Time-Frequency Feature Decomposition Based on Sound Duration for Acoustic Scene Classification

Yuzhong WU and Tan LEE

Department of Electronic Engineering
The Chinese University of Hong Kong

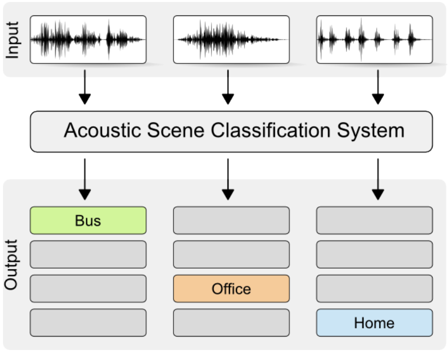
April 16, 2020

ICASSP2020 Barcelona

Task Definition

Acoustic scene classification (ASC) is the task of identifying the type of acoustic environment in which a given audio signal is recorded.

Figure: Overview of acoustic scene classification system. *(Image source: http://www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification)*



Characteristics of Acoustic Scene Signal

An acoustic scene signal is a mixture of sounds of diverse properties. It could contain

- long-duration or short-duration sound events in time domain
- wide-band or narrow-band sound events in frequency domain

Sound events are commonly overlapped in time and/or in frequency.

For example, acoustic scene signal recorded in a bus may simultaneously contain

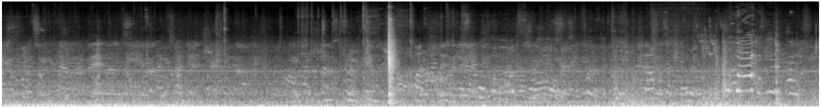
- bus engine sounds
- human speech
- sounds of nearby car horn

Audio Signal Represented as TF Feature

An audio signal can be represented as a time-frequency (TF) feature. For ASC, commonly used TF features are STFT, wavelet-based features, log-mel filter-bank, with

- x-axis representing time
- y-axis representing frequency

Figure: An example of TF feature representing an audio recorded in tram.



Why Feature Decomposition

In the DCASE challenges:

- Convolutional Neural Network (CNN) model has been widely adopted in the ASC task.
- Ensemble of models were found more accurate than a single model.
  - Remain unclear what specific aspects of scene information are addressed by individual component models.

We propose to decompose TF features based on sound duration.

- Facilitating detailed analysis on different types of acoustic scene information.
- Leveraging ensemble models with decomposed TF features.

Median Filtering for Images

In image processing, median filter is used to suppress impulse noise.

- Impulse noise: high positive pixel values concentrated locally in a small region.
- Moving-window median filter can suppress impulse events that are **narrower than half of the filtering window**.

Median Filtering for Time-Frequency Images

In a time-frequency image of an audio,

- Each pixel value indicates signal intensity at the respective time and frequency.
- Aggregations of pixels produce the acoustic patterns that can be perceived by human listeners as sound events.

The proposed feature decomposition method is based on the fact that

- Applying a median filter along the time axis would suppress impulse events of "short" duration (shorter than half of filtering window).
- Subtracting the filtered image from the original image results in an image that contains only "short" impulse events.

Proposed TF Feature Decomposition Method

- $\mathbf{S} = \mathbf{S_{long}} + \mathbf{S_{medium}} + \mathbf{S_{short}}$.
- Modify kernel sizes of median filters to control the sound information in each component image.

Algorithm 1 Proposed feature decomposition method with 2 median filters on time-frequency image.

