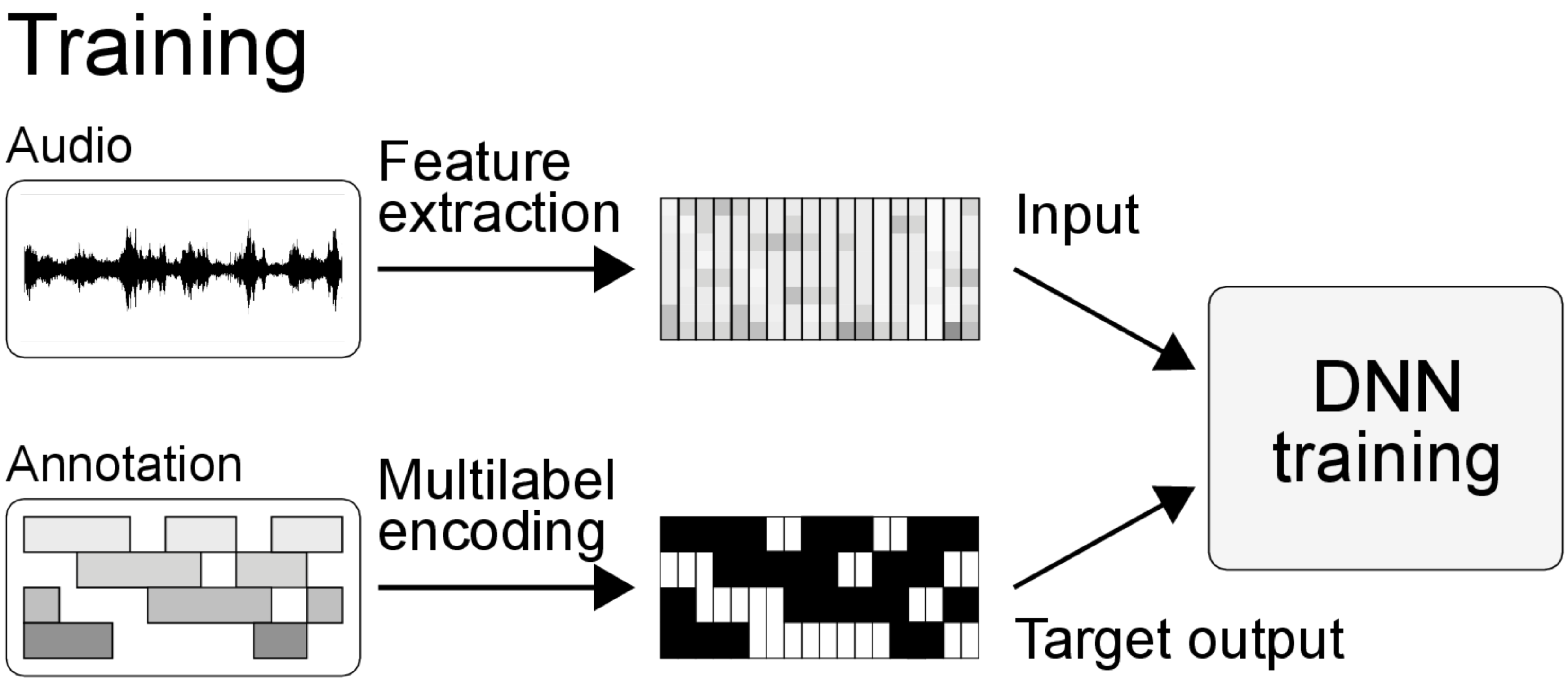


Segment-wise multilabel classification

- Binary encoding of class activities
- Predict the activity of each class in each frame

