Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023420

Sound-based animal recognition model

Hengjian Zhang

Computer and network security college (oxford brookes college) Chengdu University of Technology, No.1, Erxianqiao East 3rd Road, Chengdu China

zhang.hengjian@student.zy.cdut.edu.cn

Abstract. Conservation of biological and biological diversity is the goal of joint efforts worldwide, and protecting wild creatures requires various monitoring equipment with identification capabilities. In this case, this article will introduce a model with animal call recognition ability designed using edge impulse. A two-layers convolutional neural network (CNN) is selected and trained with the input features generated by Mel Frequency Cepstral Coefficient, Spectrogram and Mel-Spectrogram. The model hoped not only to focus on the recognition accuracy rate of the model but also on the inferencing time, peak RAM usage and other performance to cope with the poor performance of the tiny equipment used in the real wild. Using five classes consisting of over 4,000 audio samples as the input and after several round experiments. Finally, the MFCC&CNN model is chosen because it has an accuracy rate of over 85.6%, the swiftest inferencing time and the lowest peak RAM usage.

Keywords: Animal call recognition, MFCC, Mel-spectrogram, spectrogram, two-layers CNN, low peak RAM usage.

1. Introduction

In recent decades, biodiversity has decreased sharply due to human hunting, encroachment on animal habitats, and other human behaviors. Nowadays, human beings have strengthened their sense of crisis and begun to protect animals. In the wild and nature reserves, professional teams usually monitor the behaviors of animals and provide the necessary assistance by setting up infrared cameras and wearing locators for animals. However, these two methods have some disadvantages. For example, animals can rarely be photographed directly due to the limited field of vision of the infrared camera. In addition, wearing a locator will make animals who wear it for the first time produce a sense of rejection which may affect their behaviors. Therefore, an animal voice recognition model that can capture animal calls and recognize their species in the wild is designed to calculate biodiversity and species density and monitor animal behaviors to help protect animals.

Mane, Rashmi, and Tade [1] introduced a useful method that can classify animals based on their sound by combining ZeroCross-Rate (ZCR), used to remove the silence part, Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) algorithms to classify specific animal calls. They proved their model could be used to do the recognition work and predicted their model could detect the state of the particular animal like normal, hunger, sleep, and heat. However, they did not give any performance results like the accuracy and the size of datasets. Thence, it is hard to convince their model is adequate.

DOI: 10.54254/2755-2721/4/2023420

In Piczak's [2] study, they introduced a convolutional neural network (CNN) consisting of two convolutional layers with max pooling and two fully connected (FC) and proved their model, given the limit datasets, can attain the best accuracy of 64.5% when their CNN cooperate with log Mel-Spectrogram. However, the issue is that the accuracy of their model is poor, and once it is placed in a natural wild environment, its performance could be worse.

Besides, several researchers also preferred CNN. Valenti, Diment, and Parascandolo [3] selected a two-layers CNN with a max-pooling layer and used short sequences of audio as data input. Finally, they obtained an accuracy of 79% on the DCASE 2016 development dataset. Although their topic focuses on detecting and classifying acoustic scenes and events, not animal sounds, they provided valuable ideas for this model development. For example, Sasmaz and Tek [4] constructed a CNN which owns three convolutional layers and dense layers as well as one max pooling layer. With the help of MFCC, their model finally achieved the best accuracy of 75% with the help of optimizer Nesterov-accelerated Adaptive Moment Estimation. However, there was a defect in their experiment as their dataset was minimal. They have ten classes in their dataset but have only 875 sound samples in total. In addition, the number of recordings in each class is imbalanced. For instance, there were three classes have 200 samples, while another two only had 25 samples. These two issues restricted the recognition performance of their model badly.

Based on the experience of the above authors, a two-layer convolutional neural network with max-pooling is decided to be used as the learning method and combined with MFCC, spectrogram and Mel-spectrogram to build up three models and compare their best performance. A dataset containing five classes has been collected, and audio samples are denoised and cut into short segments. Different from the previous experiments, because the ultimate goal of this model is to be used in the wild, and the equipment used in the wild is usually tiny and has general performance, the model is required to have not only high accuracy but also fast inferencing time and low peak RAM usage. Finally, a model combining MFCC and two-layers CNN is chosen. The model has an accuracy rate of 85.6%, which is a comparatively good result， an inferencing time of one millisecond and most importantly a very low peak RAM usage of 9KB that can satisfy the performance requirements of small devices used in the wild.

