Animal Sounds Classification Scheme Based on Multi-Feature Network with Mixed Datasets

Chung-Il Kim, Yongjang Cho, Seungwon Jung, Jehyeok Rew, and Eenjun Hwang*
School of Electrical Engineering, Korea University
Seoul, South Korea
[e-mail: cilkim1@korea.ac.kr, jsw161@korea.ac.kr, dydwkd486@korea.ac.kr, rjh1026@korea.ac.kr, ehwang04@korea.ac.kr]
*Corresponding author: Eenjun Hwang

Received February 15, 2020; revised April 24, 2020; accepted June 13, 2020; published August 31, 2020

Abstract

In recent years, as the environment has become an important issue in dealing with food, energy, and urban development, diverse environment-related applications such as environmental monitoring and ecosystem management have emerged. In such applications, automatic classification of animals using video or sound is very useful in terms of cost and convenience. So far, many works have been done for animal sounds classification using artificial intelligence techniques such as a convolutional neural network. However, most of them have dealt only with the sound of a specific class of animals such as bird sounds or insect sounds. Due to this, they are not suitable for classifying various types of animal sounds. In this paper, we propose a sound classification scheme based on a multi-feature network for classifying sounds of multiple species of animals. To do that, we first collected multiple animal sound datasets and grouped them into classes. Then, we extracted their audio features by generating mixed records and used those features for training. To evaluate the effectiveness of our scheme, we constructed an animal sound classification model and performed various experiments. We report some of the results.

Keywords: Environmental monitoring, Animal sound classification, Convolutional neural networks

This paper was supported by Korea Environment Industry & Technology Institute (KEITI) through Public Technology Program based on Environmental Policy, funded by Korea Ministry of Environment (MOE) (2017000210001).

http://doi.org/10.3837/tiis.2020.08.013 **ISSN : 1976-7277**

1. Introduction

In recent years, as the environment has become an important issue in dealing with food, energy, and urban development, diverse environment-related applications such as environmental monitoring and ecosystem management have emerged. In such applications, automatic classification or recognition of animals using video or sound is very effective because manual classification is expensive and time-consuming and requires specialized domain knowledge.

Even though animal sound classification (ASC) has been an interesting topic in signal processing area, it can also be used effectively for environmental monitoring or biodiversity studies [1]. For example, birds sound classification is effective for checking environmental changes quickly because birds have been widely used as biological indicators in the ecological research [2]. Moreover, insects sound classification can be used for tracking pests and infectious diseases [3] and for estimating food production around the world [4]. Further, anurans sound classification is a useful tool for assessing the condition of wetlands, which are very important for biodiversity [5].

Traditional ASC models usually depend on handcrafted features such as log-mel feature, mel frequency cepstral coefficient (MFCC), delta MFCC and delta-delta MFCC [6, 7, 8]. However, these features showed poor classification performance on support vector machines (SVMs) [9] and k-nearest neighbor (kNN) classifiers [10]. Recently, several convolutional neural network (CNN)-based models have been proposed to accurately classify animal sounds [11, 12, 13]. Unlike traditional ASC models, these deep learning models can extract higher-level features that are invariant to local spectral and temporal shifts [13].

Nevertheless, CNN-based classification models suffer from a lack of generality. That is, such models showed good performance for their own dataset, but for other datasets, they often showed poor performance. For instance, even though Warblr and TREE datasets [14] are usually used to train and evaluate animal sound classification models, they only contain data records of birds living in the United Kingdom and the Chernobyl regions, respectively. Hence, they are not suitable for training models for classifying various kinds of animal sounds.

The most intuitive way to solve this problem is to train a classification model with multiple datasets [15]. However, in ASC, this methodology has not attracted much attention so far due to diverse reasons such as lack of dataset for such purpose. In this paper, we propose a new model that can classify various types of animal sounds using multiple datasets. To do this, we first collected diverse open datasets to increase the generality of the sound classification model and defined new classes by merging or separating existing classes in the datasets. Second, we applied the oversampling strategy to mitigate the data imbalance problem caused by merging and separating classes. Finally, we built a multi-feature network-based model for various types of animal sound classification. To improve the classification performance, we trained the oversampled raw sounds and short-term Fourier transformed (STFT) spectrograms feature together and evaluated the sound classification performance of our proposed model through various experiments.

The rest of this paper is organized as follows. In Section 2, we present a literature review. In Section 3, we briefly describe animal sound datasets, data preprocessing for constructing a classification model, and the structure of our proposed model. We present various experiments we conducted and some of the results in Section 4. Lastly, in Section 5, we discuss the conclusion.

2. Related Work

Many previous works for animal sound classification are related to public challenges such as bird sound detection because high-performance models in public challenges are believed to be useful in ecological applications [14, 15]. In the public challenges, participants were informed of the problem to solve with some preselected datasets for training and evaluating their algorithm. Here, we introduce some of the challenges.

