ACOUSTIC SCENE CLASSIFICATION BY ENSEMBLE OF SPECTROGRAMS BASED ON ADAPTIVE TEMPORAL DIVISIONS

Technical Report

Yuma Sakashita, Masaki Aono

Toyohashi University of Technology
Computer Science and Engineering, Aichi, Japan
sakashita@kde.cs.tut.ac.jp, aono@tut.jp

ABSTRACT

Many classification tasks using deep learning have improved classification accuracy by using a large amount of training data. However, it is difficult to collect audio data and build a large database. Since training data is restricted in DCASE 2018 Task 1a, unknown acoustic scene must be predicted from less training data. From the results of DCASE 2017[1], we determine that using a convolution neural network and ensemble multiple networks is an effective means for classifying acoustic scenes. In our method we generate mel-spectrogram from binaural audio, mono audio, Harmonic-percussive source separation (HPSS) audio, adaptively divide the spectrogram into multiple ways and learn 9 neural networks. We further improve ensemble accuracy by ensemble learning using these outputs. The classification result of the proposed system was 0.769 for Development dataset and 0.796 for Leaderboard dataset.

Index Terms— DCASE 2018, acoustic scene classification, convolutional neural network, mixup, harmonic- percussive source separation, ensemble, stacking, Random Forest, SVM

1. INTRODUCTION

Audio information obtained by auditory sense plays a very important role for human behavior. Human ears are trained by everyday life and can grasp surrounding circumstances even from fine sounds. For example, if you hear the birds singing in a quiet environment, the place is outside, where there are many foods, you can see that there are easy-to-stop trees of birds. If you have more knowledge you can also distinguish seasons and time from bird types. If the computer can automatically recognize the acoustic scene at the same level as a human being, it can be applied to various fields. For example, autonomous robots are currently mostly those that recognize information obtained from cameras and people's words. In addition to these, if it is possible to recognize the acoustic scene, it can be considered that it is possible to change the behavior of the robot and to give variations to the dialogue. However, the environmental sound continues to change over time, and the same sound will not necessarily occur again. Humans can respond flexibly to trivial changes in sound depending on experience, but it is extremely difficult to automate with computers. The acoustic scene classification (ASC) is one of the research subjects which is currently actively undertaken and DCASE hosted by IEEE Audio and Acoustic Signal Processing (AASP) is one of the large tasks of ASC research. The method that achieves top rank in DCASE has been changing year by year. DCASE 2016[2] achieved the top ranking method using the conventional dictionary learning method i-Vector[3] and NMF (Non-Negative Matrix factorization)[4]. In DCASE 2017, most of the top is a method using a convolution neural network (CNN). Han et al.[5] ensemble their outputs after learning neural networks with spectrograms generated from binaural audio, HPSS, and Background Subtraction. In DCASE 2018, the number of data increased compared with DCASE 2017, but it can not be said that it is still satisfactory. We divided mel-spectrogram by plural division methods and learned CNN respectively. Furthermore, we confirmed that the output was ensemble, and further improvement of classification accuracy was attempted. The following section explains the details of the system we proposed, experimental results, and conclusion.

2. SYSTEM ARCHITECTURE

This section describes the audio preprocessing method used in this experiment. It also describes the architecture of the neural network.

2.1. Audio Preprocessing

We use mel-spectrogram as audio feature. Mel-spectrogram is used by most of the top teams of DCASE 2017 and is considered to be most suitable for acoustic scene classification.
Because the DCASE 2018 data set is sampled at 48 kHz, downsample to 44.1 kHz. Next, frequency and phase are analyzed by short-time Fourier transformation(STFT). STFT can calculate the spectrum at each time by looking at the time change by multiplying the window function little by little. The window function of STFT is a hann window, the window size is 2,048 samples (46 ms), and the hop size is 1,024 samples (23 ms). Finally, mel-spectrogram is obtained by applying Mel filter bank. The number of bandpass filters was 128, and the HTK method was used. n the HTK method, Hz is converted to mel by using the following equation.

$$\mathrm{mel} = 2595.0 \log 10(\frac{1.0 + \mathrm{frequencies}}{700.0}) \tag{1}$$

Mel spectrogram was converted to a logarithmic scale, normalized by dividing by the standard deviation subtracting the mean value.