In the following section, the process of building the model will be described in detail, including the collection and processing of datasets and how the model is trained with the help of edge impulse [5]. Finally, the model's performance will be evaluated, and the causes of possible errors will be analyzed.

2. Methodology

2.1. Dataset introduction

Multiple publicly available free datasets are downloaded from kaggle.com [6] and picked the calls of 5 common animals (cats, dogs, frogs, ducks and sparrows). The raw recordings in those datasets are not suitable for training the machine directly because most of them are clipped from videos rather than recorded in a professional situation. Therefore, these recordings are processed as follows:

a) Recordings are split into several clips, and each part is one second long.
b) Most of the background noise is masked by audio processing software.
c) Unqualified clips like those that do not have the sound of the target animals and those with other interfering sounds are dropped.

Finally, a dataset with five classes with over 700 recordings in each class has been obtained (Table 1).

DOI: 10.54254/2755-2721/4/2023420

Table 1. Dataset Description.

Class Name	Number of Recordings	Recording Length
Cat	864	One Second
Dog	961	One Second
Duck	843	One Second
Frog	711	One Second
Sparrow	869	One Second

2.2. Spectrogram

A spectrogram illustrates the signal intensity with time at the different frequencies of a waveform. Audio spectrograms may be used to phonetically detect spoken words and study animal cries. Four stages are required to make a spectrogram (figure 1).

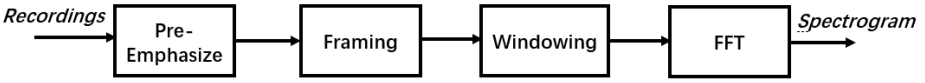

Figure 1. Spectrogram progress.

2.2.1. Pre-emphasize. Pre emphasis is a filtering technique that emphasizes higher frequencies. Pre emphasis can balance the strongly attenuated spoken spectrum in the high frequency region and eliminate some glottic effects in vocal tract features in advance [7].

$$\boldsymbol{H(Z) = 1 - \mu Z^{-1}} \tag{1}$$

U is usually set to 0.97 which make 97% of any one sample is presumed to originate from previous sample.

2.2.2. Framing. The speech signal is time-varying. In order to obtain stable acoustic characteristics, it is necessary to minute the data. Each frame represents the characteristics of sound in a very short time, and it is assumed that the signal does not change or changes very little during this time.

2.2.3. Windowing. The signal tapers approach the frame border by adding a window on each frame. When DFT is applied to the signal, this boosts harmonics, smooth edges and minimizes edge effects [7].

2.2.4. *Fast Fourier Transform.* Observing the features of a signal in the time domain is difficult, thus it is often transformed to the energy distribution in the frequency domain. Different distributions of energy may indicate distinct speech features. After doubling the Hamming window, each frame must undergo Fast Fourier Transform to determine the spectrum's energy distribution. The frequency spectrum of each frame is acquired by transforming each frame signal using the Fast Fourier transform after windowing, and the modulus square of the frequency spectrum of the speech signal is used to calculate the power spectrum of the spoken signal. The FFT for the voice signal should be:

$$\boldsymbol{X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi k/N}, 0 \leq k \leq N-1} \tag{2}$$

Where x (n) is the input speech signal and N represents the number of points of Fourier transform.

2.3. Mel-Spectrogram

Humans do not perceive frequencies on a linear scale and are more sensitive to variations at lower frequencies than those at higher frequencies. In 1937, Stevens, Volkmann, and Newmann[8] created the

DOI: 10.54254/2755-2721/4/2023420

Mel-Scale, a unit of pitch in which equal distances in pitch sound similarly far to the listener. A Mel-spectrogram is a spectrogram in which the frequencies have been transformed to the Mel-scale. Consequently, the Mel-spectrogram method involves one more step compared to the spectrogram shown in figure 2.

Figure 2. Mel-spectrogram process.

The frequency range of the FFT spectrum is expansive, and the linear scale is not followed by the speech signal [9]. Therefore, the energy spectrum is processed through a collection of Mel-scale triangle filter banks, and a filter bank with m filters is defined (the number of filters is near to the number of crucial bands). There is a triangle filter in use. As seen in Figure 3, the interval between f(m) reduces with decreasing m value and increases with increasing m value.

