SubSpectralNets – Using Sub-Spectrogram based Convolutional Neural Networks for Acoustic Scene Classification

Sai Samarth R Phaye[1], Emmanouil Benetos[2,3], and Ye Wang[1]

phaye.samarth@gmail.com, emmanouil.benetos@qmul.ac.uk, wangye@comp.nus.edu.sg

[1]School of Computing, National University of Singapore, Singapore

[2]School of EECS, Queen Mary University of London, UK [3]The Alan Turing Institute, UK

1. Introduction

- **Acoustic Scene Classification** (ASC) – problem of classifying a recording into a scene label in which it is recorded, is one of the core research problems in the field of Computational Sound Scene Analysis.
- We propose **SubSpectralNets**, a novel deep learning model which captures intricate features by incorporating frequency band-level differences to model soundscapes.
- Evaluated on the public ASC development dataset provided for the "Detection and Classification of Acoustic Scenes and Events" (DCASE) 2018 Challenge.
- Code: *https://github.com/ssrp/SubSpectralNet*
- Paper: *https://arxiv.org/abs/1810.12642*

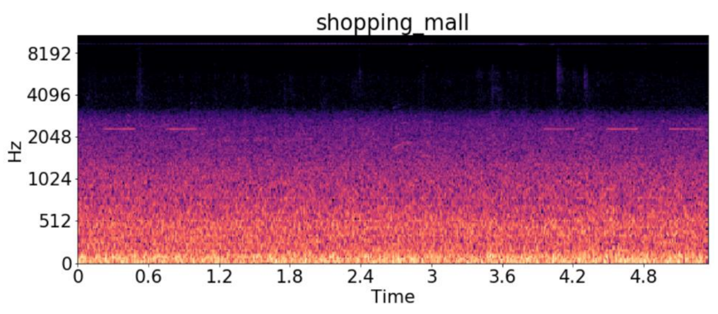

Figure 1. A spectrogram of an audio sample belonging to shopping mall class.

