Acoustic Scene Classification with Multiple Devices Based on Ensemble Learning

YeRin Lee[1], Il-Youp Kwak[2]

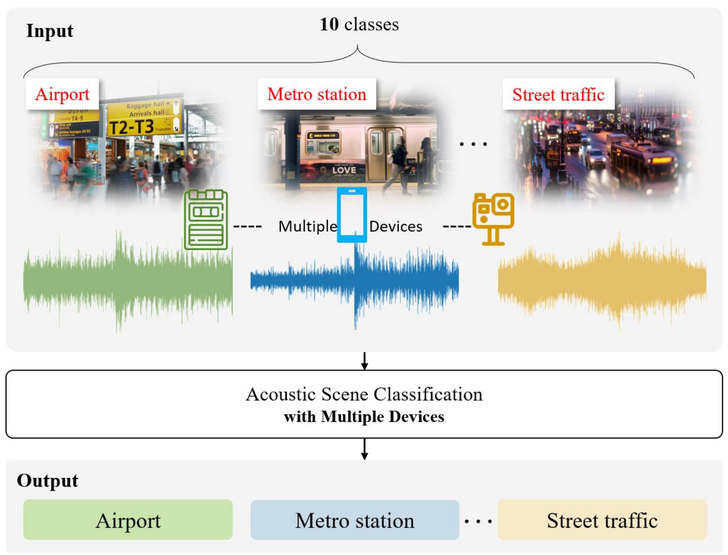
[1]Department of Statistics, College of Business & Economics, Chung-Ang University, Seoul, Republic of Korea

[2]Department of Applied Statistics, College of Business & Economics, Chung-Ang University, Seoul, Republic of Korea

Introduction

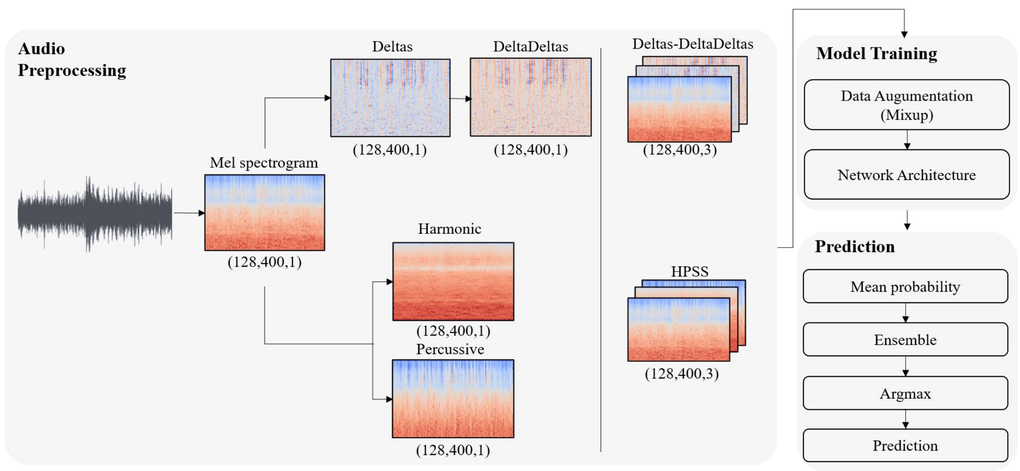
- Acoustic scene classification (ASC) categorizes an audio file based on the environment in which it has been recorded.
- ASC can be used to a smartphone that automatically changes to silent mode or adjusts the volume based on the location, or to a hearing aid that adjusts its function according to indoor or outdoor recognition.
- In addition, it can be performed as a pre-processing step to address other problems such as seperating the source of the speech signal from background noise.
- We addressed a challenge that ASC faces in real-world applications.
- It is that the audio recorded using different recording devices should be classified in general.
- We proposed a more general classification model by combining the harmonic-percussive source separation (HPSS) and deltas-deltadeltas features with four different convolutional neural network models.
- Moreover, we used the ensemble technique to learn not just one classifier but a set of classifiers.

Figure 1. Overview of the ASC system.