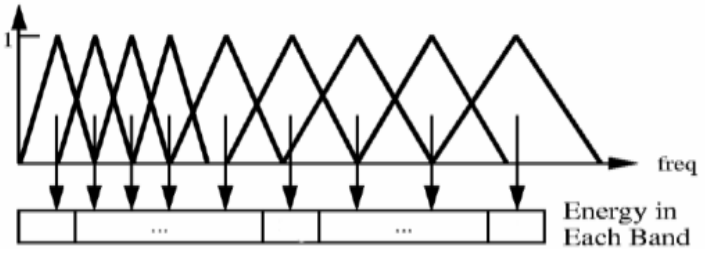

Figure 3. Mel scale filter bank [9].

The process is calculated as:

$$H_m(k) == \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k \ge f(m+1) \end{cases} \quad (3)$$

Where f (x) donates the Mel-spectrum of given frequency:

$$f_{mel}(f) = 2595 \times \ln(1 + \frac{f}{700HZ}) \quad (4)$$

2.4. MFCC

The human ear's critical bandwidth varies with frequency, and this is the basis for MFCC. The two types of filters used by MFCC technology are linear interval filters and logarithmic interval filters. The signal is encoded in the Mel frequency scale to capture significant speech elements. This scale's linear frequency range is less than 1000 Hz, while its logarithmic range is more significant than 1000 Hz. The typical speech waveform may occasionally alter depending on the health of the speaker's vocal cords. Compared to the speech waveform itself, the MFCC is less sensitive to the alterations mentioned above [10-11].

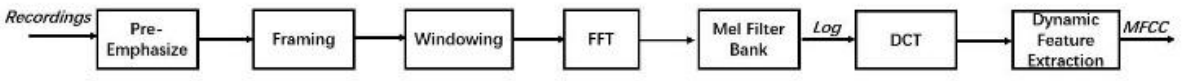

Figure 4. MFCC Process [2, 3].

As illustrated in Figure 4, a total of 7 steps are needed to implement MFCC. It can be seen that MFCC can be generated from Mel-spectrogram after implementing Discrete Cosine Transform (DCT).

DOI: 10.54254/2755-2721/4/2023420

1.1.1. Discrete Cosine Transform. Before performing DCT, it is necessary to calculate the logarithmic energy output by each filter bank.

$$\boldsymbol{s(m) = \ln(\sum_{K=0}^{N-1} |X_a(k)|^2 H_m(k)), 0 \le m \le M} \quad (5)$$

Fourier transform is connected to discrete cosine transform (DCT). It is comparable to the discrete Fourier transform, but employs only real integers. Signal processing and image processing often use this variant for lossy compression of signals and pictures. Finally, MFCC is calculated as:

$$\boldsymbol{c(n) = \sum_{m=0}^{M-1} s(m) \cos(\frac{\pi n(m-0.5)}{M}), n = 0, 1, 2, \dots, C-1} \quad (6)$$

2.4.1. Dynamic Feature Extraction. The combination of dynamic and static features may increase the system's recognition ability, while the typical cepstrum parameter MFCC solely represents the static properties of speech parameters. The spectrum of these static qualities may be used to define speech's dynamic features.

For the data obtained after the DCT transformation, 12 dimensions (more information is reserved) are taken, added short-time energy, and a total of 13 dimensions, and then perform delta and delta-delta differences to obtain 39-dimensional features. These 39 features represent the dynamic change between frames in the corresponding features. The dynamic parameter can be calculated using [7]:

$$\boldsymbol{\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|}} \quad (7)$$

T stands for number of subsequent frames required to be computed, k_i is the i_{th} weight, and cm(n) is the m_{th} feature for the nth time frame. T is often assumed to be 2. The delta-delta coefficients are computed by calculating the first-order derivative of the delta coefficients [7].

2.5. Convolutional Neural Network

Two convolution layers, one dense layer and one maximum pooling layer comprise the deep learning network. Rectified linear units (ReLU) serve as the activation function for these convolution layers. The last dense layer's activation function is softmax. The kernel size used in corresponding layers is 3 (figure 5).

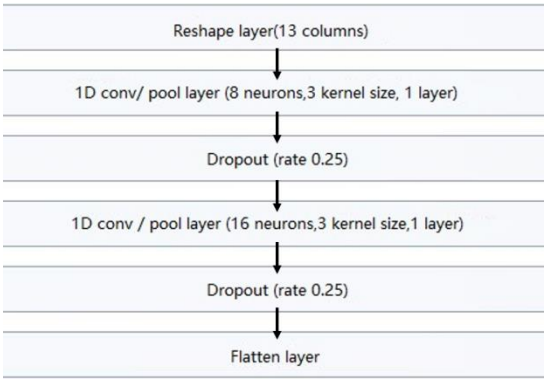

Figure 5. CNN Architecture.