2.1.1. Binaural audio feature

The DCASE 2018 dataset is recorded using binaural microphone. So you can use binaural audio data of 2 channels (Left and Right). From the experimental results of Han et al.[5], It turns out that using 2-channel audio leaves better results than using mono audio. It is

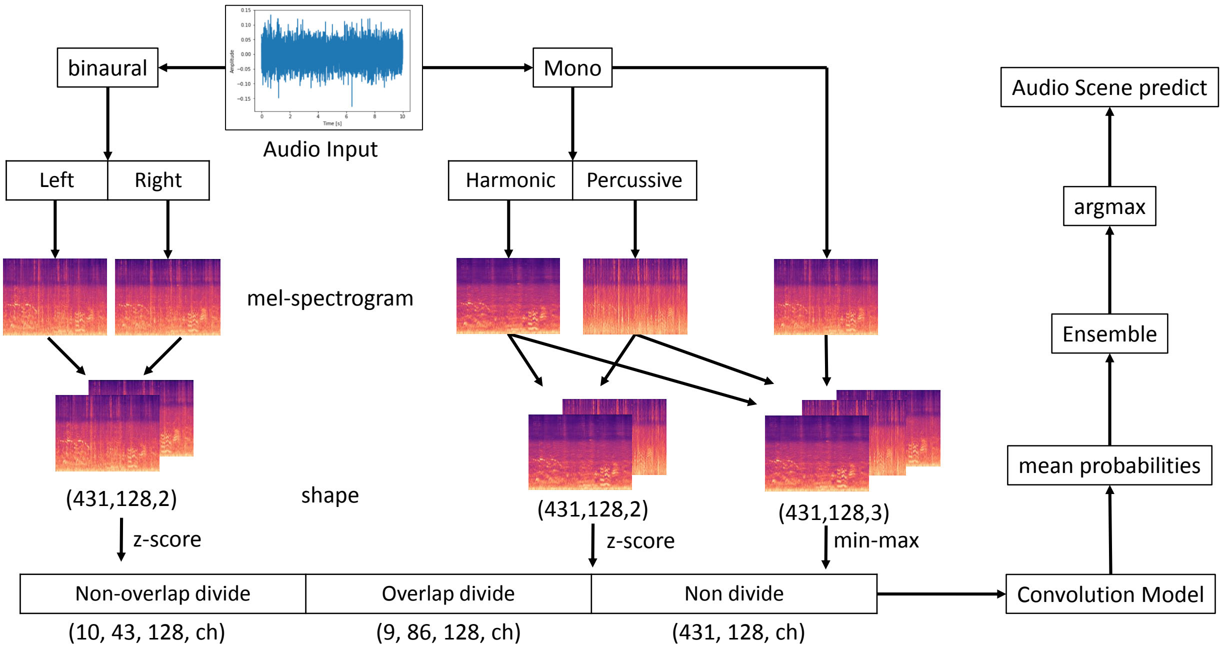

Figure 1: Architecture of the proposed system.A plurality of mel-spectrograms are generated from one piece of audio data, and the spectrogram is divided by three dividing methods. We learn Network using these, and finally Ensemble learning.

presumed that it is a factor that holds more spatial information than mono audio. (Example: cars and trains move from right to left) By calculating the mel-spectrogram with the parameters of section 2.1, you can obtain data of $(431, 128, 2)$ shape from one audio clip.

2.1.2. *Harmonic-percussive source separation*

Similarly to Han et al.[5], Mel-spectrogram is also obtained from Harmonic-percussive source separation (HPSS) applied to mono audio. For HPSS, librosa which is a Python package for music and audio analysis is used, and initial values are used for parameters. In order to separate it into Harmonic audio and Percussive audio, it is possible to calculate the mel-spectrogram of two channels as in **??**. By calculating the mel-spectrogram with the parameters of section 2.1, you can obtain data of $(431, 128, 2)$ shape from one audio clip.

2.1.3. *Proposed feature*

We propose new features inspired by image features. Many tasks of image classification use data of three channels of R, G, B. Among the studies using deep learning, the task of image classification has been developed particularly, and many methods have been published in recent years. Create features of 3 channels so that the method of image classification task can be applied to sound classification. In addition to the HPSS in section 2.1.2, add the calculated mel-spectrogram of mono audio and create the features of 3 channels of (431, 128, 3). Binaural feature and hpss feature are normalized by z-score, but proposed features are normalized using min-max normalization.

