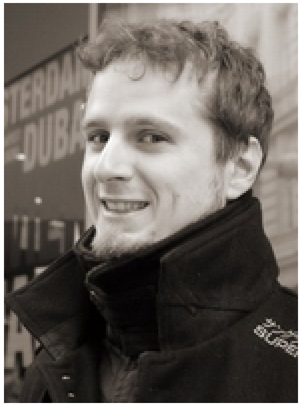
Music Information Retrieval:
Feature Extraction, Evaluation, Applications

http://www.ifs.tuwien.ac.at/mir



Alexander Schindler
Research Assistant
Information & Software Engineering Group
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~schindler

FACULTY OF !INFORMATICS

Outline

- Acoustic Scene Classification
  - Defintion
  - Approaches
- Sound Event Detection
  - Approaches
- Examples

History

1983 Bregman: ‚Auditory Scene Analysis'

1993 Computational Auditory Scene Analysis (CASA)

Development of digital hearing aids pushed CASA

2003 MFCC + Hidden Markov Models

2009 Negative Matrix Factorization, Image Features

2012 Detection and Classification of Acoustic Scenes and Events (DCASE) – *by IEEE Audio and Acoustic Signal Processing Technical Committee*

- *recognizing the general environment type (the acoustic "scene")*
- *detecting and classifying events occurring within a scene*

2016 DNN based approaches dominating

- Computational Auditory Scene Analysis (CASA)
  - Terminology based on
    - A.S. Bregman, ‚*Auditory Scene Analysis*'
    - D.L.Wang, G.J.Brown, ‚*Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.*'
  - CASA is human-centric
  - Often taken to imply an approach which aims to
    - parallel the stages of processing in human audition
    - mimic observed phenomeny of human audition
      - Including illusions and phantasms

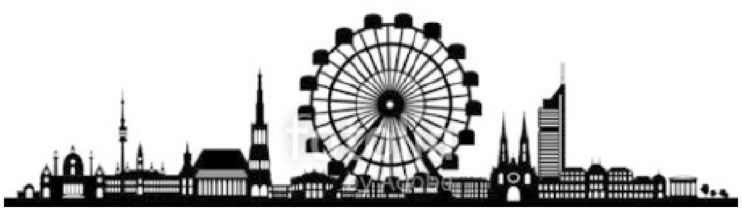
- **Acoustic Scene Classification (ASC)**
  - characterize the acoustic environment of an audio stream
  - by selecting a semantic label for it
- Machine Learning Task
  - Single-Label Classification problem
  - Similar to
    - Music Genre Recognition
    - Artist Identification
    - Speaker Recognition

Challenges

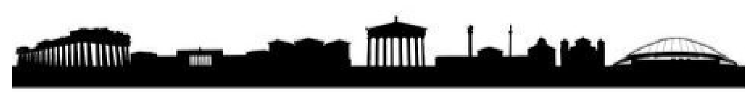
- Different Sound Scapes
  - Same acoustic scene / different city

- Different recording devices
  - Professional microphone
  - Smartphone

Vienna!

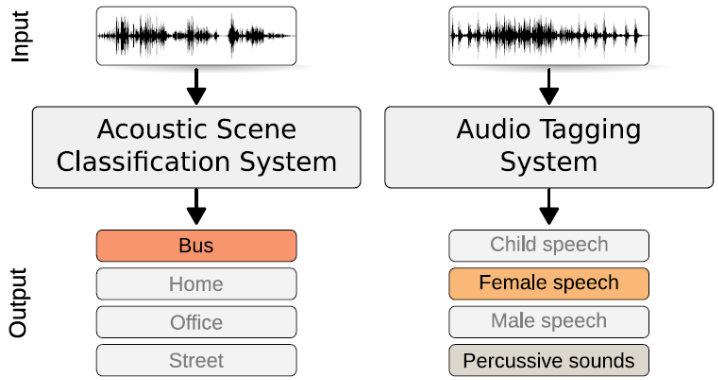


Athens!



Task Differences

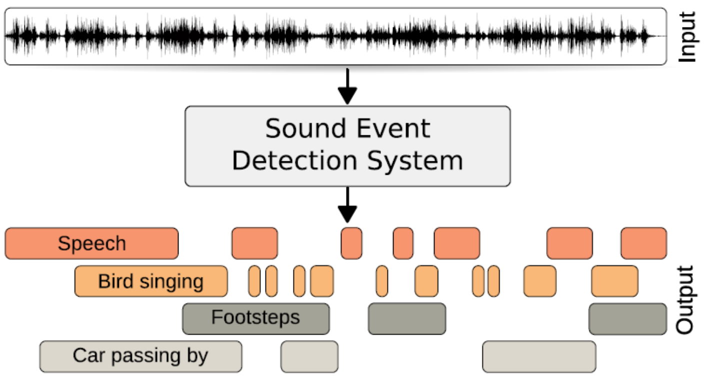
Acoustic Scene Classification



- Single Label
- No Onsets
- Entire track

- Multi Label
- No Onsets
- Entire track

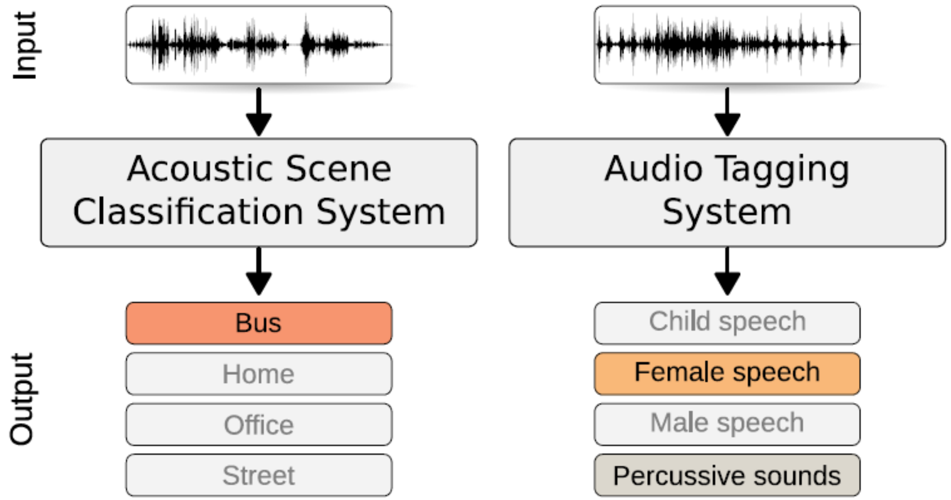
Sound Event Detection



- Multi Label
- With Onsets and length

A. Mesaros *et al.*, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379-393, Feb. 2018.

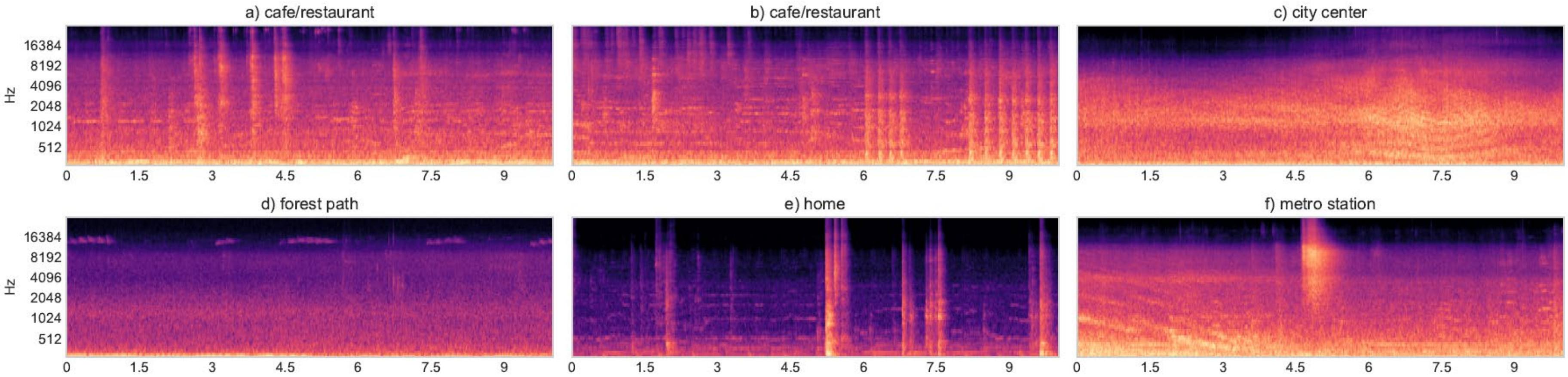
Acoustic scene classification and **Audio Tagging**