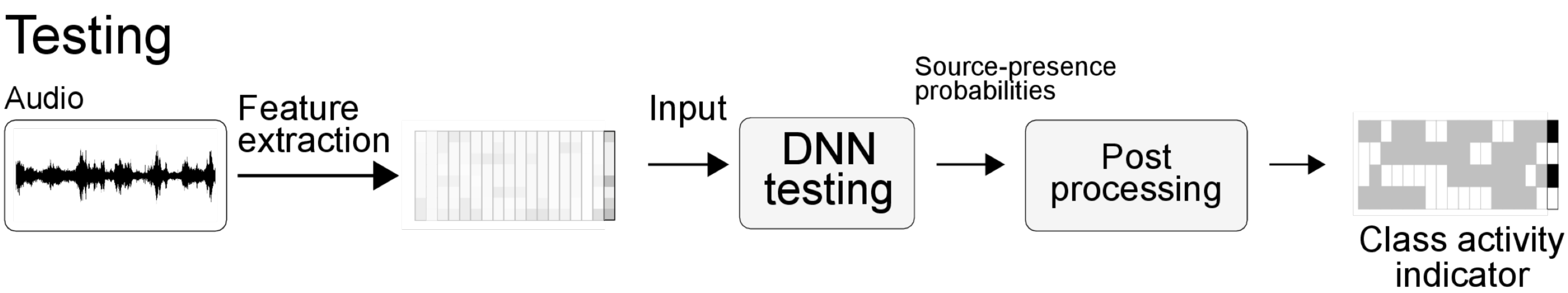


The multilabel deep neural network (DNN) approach



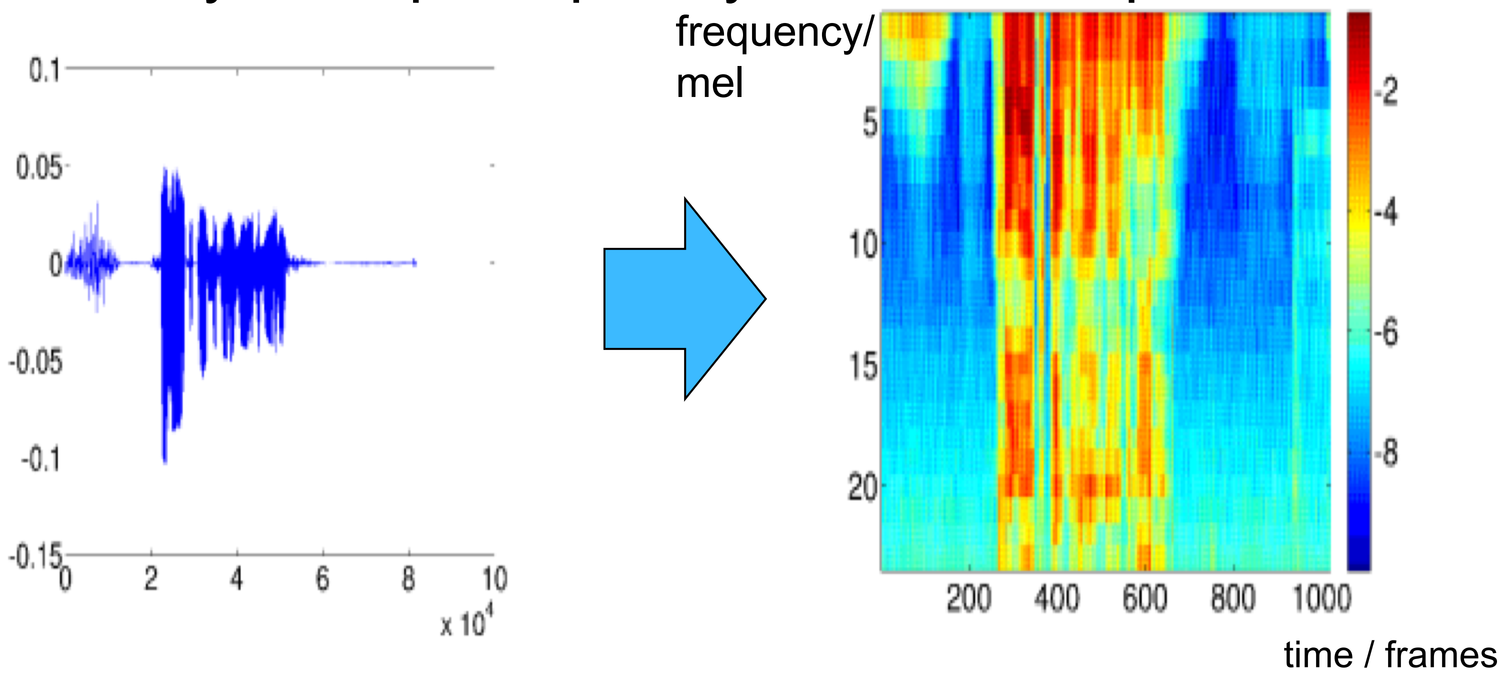
TAMPERE UNIVERSITY OF TECHNOLOGY

Multilabel DNN approach

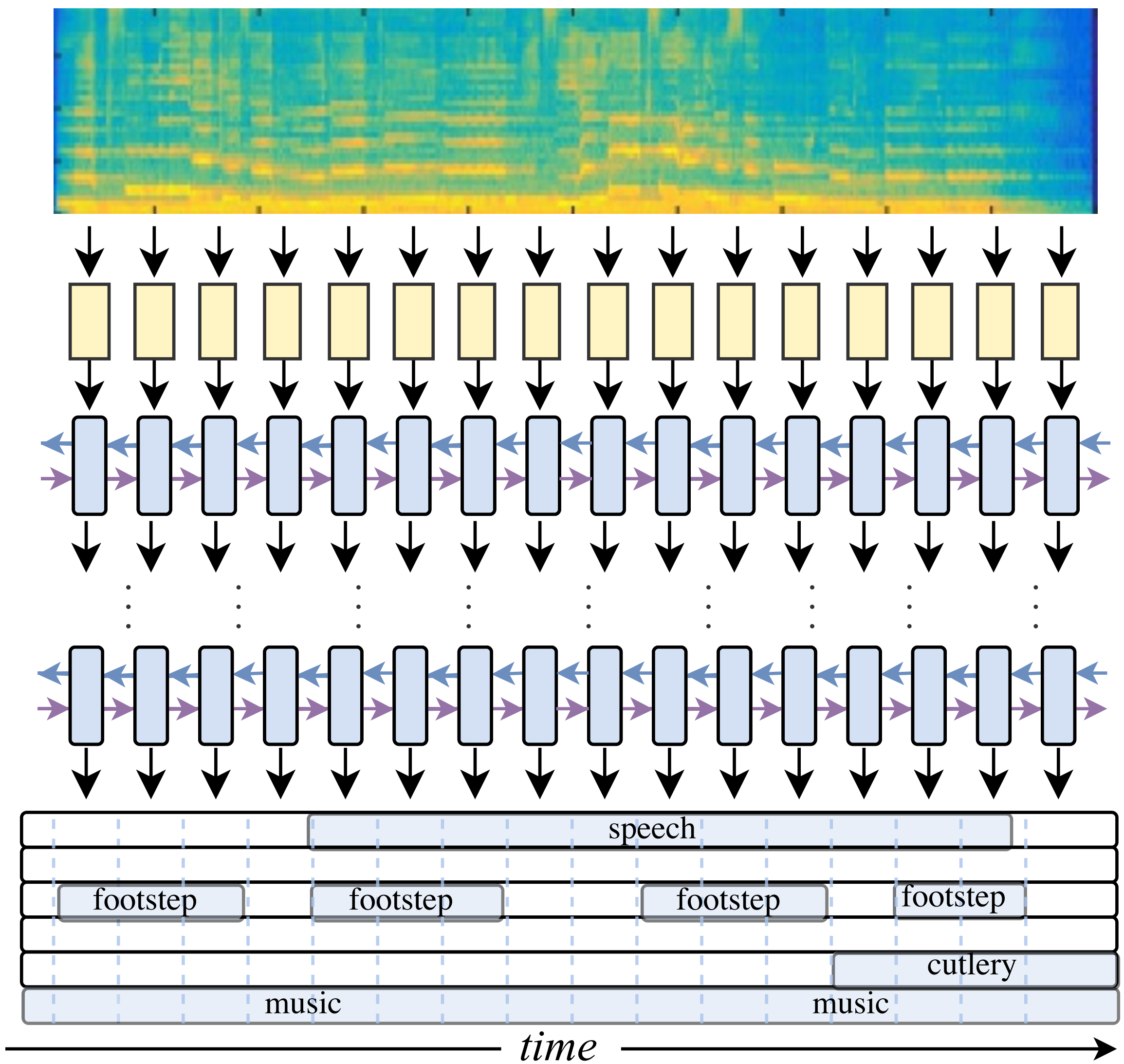


Acoustic features

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

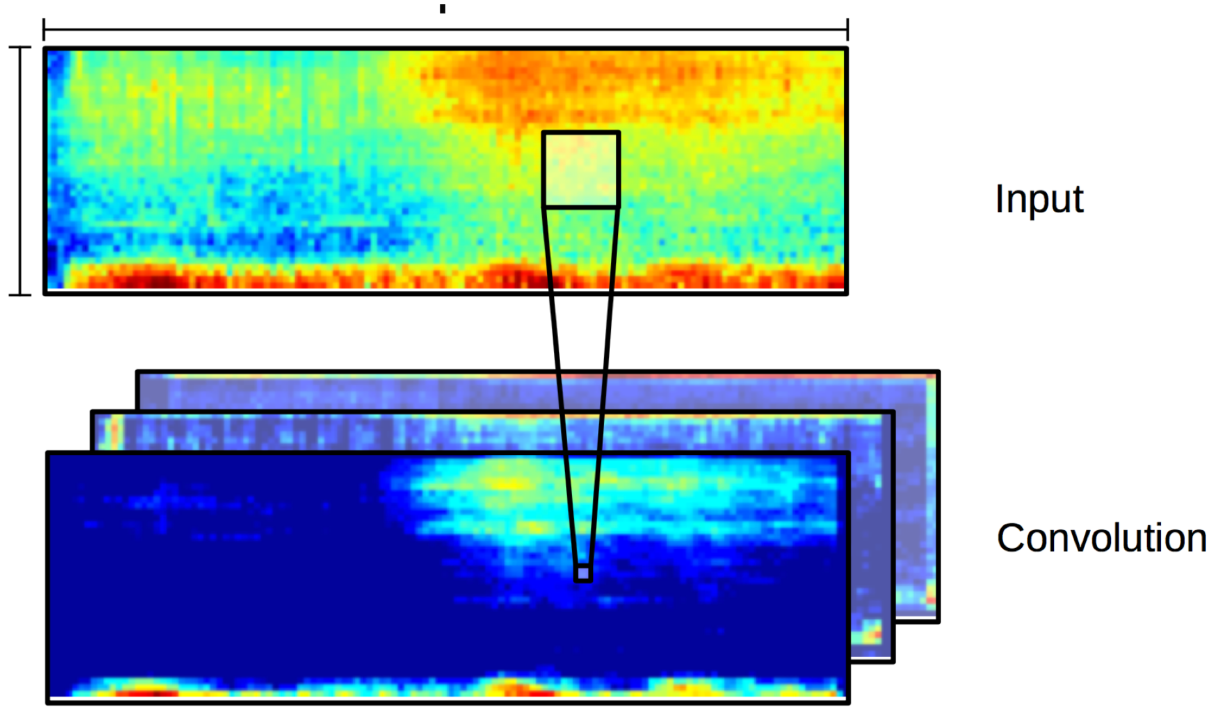