The Bird Detection in the Audio 2016 (BDA) challenge was to distinguish whether there was a bird sound in a given sound file or not [14]. In the challenge, the Warblr and FreeField1010 datasets were given as training datasets and the Chernobyl dataset as an evaluation dataset. The Warblr dataset is a crowd-sourced bird sound dataset biased towards population centers across the UK in 2015-2016. This challenge extracted only the birds' sound tagged data from the Freefield1010 dataset, which is a subset of the crowd-sourced dataset, Freesound. On the other hand, the Chernobyl dataset has sounds collected from various Chernobyl Exclusion Zones and only bird sounds in the dataset were used in the challenge.

DCASE 2018 Bird Audio Detection (BAD) challenge was the most recent public challenge dealing with mixed datasets [15]. This challenge aimed to improve the generality of the classification algorithms compared to BDA 2016 challenge by using five datasets. In this challenge, the Warblr, FreeField1010, and BirdVox DCASE 20k datasets were given as training datasets, and the Chernobyl and PolandNFC datasets as evaluation datasets.

Even though the BDA 2016 and BAD 2018 challenges used more than two datasets to improve generality, they both focused on detecting bird sounds only. Hence, they might not be suitable for classifying comprehensive animal sounds.

In a recent study, they attempted to classify various types of animals by using real recordings, eBird dataset, and Korea Wild Animal Sound Dictionary [16]. They presented a model that combines three CNN models, each trained to classify only one of the anurans, insects, and birds. They did not consider the absence of animal sounds in a given input sound.

To the best of our knowledge, few works have been conducted to classify the sounds of various animal species. In this paper, we present a sound classification model that distinguishes between different animal species. In particular, by considering additional non-

Table 1. Animal sound classification works based on multiple datasets.

Works	Training datasets	Test datasets
The Bird Detection in the Audio 2016 (BDA) [14]	Warblr FreeField1010	Chernobyl
DCASE 2018 Bird Audio Detection [15]	Warblr FreeFiled1010 BirdVox DCASE 20k	Chernobyl PolandNFC
Ko et al. [16]*	Anuran recording Bird recording eBird Korea Wild Animal Sound Dictionary	
Ours*	Anuran recording Bird recording eBird Korea Wild Animal Sound Dictionary Freesound	

*: These studies used a k-fold cross validation method in the experiments.

bio-acoustic dataset, our model can determine whether there is an animal sound in the input or not. Aforementioned works and their datasets including ours, are organized in **Table 1**.

3. Methodology

In this section, we first describe the datasets we used for constructing our model. Then, we present the classes that we made for animal sound classification and the preprocessing steps to combine those datasets. Finally, we describe the overall structure of our classification model.

3.1 Datasets

In this paper, we consider four datasets, each containing anuran sounds, bird sounds, insect sounds, and the Freesound [17]. **Figs. 1 and 2** show the number of classes (species) contained in each dataset and their total sound length in seconds, respectively. The first dataset contains anuran sounds recorded at 44.1 kHz, mono, 16-bit resolution in the natural habitat of each species. In the recorded sound, the sections in which the target sound was distorted by other sounds such as wind and non-target species, were removed for data integrity.

The second dataset contains sounds of birds and distorted sections were also removed in the same way as in the first dataset. Further, we collected additional bird sounds from eBird [18]. However, as all the data on the website were labeled with the collector's voice, we manually removed the human voices in the sound data for data integrity.

The third dataset contains insect sounds that were collected from the Korea Wildlife Dictionary published by the National Institute of Biological Resources. In particular, we used the orthoptera including crickets and grasshoppers.

The last dataset is the Freesound dataset used in the DCASE 2018 task 2 challenge [19]. This dataset consists of 39 classes of non-bio acoustic sounds and 2 classes of bio-acoustic sounds (bark, meow). The purpose of non-bio acoustic sounds is to filter out user inputs that are not animal sound. To summarize, we collected a total of 113 classes (species) of animal sounds from four datasets and 39 classes of non-bio-acoustic sounds.

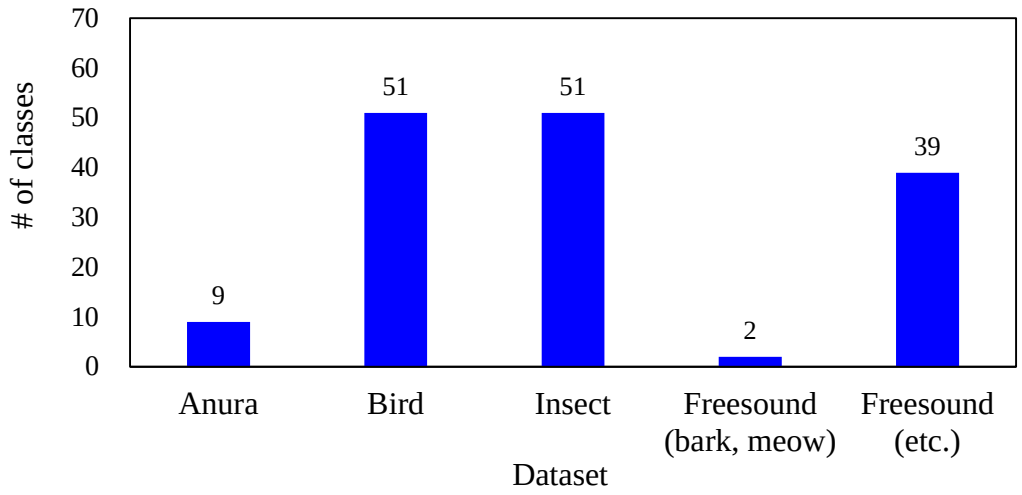


Fig. 1. Number of classes in each dataset.