3. Results

Edge impulse [5] is an online machine learning tool. Provided ready-prepared machine learning methods, developers can straightly upload their datasets, build up models, adjust parameters and train their model without coding.

All the data were uploaded to edge impulse, of which 20% were selected as test data randomly and the remaining is used as training data. Frame stripe, frame length, learning rate and other parameters are constantly adjusted during the experiment to obtain the best performance of the

DOI: 10.54254/2755-2721/4/2023420

model. Three indicators were used to evaluate the model's performance: accuracy, inferencing time and RAM usage.

3.1. Training details and recognition performance

CNN's learning epochs are set to 100. After several rounds of parameter adjustment and training, the best performance of the three models is found under the parameters shown in Table 2, 3 and 4.

Table 2. Parameters setting of MFCC&CNN model.

Number of coefficients	Frame Length	Frame Stride	Filter number	FFT Length	Low Frequency	Learning Rate
13	0.025	0.015	40	256	400	0.003

Table 3. Parameters setting of Mel-Spectrogram &CNN model.

Frame Length	Frame Stride	Filter number	FFT Length	Low Frequency	Noise Floor	Learning Rate
0.02	0.015	40	256	400	-40	0.003

Table 4. Parameters setting of Spectrogram&CNN model.

Frame Length	Frame Stride	Frequency bands	Noise Floor	Learning Rate
0.02	0.015	128	-40	0.003

The reason why the lowest frequency is 400 is that among the five types of animals, dogs have the lowest frequency, which is also greater than 450Hz. In addition, the noise floor is -40 because the background noise is wished to be remove as much as possible without affecting the original animal calls. The performance of these three models is shown in table 5.

Table 5. Models' Performance.

Model Name	Inferencing Time	Peak RAM Usage	Accuracy
MFCC&CNN	1ms	5.4k	85.6%
Mel-Spectrogram&CNN	1ms	9.2k	85%
Spectrogram&CNN	1ms	12.1k	86%

The feature extraction method determines the model's performance when the learning methods are all two-layers CNN with the same settings. It can be seen that the three models have little difference in the accuracy of sample recognition. Since Mel spectrogram and MFCC are both improved based on spectrogram for better recognition of human voice, the recognition accuracy of other sounds is slightly decreased. However, it has been widely confirmed that MFCC is more capable of recognizing bird calls, and there are two bird classes in the data set used in this experiment. Therefore, the performance of MFCC in accuracy is slightly higher than that of the Mel spectrogram. For example, in the experiment with the best performance, the recognition accuracy of MFCC for duck is 91%, and the recognition accuracy of song sparrow is 97.8% (table 6), which is much higher than the other two models. In addition, because the audio length used in the data set is only one second, the three models perform very well in the inferring time. Finally, the gap between the three indicators in peak RAM usage is the largest. Although the calculation steps of MFCC are more than those of the other two methods, the number of frames is smaller due to the longer frame length and the larger frame spacing in the framing step of MFCC, and the calculation amount and the peak RAM usage are reduced.

DOI: 10.54254/2755-2721/4/2023420

Table 6. Recognition detail of MFCC&CNN model.

	Song Sparrow	Cat	Dog	Duck	Frog
Song Sparrow	97.8%	0%	0.7%	1.5%	0%
Cat	4.4%	81.0%	6.7%	7.9%	0%
Dog	1.9%	19.1%	71.6%	6.2%	1.2%
Duck	5.3%	1.1%	2.7%	91.0%	0%
Frog	0%	0%	0%	0%	100%

Combining the performance of accuracy, peak RAM usage, inferencing time, Etc., the model of MFCC & Classification is finally chosen because it has comparatively high accuracy and short inferencing time and uses the least RAM in peak, making it ideal for use on small devices in the wild.