A. Mesaros *et al.*, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379-393, Feb. 2018.

Approaches

- Two main strategies
  - „Bag-of-frames“ approach using sets of low-level features
  - Set of High-Level Features
    - Vocabulary of acoustic atoms

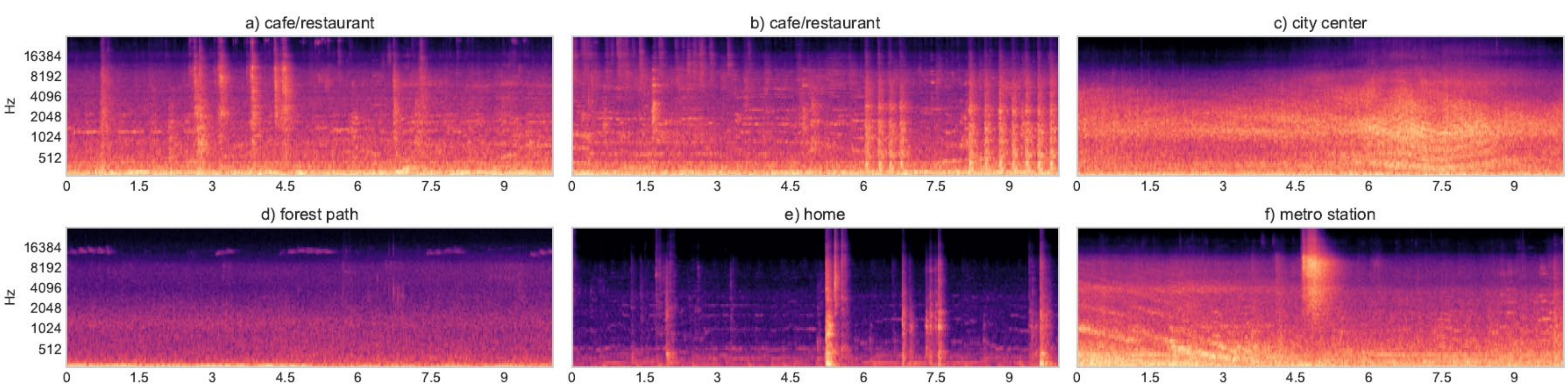


Bag-of-frames Approach

- Scene/Object is represented as
  - long-term statistical distribution of local spectral features
  - Most common: Mel-frequency Cepstral Coefficients (MFCCs)
- Standard Approach
  - Constructing a Gaussian Mixture Model (GMM) for each class

Set of High-Level Features

- Vocabulary of acoustic atoms is learned
- Non-negative Matrix factorization (NMF)
  - Extract bases
  - Convert to MFCC
  - Use for classification

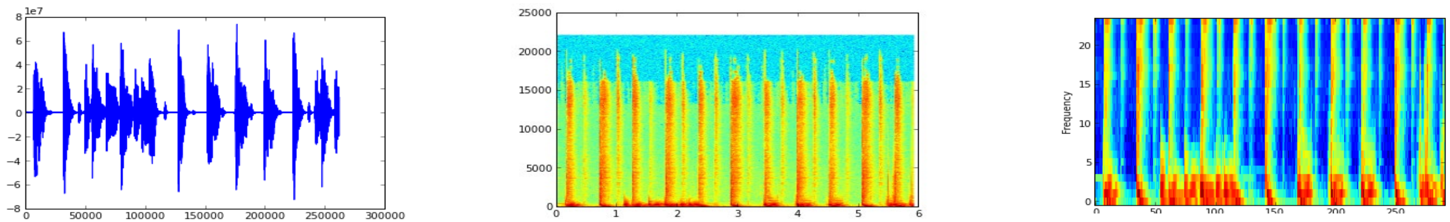


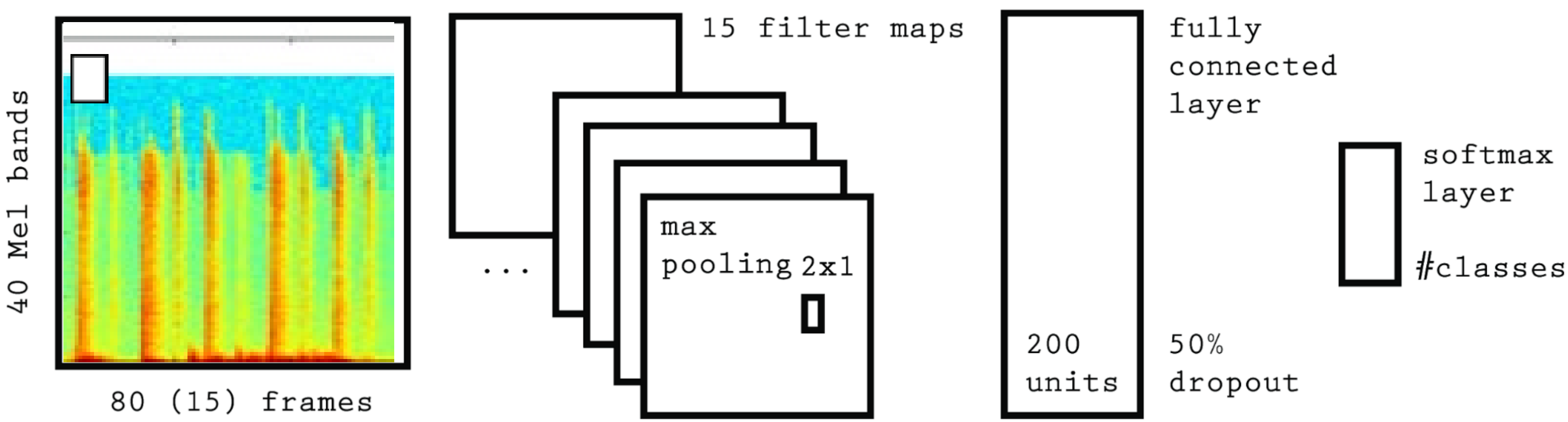
Deep Learning

- Pros:
  - A powerful method for supervised learning
  - Convolutional Neural Networks (CNNs)
  - Spectrograms as images
  - Feature Learning
  - Successfully applied on images, speech and music
- Cons:
  - Confusion of classes when dealing with noisy scenes and blurry spectrograms
  - Lack of generalization and overfitting if the training data does not contain various sessions
  - General tendency for overfitting in audio due to high self-similarity and low variance in spectrograms