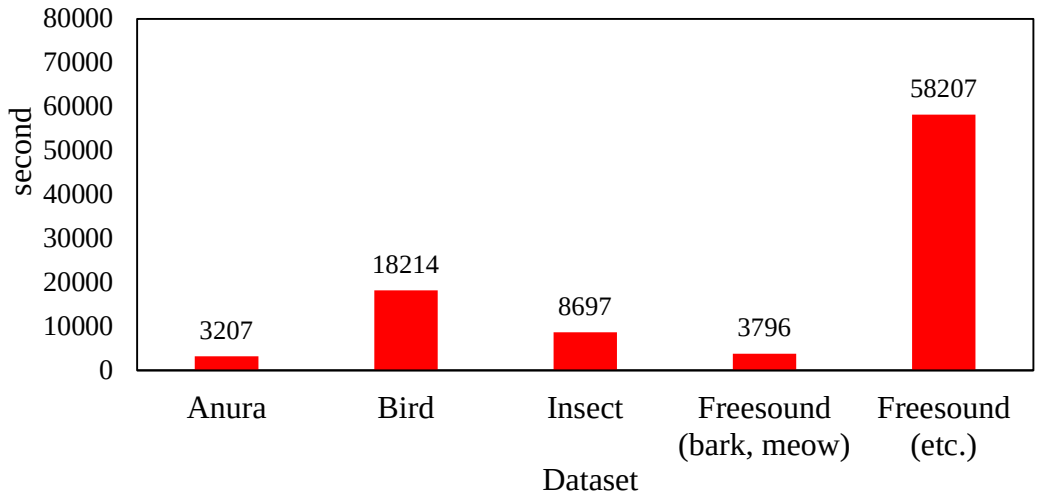


Fig. 2. Total recording length of each dataset.

3.2 Class Definition and Loss

To classify various types of animal sounds, we perform two different classifications in sequence. The first is the binary classification to filter out user input without animal sound by using bio and non-bio classes. The second is the animal sound classification for user input by using five different animal species. We first describe these two classifications briefly with their loss functions.

One of the basic tasks in ASC is to determine whether the animal sound of interest exists in the sound collection or not [20]. The most common method of determining the presence of bio-acoustic sounds is to consider this problem as a binary classification task that decides whether the given sound contains bio-acoustic sounds or not. More formally, let C_i be the class indicator that represents whether the input sound is contained in class i and l_i be the classifier's output which represents the probability that the input sound is contained in class i. In this case, the loss of the binary classification task can be defined as follows:

$$\mathcal{L} = -\sum_{i=1}^{2} C_i * \log(l_i) \tag{1}$$

To perform the binary classification task, we consider 113 classes of bio-acoustic sounds as a bio-acoustic class and 39 classes of non-bio-acoustic sounds as a non-bio-acoustic class. Based on these two distinct classes, we determine whether the input sound is bio-acoustic or not.

In the second classification, we determine the animal species (dataset) to which the input sound corresponds. In this paper, we consider five different animal species even though two of them belong to the same dataset. As mentioned earlier, each dataset has regional characteristics. In particular, anuran, bird, and insect datasets contain their unique regional characteristics and animal-class information. Therefore, unlike the binary classification in the first step, it is necessary to merge or separate dataset groups. Since we need to recognize one dataset only for the input sound, we apply a softmax function.

$$\mathcal{L} = -\sum_{i=1}^{5} C_i * \log(\frac{e^{l_i}}{\sum_{j=1}^{5} e^{l_i}}) \quad (2)$$

3.3 Preprocessing

As you can see in **Figs. 1 and 2**, four datasets have different total recording length, and sound files in the dataset also have a different recording length. According to [21], we generated the records of each class by the record length of the class with the longest record to avoid overfitting caused by the data imbalance problem. **Fig. 3** shows the procedure of oversampling.