2.2. Spectrogram Division

The features obtained in 2.1 are divided by three methods. The first is non-overlapping division. Spectrogram is divided every 1 second and ten $(43, 128, channel)$ features can be obtained from one audio clip. The second is overlapping division. Divide features every 2 seconds with half overlapping. Nine (86, 128, channel) features can be obtained from one audio clip.

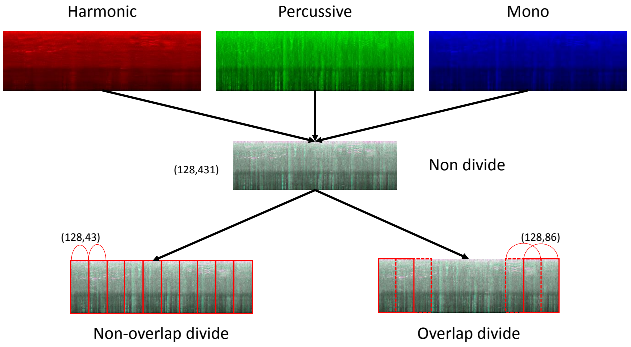

Figure 2: Spectrogram division.Here we use the propose feature as an example.

2.3. Network Architecture

The Network we used is 1.Conv medel proposed by Han et al[5]. This neural network is a convolution model constructed by inspired by VGGNet[6]. It is characteristic that Batch Normalization

(BN)[7] is used instead of Dropout. In contrast, we combine the features of the two channels and input them into one 1.Conv model. We also considered using Recurrent Neural Network (RNN) in addition to CNN, but judged that it is not suitable for acoustic scene classification from the results of preliminary experiments. In music and conversation, time series data is an important role, but in the acoustic scene it can be inferred that spatial information (echo of sound etc.) is more important.

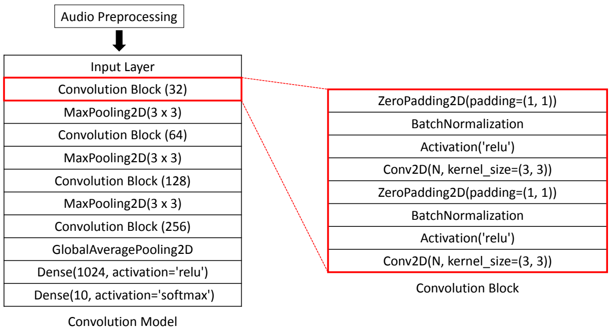

Figure 3: Convolution model used for experiment.Numbers in parentheses of Convolution Block indicate the number of output filters of Conv 2D layer.

3. DATA AUGMENTATION

Since training data is restricted in DCASE2018 Task1a, it is necessary to perform data augmentation to increase the flexibility to unknown data. Mun et al.[8] performed data augmentation using Generative Adversalial Network (GAN) and achieved Top Rank in DCASE2017 Task1. From this, it can be seen that data auditionation greatly affects the sound scene classification. As a major data augmentation method in the speech field, there are addition of Noise, pitch shift, time delay, but neither method is considered to be effective from the result of DCASE 2017.
We use mixup[9] for the data augmentation method. Mixup creates a new training sample by mixing a pair of two training samples. Create a new training sample (X, y) from the data and label pair (X_1, y_1),(X_2, y_2) by the following equation.

$$\begin{aligned} X &= \lambda X_1 + (1-\lambda)X_2 \\ y &= \lambda y_1 + (1-\lambda)y_2 \end{aligned} \tag{2}$$

Here, $\lambda \in [0, 1]$ is acquired by sampling from the beta distribution $Be(\alpha, \alpha)$, and α is a hyper parameter. Besides the data X_1 and X_2, it is characteristic to mix the labels y_1 and y_2.

4. EXPERIMENTS

4.1. Datasets

The dataset for this task is the TUT Urban Acoustic Scenes 2018 dataset, consisting of recordings from various acoustic scenes. The dataset was recorded in six large European cities, in different locations for each scene class. For each recording location there are 5-6 minutes of audio. The original recordings were split into segments with a length of 10 seconds that are provided in individual files. The dataset includes 10 scenes which are Airport，Indoor, shopping mall，Metro station，Pedestrian street，Public square，Street with medium level of traffic，Travelling by a tram，Travelling by a bus，Travelling by an underground metro，and Urban park. The dataset was collected by Tampere University of Technology between 01/2018 - 03/2018. TUT Urban Acoustic Scenes 2018 development dataset contains only material recorded with single recording device, having 864 segments for each acoustic scene(144 minutes of audio). The dataset contains in total 8640 segments, i.e. 24 hours of audio. Compared to DCASE 2017 dataset, the scene is decreasing, but the data has increased significantly.