Deep Learning for Music IR

Pre-Processing: Waveform → Spectrogram → 40 Mel bands → Log scale

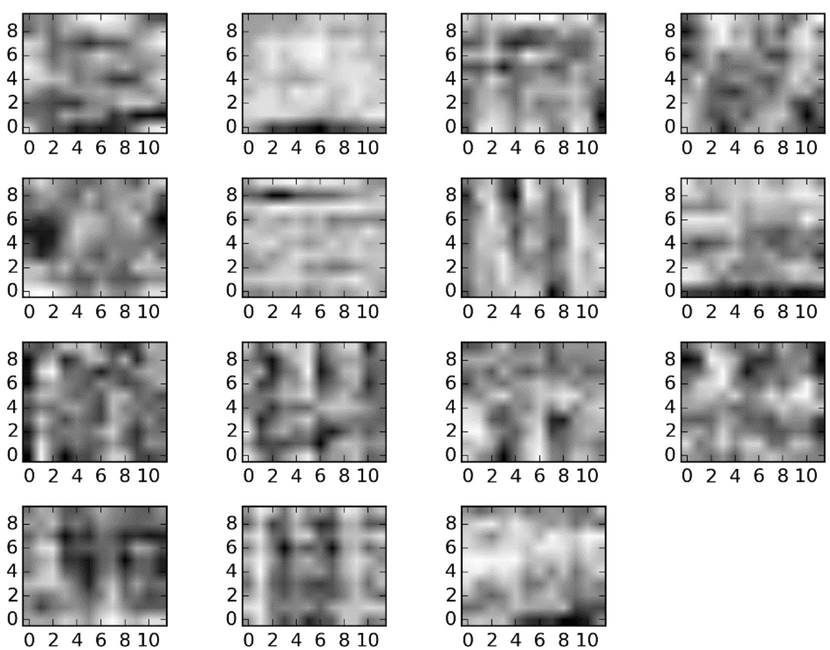




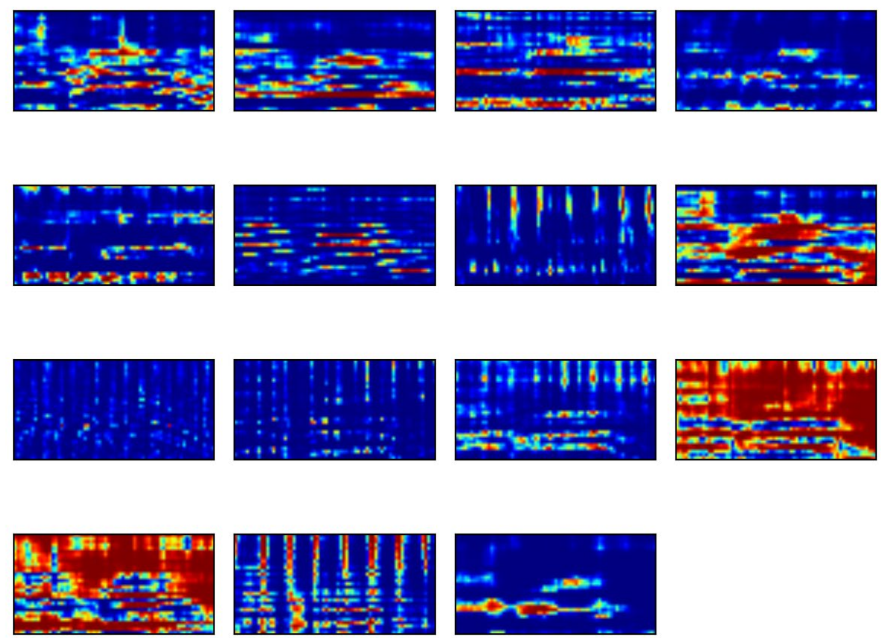
Winning algorithm MIREX 2015 music/speech classification task (99.73%) by Thomas Lidy