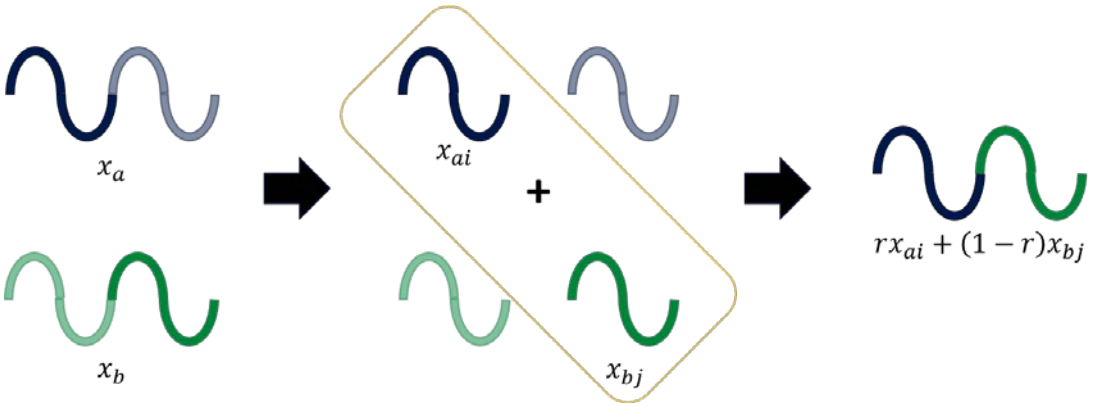


Fig. 3. Oversampling process.

That is, for two record files x_a and x_b in the same class, two randomly selected starting points in time i and j, and a random variable r with a uniform distribution in the range (0, 1), an augmented data mix(x_{ai}, x_{bj}, r) can be defined by Equation 3.

$$\mathrm{mix}(\mathrm{x_{ai}}, \mathrm{x_{bj}}, \mathrm{r}) = rx_a + (1-r)x_b \quad (3)$$

After generation, we extracted multi-features from the record files. In [21], they observed that raw data and log-mel based features such as log-mel, MFCC, and delta-MFCC could capture diverse patterns of a given sound. Based on this observation, these Fourier-based mel features were used in several state-of-art sound classification schemes [22,23,24]. Moreover, in [25], the authors proposed a classification model that utilizes raw data and log-mel STFT spectrogram as multi-feature inputs and compared their model with other CNN-based classification models that used one of them only. Since then, additional multi-feature based models were proposed for sound classification with good performance [22, 23]. In this study, we experimented with all the four features for constructing and training the classification model [22].

Our model requires two input variables: waves and STFT spectrograms. Hence, to obtain two input variables, we performed the following preprocessing steps:

1. Chunk the audio waveform into non-overlapping segments of 3.84s, which gives 44.1k × 3.84 samples.

2. Calculate both the STFT spectrogram with 1,024 FFT points and hop-lengths of 10ms.
3. Scale all the STFT as (256 × 192).

More details can be found in [22].

3.4 Architecture of Our Model

Fig. 4 shows the overall architecture of our classification model. Our model utilizes both raw audio and its STFT spectrogram. Blue boxes and green boxes represent a 1-dimensional (1-D) convolutional neural network and a 2-dimensional (2-D) convolutional neural network, respectively. The end of our model consists of two 2-D convolutional neural networks and two fully connected networks to efficiently infer each class.

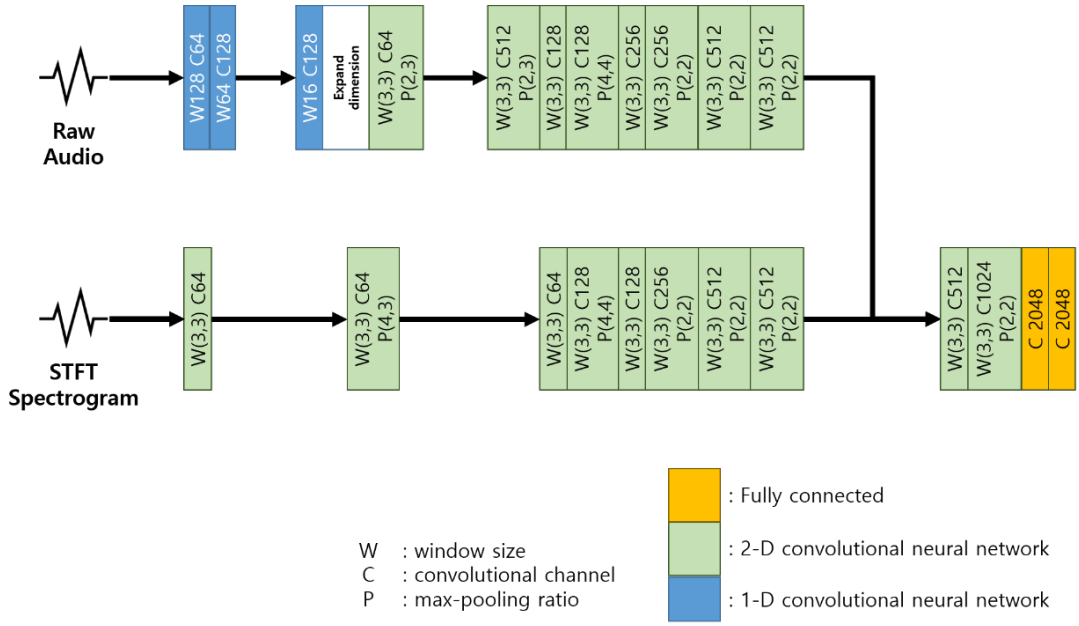


Fig. 4. The architecture of our model.