Recurrent neural network



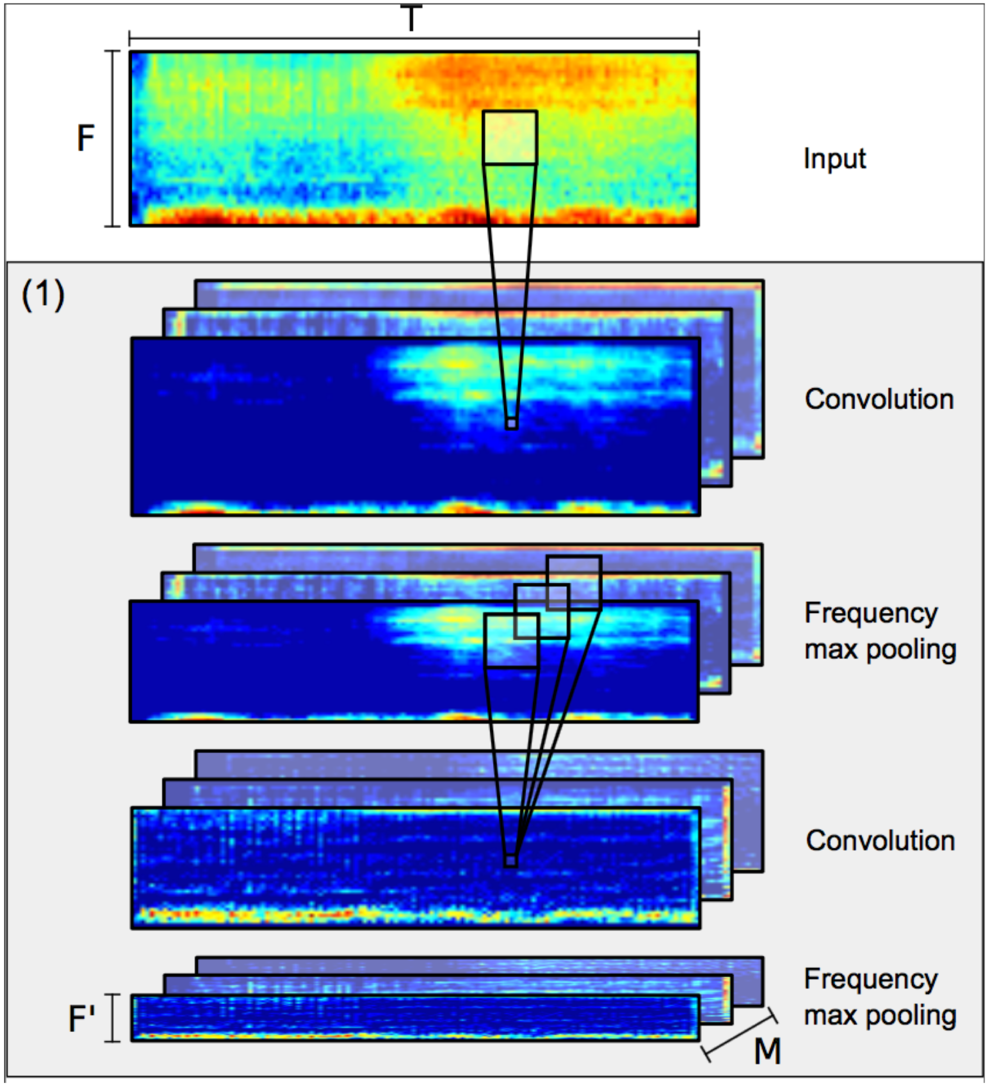
Convolutional neural networks

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



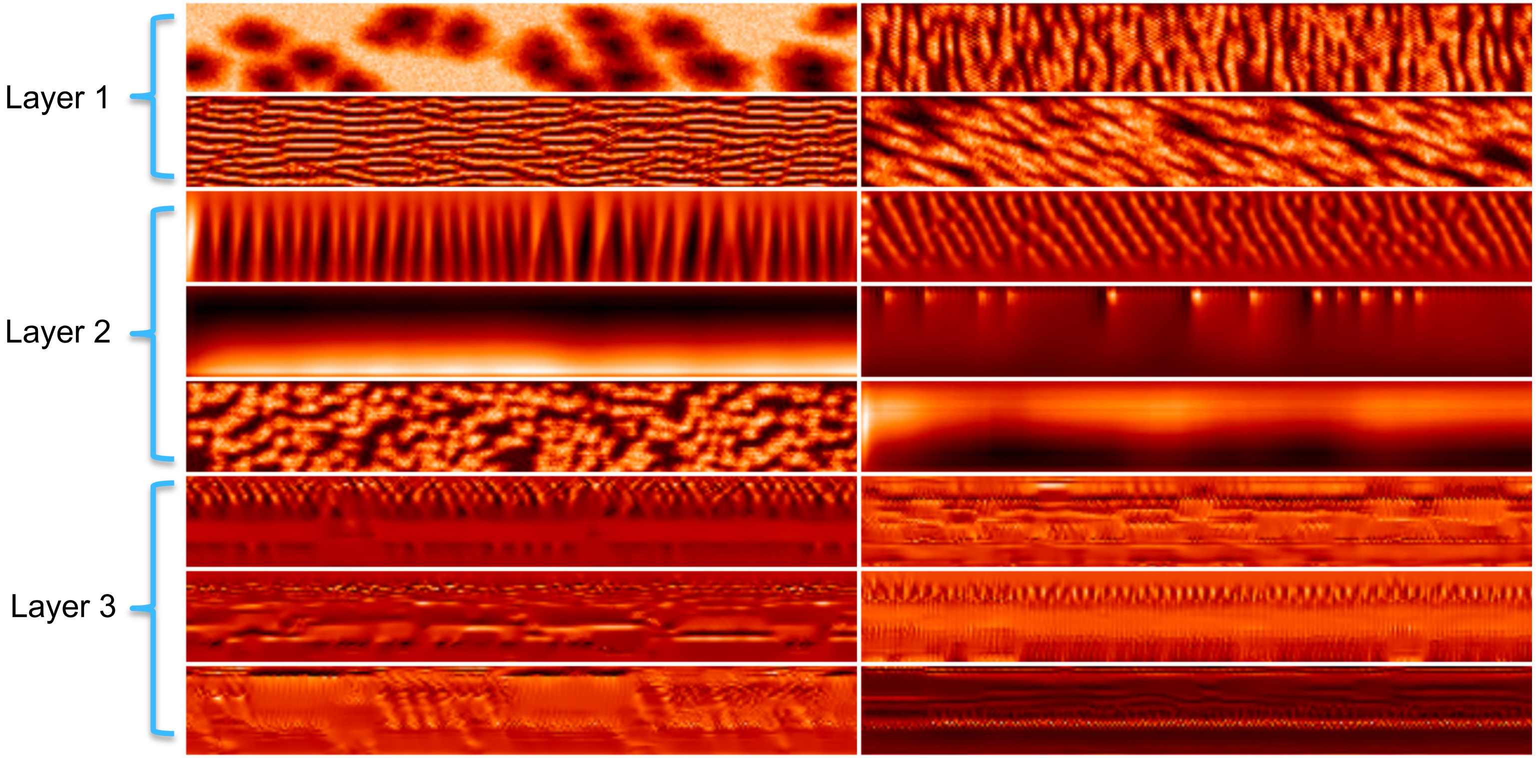
CNN

- Pooling allows learning shift-invariant features
- Multiple CNN layers allows learning higher-level features