Visualizing CNN Filters learned for Music/Speech Classification

Learned Filter Weights

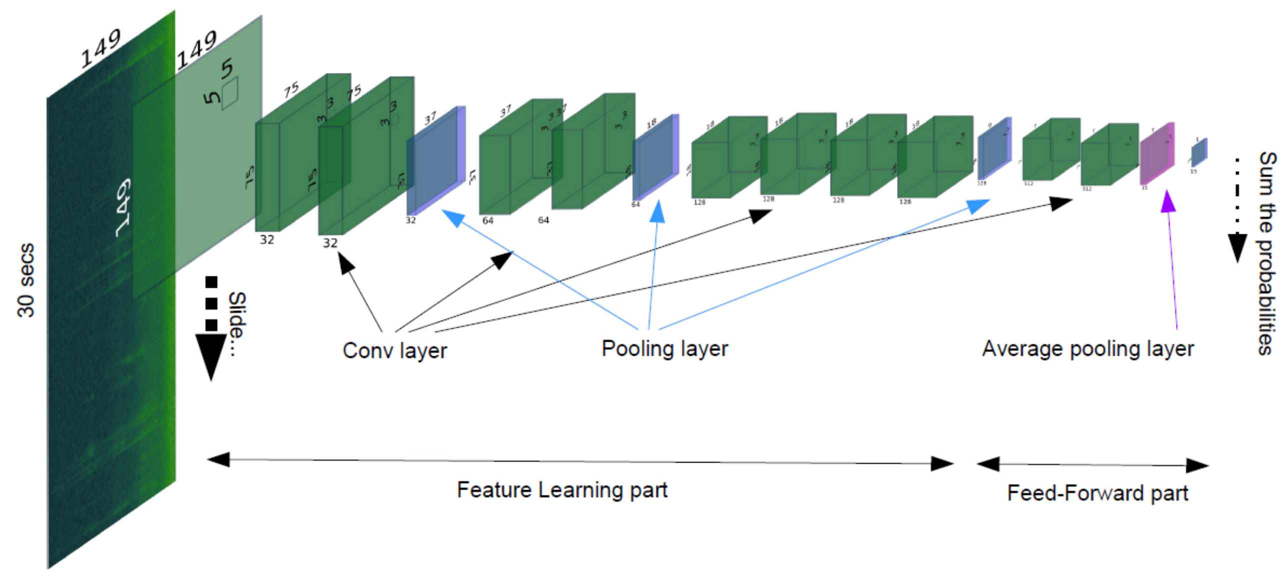


Convolved Spectrograms

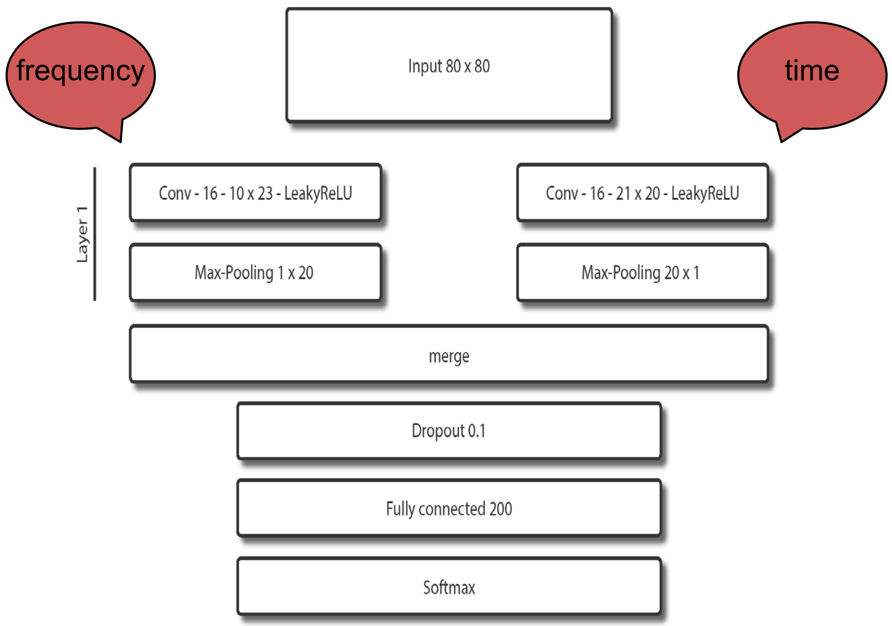


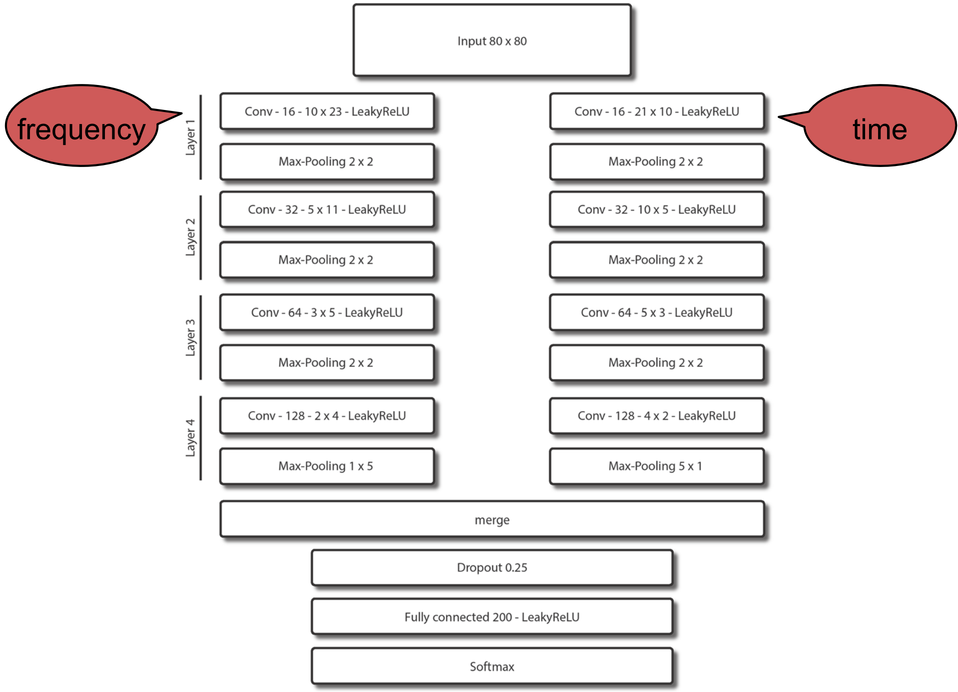
VGG Style CNN

- Most Common Convolutional Neural Network (CNN) Architecture
- Also very common in Audio Analysis



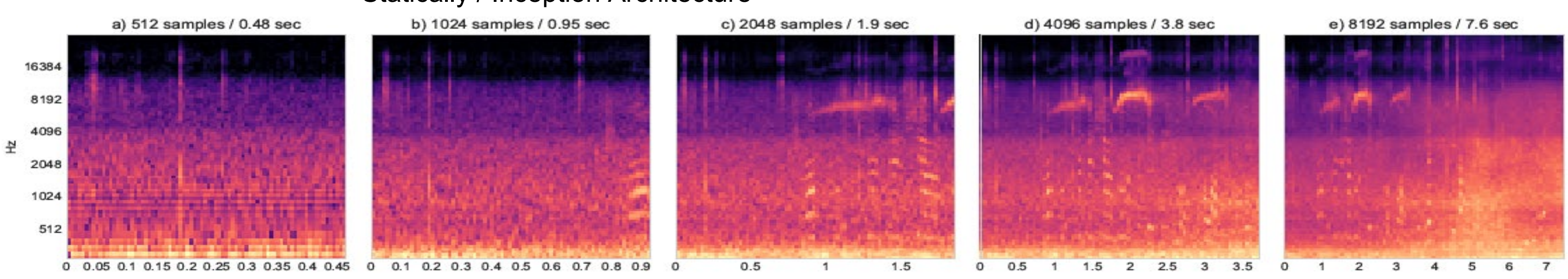
Parallel Architecture



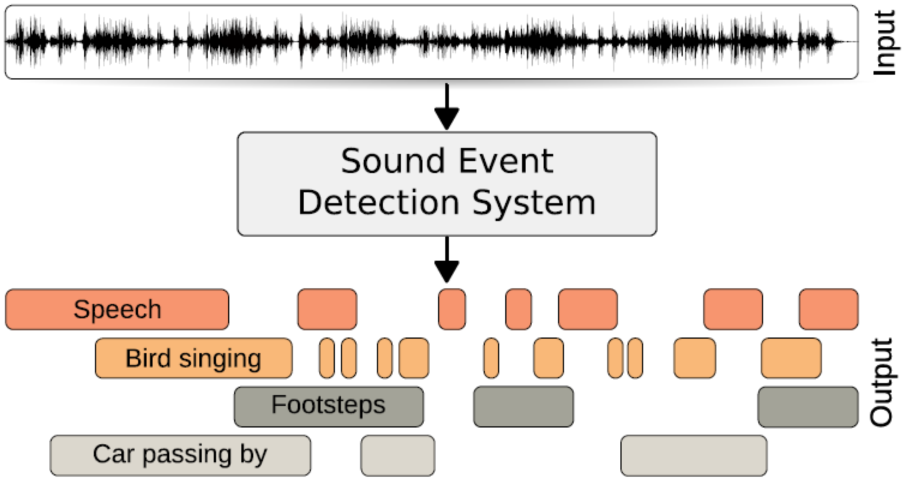


	100 epochs		200 epochs	
	Shallow	**Deep**	**Shallow**	**Deep**
GTZAN	78.1	**78.6**	**80.8**	80.6
ISMIRgenre	**85.5**	84.1	84.9	**85.1**
Latin	92.4	**94.4**	93.5	**95.1**
MSD	63.9	**67.2**	/	/

- Temporal resolution is critical
  - High temporal resolution (zooming out)
    - pro: sound structure, structured acoustic events
    - con: Fluctuation patterns get lost
  - Low temporal resolution (zooming in)
    - pro: phase and flucuations (e.g. difference between Truck and Car)
    - con: structure missing
  - Solution
    - Find a compromise (tune resulution as parameter)
    - Use multiple samples per track (random/structured) + aggregation (majority vote, max, sum, avg)
    - Use multiple resolutions
      - Statically / Inception Architecture



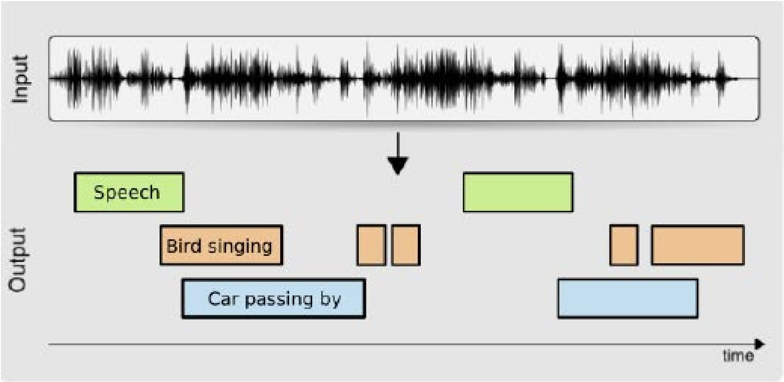
Sound event detection



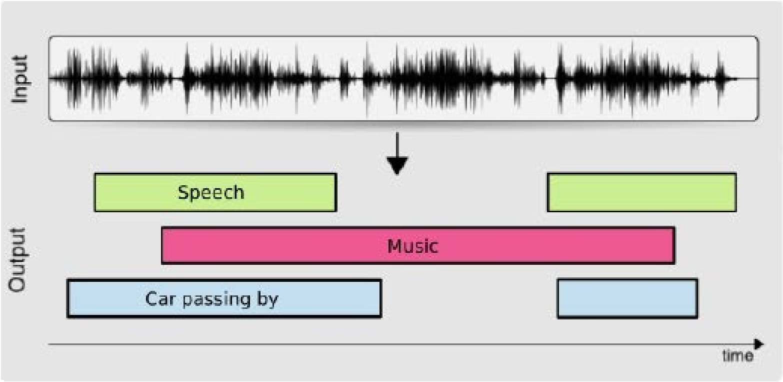
A. Mesaros *et al.*, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379-393, Feb. 2018.

