Deep Learning Based Feature Extraction Methods for Coyote Howling Detection

1st Bokyung Kwon
dept. Computer Information Engineering
Kwangwoon University
Seoul, South Korea
bbo1209@gmail.com

2nd Youngbin Kim
dept. Computer Information Engineering
Kwangwoon University
Seoul, South Korea
binny9904@naver.com

3rd Jihyeon Park
dept. Computer Science and Statistics
Jeju National University
Jeju, South Korea
wpfflzld325@jejunu.ac.kr

4th Yejin Lee
dept. Computer Science
Hallym University
Chuncheon, South Korea
leeye0616@naver.com

5th Heesun Jung
dept. Computer Science
Hallym University
Chuncheon, South Korea
glee623@naver.com

***Abstract*—The attacks on livestock, human, and crops by coyotes are occurring over the United States, while traditional simple management such as public education about the method of avoiding coyotes and coyote hunting contests to reduce their numbers are executed. There are not sufficient cases of technical approaches or research about the damage to coyotes. The method of coyote howling sound classification using Convolutional Neural Network (CNN) to reduce the damage of coyotes is needed. This paper suggests using a network connection in order to prevent the damage by informing the neighborhood farms when coyotes appear and chasing coyotes through a coyote alert system. It is expected that additional technical approach to current coyote damage prevention can improve the accuracy and make the previous management more practical.**

***Index Terms*—audio classification, feature extraction, spectrogram features, tempogram features, deep learning**

I. Introduction

In the United States, substantial damage to livestock, human costs, and crops occurs in countless farms by coyotes. According to reported data from the United States Department of Agriculture (USDA), the damage to livestock by predators is severe in 2015 [1]. Fig. 1 shows specific number of livestock that was killed by coyotes among those killed by predators. However, rather than reducing the direct damage of coyotes, country and state officials focusing on preventing damages due to coyotes entirely. The above method is temporal solution, therefore it is essential to make the method of managing the coyote technically to solve the fundamental problem of coyote.

This paper suggests a technical approach using the method of deep learning based feature extraction to prevent and reduce the damage of coyotes. In the acoustic sound classification, simply classifying the animal howling sound is a field of being actively researched. Several research showed outstanding performance of audio classification using CNN [2], [3]. Feature extraction makes signal appropriate for machine learning by reducing the size of audio signal and improving computational and time complexity of machine learning algorithm. It is crucial for determining performance during machine learning process in audio classification. Therefore, the goal of this paper is to derive the state of the art performance through the comparison experiment that combine deep learning model with various feature extraction method such as Mel Frequency Cepstral Coefficient (MFCC) and Gammatone Cepstral Coefficient (GTCC) [4]. Additionally, this study not only classifies the coyote using the coyote howling but also proposes a platform for detecting the coyote which is connected with the network. Through this platform, reducing damage to livestock and private properties, furthermore protect human life is possible.

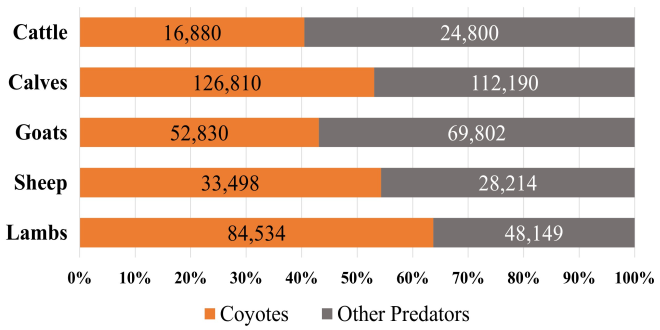

Fig. 1. The number of livestock killed by coyotes among those killed by predators [1]

II. Literature review

The research of audio classification is one of the topics currently being actively studied, and techniques for performance improvement are suggested. In those studies, various feature extraction techniques and models are used, and experiments are conducted by combing features and models. The combinations of features and models that show good performance in each case were shown in following studies.

Ramashini *et al.* [5] classifies using Linear Prediction Cepstrum Coefficients (LPCC), MFCC, GTCC and showed that GTCC outperformed LPCC and MFCC-based classification with Support Vector Machine (SVM). In addition, although the 3 cepstral features were combined, the accuracy was not increased compared to using only the GTCC function. Hence, GTCC showed the highest accuracy of 93.3% in bird sound classification.

