Waveform Analysis using Sample-Level Convolutional Neural Networks

Fadi Moukayed
Product Development, Scrum-Master (PRISMON)
Rohde & Schwarz Broadcast and Media

ROHDE&SCHWARZ

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

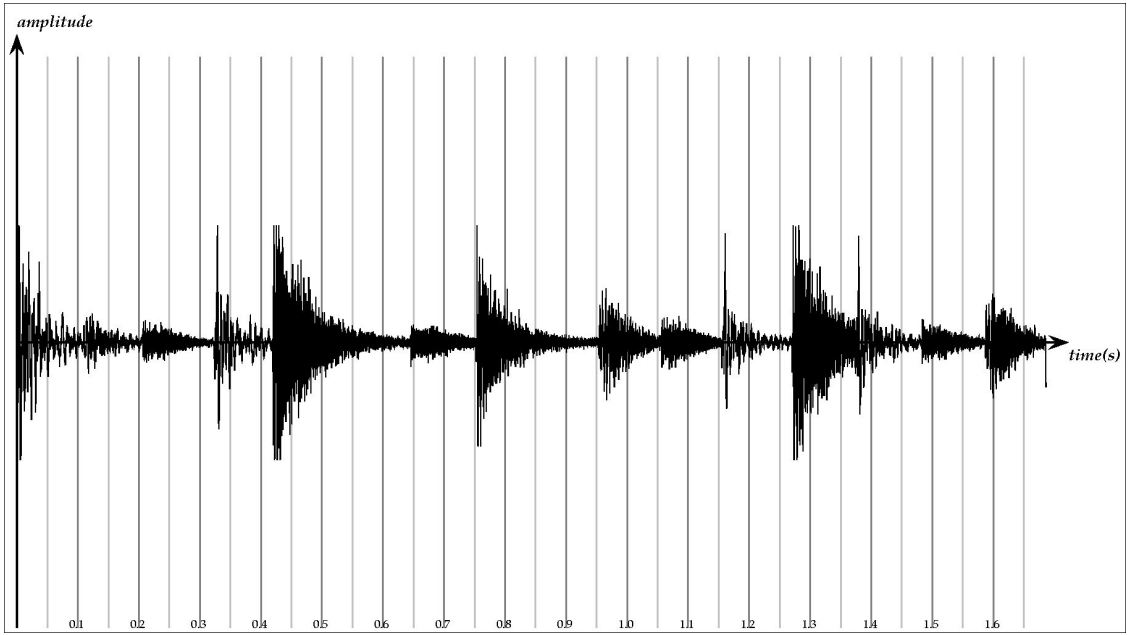
Speaker Introduction

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

- HTW Alum (MA Applied Informatics, 2019)
  - MA Thesis: “Referenceless Detection and Measurement of Artifacts in Digital Video Using Machine Learning”
- Student Researcher at HTW (2016 - 2018)
- Software Engineer @ Rohde & Schwarz (2016-)
  - Focus: Audio/Video Transcoding & **Analysis**
- ML increasingly important in A/V Broadcast Industry
  - Customers *expect* content-level analysis

Audio

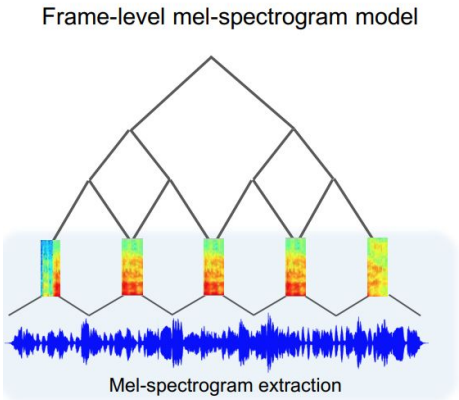
- Sound = Variation in air pressure **over time**
- Time/Amplitude representation
- Essentially, a time series

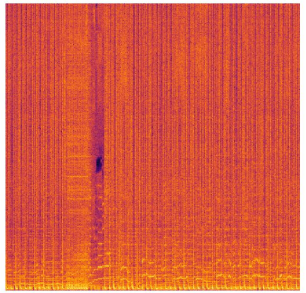


Frame-Level Spectrogram Models

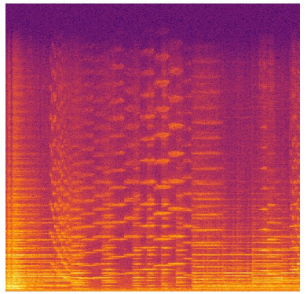
Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