Require:

The original time-frequency image, S;

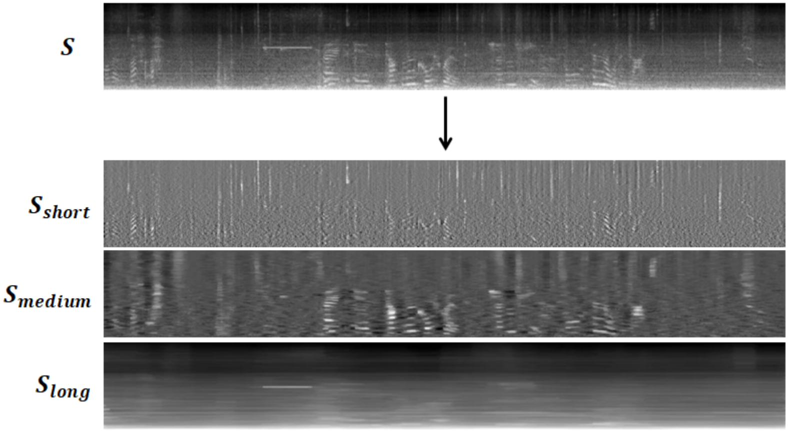
Median filtering function with small kernel size, M_s;

Median filtering function with large kernel size, M_l;

Procedure:

1: $S_r = M_s(S)$;
2: $S_{short} = S - S_r$;
3: $S_{long} = M_l(S_r)$;
4: $S_{medium} = S_r - S_{long}$;
5: **return** $(S_{long}, S_{medium}, S_{short})$;

Example of decomposing a TF image S



ASC System Design

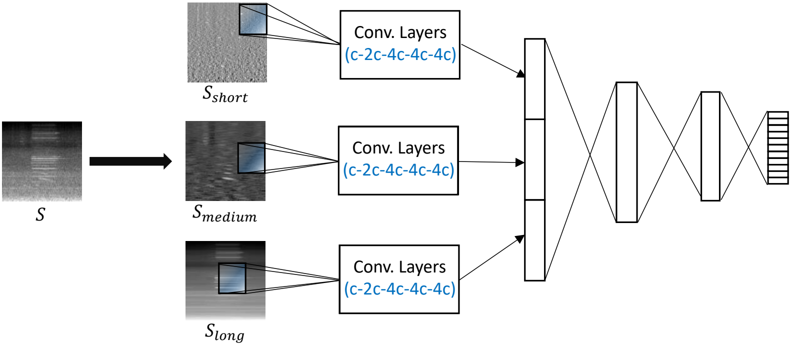


Figure: An illustration of our CNN model with 3 input channels. Independent feature extractor (convolution layers) is applied on each input channel. The "c-2c-4c-4c-4c" means the corresponding number of filters for the 5 convolution layers.

Structure of CNN Model

n is the number of input channels. c is used to control the number of filters in convolution layers.

1	Input $n \times 128 \times 128$
2	3x3 Convolution-BN-ReLU ($c \times n$ filters)
3	2x2 Max Pooling
4	3x3 Convolution-BN-ReLU ($2c \times n$ filters)
5	2x2 Max Pooling
6	3x3 Convolution-BN-ReLU ($4c \times n$ filters)
7	2x2 Max Pooling
8	3x3 Convolution-BN-ReLU ($4c \times n$ filters)
9	3x3 Convolution-BN-ReLU ($4c \times n$ filters)
10	2x2 Max Pooling
11	Flattening
12	Fully Connected (dim-1024)-BN-ReLU
13	Fully Connected (dim-256)-BN-ReLU
14	10-way Sigmoid

Dataset

The TAU Urban Acoustic Scenes 2019 development dataset (Mesaros et al. [2018])

- Used for subtask A of the DCASE 2019 ASC challenge.
- Each audio clip is 10-second long.
- 40-hour binaural audios from 10 different acoustic scene classes.
- Audios recorded with the same device.

We follow the training/test setup officially provided in the DCASE 2019 ASC challenge.

- Training set contains 9185 audio clips
  - covering about 70% of recording locations from 9 cities
- Test set contains 5215 audio clips
  - 4185 audio clips from the 9 cities (seen cities in training set)
  - 1030 audio clips from the 10th city Milan (unseen city).

