LEARNING ENVIRONMENTAL SOUNDS WITH END-TO-END CONVOLUTIONAL NEURAL NETWORK

Yuji Tokozume and Tatsuya Harada, The University of Tokyo

2017 IEEE International Conference on Acoustics, Speech and Signal Processing
ICASSP 2017
March 5-9, 2017, New Orleans, USA

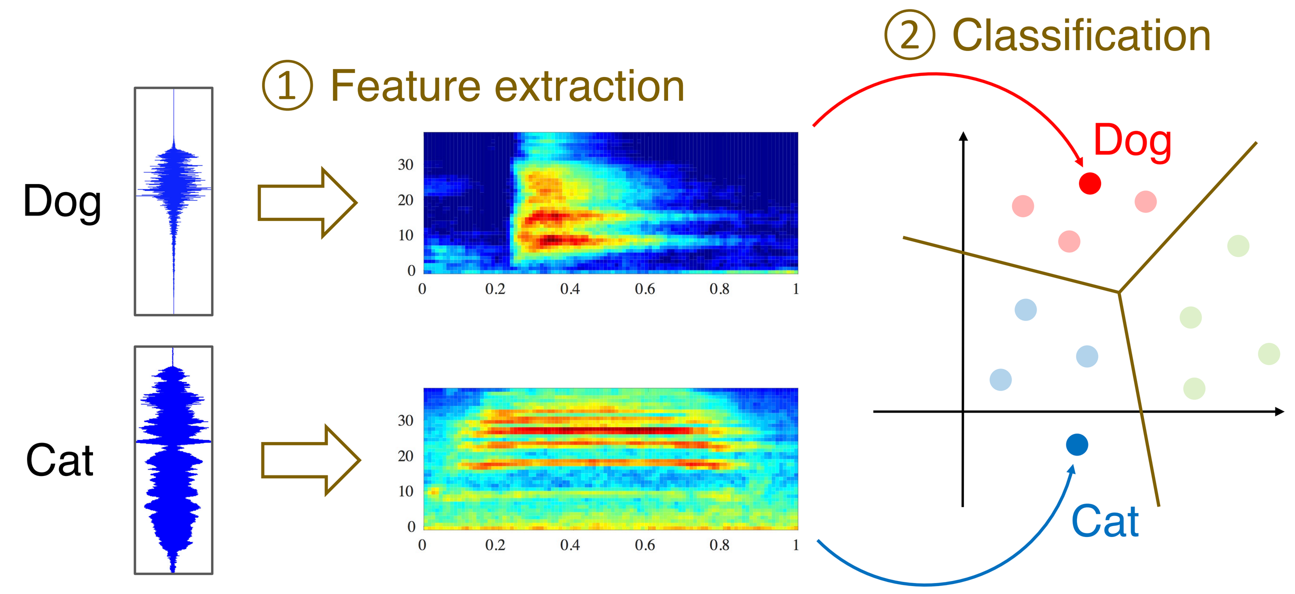
Summary

- We proposed an end-to-end environmental sound classification (ESC) system with a CNN
- We achieved a 6.5% improvement in classification accuracy over the state-of-the-art logmel-CNN, simply by combining our system and logmel-CNN
- We analyzed the feature learned with our system, and showed that our end-to-end system is capable of extracting a discriminative feature that complements the log-mel feature

Introduction

Background & Goal

- ESC is usually conducted based on spectral features such as the log-mel feature
- These features are designed by humans separately from other parts of the system
  → There could be other effective features of ESC



