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

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

Overview

Input: fixed T-5 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
Processing on feature-map (b)
• 2-D convolutional and pooling layers
Finally, classify sounds with fully connected layers

Main results

logmel-CNN
EnpNet (ours)
Accuracy [%]
static delta (ours)
58.9 ± 2.6
66.5 ± 2.8
64.0 ± 2.4
69.3 ± 2.2
71.0 ± 3.1

Piczak logmel-CNN
Human
64.5
81.3

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

Analysis on learned feature

Frequency response of pool2

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.

EnpNet 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 EnpNet feature is effective and has the ability to complement the log-mel feature.