Valero *et al.* [6] adjusted GTCC, which was used in previous speech research fields for non-speech audio classification purposes. Experiments were employed with general sound and audio scenes for 4 hours each, according to 2 cross-validation methods and 4 machine learning methods. As a result, GTCC outperformed MFCC in collecting the spectral feature of non-speech audio signals at low frequencies.

Piczak *et al.* [2], acoustic classification CNN based model has outstanding performance in audio classification fields. This CNN method outperforms baseline implementations models relying on MFCC. Even if with the limited dataset it showed not only CNN can be applied effectively but also substantial possibility of performance improvement on the environment sound classification.

Valenti *et al.* [3] suggested the model of CNN for acoustic scene classification. In order to validate whether CNN model is suitable for acoustic classification, researchers compared CNN model with other systems such as Multi-Layer Perceptron (MLP), one-layer CNN, and Gaussian Mixture Model (GMM). In the experiment, 2 different feature extraction technologies which are log-mel spectrogram and MFCC were used with each systems. For the result, 2-layer CNN with log-mel spectrogram showed the best and highest accuracy of 79%.

Xu *et al.* [7] suggests multi-view CNN model and evaluation results show that the proposed model outperforms the conventional CNN for automatic animal species identification. In total, compare their method with six classifiers: CNN, Sparse Representation-based Classification (SRC), SVM-MFCC, SVM-Spectrum, k-Nearest Neighbors (kNN)-MFCC and kNN-Spectrum. Comparing the performance of different methods using frog dataset, CNN was the best with 82.7% accuracy, and SVM-Spectrum obtained 40.5% accuracy.

Daniel *et al.* [8] research about multiple combinations of classical machine learning algorithms such as kNN, Neural Network (NN), and GMM with feature extraction methods. Except one case birds dataset, GTCC outperforms the MFCC most of the feature extraction method. Pairing of GTCC and kNN gives the best performance in the all experiment.

As can be seen in the papers that mentioned above, there are several methods of combination of feature extraction technologies and machine learning or deep learning models for audio classification. Through related research, it was confirmed that the method of deriving optimal results was different for each situation. Therefore, the aim is to derive optimal performance of coyote audio classification through performance comparison of each combination feature extraction technology and deep learning or machine learning model.

III. Methodology

Audio data is sent by the network into Raspberry Pi. A pre-trained machine learning model is put into the Raspberry Pi to determine whether the received sound data is a coyote or not. It is sent back to the server through the network. In consideration of transmission efficiency, not wav files but the extracted audio features are delivered to Raspberry Pi.

For the model, it uses coyote's howling sound because coyotes have a habit of howling frequently and the howling sound's transmission distance range is much longer than the image. Also, it is better to use a 1-dimensional audio file than using a 2-dimensional image file from a calculation cost point of view because of using Raspberry Pi.

A. Feature Extraction

The feature extraction part is essential when training audio file in a machine learning model. This feature is obtained using the Librosa library in python. The features are input to the model. The trained model is deployed to Raspberry Pi and it detects the coyote.

In this research, three feature extraction methods are considered. First, mel-spectrogram is a spectrogram where the frequencies are converted to the mel-scale. Second, MFCC which applies filter bank based on mel-scale on spectrum is one of the most broadly used method in audio recognition. Last, GTCC use a gammatone filer to extract a audio feature and use Discrete Cosine Transform (DCT) to reduce the dimension of audio feature.

The mel-spectrogram used in MFCC has correlations between variables within features. Because the energy in the reference frequency band is gathered in one place. This problem adversely affects the GMM, which assumes independence between variables and modeling. The MFCC applies inverse Fourier transforms to the mel-spectrogram to resolve the correlation between variables. MFCC is widely used in several audio classification fields because it highly efficient in general situations by removing correlation and with low computing power due to low amount of computation. For this reason, extraction using MFCC was carried out.

In addition, feature extraction was performed using mel-spectrogram. Mel-spectrogram performs better in limited domain problem because the large correlation between frequencies. Thus, extraction using mel-spectrogram was experimented to compare performance depending on whether the correlation between the features was removed. However, recent research shows that a combination of multiple features provides better performance. For the purpose of this research is to find the best combination of features.