We modified the model mentioned in Section 3.3. Given a single 1-D waveform x_{raw}, a single STFT x_{STFT}, 1-D convolution *1D_Conv*, and 2-D convolution *2D_Conv*, features of waveform $F_{waveform}$ and features of STFT F_{STFT} are calculated by the equations (4) and (5).

$$F_{waveform} = 2D\_Conv\big(1D\_Conv(x_{raw})\big) \tag{5}$$

$$F_{STFT} = 2D\_Conv(x_{STFT}) \tag{6}$$

We used a window size of 3×3 for 2-D convolution, and larger filters (128, 64, and 16) for 1-D convolution. All convolutions were used with batch normalization and rectified linear unit (ReLU) activation function. Algorithm1 shows the pseudo-code for training our network model.

Require: α learning rate, b batch size, w parameters, m the number of index.
While w has not converged **do**
  For i=0,…, 1000 do
  Sample $\left\{\boldsymbol{x}^{(j)}, \boldsymbol{C}_{\boldsymbol{k}}^{(j)}\right\}_{\boldsymbol{j=1}}^{\boldsymbol{b}} \sim \mathbb{P}$ a batch from the training data and its label.
  $\left\{l_k^j\right\}_{k=1}^{m} \leftarrow f_w\left(x^{(j)}\right)$ outputs of index m from our network given a batch of training data.
  if k is 2 **do**
    $\mathrm{g_w} \leftarrow \nabla_{\mathrm{w}}\left[\frac{1}{\mathrm{b}}\sum_{j=1}^{b}\left\{-\sum_{k=1}^{2} C_m^{(j)} * \log\left(l_k^{(j)}\right)\right\}\right]$
  else do
    $\mathrm{g_w} \leftarrow \nabla_{\mathrm{w}}\left[\frac{1}{\mathrm{b}}\sum_{j=1}^{b}\left\{-\sum_{k=1}^{m} C_k * \log\left(\frac{e^{l_i}}{\sum_{j=1}^{m} e^{l_i}}\right)\right\}\right]$
  $\mathrm{w} \leftarrow \mathrm{w} + \alpha \cdot \mathrm{Adam}(w, \beta_1, \beta_2)$
  end For
  $\alpha \leftarrow \alpha * 0.9999$
end While

Algorithm 1. Pseudo-code for training our network model.

4. Experiments and Results

4.1 Experimental Setting

To evaluate the performance of our scheme, we constructed a classification model using Tensorflow under a Python environment. We used a desktop computer with an Intel Core i5-4440 3.1GHz CPU, 24GB RAM, and NVIDIA GeForce GTX 1080ti GPU, under the Windows 10 operating system. In the implementation, we used Adam optimizer with an initial learning rate of 0.001, β_1 of 0, and β_2 of 0.9, and the learning rate was decayed by a factor of 1/0.9999 for every one hundred iterations. In addition, we used the cross-entropy loss mentioned in Section 2.2, with a batch size of 8 and evaluated the performance using the area under the ROC curve (AUC) and F1-score through 5-fold cross-validation. For dataset classification, all classes other than the selected class when calculating AUC and F1-scores are considered negative classes.

As we removed sound segments in which the target sound was distorted by other sounds or non-target species, significant performance degradation could occur in very noisy environments. Therefore, to make our model robust in the noisy environments, we added random noises to the SNR ratio to 1 for all datasets during training and validation.

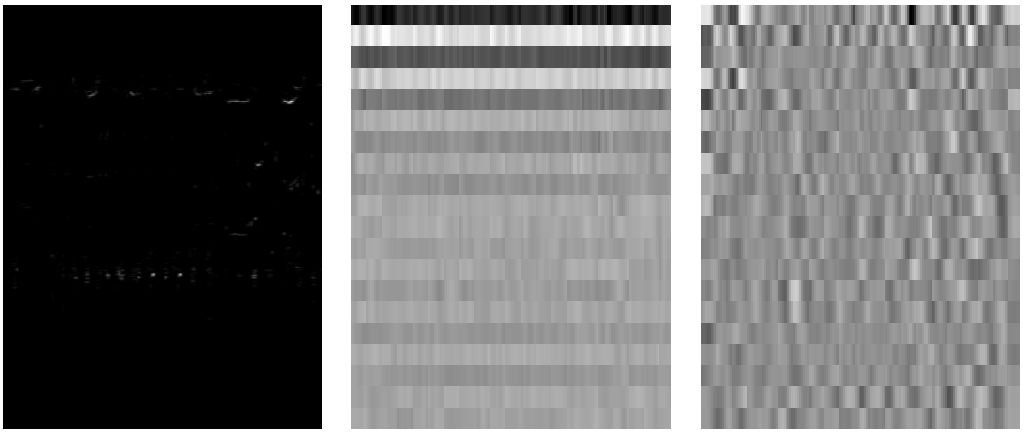
In [22], Li et al. used log-mel spectrograms for STFT spectrogram. In this paper, to find other spectrogram features that may give better classification performance, we considered two more popular features for sound classification, which were MFCC and delta-MFCC. **Fig. 4** shows the log-mel spectrogram, MFCC spectrogram, and delta-MFCC spectrogram for a bird sound.