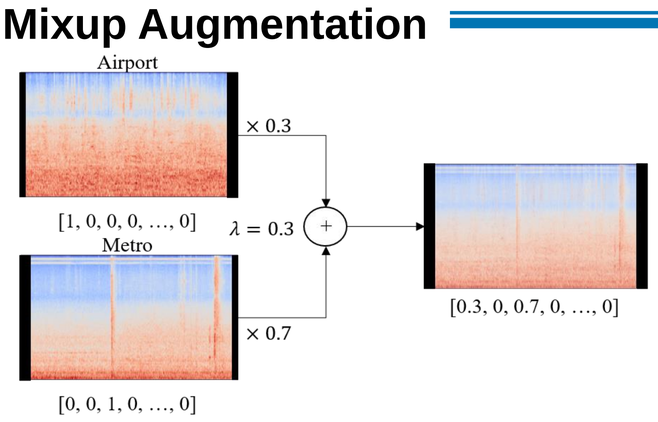
- We aim to classify audio into 10 classes.
- Related audio files are recorded or simulated using 10 devices.
- But 3 devices are not used for training, only for testing.
- The dataset includes 40h of data from device A and and smaller amounts of data from other devices.

Figure 2. Proposed system architecture



- We used log-mel spectrogram.
- The audio files are monaural and have a sampling rate of 44.1 kHz.
- To generate each spectrogram, we used 2048 fast Fourier transform basis functions, a hop-length of 1024 samples, 128 frequency bins, and the hidden markov model toolkit (HTK) to convert Hz to mel. Subsequently, we extracted a log-mel spectrogram.
- The HPSS decomposes monaural audio into two channels: (1) harmonic and (2) percussive,. Harmonic sound is perceived as pitched sound and enables us to hear melodies and chords.
- In contrast, percussive sound is more related to noise and usually stems from instrument onsets, e.g., hitting on a drum.
- An important characteristic of percussive sounds is that they do not have pitch but have clear localization in time.
- The deltas and delta-deltas indicate the first and second temporal derivatives of the spectrogram, respectively

Figure 3 . Mixup Augmentation

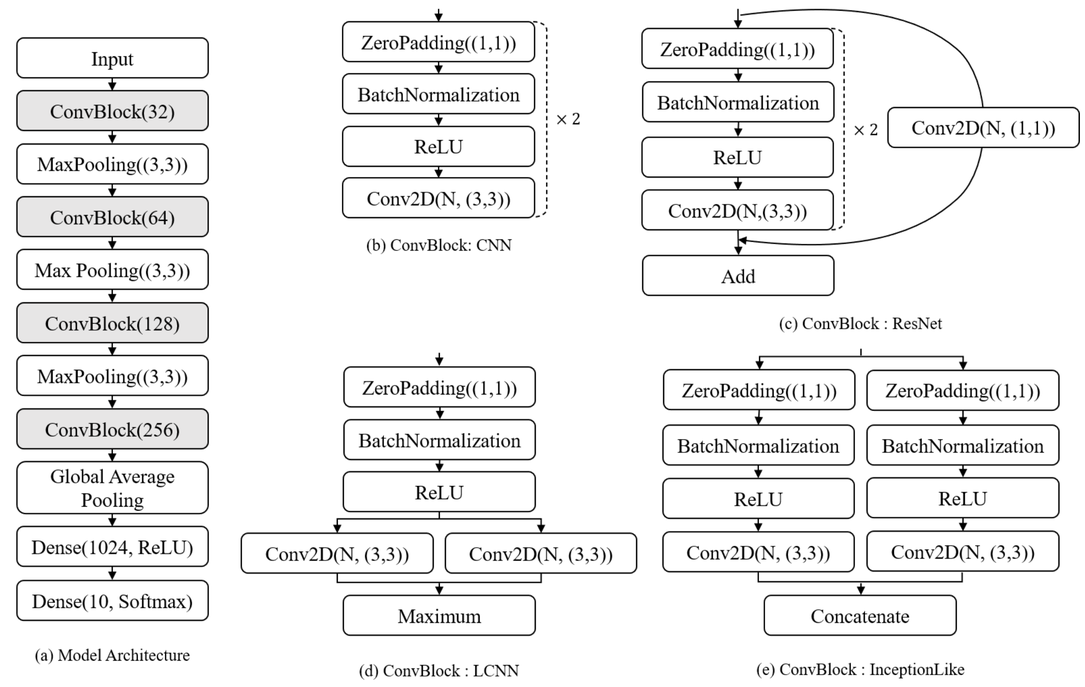


- Mix- up is an effective data agumentation method.
- We used a general augmentation approach: we mixed different samples of the training set according to their weights.
- The method is a s follows:

$$X = \lambda X_i + (1-\lambda)X_j$$
$$y = \lambda y_i + (1-\lambda)y_j$$

- Where $\lambda \in [0,1]$ and is acquired by the sampling of the beta distribution with parameter α, $\beta(\alpha,\alpha)$, $\alpha \in (0,\infty)$. X_i and X_j are different data samples; y_i and y_j are their corresponding labels.
- In our experiment, we used the mixup to augment the log-mel spectrograms.
- We set α at 0.4 used crop augmentation as 400 on the temporal axis before the mixup augmentation.