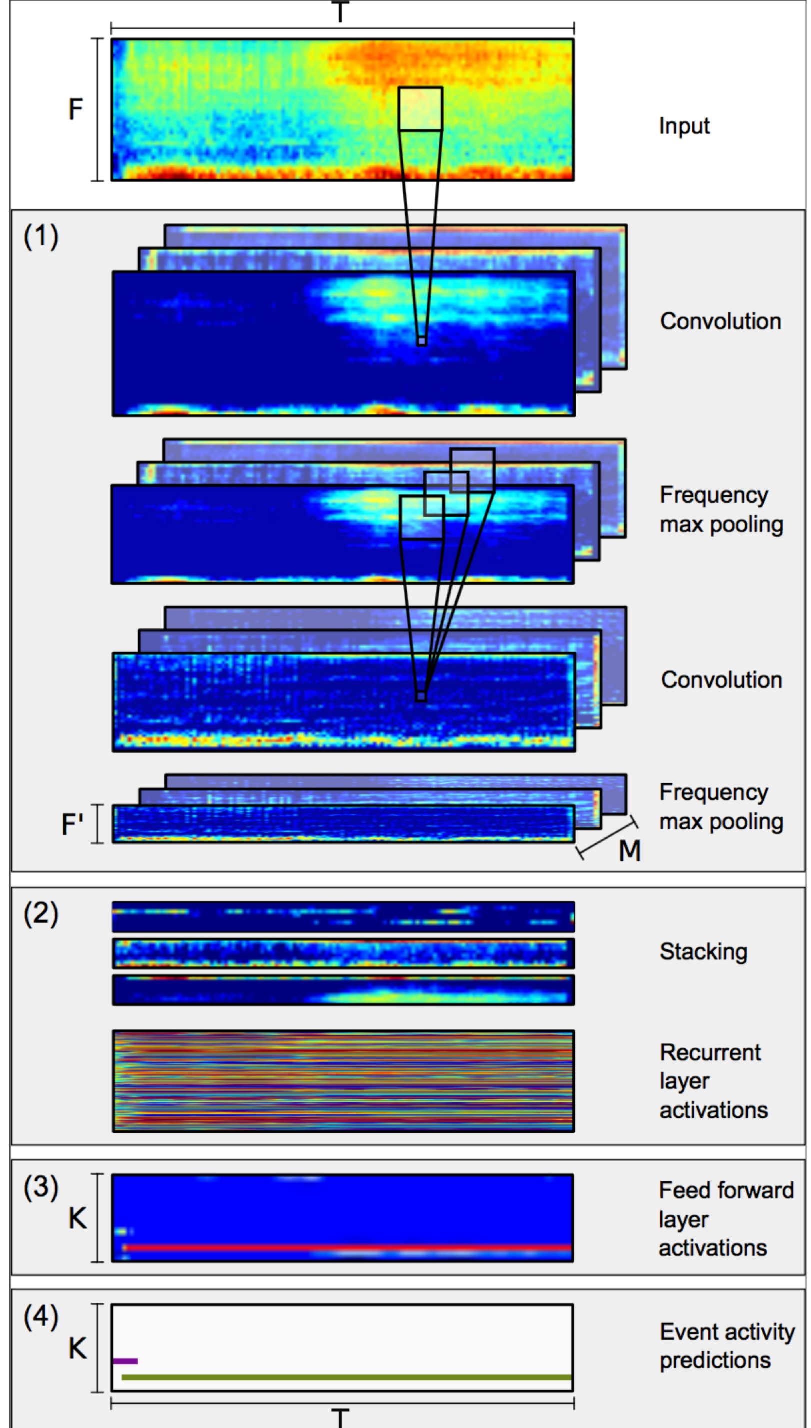
What do the CNN filters represent?

- Synthetic input maximizing the activivation of selected neurons



CRNN

- Convolutional recurrent neural network
- Convolutional layers learn features
- Recurrent layers model longer temporal context

