Karol J. Piczak

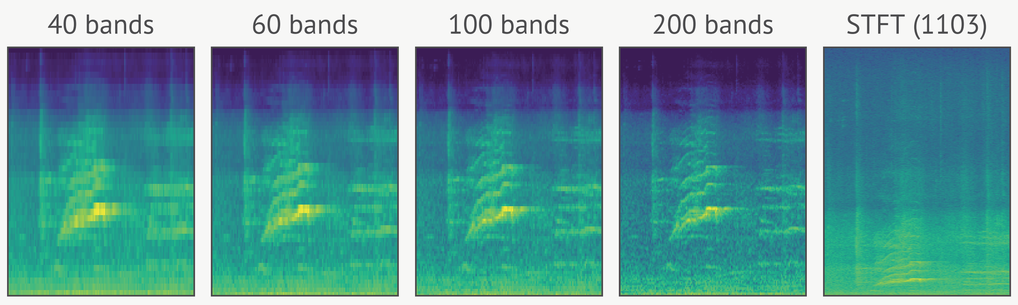
Institute of Computer Science, Warsaw University of Technology

# The details that matter: Frequency resolution of spectrograms in acoustic scene classification

## Overview

This work presents a submission to the acoustic scene classification task of the DCASE 2017 challenge. The study is based on the following premises:

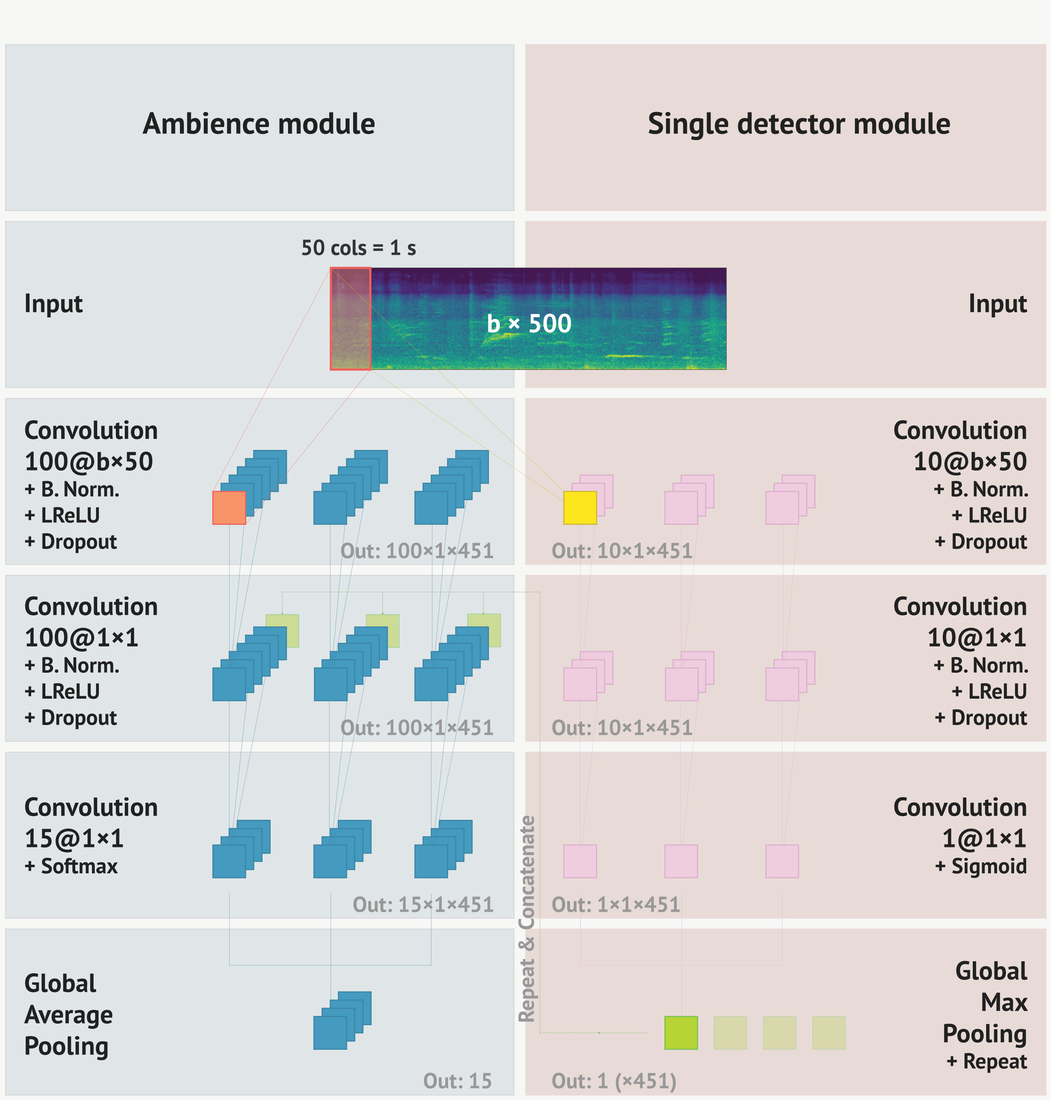
- The prevailing tendency of convolutional neural network models employed in audio classification tasks is to utilize spectrogram representations limited to 40-60 mel frequency bands. Higher values are rarely chosen despite the additional granularity they introduce:




  Therefore, the main goal of this paper is to check **how using spectrograms with different frequency resolution could impact the accuracy in this task.**

- Most acoustic scenes can be described as a combination of a recurring background (ambient noise) mixed with distinct foreground elements (sound events). Still, *Mafra et al.* **(2016)** have shown that even a single averaged frame can have a good predictive capacity in acoustic scene classification tasks, it is thus likely that a good model should not be overly complicated. **Could both these assumptions be introduced a priori into the architecture of the employed network?**

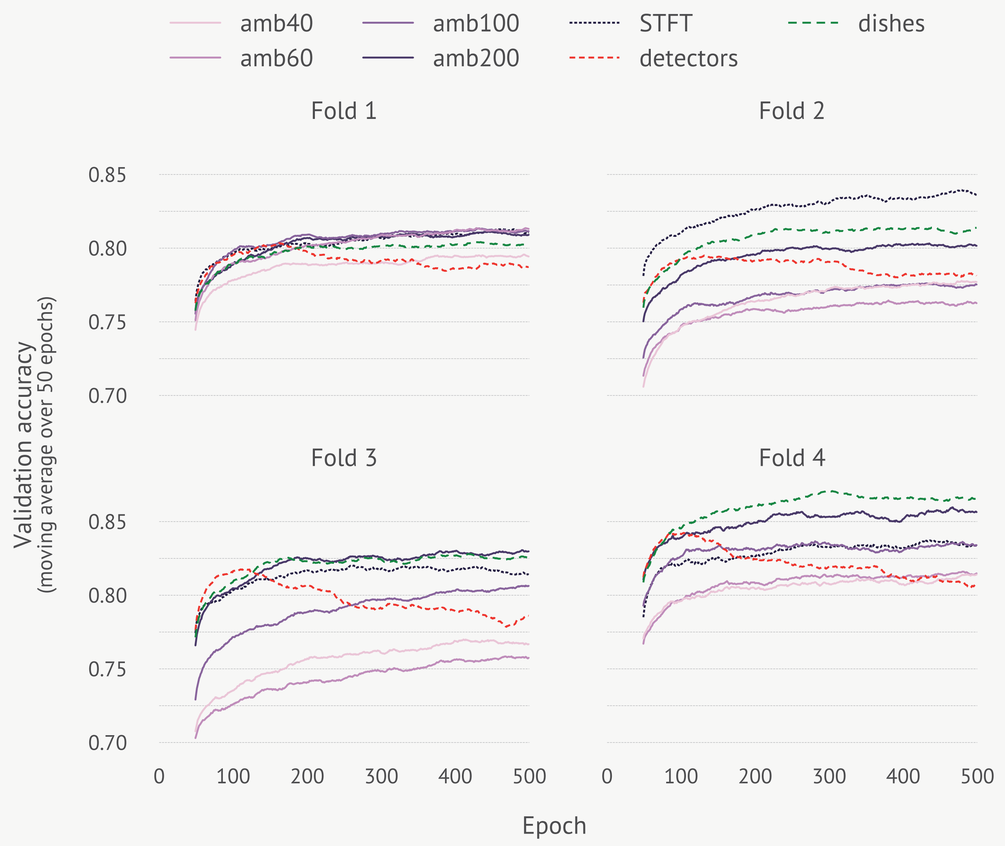
## Model structure




To this end, the proposed system has a simple design coming in two flavors - ambient only (left side) or extended with sound event detectors that signal if a template match has occurred anywhere in the whole recording (right side).

The ambience processing variant is evaluated with different frequency resolutions (*amb** and *STFT*) and compared with models extended with 15 detector modules trained from scratch (*detectors*) or with a single detector pre-trained on hand-annotated fragments of *cafe/restaurant* recordings containing sounds of cups, plates, kitchenware, etc. (*dishes*).

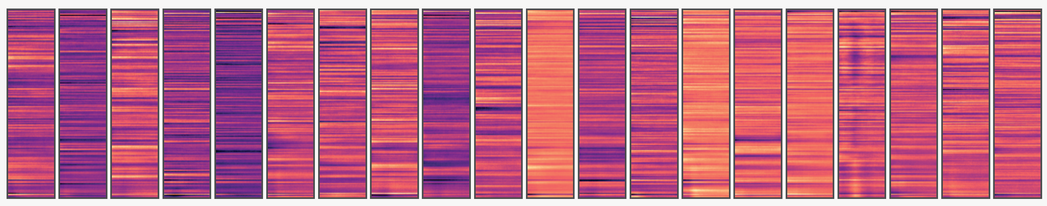
## Results of the proposed systems




| System | Development | | | | | Final |
|---|---|---|---|---|---|---|
| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | 1–4 | |
| amb40 | 79.4 (0.5) | 77.7 (0.8) | 76.7 (1.0) | 81.4 (1.0) | 78.8 | – |
| amb60 | 81.3 (0.6) | 76.3 (1.0) | 75.8 (0.9) | 81.5 (1.0) | 78.7 | 62.0 |
| amb100 | 81.1 (0.6) | 77.5 (0.9) | 80.6 (0.7) | 83.4 (1.3) | 80.7 | 67.7 |
| amb200 | 80.9 (0.8) | 80.2 (0.8) | 83.0 (0.9) | 85.6 (1.3) | 82.4 | 70.6 |