Sound Event Detection

- Identify Sounds by a predefined set of classes
  - Detect Events
  - Categorize Events
- Two main Approaches
  - Detect Onsets => Classification
  - Moving Window Classification => Interpret peaks in classification results
  - New: Integrated Neural Network based approaches
- More Complicated
  - Multi-event Detection



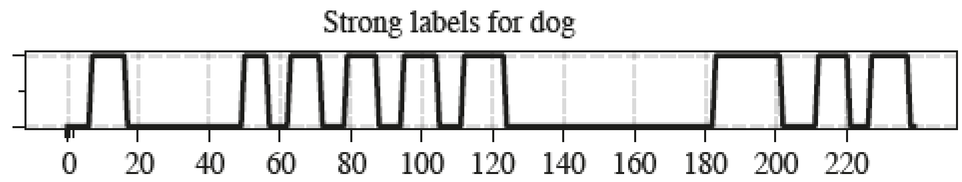
Park



City Center

Ground Truth Label Types

- **Strong Labels**
  - High precission (~0.5s)
  - Per class labelling
  - Expensive
  - Datasets usually small



- **Weak labels**
  - Low precission (~10s)
  - Multi class labelling
  - „cheap“
  - Large Datasets

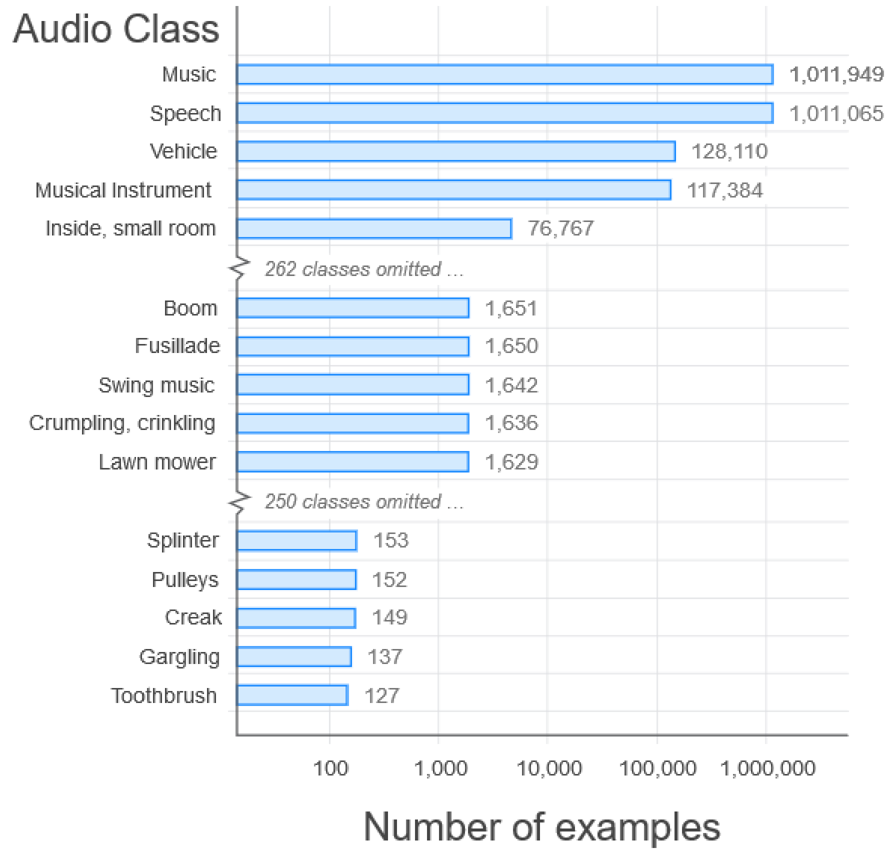
Dog, Wind, Speech

0s 10s

Harb, Robert, and Franz Pernkopf. "Sound event detection using weakly-labeled semi-supervised data with GCRNNS, VAT and Self-Adaptive Label Refinement." *arXiv preprint arXiv:1810.06897* (2018).

Google Audio Set

- 2M Videos
- 632 audio events
- annotaded according acoustic categories
- Weakly labelled (10s)
- Currently largest source of data



Human sounds
- Human voice
- Whistling
- Respiratory sounds
- Human locomotion
- Digestive
- Hands
- Heart sounds, heartbeat
- Otoacoustic emission
- Human group actions

Source-ambiguous sounds
- Generic impact sounds
- Surface contact
- Deformable shell

Animal
- Domestic animals, pets
- Livestock, farm animals, working animals
- Wild animals

Sounds of things
- Vehicle
- Engine
- Domestic sounds, home sounds
- Bell
- Alarm
- Mechanisms
- Tools
- Explosion

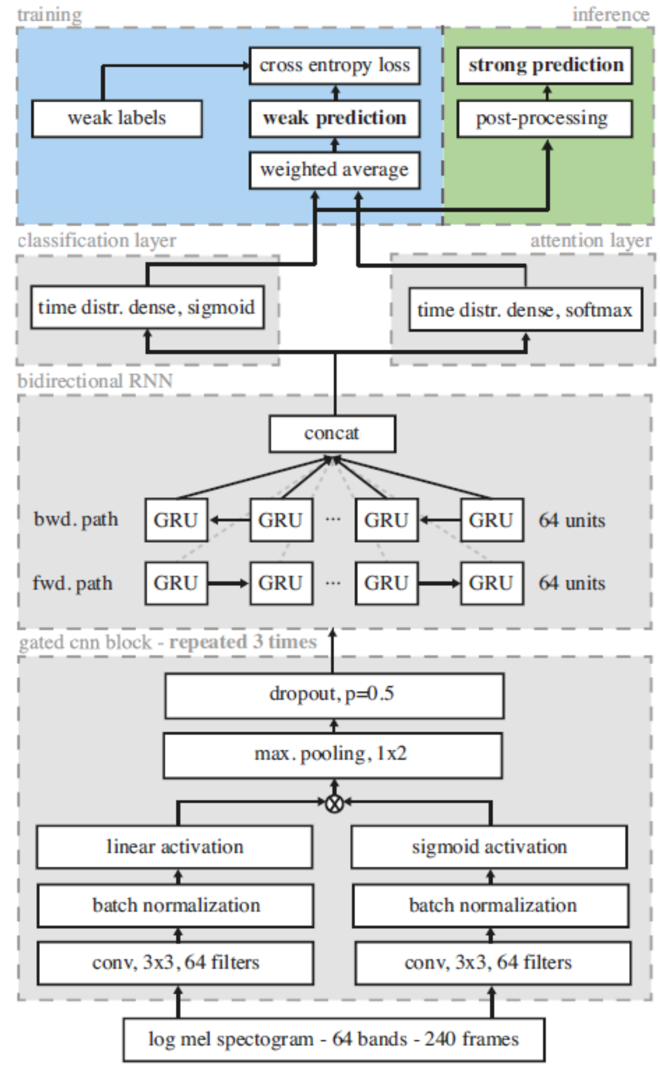
Music
- Musical instrument
- Music genre
- Musical concepts
- Music role
- Music mood