The steps for extracting those features are as follows. First, we can obtain log-mel features by taking the logs of the powers at each of the mels. The mels are calculated by Equation (4) which is a formula for converting a frequency f (hertz) into a mel m [26].

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

On the other hand, MFCC and delta-MFCC features are commonly derived as follows [27]:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.
6. The delta-MFCCs are calculated by differentiating the MFCCs in terms of time.



(a) Log-mel (b) MFCC (c) Delta-MFCC

Fig. 5. STFT spectrograms for a bird sound.

A CRNN-based classification technique that combines CNN and RNN was recently used in deep learning-based sound classification. Thus, we implemented the attention based convolutional recurrent neural network (ACRNN) model [24] and compared it with our model under the same experimental setting.

4.2 Results and Discussion

We considered five different models: one model without any STFT spectrogram as a baseline and three models based on the log-mel spectrogram, MFCC spectrogram, delta-MFCC spectrogram, and ACRNN model [24]. We compared their performance in terms of AUC and F1-score.

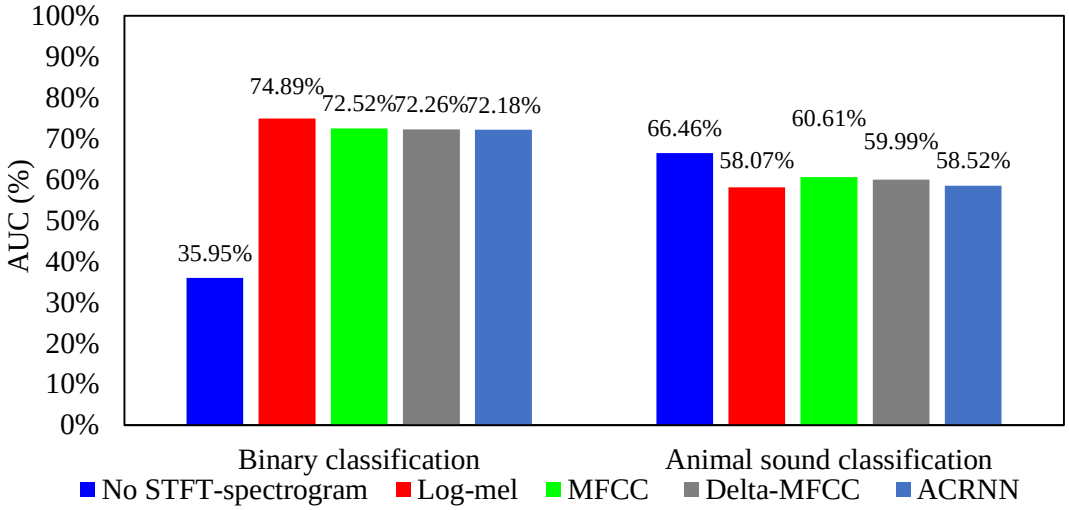


Fig. 6. AUC comparison of five models.

Fig. 6 shows the classification performance of five models in terms of AUC. In the binary classification, using features like log-mel, MFCC, or delta-MFCC gives better results than not using any STFT-spectrogram. Because the STFT-spectrogram is a feature that visualizes the frequency better than the original wave-feature, using one of the STFT features gives better results in binary classification. In particular, the log-mel spectrogram showed the best performance. This is because MFCC and delta-MFCC are cepstral analysis of log-mel. That is, while the frequency feature undergoes an inverse Fourier transform into a time feature, some of the features in the frequency domain disappear and the performance may decrease. ACRNN showed a slightly lower AUC value compared to our model. ACRNN consists of several convolution layers and recurrent neural networks, which include Gated Recurrent Units (GRU) [28]. Reset gates in GRU work to ignore the previous state and reset with the current input only when the weights of the previous state are close to 0. Because of this reset structure of ACRNN, ARCNN showed lower performance than Log-mel, MFCC, Delta-MFCC, and ours.

On the other hand, in the classification of animal sounds, those three STFT-spectrogram features showed poor performance. This is because STFT spectrograms of bio-acoustic sounds may share common frequency characteristics and classifying them becomes more complicated.

Table 2. AUC comparison of five models by class.

		AUC				
		no-STFT spectrogram	log-mel	MFCC	delta-MFCC	ACRNN
Binary classification	Bio-acoustic sound	**38.37**	**78.41**	**75.63**	**76.3**	77.46
	Non-bio-acoustic sound	34.54	72.89	70.72	69.91	69.11
Animal sound classification	Anuran	**69.28**	**61.32**	**64.18**	**63.43**	61.19
	Bird	66.71	59.07	60.69	60.78	58.68
	Insect	67.78	59.01	61.92	61.59	59.15
	Freesound (bark, meow)	68.36	60.12	63.07	62.77	60.31
	Freesound (etc.)	64.33	57.31	60.04	59.14	57.69

Fig. 7 compares the classification performance of the five models in terms of F1-score. In the comparison, F1-score was calculated based on the number of test data in each class. The table shows that using log-mel, MFCC, and delta-MFCC gives better results than not using any STFT-spectrogram in both classifications. However, all the models showed very low F1-scores in the classification of animal sounds compared to the binary classification. This can be explained by the fact that whereas the number of negative classes in the binary classification is one, it is four in the animal sounds classification. Our model was slightly better than ACRNN, except when using the log-mel feature. MFCC and delta-MFCC eliminate unnecessary features in extracting some significant coefficients from the log-mel feature. However, for log-mel, our model trained unnecessary features and showed worse performance compared to MFCC or delta-MFCC features.