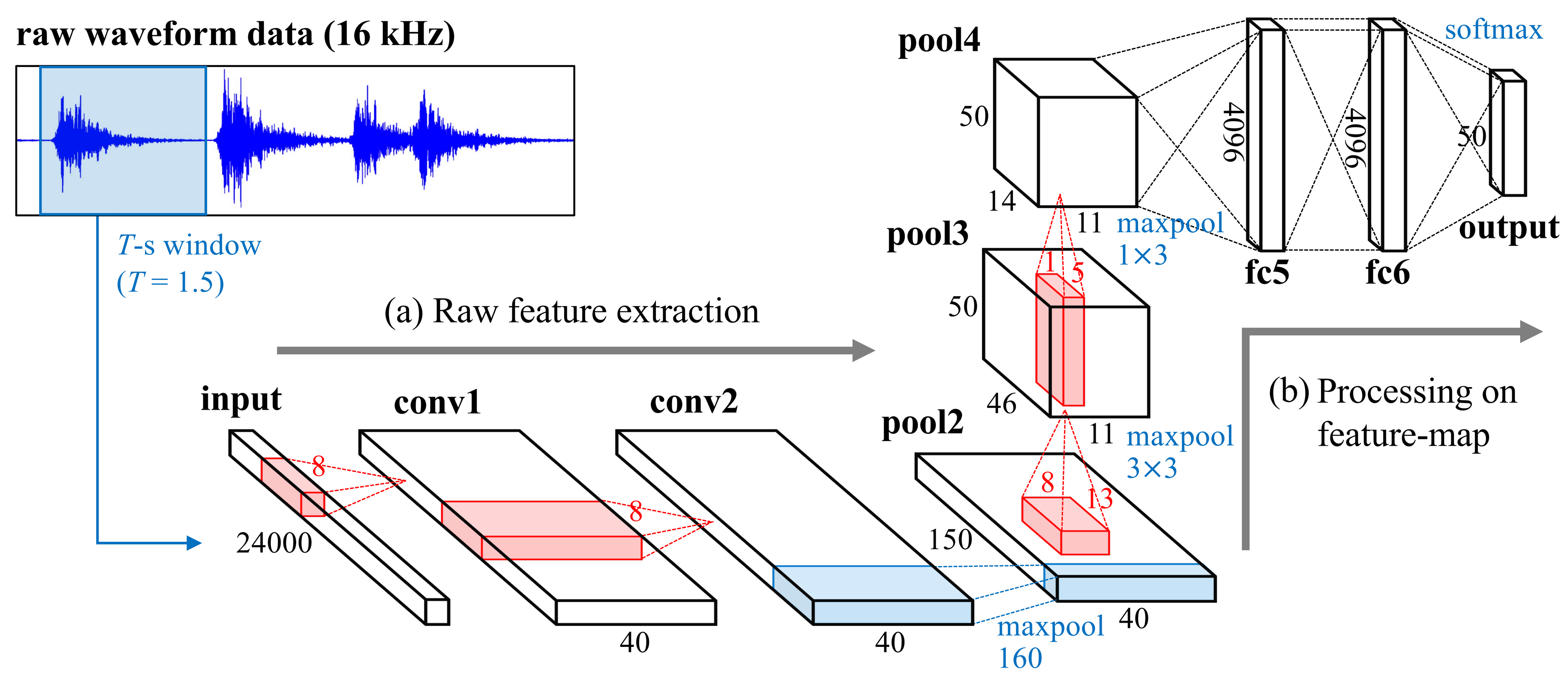
- If environmental sounds could be directly learned from the raw waveform,
  - We would be able to extract a new feature representing information different from the log-mel feature
  - This new feature could contribute to the improvement of classification performance

➢ **Goal: End-to-end ESC system**

Related work

- Log-mel feature + CNN [Piczak, 2015]
  - State-of-the-art method of ESC
- End-to-end speech recognition [Sainath et al., 2015]
  - Performance matches the static log-mel feature

End-to-end ESC system



EnvNet: End-to-end convolutional neural network for environmental sound classification

Overview

- Input: fixed T-s raw waveform
  - 16 kHz, range from -1 to 1
- Output: class probabilities
- Data augmentation
  - Training: random cropping (max amplitude > 0.2)
  - Test: probability voting (create a sliding window and take the average of all the softmax outputs)

Network architecture

- Raw feature extraction (a)
  - 1-D convolutional and pooling layers
  - Pool2: 40 types of frequency features per 10 ms
- Processing on feature-map (b)
  - 2-D convolutional and pooling layers
  - Finally, classify sounds with fully connected layers