Figure 4. CNN architecture.



- (a) General model architecture.
- (b--e): Four different modeling architectures for ConvBlock inspired by VGGNet, ResNet, LCNN, and InceptionNet, respectively.
- The first architecture for ConvBlock (b) is a convolutional model based on VGGNet.
- The second architecture, (c) ,uses the skip-connection network in ResNet.
- The third one, (d), is an LCNN module that uses Max-Feature-Map (MFM) activation inside the skip-connection network.
- The fourth one, (e), is based on InceptionNet that concatenates two tensor.

Table 1, 2, 3. Results.

Model	Model Configuration (Subtask A)	Accuracy (%)	Recall (%)	Precision (%)
1	HPSS-CNN	59.36	59.36	60.27
2	HPSS-ResNet	58.55	58.55	59.93
3	HPSS-LCNN	58.08	58.08	58.35
4	HPSS-InceptionLike	58.69	58.69	58.97
5	Deltas-DeltaDeltas-CNN	63.43	64.43	63.60
6	Deltas-DeltaDeltas-ResNet	64.21	64.21	65.23
7	Deltas-DeltaDeltas-LCNN	64.11	64.11	64.58
8	Deltas-DeltaDeltas-InceptionLike	63.94	63.94	63.94
9	HPSS-Ensemble	64.04	64.04	64.33
10	Deltas-DeltaDeltas-Ensemble	69.16	69.16	68.86
11	All-Ensemble	68.62	68.62	68.61

- Each model appilied to Deltas-DeltaDeltas had an accuracy of over 60%, while the same model applied to HPSS had a lower accuracy.
- The accuracy of Deltas-DeltaDeltas-Ensemble is 69.16%, which is 5.12% higher than that of the HPSS-Ensemble. In addition, it has higher accuracy than the all-ensemble that ensembles all eight models.

Classes	HPSS-CNN (%)	Deltas-DeltaDeltas-ResNet (%)	Deltas-DeltaDeltas-Ensemble (%)
Airport	48.48	50.17	56.57
Bus	62.96	78.45	80.47
Metro	57.58	64.31	69.70
Metro Station	62.63	68.01	69.02
Park	69.36	80.13	85.19
Public Square	46.13	45.45	52.86
Shopping Mall	66.00	65.32	67.34
Street Pedestrian	33.00	42.42	48.48
Street Traffic	78.79	81.82	84.18
Tram	68.69	66.00	77.78

- Class-wise results.
- With Deltas-DeltaDeltas-ResNet, the accuracy improved for eight classes, excluding Shopping mall and Trame, compared with HPSS-CNN, the baseline model.
- In particular, the accuracy significantly increased for Bus, Park, and Streed Pedestrian.
- Deltas-DeltaDeltas-Ensemble was used, the accuracy for the Tram significantly increased compared to when the single Deltas-DeltaDeltas-ResNet model was used.

Devices	HPSS-CNN (%)	Deltas-DeltaDeltas-ResNet (%)	Deltas-DeltaDeltas-Ensemble (%)
A	74.45	72.42	80.00
B	64.24	68.79	73.03
C	66.67	69.39	76.67
S1	62.42	64.85	70.90
S2	53.94	58.79	63.03
S3	59.70	61.62	70.60
S4	52.42	58.18	64.24
S5	53.64	56.06	62.42
S6	46.67	56.06	61.51

- Device-wise results.
- Devices A, B, C, and S1—S3 are seen devices, and S4—S6 are unseen devices.
- In all the devices except device A, when Deltas-DeltaDeltas-ResNet was used, the classification accuracy was higher than when HPSS-CNN.
- In particular, the accuracy of the unseen devices (A, B, C, S4—S6), which were not used for training, relatively increased compared with the seen devices (S1—S3).
- Deltas-DeltaDeltas-Ensemble is more accurate than HPSS-CNN and Deltas-DeltaDeltas-ResNet for all devices.

Summary

- We addressed Acoustic scene classification in complex environments with multiple devices.
- This paper presents our novel architectures.
- We considered two features, Deltas-DeltaDeltas and HPSS, and four models inspired by VGGNet, ResNet, LCNN, and InceptionNet.
- In all the four models, the use of Deltas-DeltaDeltas surpassed the performance of HPSS, and among them the use of ResNet exhibited the highest accuracy.
- And We proposed ensemble techniques to learn a set of classifiers as well as one classifier.
- This model is a modification of our model that ranked 9th in DCASE 2020.

This work was supported by the National Research Foundation of Korea (NFR) grant funded by Ministry of Science and ICT (2020R1C1C1A01013020)