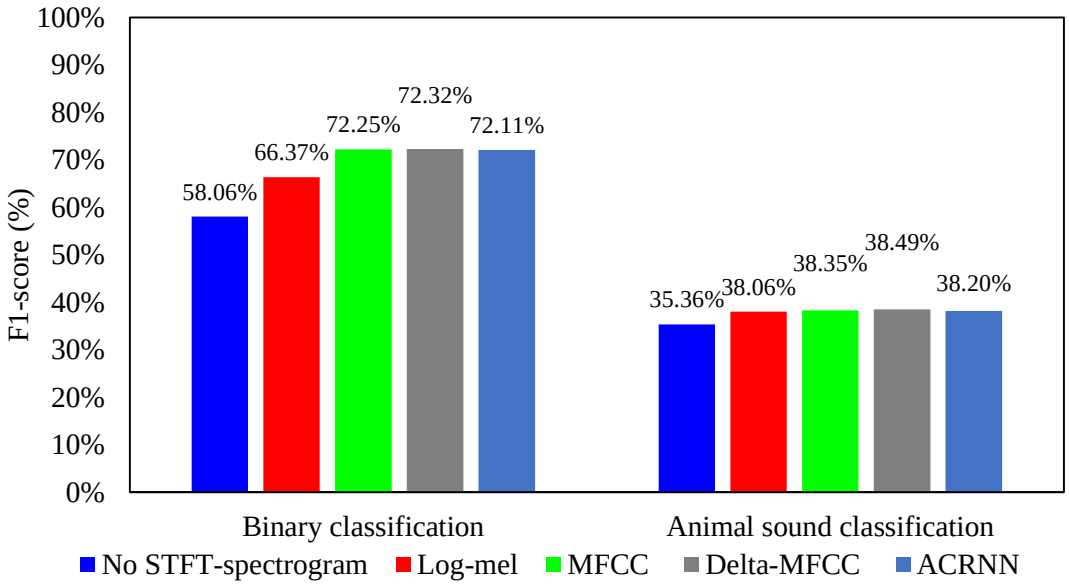


Fig. 7. F1-score comparison of five models.

Table 3 shows F1-scores for each class in the same way as **Table 2**. The results were similar to those in **Table 2**. However, the model that achieved the highest F1-score was different depending on the class. For instance, delta-MFCC gave better F1-scores in the anuran and the Freesound classes. Even though log-mel gave the best F1-scores in the bird and insect class, their differences from delta-MFCC are almost negligible. From this experiment, it can be seen that delta-MFCC is most effective overall in animal sounds classification.

Table 3. F1-score comparison of five models by class.

		F1-score				
		no-STFT spectrogram	log-mel	MFCC	delta-MFCC	ACRNN
Binary classification	Bio-acoustic sound	**60.81**	**70.23**	**74.13**	**76.31**	75.13
	Non-bio-acoustic sound	56.46	64.12	71.15	70.02	70.51
Animal sound classification	Anuran	**38.21**	**40.33**	**39.82**	**41.52**	41.08
	Bird	35.65	38.84	38.48	38.79	38.31
	Insect	36.48	39.79	39.29	39.71	39.27
	Freesound (bark, meow)	37.29	38.37	39.21	40.93	39.76
	Freesound (etc.)	34.82	37.42	38.04	38.04	37.75

5. Conclusion

In this paper, we proposed a sound classification scheme for classifying a wide range of animal sounds using a multi-feature network. To do that, we collected diverse datasets, including non-bio acoustic data, and decomposed them into classes. Then, we extracted various STFT spectrogram features including log-mel, MFCC, and Delta-MFCC and used them for building sound classification models. Our model works in two steps. In the first step, we performed a binary classification to determine whether the user input contains an animal sound. If so, in the second step, we performed animal sound classification to determine the animal species of the input sound. We evaluated the classification performance in terms of AUC and F1-score. The experimental results showed that in most cases, using STFT-spectrogram features in addition to wave-features improved the classification performance. In the future, we plan to expand our model to cover a larger number of animal species.

References

[1] G. M. Lovett, D. A. Burns, C. T. Driscoll, J. C. Jenkins, M. J. Mitchell, L. Rustad, J. B. Shanley, G. E. Likens, and R. Haeuber, “Who needs environmental monitoring?,” *Frontiers in Ecology and the Environment*, vol. 5, no. 5, pp. 253-260, 2007. Article (CrossRef Link)

[2] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, and X. Z. Fern, "The 9th annual MLSP competition: new methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *Proc. of 2013 IEEE international workshop on machine learning for signal processing (MLSP)*, pp. 1-8, 2013. Article (CrossRef Link)

[3] D. Pimentel, and M. Burgess, "Environmental and economic costs of the application of pesticides primarily in the United States," *Integrated pest management*, pp. 47-71, 2014.
Article (CrossRef Link)

[4] M. Q. Benedict, and A. S. Robinson, “The first releases of transgenic mosquitoes: an argument for the sterile insect technique,” *Trends in parasitology*, vol. 19, no. 8, pp. 349-355, 2003.
Article (CrossRef Link)

[5] A. D. Garg, and R. V. Hippargi, “Significance of frogs and toads in environmental conservation,” 2007.

