TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

AUSTRIAN INSTITUTE OF TECHNOLOGY

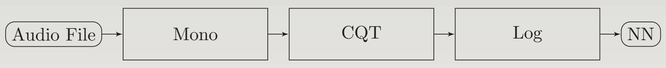
CQT-based Convolutional Neural Networks for Audio Scene Classification and Domestic Audio Tagging

Thomas Lidy (lidy@ifs.tuwien.ac.at), Alexander Schindler (alexander.schindler@ait.ac.at)

Motivation and Goal

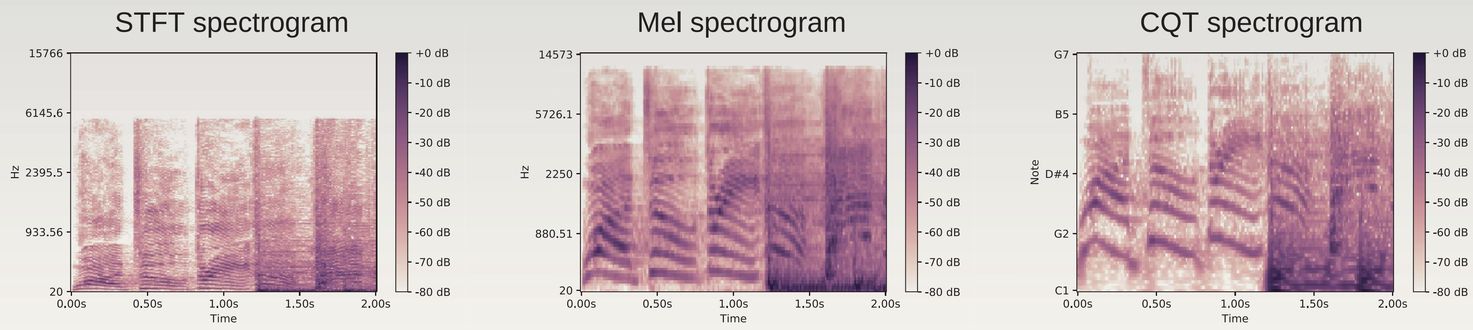
Identification of urban and domestic audio events in the IEEE AASP DCASE 2016 Challenge: Detection and Classification of Acoustic Scenes and Events (Tasks 1 and 4)

Audio Pre-processing



- mono conversion
- apply Constant-Q-Transform (CQT):
  - adapted wavelet transform
  - geometrically spaced frequency bins
  - 12 bands per octave
  - we use a total number of 80 bands (4 highest cut off)
  - analyze 82 consecutive CQT frames (0.94 seconds) as one audio segment
- log10 transform

- process a multitude of short-term segments from an audio example to be learned by the neural network (30 seconds audio results in 31 CQT excerpts)



CQT provides a better resolution in the low frequencies and achieves better accuracies on the development set

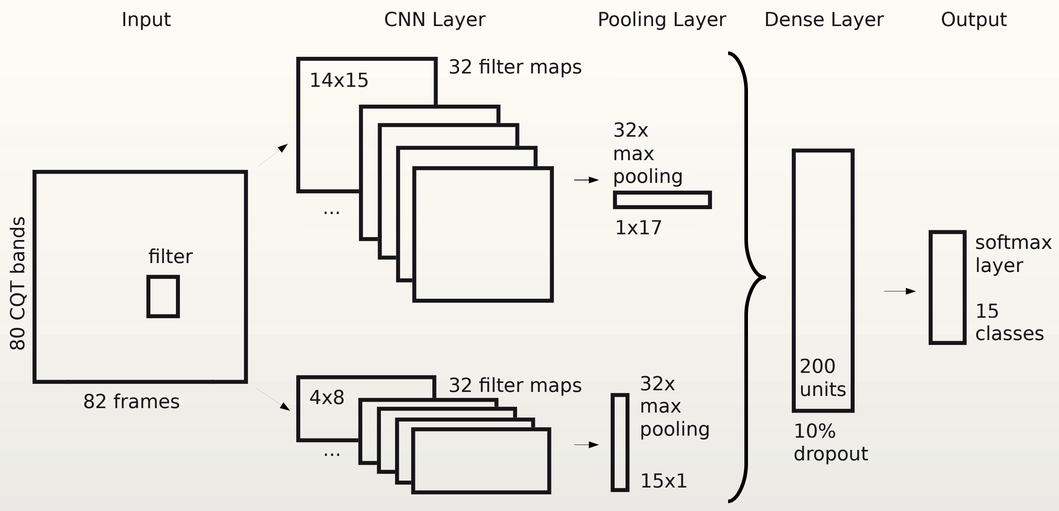
Transform	Bands/Frames	Acc max prob	Acc maj vote
Mel	40 x 80	76,23%	75,62%
Mel	80 x 80	76,55%	76,38%
CQT	80 x 82	**80,25%**	**80,07%**
CQT	84 x 82	78,11%	77,59%
CQT	126 x 82	79,39%	79,14%

Neural Network Architecture

parallel Convolutional Neural Network (CNN)

- two pipelines of CNN Layers capturing:
  - frequency domain relations
  - temporal relations
- both layers
  - use the same input segment
  - are followed by MaxPooling
  - flattened
- merge layers by concatenation
- add fully connected layer and Softmax output

- classification of an audio file by:
  - maximum probability summing all segments
  - majority Vote of decision by segments



Models

Task 1:

- Model 1: as depicted left
- Model 2: improved by 2-class models to overcome biggest class confusions between:
  - park vs. residential area
  - cafe/restaurant vs. train
  - home vs. library

Task 4:

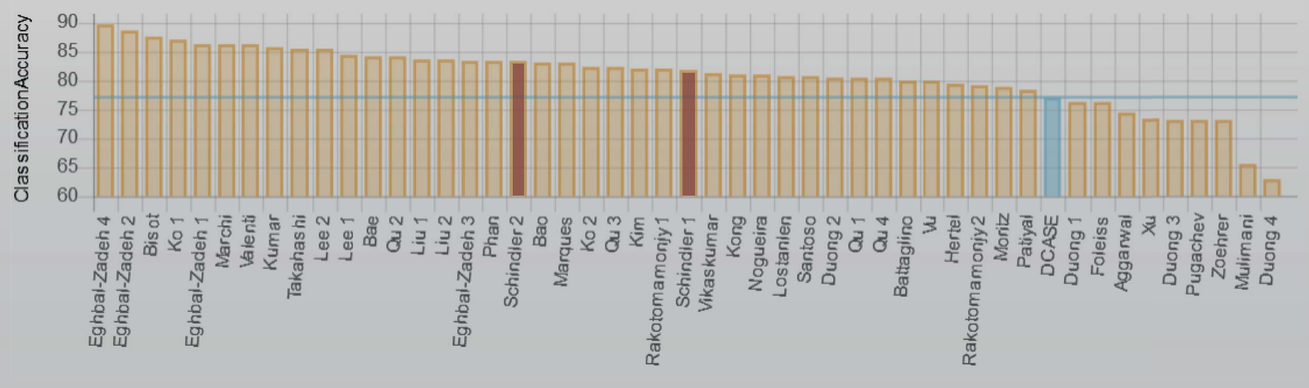
- Model 1: as depicted left, with Sigmoid output

Challenge Results

Task 1: Acoustic Scene Classification

categorize an audio sample into *one* of these 15 classes:

- beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, park, residential area, train, tram



Task 4: Domestic Audio Tagging

predict the probabilities of *all* of these 6 classes:

- child speech, adult male speech, adult female speech, video-game/TV, percussive sounds, other identifiable sounds