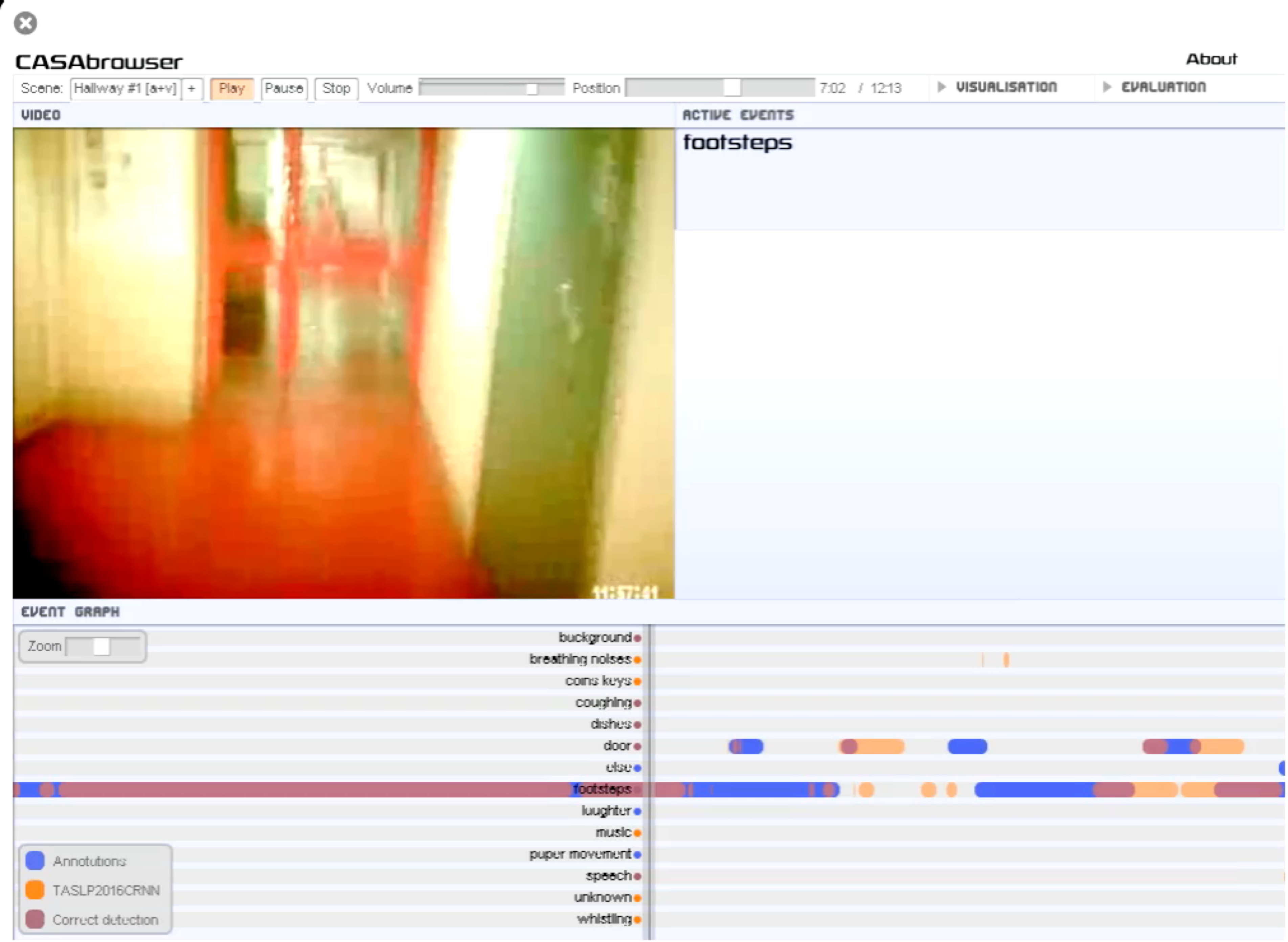


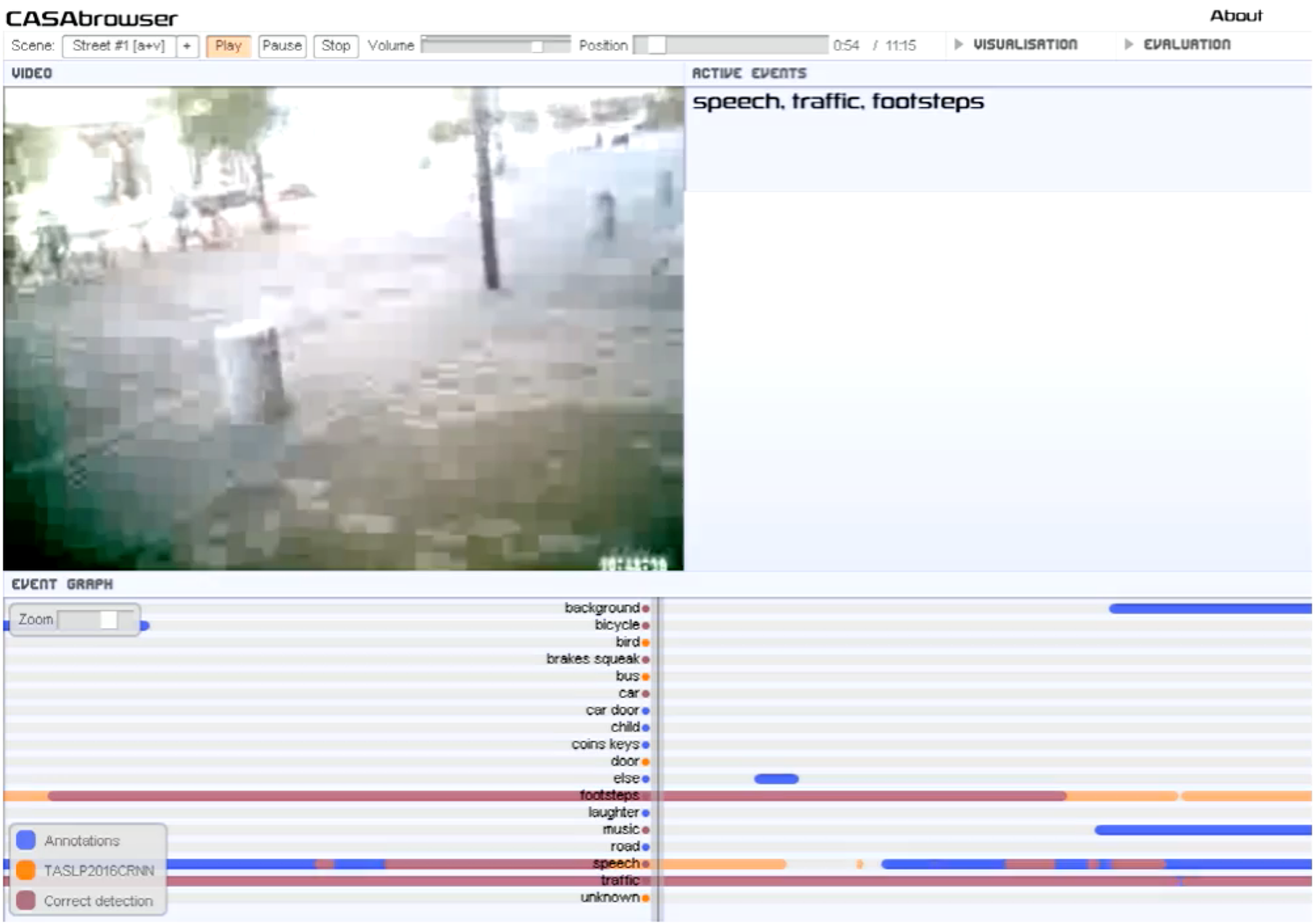
Typical CRNN parameters

- 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


CASAbrowser
About
Scene: Hallway #1 [a+v] +
Play
Pause
Stop
Volume
Position
7:02 / 12:13
VISUALISATION
EVALUATION
VIDEO
ACTIVE EVENTS
footsteps
EVENT GRAPH
Zoom
background
breathing noises
coins keys
coughing
dishes
door
else
footsteps
laughter
music
paper movement
speech
unknown
whistling
Annotations
TASLP2016CRNN
Correct detection

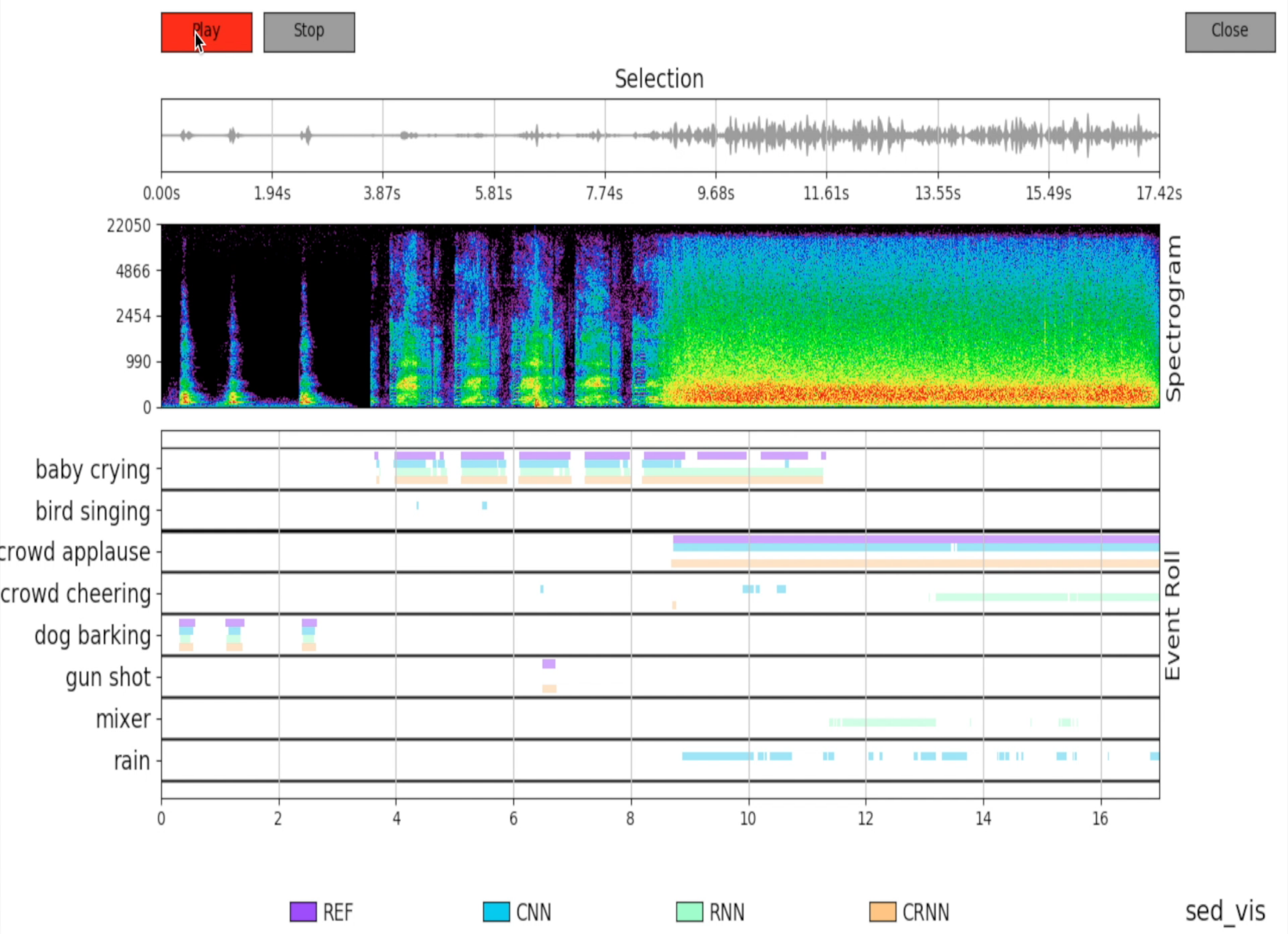


CASAbrowser
About
Scene:
Street #1 [a+v]
+
Play
Pause
Stop
Volume
Position
0:54 / 11:15
VISUALISATION
EVALUATION
VIDEO
ACTIVE EVENTS
speech, traffic, footsteps
EVENT GRAPH
Zoom
background
bicycle
bird
brakes squeak
bus
car
car door
child
coins keys
door
else
footsteps
laughter
music
road
speech
traffic
unknown
Annotations
TASLP2016CRNN
Correct detection

Objective evaluation with synthetic data

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

http://www.cs.tut.fi/sgn/arg/taslp2017-crnn-sed/tut-sed-synthetic-2016


Play
Stop
Close
Selection
0.00s
1.94s
3.87s
5.81s
7.74s
9.68s
11.61s
13.55s
15.49s
17.42s
22050
4866
2454
990
0
Spectrogram
baby crying
bird singing
crowd applause
crowd cheering
dog barking
gun shot
mixer
rain
Event Roll
0
2
4
6
8
10
12
14
16
REF
CNN
RNN
CRNN
sed_vis