2. Motivation

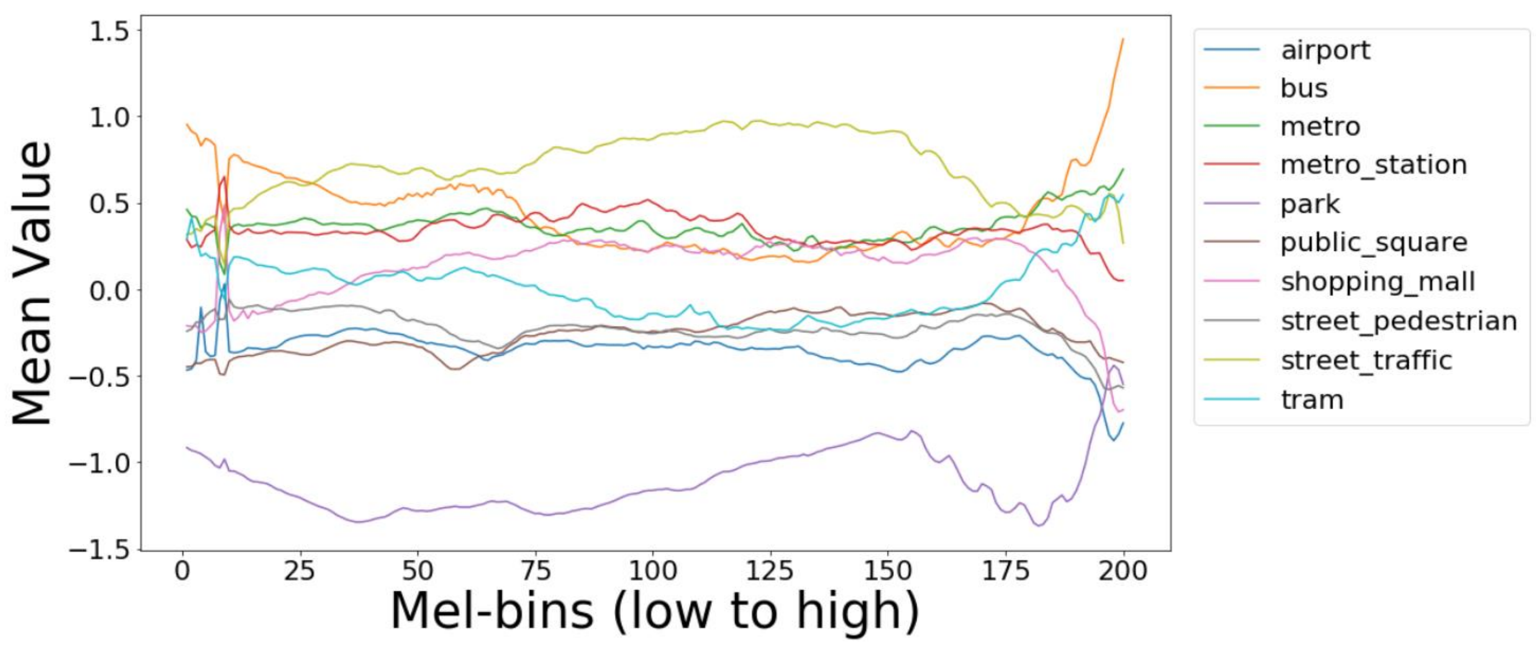

Figure 2. Distribution of average mel-bins for each class in the dataset

- **Magnitude spectrograms** are two-dimensional representations over time and frequency – *very different* from real life images.
- Definitive local relationships in the time dimension, but *not in the frequency dimension*. *Clear variation* in the frequency axis (example shown in Figure 1).
- Frequency dimension for different sounds might have either:
  - local relationships (e.g. noise-like sounds),
  - non-local relationships (e.g. harmonic sounds),
  - no local relationships at all.