| STFT | 81.1 (0.9) | 83.6 (0.8) | 81.4 (0.9) | 83.4 (1.3) | 82.4 | – |
| detectors | 78.7 (0.9) | 78.1 (1.1) | 78.6 (1.3) | 80.8 (1.4) | 79.1 | – |
| dishes | 80.3 (0.9) | 81.4 (0.7) | 82.6 (0.6) | 86.6 (1.0) | 82.7 | 69.6 |

Mean (standard deviation) of validation accuracies across 50 final epochs of training on the development set and official evaluation results for submitted models. Values in percentages.

Results obtained in the experiments indicate that **increasing the number of mel frequency bands improves accuracy of the ambience model**. The *detectors* variant unfortunately shows signs of significant overfitting combined with higher training times. This effect is constrained in the *dishes* model when the detector array is limited to fine-tuning on one pre-trained module.



Visualization of filters in the first layer shows that the ambience network essentially learns to discriminate frequency patterns. It is therefore a plausible explanation why higher frequency resolution of input data could be beneficial for classification.

## Conclusion

Further examination would be needed to extrapolate this claim to other datasets and architectures, but preliminary results show that **spectrogram resolution could be an important factor influencing the behavior of models in audio classification tasks**. Another interesting extension would be to validate the concept of individual detector pre-training with more abundant annotation data and see if incorporating such a priori domain knowledge about the problem could enhance the capabilities of the model.

## References

- G. Mafra, et al., "*Acoustic scene classification: An evaluation of an extremely compact feature representation*", DCASE, 2016.
- **Source code:** https://github.com/karoldvl/paper-2017-DCASE