- Spectrogram: Frequency/Time Representation
- Idea: Convert the data to a spectrogram (image)
  - Usually: Mel Spectrogram (More Information about Tones)
- Then, use an ImageNet Model (e.g. ResNet)



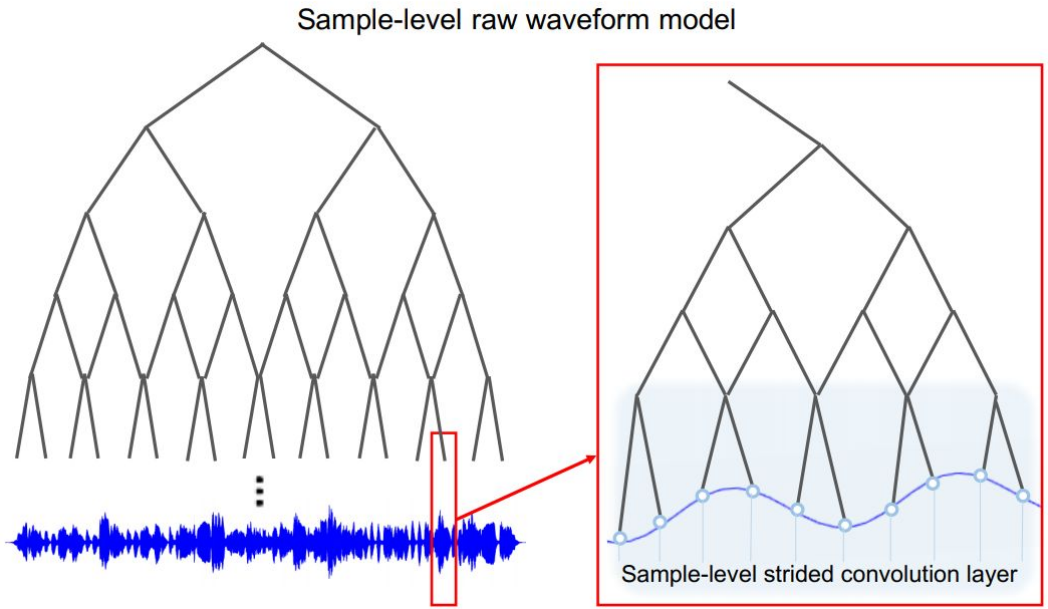


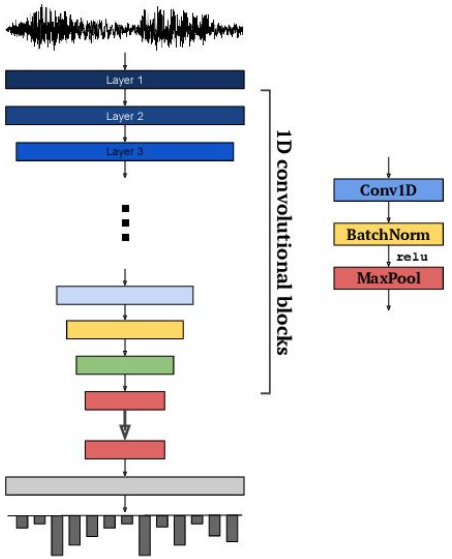
Pop



Classic

Sample-Level Models: SampleCNN

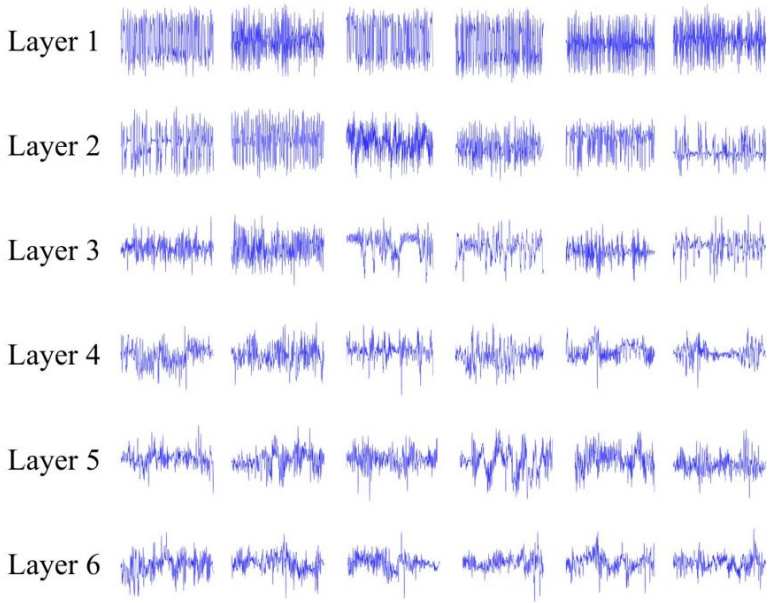




- Raw samples as CNN input
- 1D Convolutional Layers
- No pre-processing

Why SampleCNN?

- No preprocessing
- End-to-End approach
- (At least) as good as frame-level approach
- No need to fiddle with spectrogram parameters
- *Conv1D-BatchNorm-MaxPool* blocks are (effectively) bandpass/peak extraction filters



Experiments / Context

- Project 1: Audio Segmentation
  - Customer A (Radio Broadcast Provider) wants to know type of aired content (Talk Radio, Music)
  - Classification Task

- Project 2: Audio Language Detection
  - Customer B wants to know whether advertised track language (within stream metadata) matches actual content language
  - Multiclass classification task (DE, EN, FR, IT, RU, ...)

Project 1: Audio Segmentation using SampleCNN

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

- Trained on GTZAN Music-Speech Dataset
- Tested on ripped *rbb radio1* broadcasts
- SGD+CyclicLR, Dropout=0.5
- 59049 Inputs (~3.7s @ 16 kHz)
- https://github.com/kochab/samplecnn-speech-detection
  (PyTorch implementation)

```
$ python3 -B discriminator st_20191010.wav 2>/dev/null
00:00:00 🎶 (0.000)
00:00:01 🎶 (0.000)
00:00:02 🎶 (0.000)
00:00:03 🎶 (0.000)
00:00:04 🎶 (0.000)
00:00:05 🎶 (0.000)
…
00:56:21 🎤 (0.964)