4.2. Experiment Settings

This experiment uses the Development Dataset setup provided by the organizer. Stochastic Gradient Descent (SGD) using Nesterov momentum[10] was used for Optimizer of the network. Learning rate, decay, and momentum were 0.01, 0.0001, and 0.9, respectively. The mini batch size differs depending on the division method. 128 for non-overlap divide or Overlap divide and 32 for Non divide divide. It took us about three hours to train the network with NVIDIA Tesla K40 and train one network. We used 15% of Training data for Validation data so that each scene is selected equally.

4.3. Network Ensemble

We combine multiple classifiers to reduce prediction error. In this experiment we used an ensemble learning method called Stacking. Stacking learns the relationship between the output of multiple sorting machines and the true output by machine learning. Random Forest Classifier (RFC) was used for learning. The number of decision trees was set to 1000, 2000, 3000.

5. RESULT

Table.1 shows the experimental results. In all results, it exceeds the accuracy of Baseline system. Also, by taking the mean probability of all networks, you can see that the accuracy is greatly improved. Since Ensemble model uses all data of Development dataset, it describes only Accuracy of Leaderboard dataset.

6. CONCLUSION

In this paper, we described how to identify acoustic scenes using multiple spectrogram divide methods. In addition, we propose new audio features inspired by image features. We trained nine neural networks from the generated features and further improved accuracy by ensemble learning using these outputs. As a result, accuracy of 0.796 was obtained in the Leaderboard dataset. However, the method of this paper requires training of many networks, which is not an excellent method from the viewpoint of computational resources. In future, we plan to pursue various parameter adjustment and application method of image classification method.

7. REFERENCES

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/.

[2] http://www.cs.tut.fi/sgn/arg/dcase2016/.

Algorithms	Accuracy (Development)	Accuracy (Leaderboard)
Baseline	0.597	0.625
LR-NONDIVIDE	0.680	
LR-DIVIDE	0.699	
LR-OVERLAP	0.703	
HPSS-NONDIVIDE	0.689	
HPSS-DIVIDE	0.729	
HPSS-OVERLAP	0.720	
PROPOSE-NONDIVIDE	0.692	
PROPOSE-DIVIDE	0.687	
PROPOSE-OVERLAP	0.712	
Mean probability	0.769	0.771
Ensemble(RFC1000)		0.791
Ensemble(RFC2000)		0.796
Ensemble(RFC3000)		0.793

Table 1: Classification results of Development dataset and Leaderboard dataset.As for Ensemble, since we used all the data of Development dataset, it is the result of Leaderboard only.The numerical value of RFC shows the number of decision trees.

Scene label	Baseline(%)	Propose(%)
Airport	72.9	79.6
Bus	62.9	69.8
Metro	51.2	67.8
Metro station	55.4	85.3
Park	79.1	88.0
Public square	40.4	50.0
Shopping mall	49.6	70.9
Street, pedestrian	50.0	80.1
Street, traffic	80.5	91.8
Tram	55.1	83.1
Average	59.7	76.9

Table 2: Audio classification accuracy per scene.

[3] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, “CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks,” *IEEE AASP Challenge on DCASE 2016 technical reports*, 2016.

[4] V. Bisot, R. Serizel, S. Essid, and G. Richard, “Supervised nonnegative matrix factorization for acoustic scene classification,” *IEEE AASP Challenge on DCASE 2016 technical reports*, 2016.

[5] Y. Han, J. Park and K. Lee, “Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification,”, *IEEE AASP Challenge on DCASE 2017 technical reports*, 2017.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in arXiv:1409.1556, 2014.

[7] S. Ioffe and C. Szegedy, “Batch normalization:Accelerating deep network training by reducing internal covariate shift,” *International Conference on Machine Learning*, 2015, pp. 448-456.