IV. Experiment

A. CNN Model

CNN is deep learning algorithm, which widely using on image classification problem. CNN is commonly used in image data analysis and audio data. CNN is a model that trains while maintaining spatial information of an image. CNN can be divided into feature extraction part and class classification

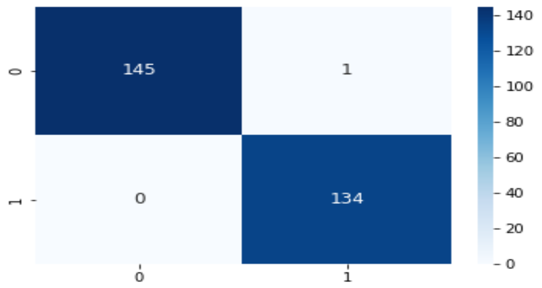

Fig. 2. Confusion Matrix

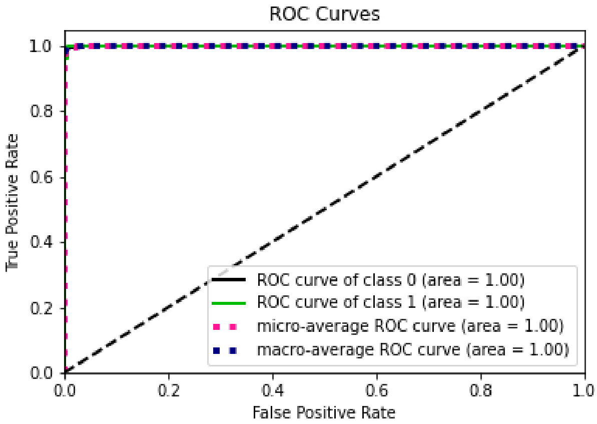

Fig. 3. Receiver Operating Characteristic curve

part. The feature extraction part consists of multiple layers of convolution layer and pooling layer. The convolution model used for training consists total of four convolution blocks, and one block consists of convolution 2-dimension, activation function ReLU, and max pooling. After passing through the last convolution block, flatten the output value in 1-dimension and put the value as input in the Fully-Connected layer. Since the task in this paper is the binary classification problem that distinguishes the Coyote class and the Non-Coyote class, the out-dimension of the last Fully-Connected layer is set to two.

B. Dataset

Since coyotes cause a lot of damage to farms, dogs and chicken that usually live on farms included in the Non-Coyote class. Foxes are also included in the Non-Coyote class because fox has similar howling sound with coyote. For the Coyote dataset, the dataset is from United States National Park Service (NPS) and Non-Coyote dataset is from Google research audio dataset.

C. Setting

The experimental setting is as follows: Adam optimizer was used as an optimization function, and the learning rate was set to 0.001, the batch size was set to 10, and the epoch was set to 100. MFCC was used as the audio feature extraction method, and the sampling rate value was set to 16,000 Hz.

There are a total of 1,160 training dataset. It consisted of 586 coyotes, 480 dogs, and 94 chickens. And the test data set is a total of 280 sheets. It consisted of 134 coyotes, 117 dogs, and 29 chickens. Training data and test data were divided in a total ratio of 8:2.

Confusion Matrix, Fig. 2 and Receiver Operating Characteristic curve, Fig. 3 were used to evaluate classification performance. The loss value of the evaluation set was 0.0324, and the accuracy was 279 out of 280.

References

[1] R. Tischaefer, "Coyotes," in *Wildlife Damage Management Technical Series.* USDA. APHIS, 2020, p. 42.

[2] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP).* IEEE, 2015, pp. 1–6.

[3] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks." in *DCASE*, 2016, pp. 95–99.

[4] H. Xu, L. Lin, X. Sun, and H. Jin, "A new algorithm for auditory feature extraction," in *2012 International Conference on Communication Systems and Network Technologies.* IEEE, 2012, pp. 229–232.

[5] M. Ramashini, P. E. Abas, K. Mohanchandra, and L. C. De Silva, "Robust cepstral feature for bird sound classification," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, p. 1477, 2022.

[6] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

[7] W. Xu, X. Zhang, L. Yao, W. Xue, and B. Wei, "A multi-view cnn-based acoustic classification system for automatic animal species identification," *Ad Hoc Networks*, vol. 102, p. 102115, 2020.

[8] D. Bonet-Solà and R. M. Alsina-Pagès, "A comparative survey of feature extraction and machine learning methods in diverse acoustic environments," *Sensors*, vol. 21, no. 4, p. 1274, 2021.