[6] F. Su, L. Yang, T. Lu, and G. Wang, "Environmental sound classification for scene recognition using local discriminant bases and HMM," in *Proc. of the 19th ACM international conference on Multimedia*, pp. 1389-1392, 2011. Article (CrossRef Link)

[7] P. Jančovič, and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 982936, 2011. Article (CrossRef Link)

[8] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, vol. 80, pp. 1-9, 2014. Article (CrossRef Link)

[9] C. M. Bishop, *Neural networks for pattern recognition*, Oxford university press, 1995.

[10] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. of 2017 22nd International Conference on Digital Signal Processing (DSP)*, pp. 1-5, 2017. Article (CrossRef Link)

[11] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. of 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 559-563, 2015. Article (CrossRef Link)

[12] T. Pellegrini, "Densely connected CNNs for bird audio detection," in *Proc. of 2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1734-1738, 2017. Article (CrossRef Link)

[13] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *Proc. of 2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1744-1748, 2017. Article (CrossRef Link)

[14] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *Proc. of 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1-6, 2016. Article (CrossRef Link)

[15] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368-380, 2019. Article (CrossRef Link)

[16] K. Ko, S. Park, and H. Ko, "Convolutional feature vectors and support vector machine for animal sound classification," in *Proc. of 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 376-379, 2018. Article (CrossRef Link)

[17] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proc. of Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017*, p. 486-493, 2017. Article (CrossRef Link)

[18] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling, "eBird: A citizen-based bird observation network in the biological sciences," *Biological conservation*, vol. 142, no. 10, pp. 2282-2292, 2009. Article (CrossRef Link)

[19] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," *arXiv preprint arXiv:1807.09902*, 2018.

[20] I. Sobieraj, Q. Kong, and M. D. Plumbley, "Masked non-negative matrix factorization for bird detection using weakly labeled data," in *Proc. of 2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1769-1773, 2017. Article (CrossRef Link)

[21] Y. Tokozume, and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2721-2725, 2017. Article (CrossRef Link)

[22] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," *Proc. Interspeech 2019*, 3604-3608, 2019. Article (CrossRef Link)

[23] Y. Su, K. Zhang, J. Wang, and K. Madani, “Environment sound classification using a two-stream CNN based on decision-level fusion,” *Sensors*, vol. 19, no. 7, pp. 1733, 2019.
Article (CrossRef Link)

[24] Z. Zhang, S. Xu, T. Qiao, S. Zhang, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," in *Proc. of Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 261-271, 2019. Article (CrossRef Link)

[25] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, “An ensemble stacked convolutional neural network model for environmental event sound recognition,” *Applied Sciences*, vol. 8, no. 7, pp. 1152, 2018. Article (CrossRef Link)

[26] S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185-190, 1937. Article (CrossRef Link)

[27] M. Sahidullah, and G. Saha, “Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition,” *Speech communication*, vol. 54, no. 4, pp. 543-565, 2012. Article (CrossRef Link)

[28] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.

Chung-Il Kim received his B.S. degree in Electronic Engineering from Korea University, Seoul, Korea, in 2018. He is currently a Master's student at the School of Electronic Engineering at the Korea University, Seoul, Korea. His current research interests image processing, signal processing, and generative model.

Yongjang Cho received the B.S. degrees in multimedia engineering from Sungkyul University, Anyang, South Korea. Currently, he is a master student at the School of Electronic Engineering, Korea University, Seoul, South Korea. His research interests include machine learning, image processing, and database.

Seungwon Jung received his B.S. degree in Electronic Engineering from Korea University, Seoul, Korea, in 2016. Currently, he is a doctoral student at the School of Electronic Engineering, Korea University, Seoul, Korea. His research interests include Data Mining, Machine Learning, Deep Learning, and Time Series Forecasting.

Jehyeok Rew received the B.S. degrees in electronics and radio engineering from KyungHee University, Korea, in 2012. From 2017 to 2018, he was a senior researcher at the HANA Micron, Korea. Currently, he is a doctoral student at the School of Electronic Engineering, Korea University, Seoul, South Korea. His current research interests include machine learning, image processing, bio-informatics, GIS and information retrieval.

Eenjun Hwang received his B.S. and M.S. degrees in Computer Engineering from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively; and his Ph.D. degree in Computer Science from the University of Maryland, College Park, in 1998. Currently, he is a member of the faculty at the School of Electrical Engineering, Korea University, Seoul, South Korea. His current research interests include database, multimedia systems, information retrieval, big data processing, and machine learning.