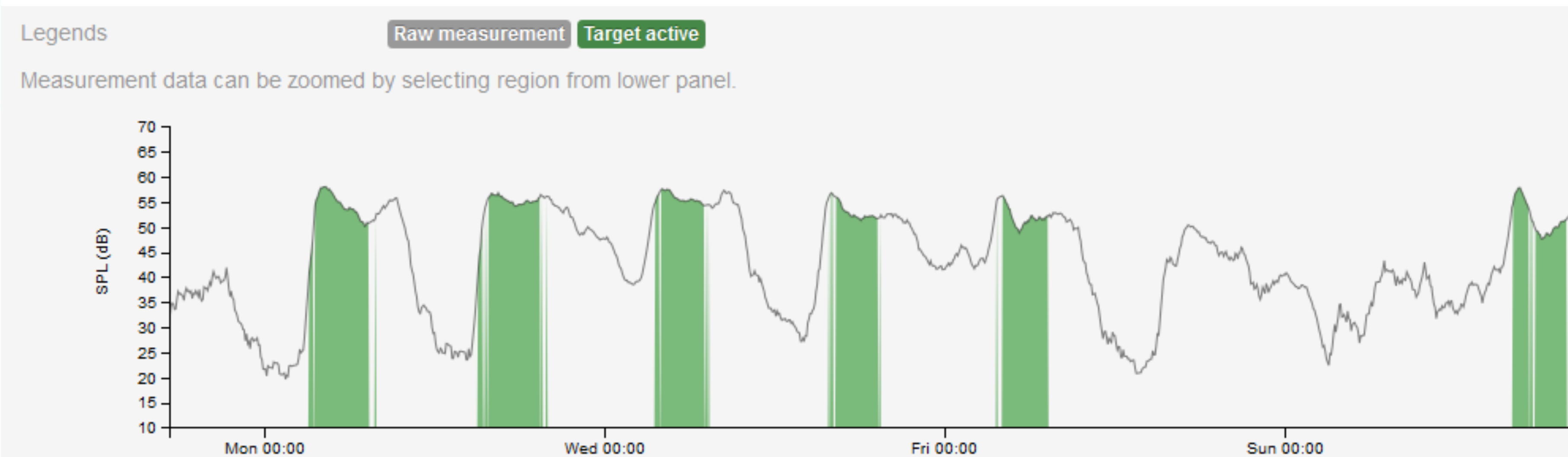
Objective evaluation with synthetic data

Classifier	F-score (framewise)
Binary GMM	40.5%
FNN	49.2%
CNN	52.8%
RNN	59.8%
CRNN	66.4%

http://www.cs.tut.fi/sgn/arg/taslp2017-crnn-sed/tut-sed-synthetic-2016

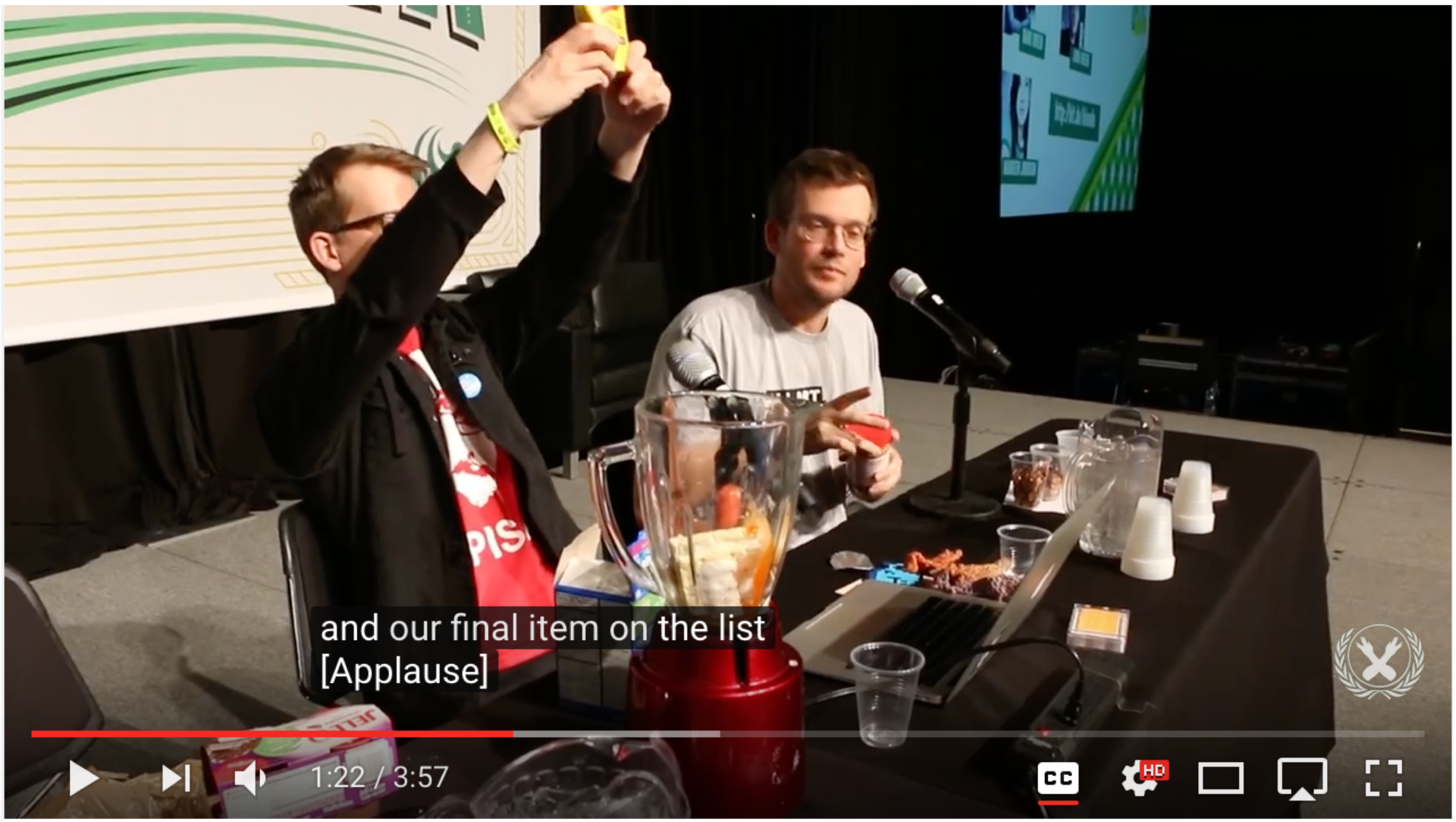
Case study: acoustic monitoring

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



Existing applications

- Automatic captioning of acoustic events in Youtube videos:



TAMPERE UNIVERSITY OF TECHNOLOGY

Existing applications

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

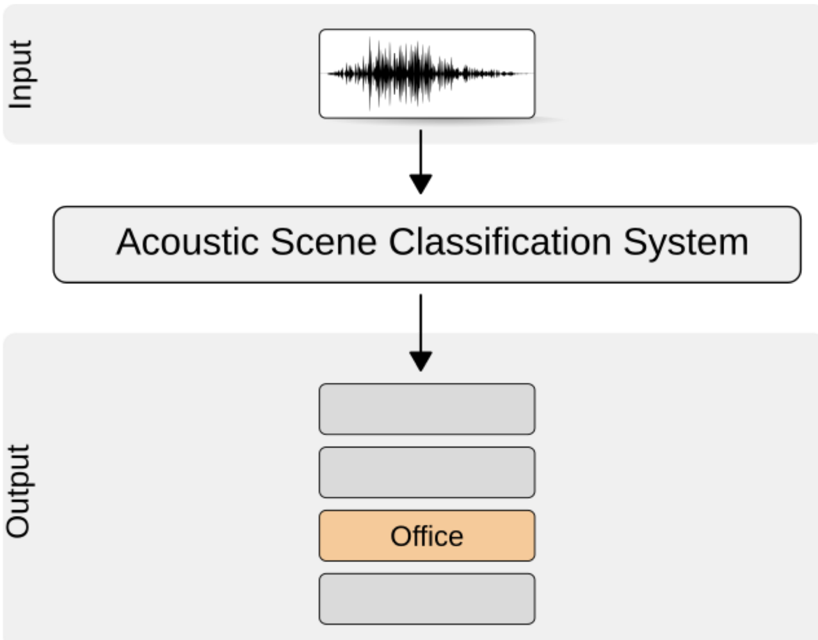
DCASE evaluation campaigns

- 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

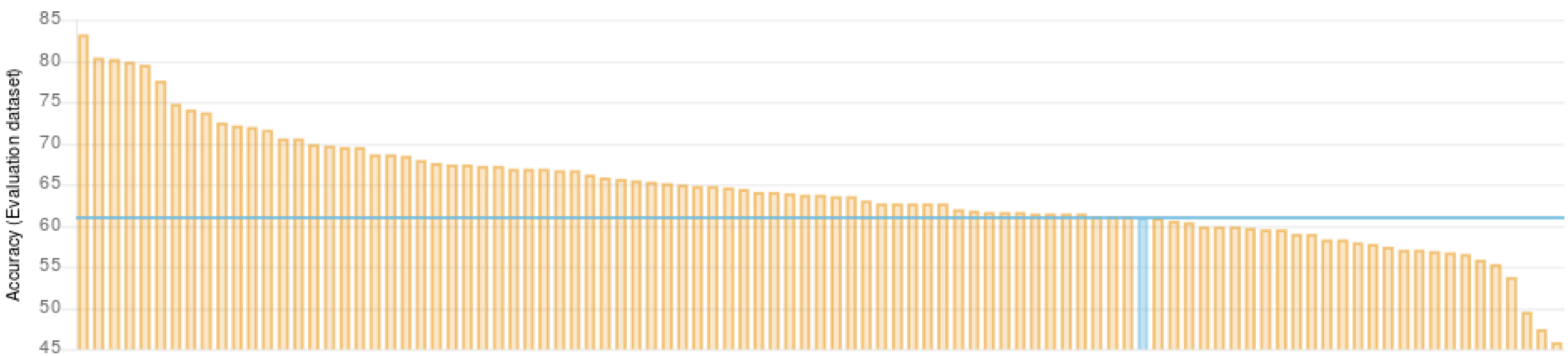


15 classes:

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

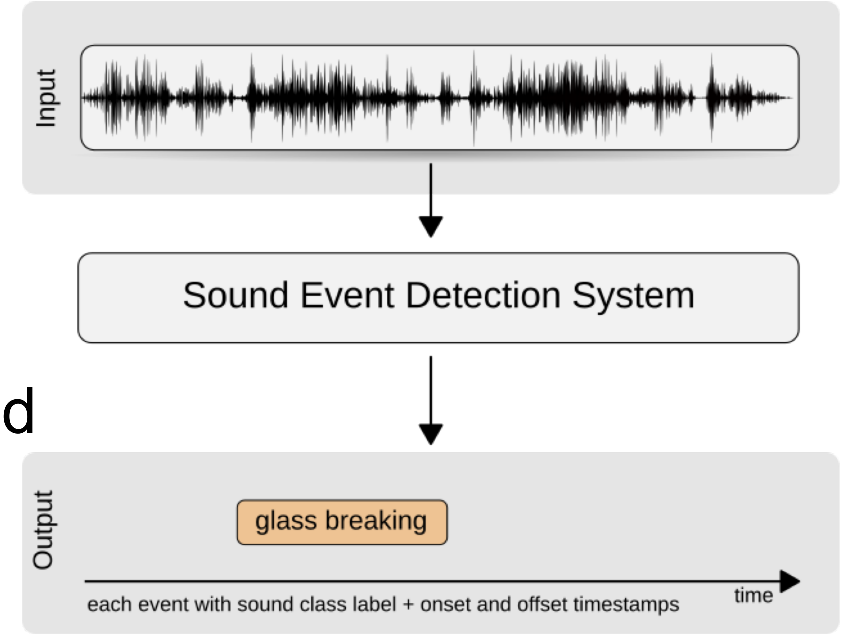
Task 1: Results

- 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



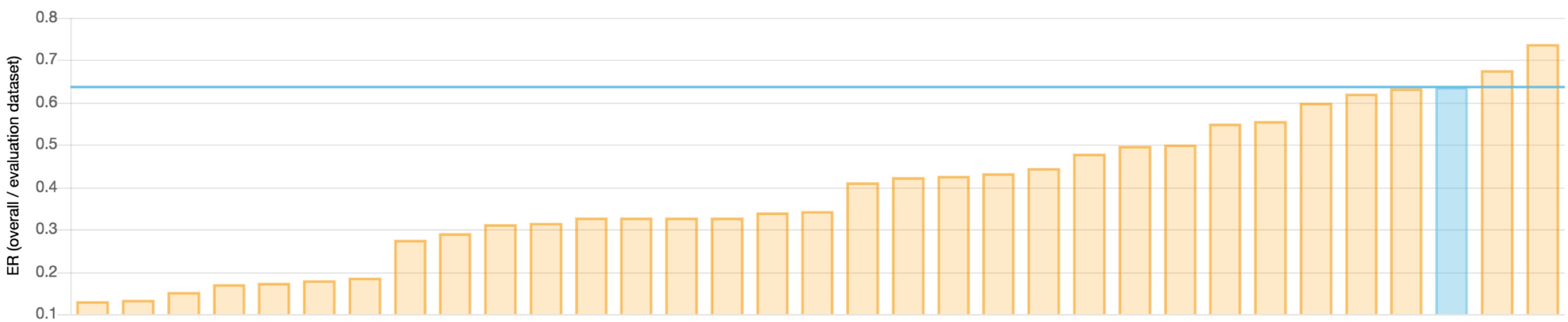
Task 2: Detection of rare sound events

- 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

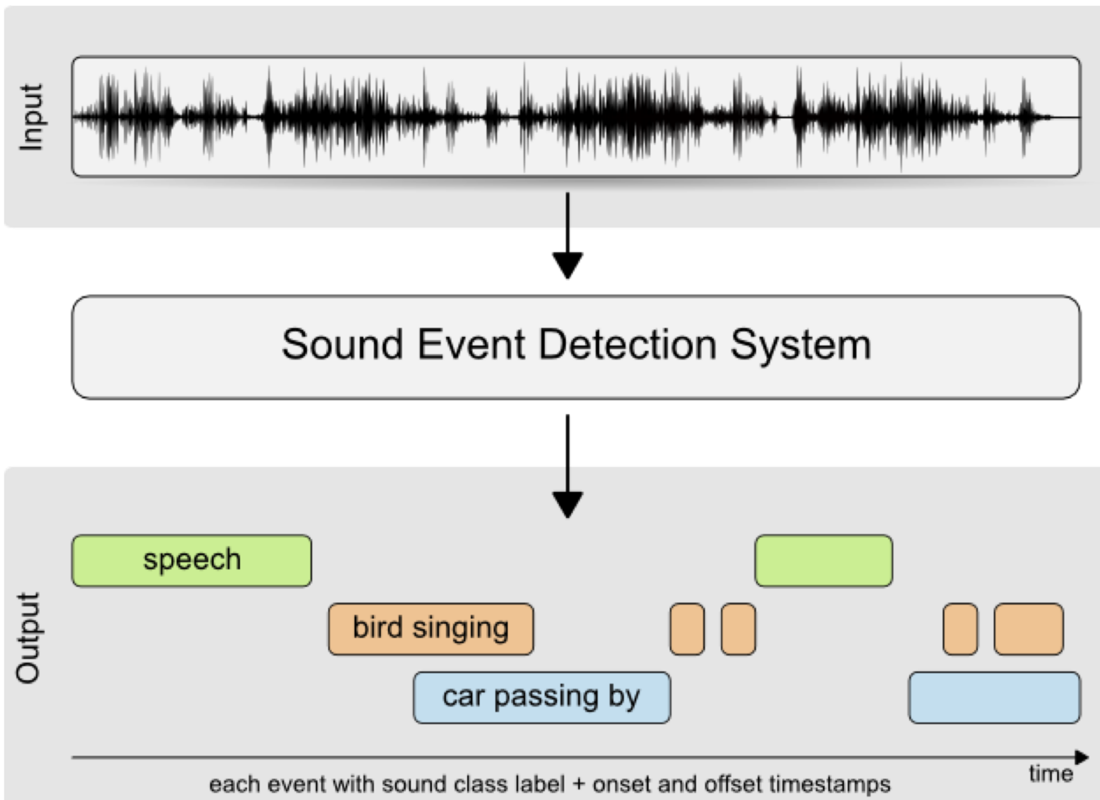


Task 2: Results

- 33 Systems (13 Teams / 38 Authors)
- Metrics:
  - event-based Error Rate (ER)
  - F1-score (secondary metric)
  - both calculated with 500ms onset collar