Natural sounds
- Wind
- Thunderstorm
- Water
- Fire

Channel, environment and background
- Acoustic environment
- Noise

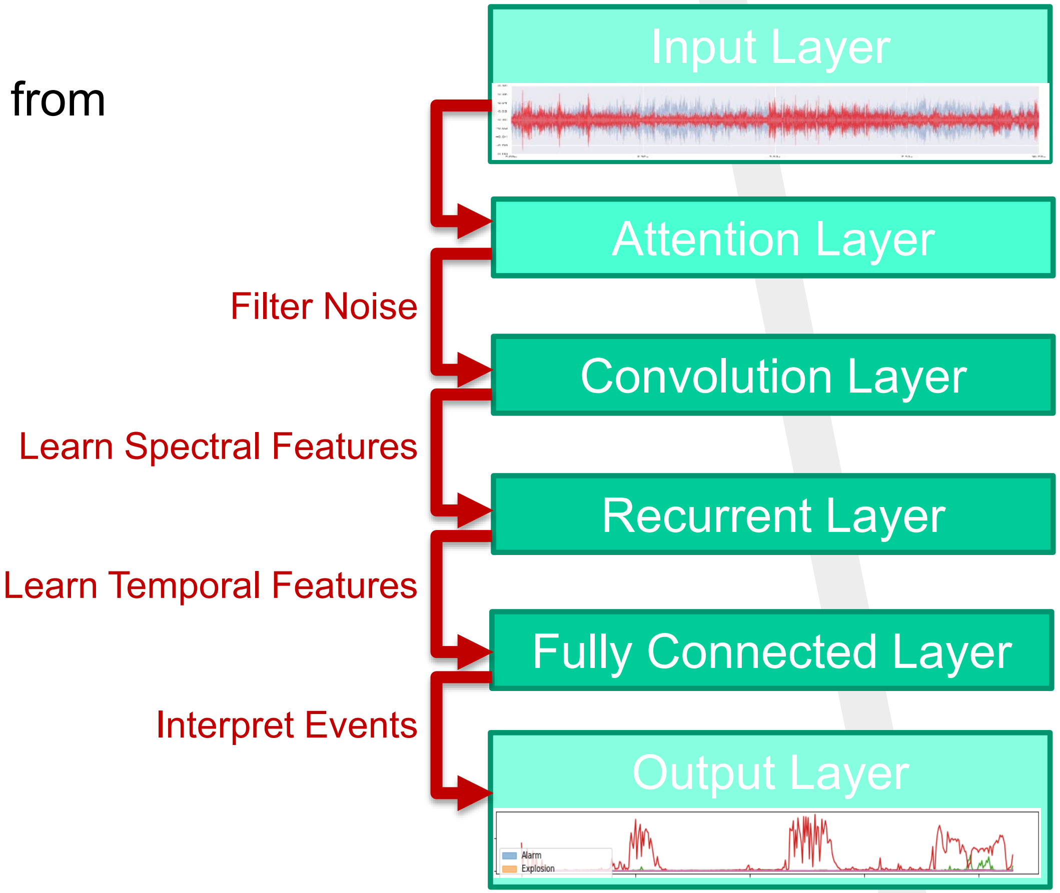
Convolutional Recurrent Neural Networks (CRNN)

1. Input representation
   - Common: Mel-Spectrograms
2. Convolutional Neural Network Block (CNN)
   - Learn audio embeddings
3. Recurrent Neural Network Block (RNN)
   - Learn Temporal dependencies of embeddings
4. Array of Fully Connected Layers
   - One Layer per temporal dimension (Time-Distributed)
   - Dimensionality of Layer = Number of classes
5. Outputs
   - **Strong Labels – Training & Inference**
     - Output of Time-Distributed Fully Connected Layers
   - **Weak Labels - Training**
     - Output Layer aggregation (e.g. avg, max)
     - Multi label prediction



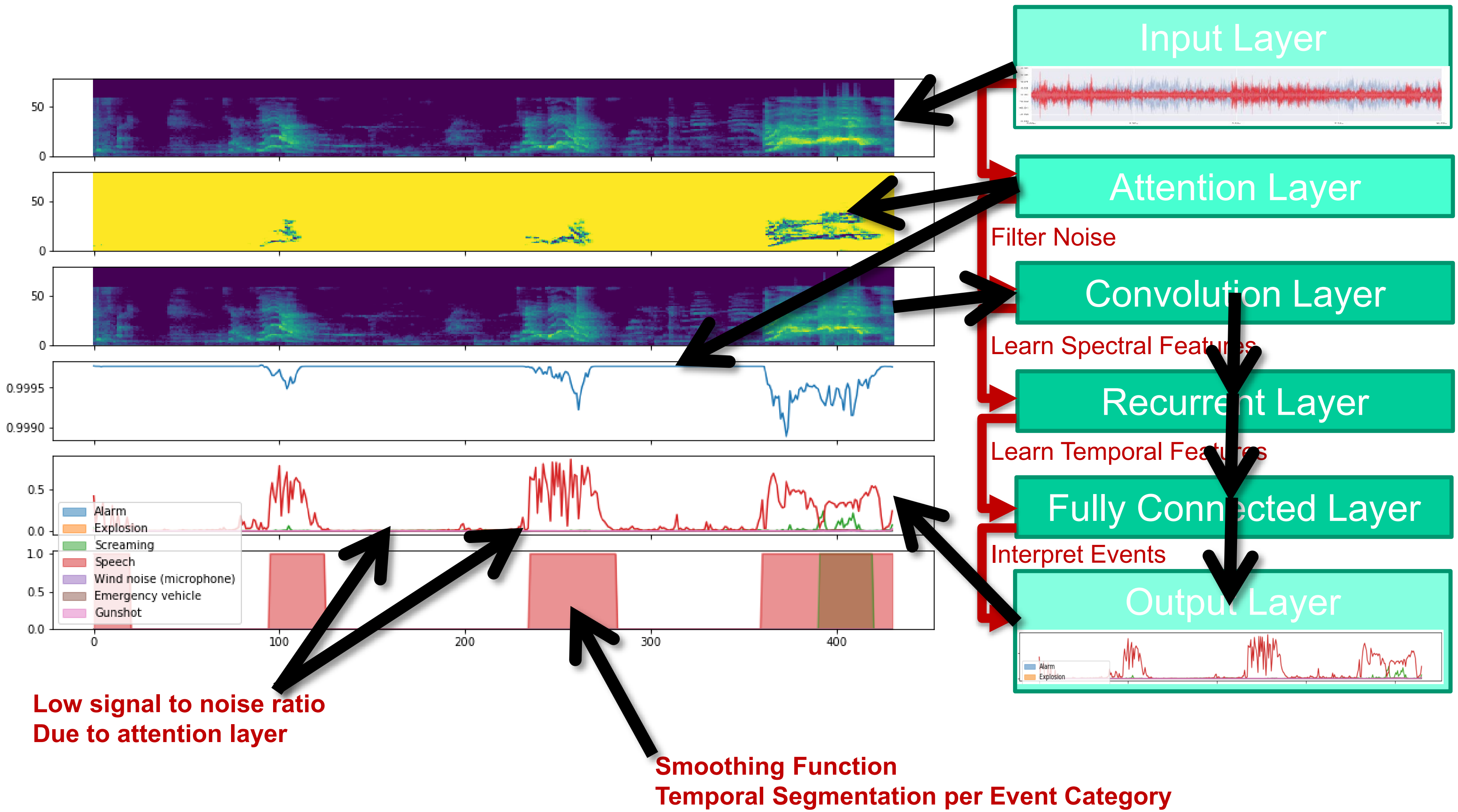
CRNNs with Attention Layers

- Attention Layer
  - Filter non-relevant information from Input
  - Help to learn faster
  - Better convergence
  - Better generalization
  - Smoother prediction signal