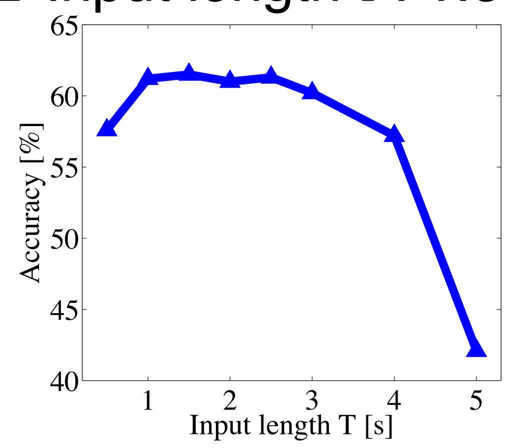
Experiments

Settings

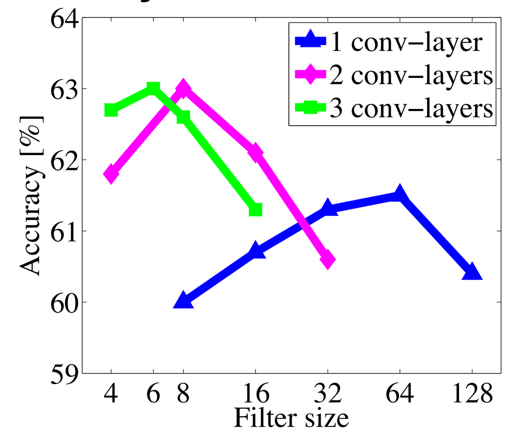
- Dataset: ESC-50 [Piczak, 2015]
  - Total: 50 classes, 2,000 samples
  - Each sample: monaural, 5 seconds, 44.1 kHz
- Evaluation: 5-fold cross-validation
  - 1,200 samples for training, 400 for validation, 400 for testing

Initial experiments

- Input length T: 1.5



- Conv-layers for raw feature extraction: 2 layers with size 8



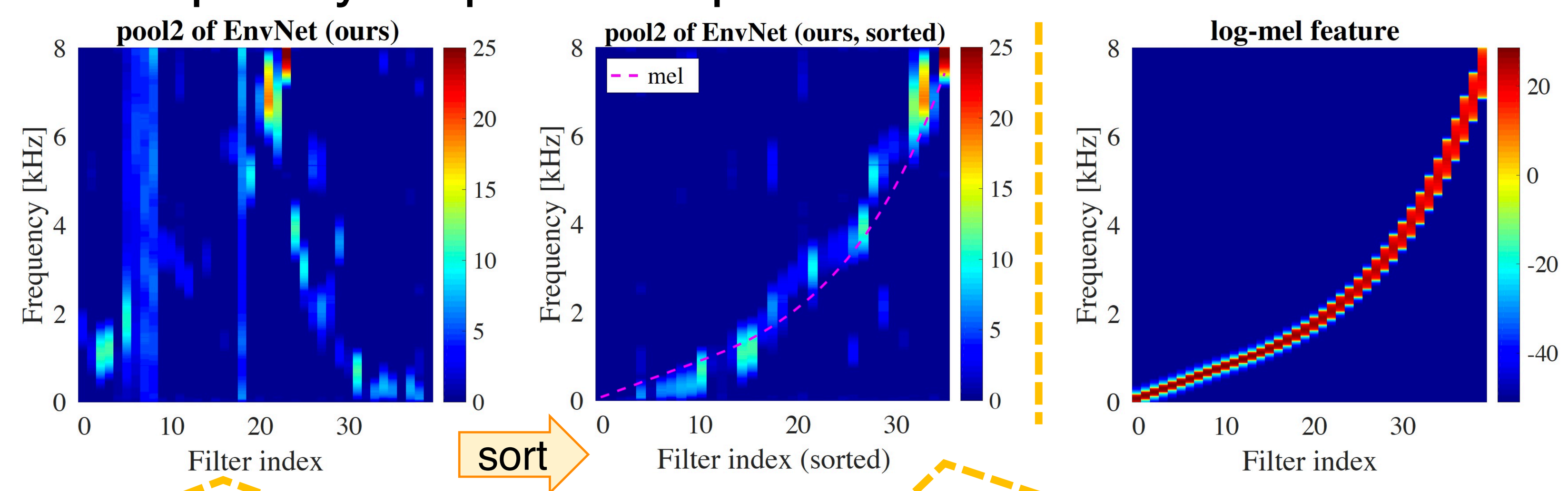
Main results

logmel-CNN static	logmel-CNN delta	EnvNet (ours)	Accuracy [%]
✓			58.9 ± 2.6
✓	✓		66.5 ± 2.8
		✓	64.0 ± 2.4
✓		✓	69.3 ± 2.2
✓	✓	✓	**71.0 ± 3.1**
Piczak logmel-CNN			64.5
Human			81.3

- The accuracy of EnvNet is higher than static logmel-CNN by 5.1 %
- We achieve a state-of-the-art accuracy by combining EnvNet and logmel-CNN (averaging)

Analysis on learned feature

- Frequency response of pool2



- Each of the 40 filters responds to a particular frequency area
- Neighboring filters have a similar frequency response

If we sort the filters based on their center frequency, the curve of the center frequency almost matches the mel-scale, i.e., how humans perceive the sound

- EnvNet learns a frequency response which is quite similar to human perception, but the order of the filters is optimized to maximize the classification performance

➢ We conjecture that is why our EnvNet feature is effective and has the ability to complement the log-mel feature

Acknowledgement

This work was funded by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan) and supported by CREST, JST.