# Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification

Yoonchang Han[1], Jeongsoo Park[1,2], Kyogu Lee[2]

[1] Cochlear.ai, Seoul, Korea [2] Music and Audio Research Group, Seoul National University, Seoul, Korea

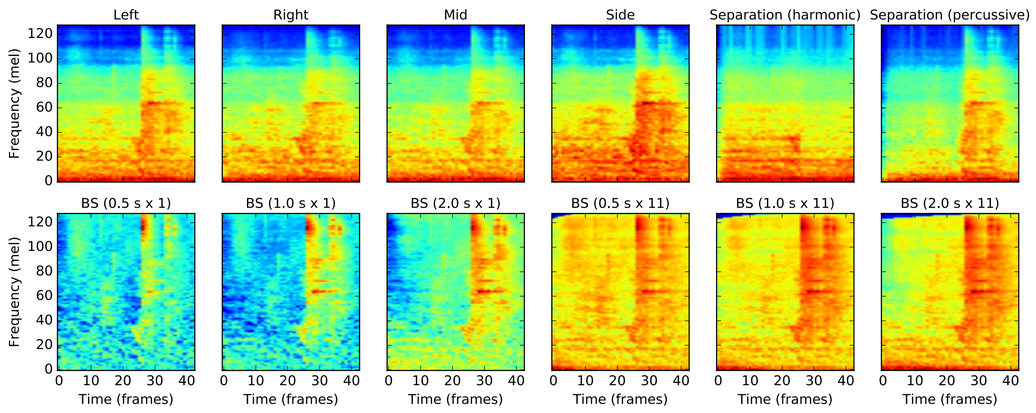
{ychan, jspark}@cochlear.ai, kglee@snu.ac.kr

## Introduction

- This poster explains how we applied convolutional neural network (ConvNet) for DCASE 2017 task 1, acoustic scene classification.
- Basically, we used mel-spectrogram for ConvNet input. However, we used many different versions of it such as left, right, mid, side, results of harmonic-percussive source separation, and background subtraction using median filtering.
- We also present a network structure designed for paired input to make the most of the spatial information contained in the stereo audio.

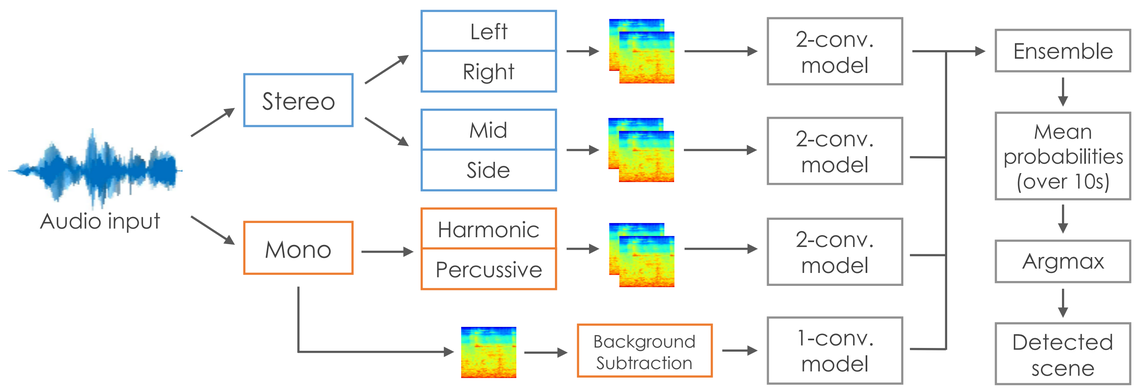
## Pre-processing

- Mel-spectrogram (128 bin)
- Total 12 versions of inputs are used in the experiment
  - Left/Right from stereo
  - Mid (L+R), Side (L-R)
  - Harmonic, percussive source separation results
  - Background subtraction with various median filtering sizes