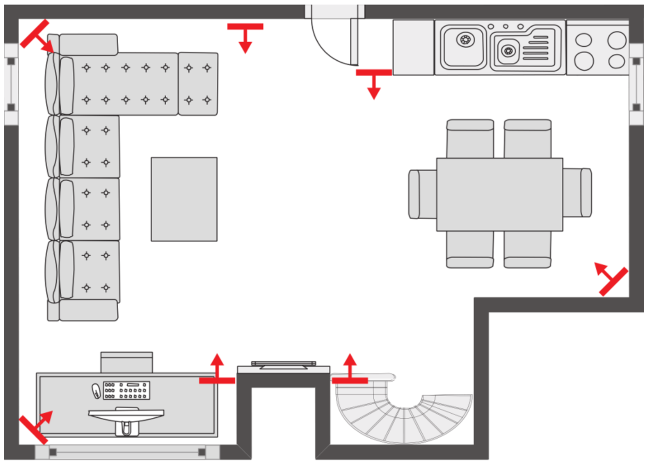


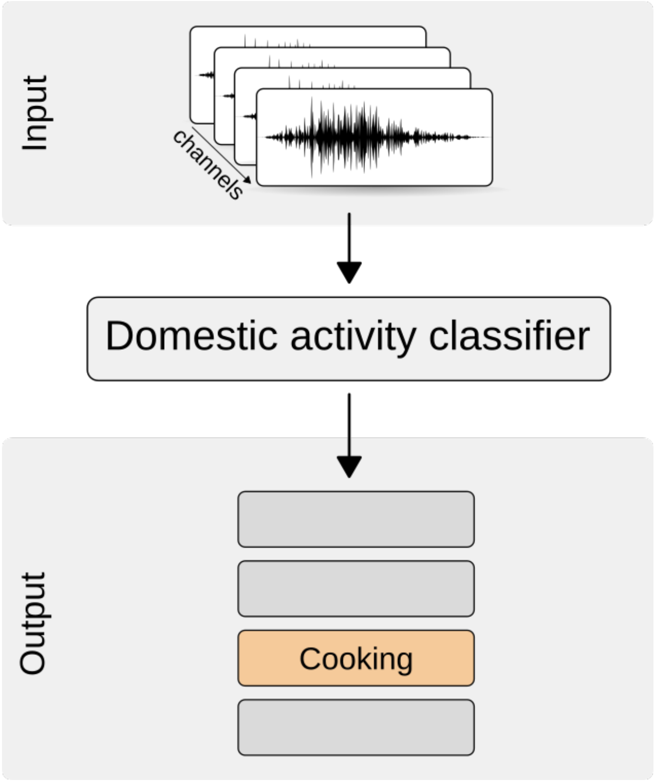
Audio Event Detection

Recurrent Convolutional Neural Networks

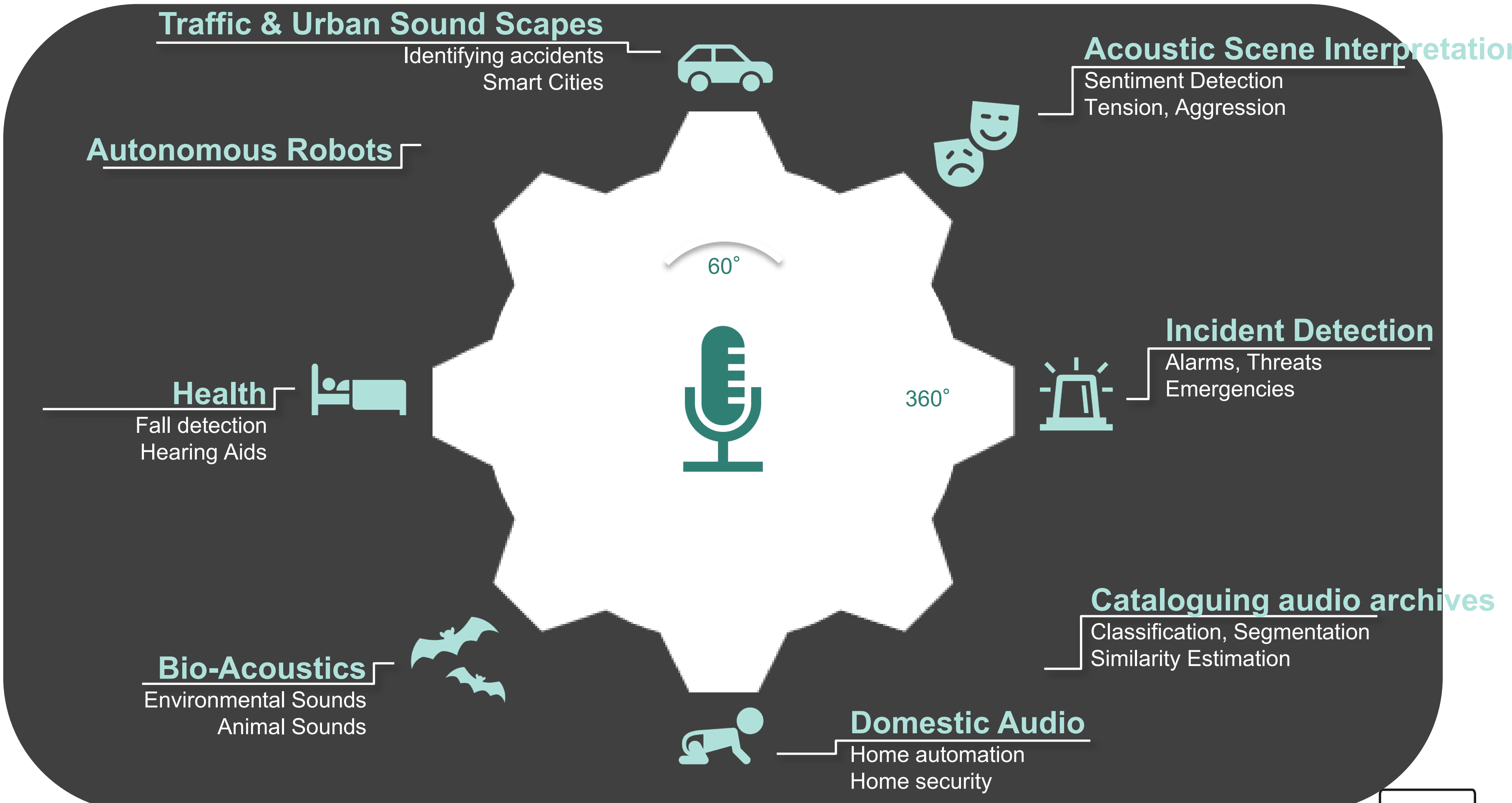


Monitoring of domestic activities based on multi-channel acoustics

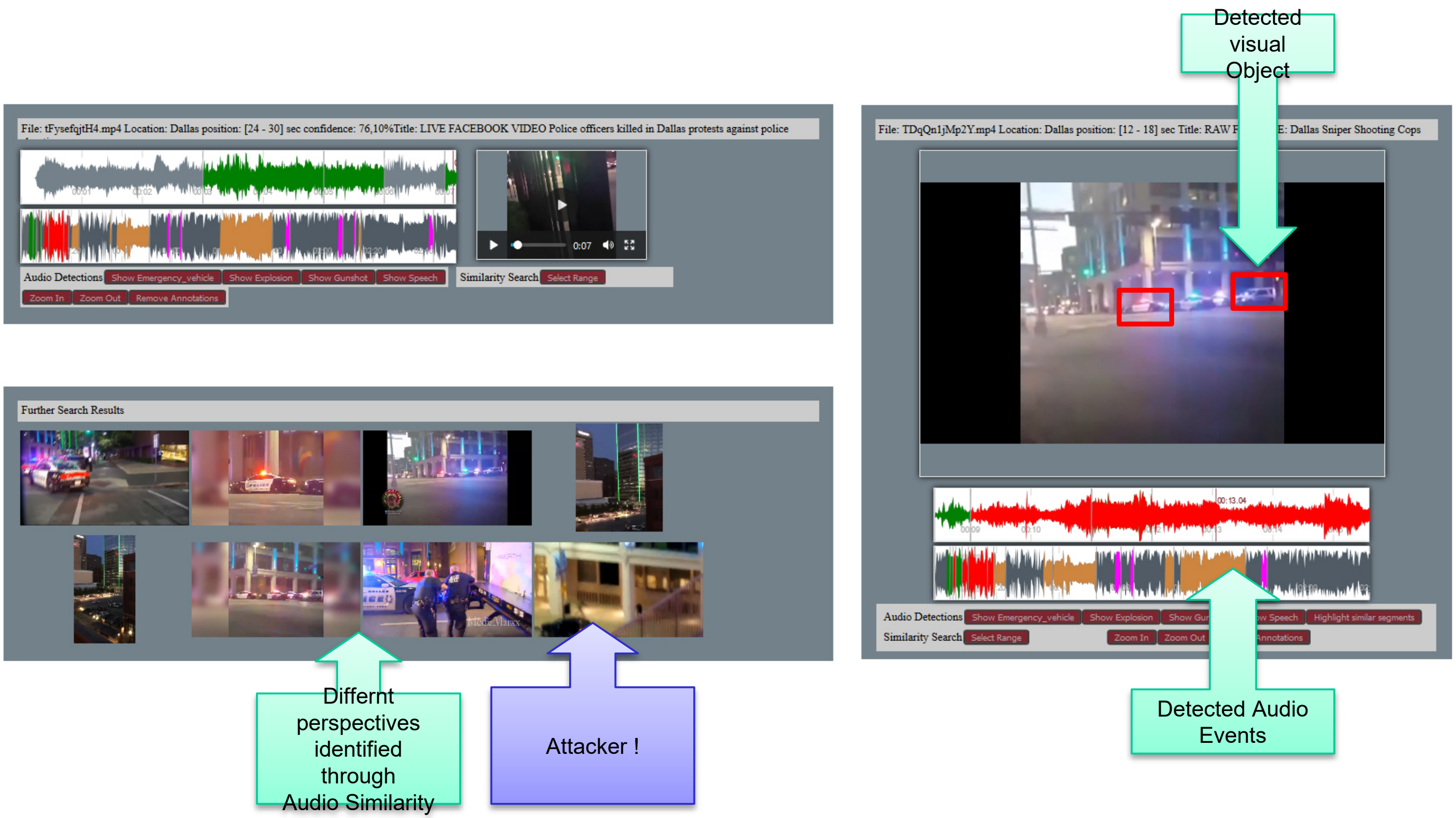




Applications



Crime Scene Investigation

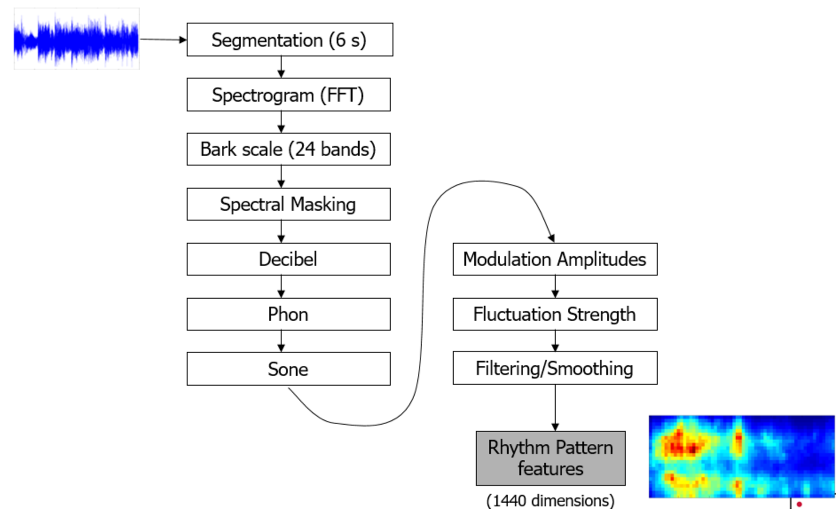


Audio Similarity – Example

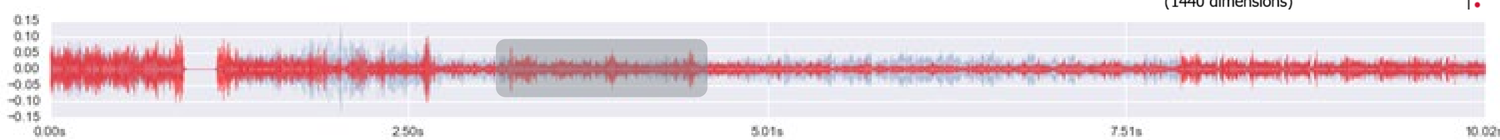
Audio Similarity Search

- Task
  - Searching for video-segments with similar audio-signature
  - Sub-Segment video-search
- Use-Case
  - Suspect could not be identified in one video
  - Select segment and search for others using audio-signature