Experiment Setup

CNN Training:

- 40 training epochs
- Initial LR is 0.0001, halved every 4 epochs
- Adam optimizer is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$
- Weight decay (coefficient $= 0.0015$) applied for regularization.

Data augmentation:

- The mixup approach (Zhang et al. [2017]).
- Temporal shifting the audio clips in training set.

Model Parameter Study

It can be seen that similar accuracy is achieved for $c \geq 16$.

- This may serve as an evidence for the following experiments that the significant performance gap between different configurations is not due to the change of model size.

Table: CNN model performance for different values of c.

Model Config	Input Feature	Accuracy
$c = 8$, $n = 1$	logmel	70.0%
$c = 16$, $n = 1$	logmel	72.5%
$c = 24$, $n = 1$	logmel	72.2%
$c = 48$, $n = 1$	logmel	72.8%

Decomposed Log-Mel Features

It can be seen from the experiment results that:

- It is helpful to explicitly learn long-lasting background sounds and transient sounds separately.
- All component images contain useful information for ASC.
- S_{long} contains the most pertinent information related to ASC.

Table: Performance of using the standard log-mel feature and the decomposed features.

Model Config	Input Feature	Accuracy
$c = 48$, $n = 1$	logmel	72.8%
$\mathbf{c = 16}$, $\mathbf{n = 3}$	**logmel-LMS**	$\mathbf{75.3\%}$
$c = 16$, $n = 2$	logmel-LM	74.3%
$c = 16$, $n = 2$	logmel-MS	70.0%
$c = 16$, $n = 2$	logmel-LS	73.7%
$c = 16$, $n = 1$	logmel-L	68.3%
$c = 16$, $n = 1$	logmel-M	60.8%
$c = 16$, $n = 1$	logmel-S	63.3%

Decomposed Log-Mel Features for ASC

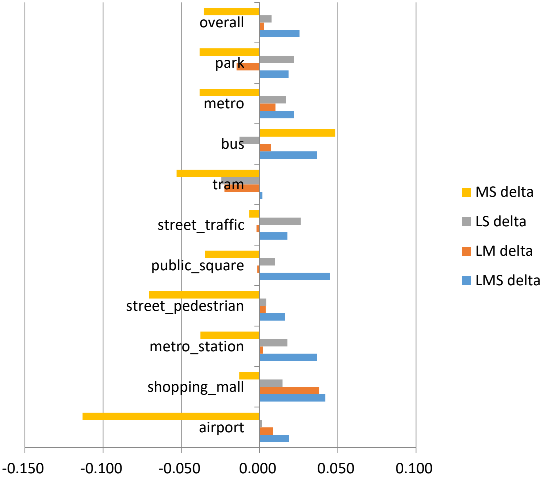


Figure: The F1 score difference between using log-mel image and decomposed log-mel images.

Decomposed Wavelet Filter-bank Features

Wavelet-based TF features were shown very effective in the best-performing system submitted to the DCASE 2019 ASC Challenge Subtask A (Chen et al. [2019]).

Table: Performance of using log-mel, wavelet filter-bank features (scalogram) and their decomposed features.

Model Config	Input Feature	Accuracy
$c = 48$, $n = 1$	logmel	72.8%
$c = 16$, $n = 3$	logmel-LMS	75.3%
$c = 48$, $n = 1$	scalogram	74.6%
$\mathbf{c = 16}$, $\mathbf{n = 3}$	**scalogram-LMS**	$\mathbf{76.7}\%$

Conclusions

A novel time-frequency feature decomposition method has been developed for audio scene classification.

- The CNN model is explicitly guided to learn long-lasting background sounds and transient sounds separately.
- Analysis of component images shows that long-duration sounds are most informative for ASC.
- Our decomposition method can be combined with wavelet based time-frequency features to obtain a further improved accuracy.

Reference I

H. Chen, Z. Liu *et al.*, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," DCASE2019 Challenge, Tech. Rep., June 2019.

A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *arXiv e-prints*, p. arXiv:1710.09412, Oct 2017.