00:56:22 🎤 (1.000)
00:56:23 🎤 (1.000)
00:56:24 🎤 (1.000)
00:56:25 🎤 (1.000)
00:56:26 🎤 (1.000)
00:56:27 🎤 (1.000)
00:56:28 🎤 (1.000)
00:56:29 🎤 (1.000)
…
01:00:04 🎤 (0.723)
```

Project 2: Audio Language Detection

- VoxForge Dataset
  - http://www.repository.voxforge1.org/downloads/
  - Focus: Indo-European Speech (EN, DE, FR, IT, ES, PT, RU)
- Status: TBD
  - *Unfortunately*, I wasn't able to finish this in time for this talk
  - BUT: It will probably work (really)
    - Will provide proof in the next talk
    - Will upload it on GitHub too

Lessons Learned

- Large Amounts of Training Data
  - Single epoch requires long processing time
  - Must downsample dataset
  - Or: Continously extract random 59049-sample windows, train for fixed # of batches
- CyclicLR + SGD = Good Results (Usually)
  - CyclicLR → Forget the LR, just select upper/lower LR bounds
  - CyclicLR will adjust LR and Momentum cyclically
  - Helps escape from “bad” local minima
- Resampling
  - Practically, resampling down to 16kHz (at least) is required
  - Input Layer = 59049 (Can fit up to ~3.7s @ 16kHz)
  - Lower bound is 8kHz (Most speech still understandable at this rate)
- Multichannel Audio: Just use mono

Fin

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

Thanks for listening

References:

https://arxiv.org/abs/1703.01789 (Original Paper explaining SampleCNN architecture)

https://github.com/Insiyaa/Music-Genre-Classification (Mel-Spectrogram+ResNet approach)

Contact Info:

fadi.moukayed@rohde-schwarz.com

fadi.moukayed@student.htw-berlin.de (Not sure when this will get deactivated)

smfadi@gmail.com