  - Instant localization (videos close to audio source)

- Technology
  - Rhythm Patterns + Statistical Pattern Descriptors



Search



90% similar

Results

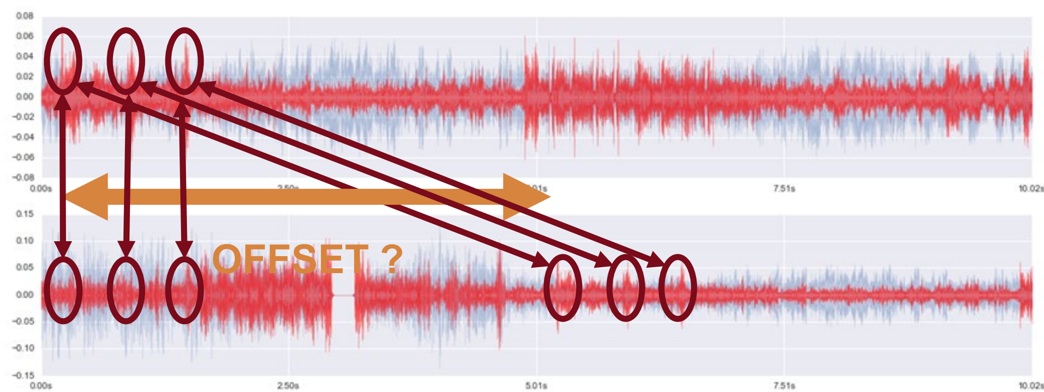
85% similar

70% similar

Audio-based Video-Synchronization

- Task
  - Synchronize various video files with unreliable time metadata
  - Use audio-signature to relatively align video files

- Technology
  - Audio-fingerprints (chromaprint)
  - Noise invariant



References

- A. Mesaros *et al*., "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379-393, Feb. 2018.
- Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013, October). Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on* (pp. 1-4). IEEE.