TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

AUSTRIAN INSTITUTE OF TECHNOLOGY

Multi-Temporal Resolution Convolutional Neural Networks for Acoustic Scene Classification

Alexander Schindler
alexander.schindler@ait.ac.at

Thomas Lidy
lidy@ifs.tuwien.ac.at

Andreas Rauber
rauber@ifs.tuwien.ac.at

Motivation and Goal

- acoustic scenes are composed of spectral texture and sequence of acoustic events
- common CNN-based approaches use single statically defined analysis window
- wrong size of this analysis window can either
  - prevent from having sufficient timbral resolution
  - fail to recognize acoustic events with longer patterns

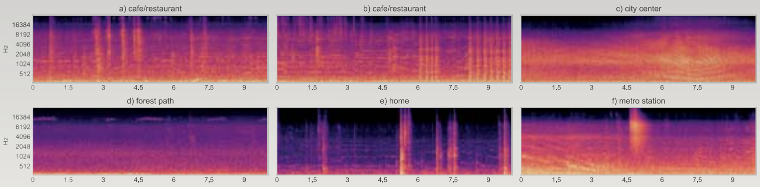


Figure 1: Example Mel-Spectrograms to visualize variances in length and shape of different acoustic events. a) dropping coins into the cash-box, b) beating coffee grounds out of the strainer, c) Doppler effect with Lloyd's mirror effect of a passing car, d) chirping bird, e) opening and closing of cupboards and drawers in the kitchen, f) arriving subway with pneumatic exhaust.

Audio Pre-processing

Audio File → Mono → Mel → Log → NN

- 10 times 80x80 log-amplitude scaled Mel-Spectrograms per audio file
- increasing STFT window sizes: 512, 1024, 2048, 4096, 8192 - (50% overlap)
- Data Augmentation
  - time-stretching
  - pitch-shifting
  - place-wise remixing
  - split-shuffle-remix

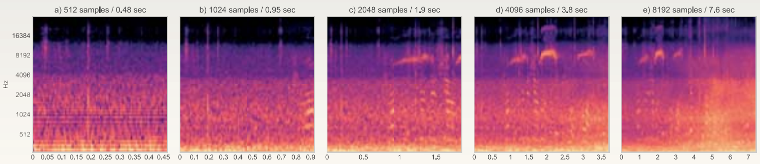
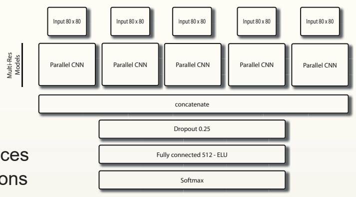


Figure 2: Input Segments for the Convolutional Neural Networks with 80 Mels spectral and five different temporal resolutions with fixed start-offset. a) spectral texture of residential area background noise, b) person saying a word (vertical wave-line), c) person talking, tweet of a bird (horizontal arc), d) person talking, bird tweeting, e) person talking, bird tweeting, car passing (light cloud to the right)

Neural Network Architecture

Multi-Resolution Model

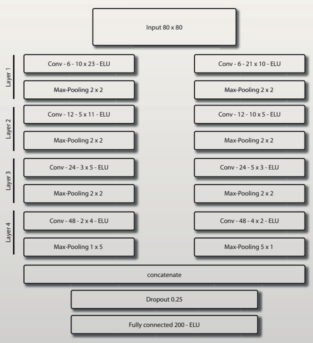
- one parallel architecture for each temporal resolution
- fully connected output layers are concatenated
- intermediate fully connected layer with 512
- learn dependencies between sequences of spectral and temporal representations of the different temporal resolutions



Parallel Convolutional Neural Network (CNN)

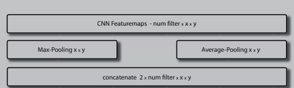
- two pipelines of CNN Layers capturing:
  - frequency domain relations
  - temporal relations
- each pipeline
  - uses the same input segment
  - consecutively filters and pools in the
    - vertical axis ('timbre')
    - horizontal axis ('rhythm')
- merge layers by concatenation
- add fully connected layer and Softmax output

- classification of an audio file by:
  - maximum probability summing



Combined Max-Average-Pooling

- Max-Pooling dominated by peek values (attack)
- Average-Pooling dominated by smoothed values (decay)
- Max-Average-Pooling
  - apply both operations and stack featuremaps
  - adds information without increasing number of trainable parameters



Evaluation and Results

fft win size	instance raw	grouped raw	instance augmented	grouped augmented
512	64.14 (2.84)	70.32 (2.96)	69.06 (4.33)	76.63 (4.44)
1024	66.32 (2.58)	71.27 (3.06)	71.70 (5.46)	77.06 (5.46)
2048	66.83 (1.52)	70.23 (1.99)	76.24 (2.53)	80.46 (3.30)
4096	69.50 (2.83)	71.92 (3.23)	79.20 (3.03)	81.66 (3.29)
8192	69.66 (2.58)	71.47 (2.95)	82.26 (2.40)	83.73 (2.63)
grouped single		73.12		83.19
multi-res	72.23 (4.15)	74.30 (4.81)	85.22 (2.11)	87.29 (2.02)
multi-res do	69.39 (2.77)	72.05 (3.26)	82.51 (2.37)	86.04 (3.03)

Table 1: Experimental results (classification accuracy with standard deviation over cross-validation folds). Single-resolution model results provided on top, multi-resolution models at the bottom.

- multi-resolution model outperforms best single-resolution model by 3.56%
- model harnesses dependencies between temporal resolutions - examples train, metro-station
- lower temporal resolutions perform better than higher
- Grouping and averaging the predictions for a file of all single-resolution models does not increase their performance
- Max-Average-Pooling improved DCASE challenge results for multi-resolution model by 3.25% (final result 90.54%)

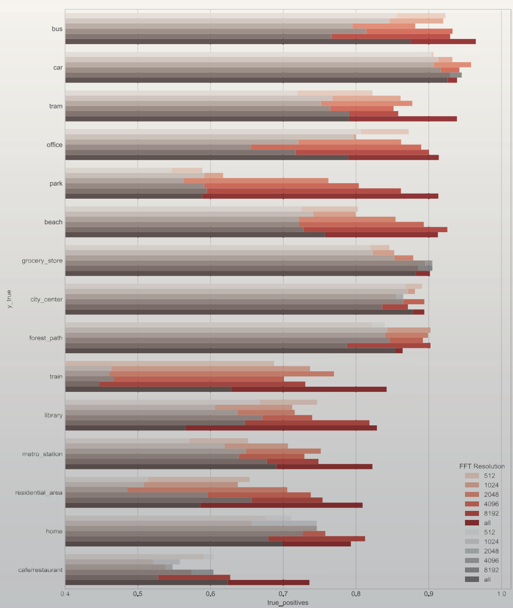


Figure 3: Results per class and FFT window size with ascending temporal resolutions. Multi-resolution results at last. Grayed bars represent un-augmented data, red bars augmented.