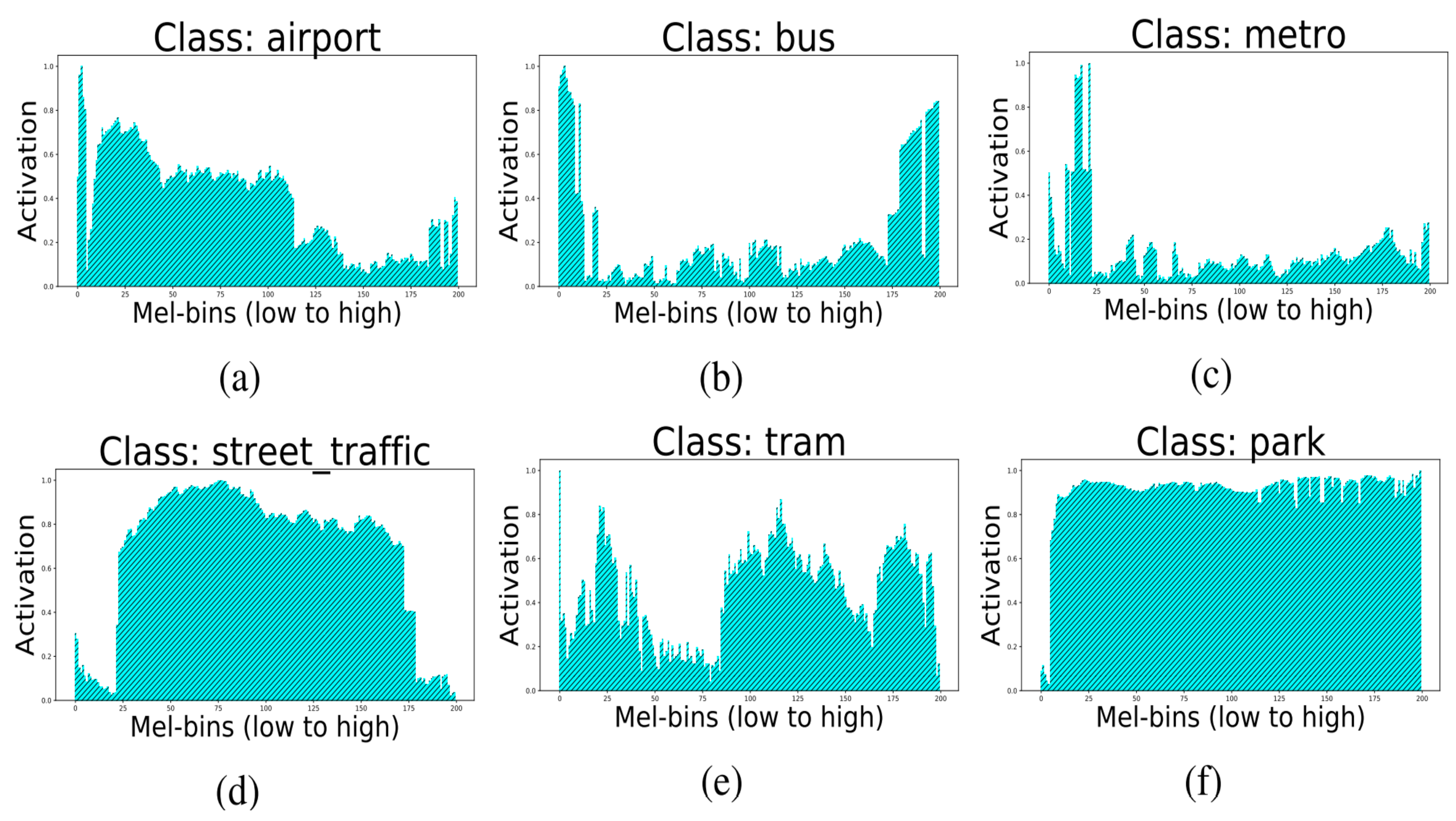

Figure 3. Distribution of average mel-bins for each class in the dataset

- We observe a definite variation of activation of mel-bins and sub-bands, which is specific to every scene.
  - For example, the "metro" class has more *activation* in lower frequency bins; the "bus" has less activation in mid frequency bins (shown in Figure 2 and 3).
- In SubSpectralNets, we exploit this property of spectrograms to leverage the performance of a CNN architecture. This has *never been done* in the literature before.