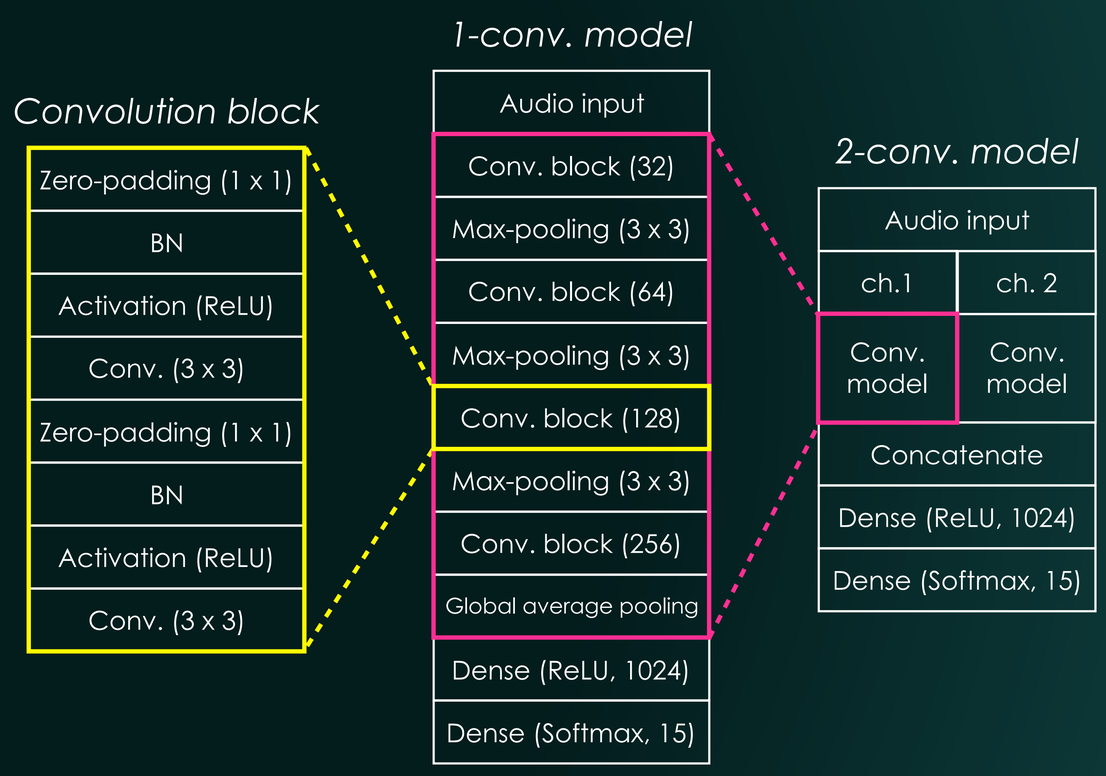



Background subtraction (BS) = original - median filtered one

## System architecture




- Multiple ConvNet models are individually trained using a various preprocessing methods and combined into an ensemble model.
- We used two different ensemble methods, mean ensemble and ensemble selection. Mean ensemble simply takes an average of ConvNet output probabilities, and ensemble selection (Caruana et. al., 2004) finds optimal weights for model combination by repeatedly adding models into ensemble model until achieving the best performance on hill-climb set.




Numbers in brackets :

"Kernel size" for padding, convolution, and pooling layers
"No. of filters" for convolution blocks
"No. of hidden units" for dense layer

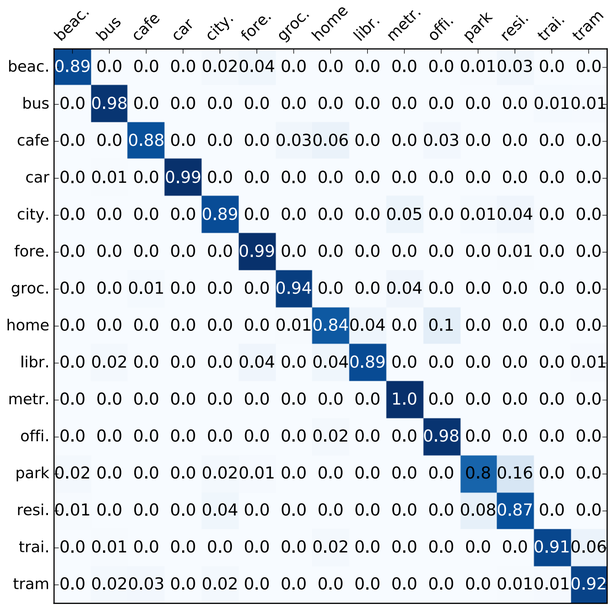
## Results

- Proposed approach achieved 2nd place in DCASE 2017 task1. For the challenge, we submitted total 4 submissions.
  - Evaluation set accuracy

| | |
|---|---|
| 4f + ME : 0.796 | full + ME : 0.803 |
| 4f + ES : 0.799 | full + ES : 0.804 |

  4f = Average probabilities of 4-fold CV models
  full = Model trained from whole development set
  ME = Mean ensemble
  ES = Ensemble selection
- 4-fold cross-validation results on development set are shown in the table, and confusion matrix of ensemble selection model is shown at right.
- Optimal weights found by ensemble selection were 26, 25, 21, 23, 29, 33, 17, 12, and 7 for LR to BS following the order listed in the table, which can be considered as a contribution of each model.

| Algorithms | Mean Acc. | Algorithms | Mean Acc. |
|---|---|---|---|
| Baseline | 0.748 | BS (2.0 s, 1) | 0.816 |
| Mono | 0.844 | BS (0.5 s, 11) | 0.861 |
| LR | 0.871 | BS (1.0 s, 11) | 0.856 |
| MS | 0.879 | BS (2.0 s, 11) | 0.843 |
| HPSS | 0.869 | Mean ensemble | 0.917 |
| BS (0.5s, 1) | 0.801 | Ensemble sel.* | 0.919 |
| BS (1.0s, 1) | 0.805 | | |




## Conclusion

- The main contribution of this paper is proposing various preprocessing methods that are useful for ConvNet, and presenting a 2-conv. model to make the most of the spatial information contained in the stereo.
- Especially, background subtraction showed an interesting result. Although each model showed just moderate identification accuracy, using background subtraction results from various median filtering kernel sizes for the ensemble model greatly improved the performance.
- Using an ensemble selection method improved accuracy for both CV and evaluation set results, but its effect was almost negligible.
- There was considerably huge gap between the performance from CV and evaluation set. We are planning to investigate further on this matter.