[8] S. Mun, S. Park, et. al. “Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane,”, *IEEE AASP Challenge on DCASE 2017 technical reports*, 2017.

[9] H. Zhang, M. Cisse, Y. N. Dauphin,and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in arXiv:1710.09412, 2017.

[10] Y. Nesterov, et. al. “Gradient methods for minimizing composite objective function,” *CORE DISCUSSION PAPER* 2007.

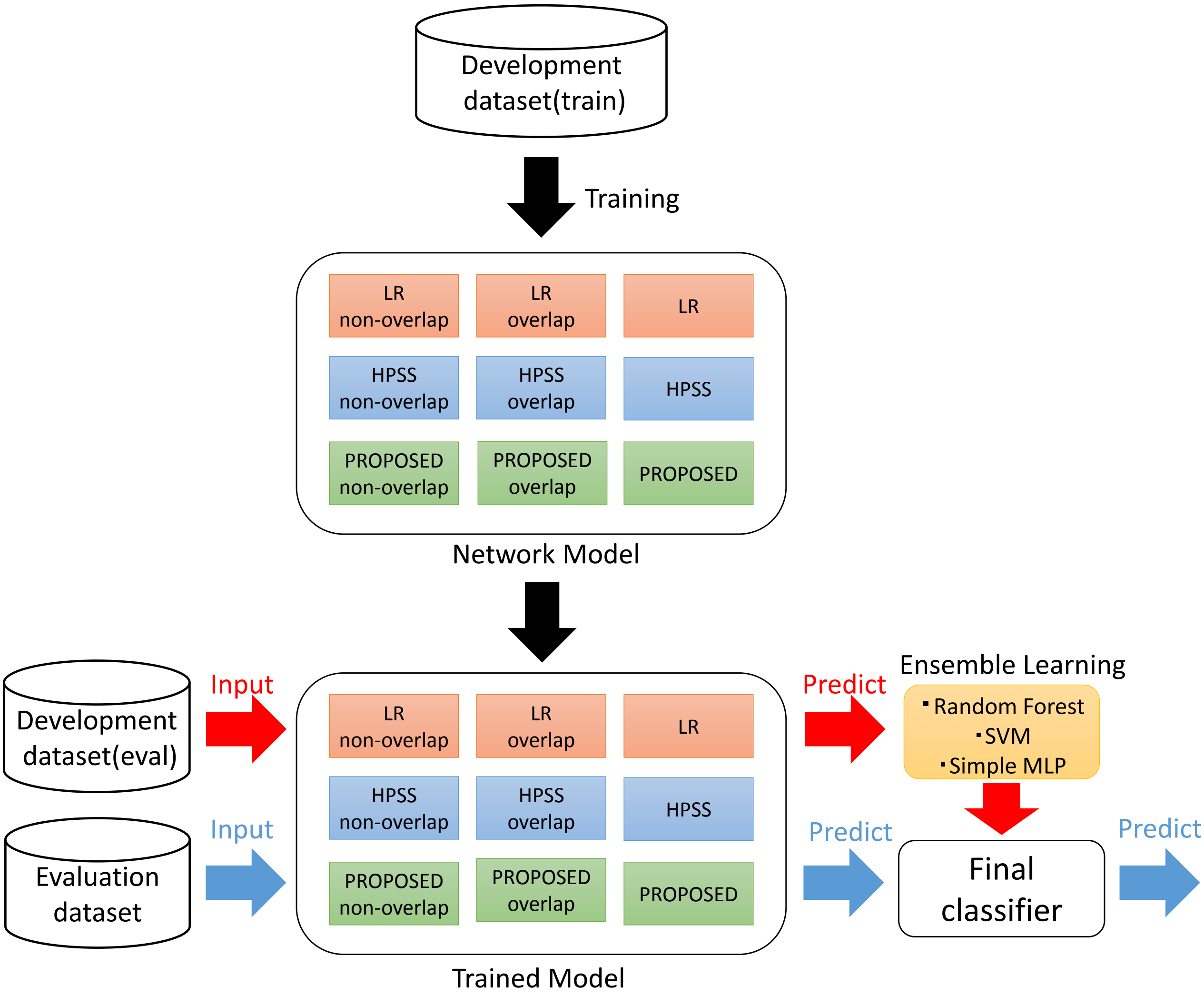

Figure 4: Network Ensemble.