Sound & Music Computing Lab
http://www.smcnus.org/

3. Proposed Architecture

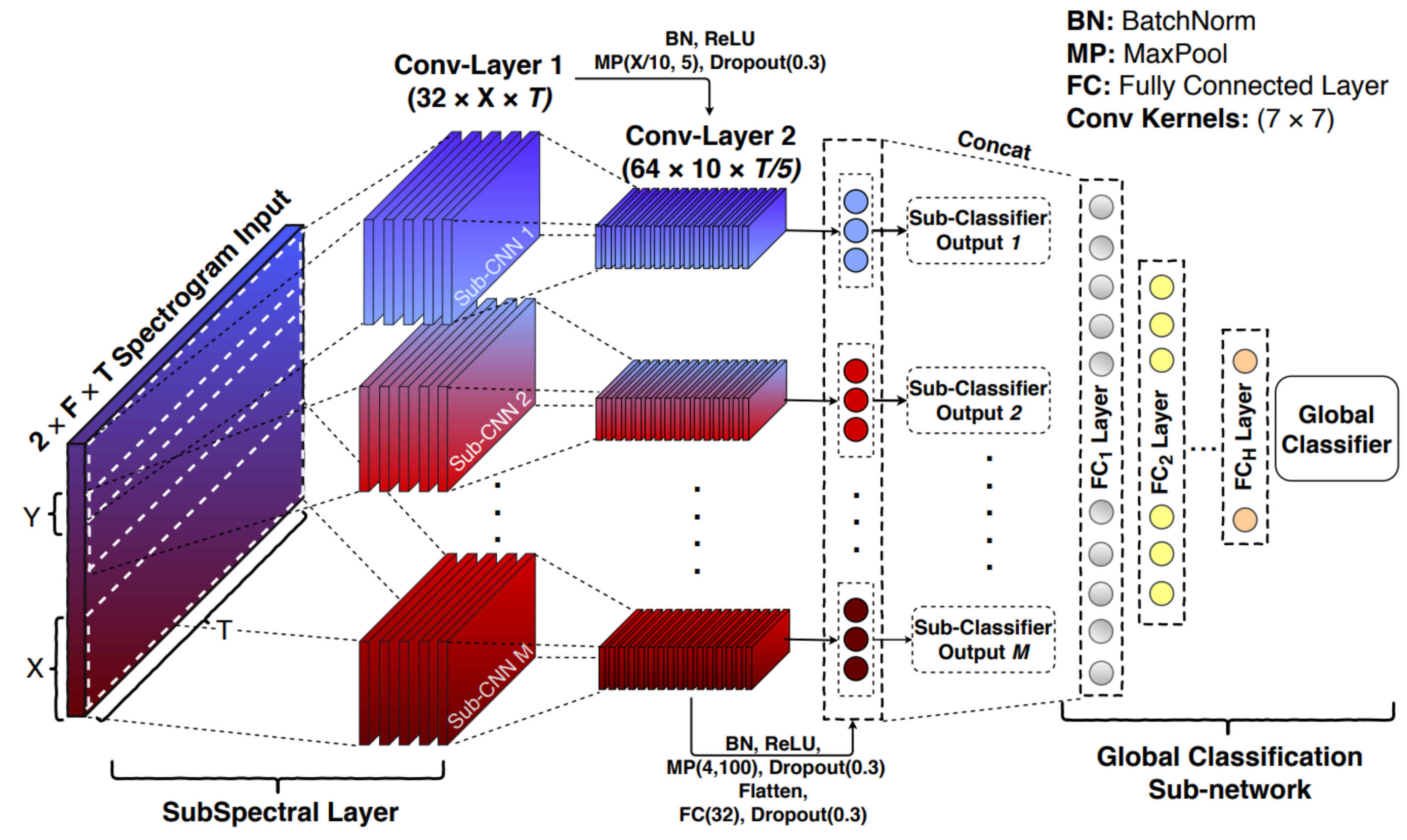

Figure 4. Proposed pipeline of SubSpectralNet

- Divide Spectrogram into various vertical splits (Figure 4):
  - ✓ Use appropriate sub-spectrogram size
  - ✓ Use a vertical mel-bin hop size
- Separate backpropagations for every weak classifier – "*sub-classifiers*".
- **Global Classifier** to extract correlations in the sub-spectrograms.

4. Experiments and Observations

- **Faster convergence** over the baseline (Figure 5).
- **74.08% best test accuracy**. +14% increase over the baseline (Figure 6).
- **Statistical Analysis is robust**. For example, for the "airport" class: statistical distribution says that lower frequencies are more effective in classification. Same trend is shown in SubSpectralNet where the low-band sub-classifier shows better results (Figure 7).

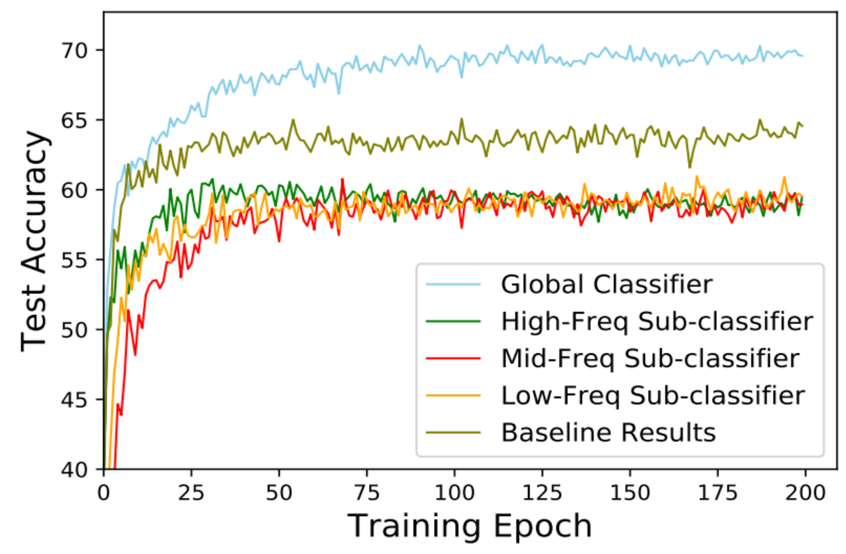

Figure 5. Comparison of performance between the DCASE 2018 baseline model and SubSpectralNet