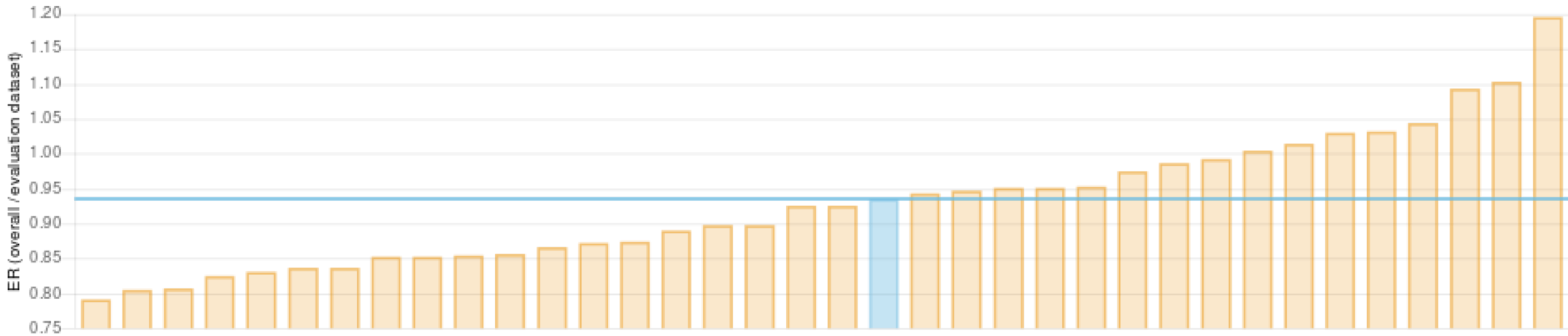


Task 3: Sound Event Detection in Real-life Audio

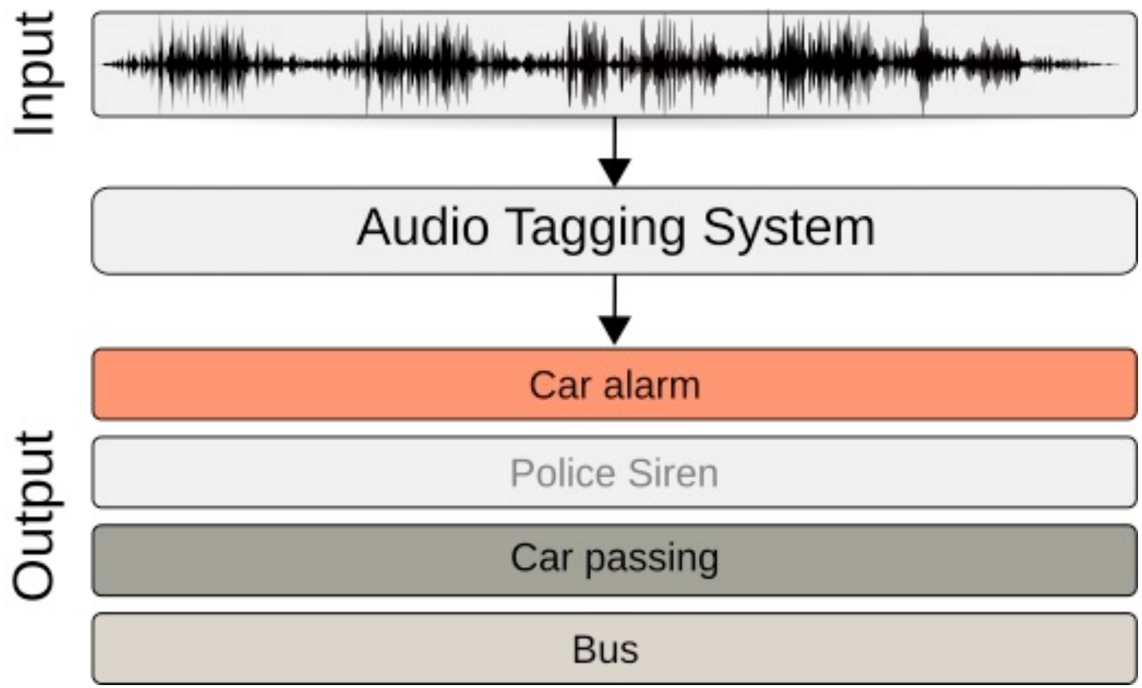


Task 3: Results

- 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%



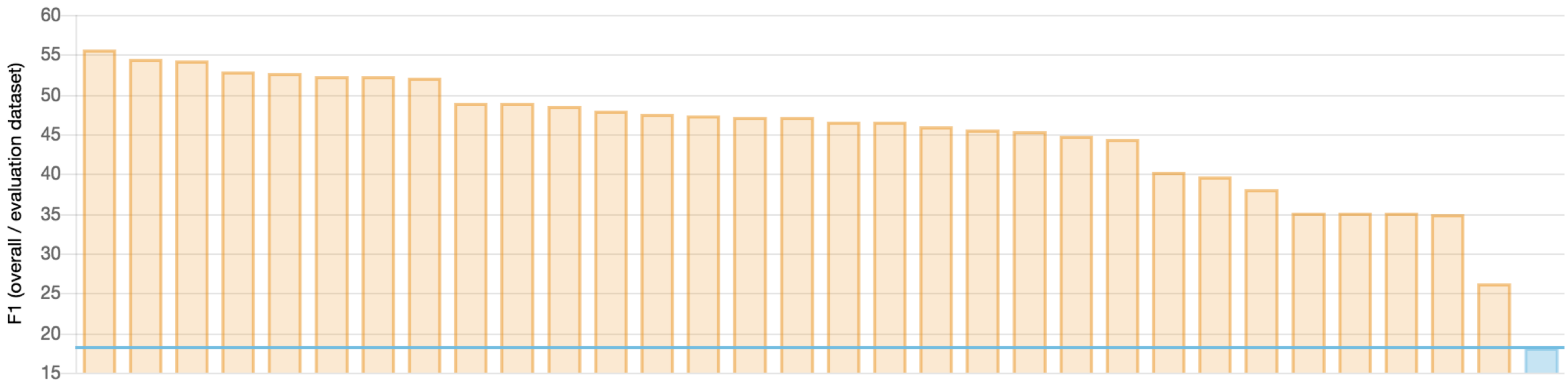
Task 4: Large-Scale Weakly Supervised Sound Event Detection for Smart Cars



Task 4: overall results

- 34 submissions / 9 Teams / 25 Authors (for both subtasks)
- Significant improvement over MLP-based baseline

Subtask A - Audio tagging



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

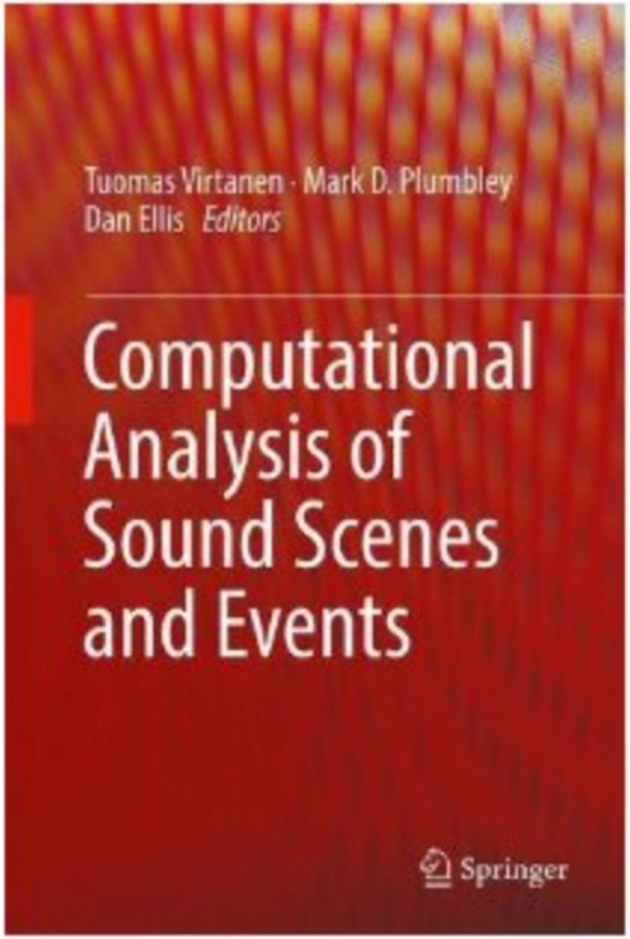
Contributors

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

References

T. Virtanen, M. D. Plumbley, D. Ellis (eds). Computational Analysis of Sound Scenes and Events. Springer, 2018.

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