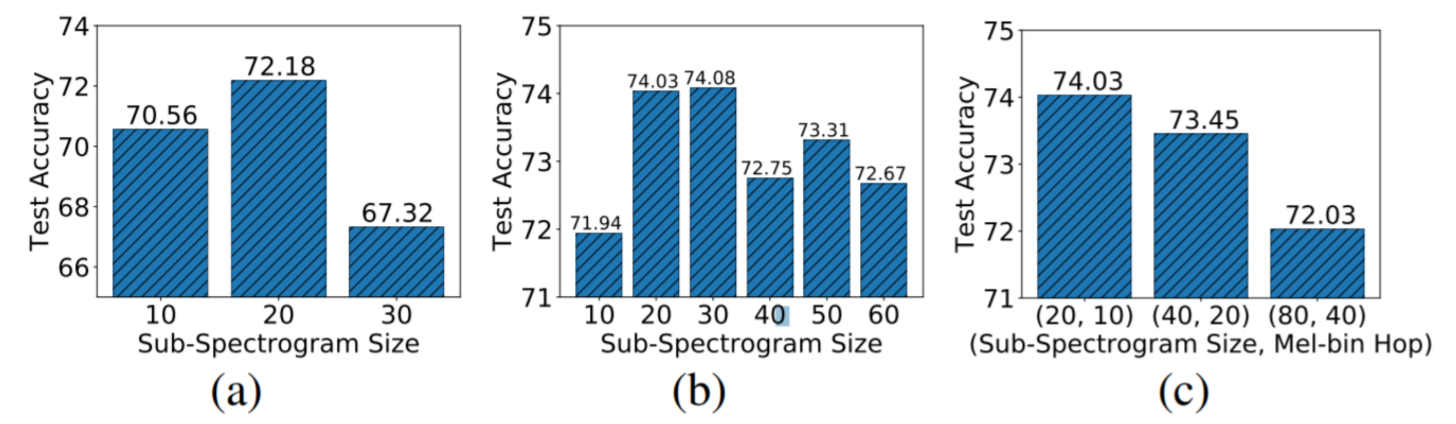

Figure 6. Results obtained by SubSpectralNet on – (a) 40 mel-bin spectrogram and 10 mel-bin hop-size; (b) 200 mel-bin spectrogram with 10 mel-bin hop-size; (c) 200 mel-bin spectrogram, varying subspectrogram and mel-bin hop-size

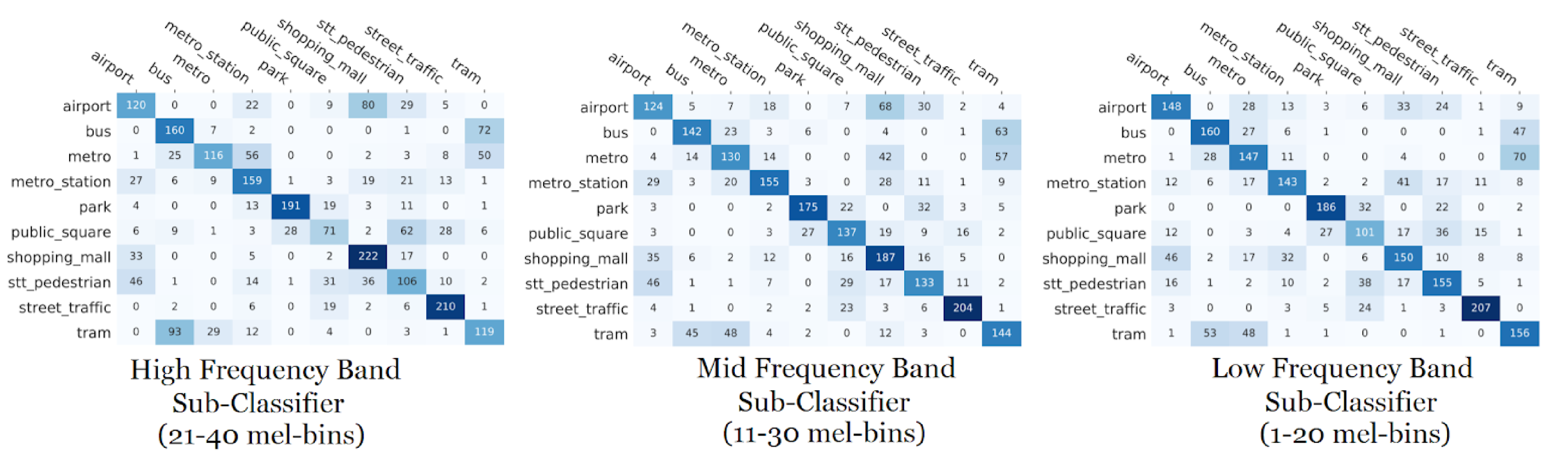

Figure 7. Confusion Matrices for specific sub-classifiers of different bands

5. Conclusion

- Specific bands of mel-spectrograms carry discriminative information than other bands, which is specific to every soundscape.
- SubSpectralNets split the time-frequency features into sub-spectrograms, then merges the band-level features on a later stage for the global classification.