3.2. Deviation Analysis

These three models all got an accuracy of between 85 and 86 percent in their best, which is a good result but still have space to improve. Therefore, the reasons that affected the accuracy rate are analyzed and came into the following conclusions:

a) Animal calls vary greatly depending on their age, especially in animals such as cats and dogs. In the dataset used, the age of the animals was not classified, and the calls of animals of the same species differed greatly confounded the model. For example, figure 6 shows the voiceprint of juvenile dog's barking and the model misrecognize it as a cat's meowing. Because the barking of juvenile dogs is not as rapid as that of adult dogs, the frequency is higher than that of adult dogs, and the voiceprint is similar to that of cats, causing the machine to misjudge.

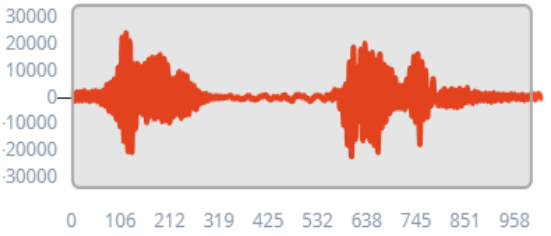

Figure 6. An example clip that a juvenile dog's call is recognized as the call of cats.

b) Animals will make different sounds when faced with different situations. For example, when cats and dogs face threats, they will make a low and long warning sound. The voiceprints are similar, and the models are difficult to distinguish.

c) When processing the audio, the original audios are automatically split into one second segments, which resulted in incomplete calls recorded in some segments. As shown in the figure 7 below, this clip only records the first half of the complete cat meow, and the voiceprint shows that the sound frequency is low and rapid, similar to that of an adult dog.

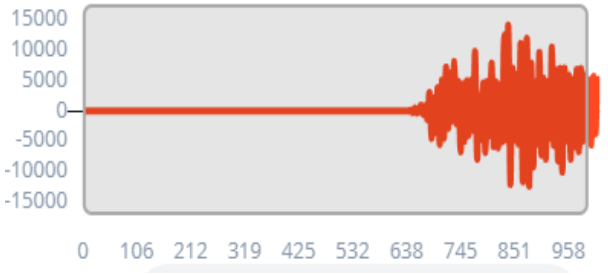

Figure 7. An example clip that only record the first half part of a "meow".

d) Because most of the calls of sparrows and ducks in the dataset were recorded in the wild, which led to the recordings often accompanied by the calls of other birds, the data could not fully reflect the call characteristics of the target birds.

DOI: 10.54254/2755-2721/4/2023420

e) In addition to the main appeal factor, some less influential factors are found either. For example, the different equipment used by the photographer resulted in a difference in the frequency of the recorded sound, and a small number of useless clips that were not eliminated are found.

Since all recordings used in training have been subjected to noise reduction processing, it is still challenging for the model to eliminate the interference of natural background noise and accurately capture and recognize animal calls when working in the field.

4. Conclusion

To sum up, a sound-based animal recognition model is hoped to be built and can be used in small devices in the wild. With the help of edge impulse, three models are designed and built at first, and only one model, the MFCC&CNN, is selected to be used in the end after comparisons of their performance in accuracy, peak RAM usage and inferencing time. The MFCC&CNN model has a short inferencing time, the lowest Peak RAM usage and an accuracy of 85.6%.

In future work, the problems mentioned in section 4.2 will be solved first, and more accurate and professional data will be collected to train and adjust the model. In the long term, the model is hoped to be trained to classify the calls of more than 100 animals with high accuracy and embedded into the equipment for field experiments in the wild. In the end, the model processed in the article is hoped to contribute to animal protection in the future.

References

[1] Mane, Anand D., R. A. Rashmi, and S. L. Tade. Identification & Detection System for Animals from their Vocalization. 2013, *Int J Adv Comput Scr* **3.3**: 352.

[2] Piczak, Karol J. Environmental sound classification with convolutional neural networks. 2015 *25th Int Mach Learn Sig Pro.* **23**.46

[3] Valenti, Michele, et al. Acoustic Scene Classification Using Convolutional Neural Networks. 2016, *Detect Classif Acoust Sce Ev* **53** 201.

[4] Şaşmaz, Emre, and F. Boray Tek. Animal sound classification using a convolutional neural network. *2018 3rd Int Conf Comput Sci Eng* **370**.77.

[5] Edge Impulse, *www.edgeimpulse.com.*

[6] Kaggle, *www.kaggle.com/datasets.*

[7] Rao, K. Sreenivasa, and Anil Kumar Vuppala. Speech processing in mobile environments, 2014, *Comput Syst Sci Eng* **22.**

[8] Stevens, Stanley Smith, John Volkmann, and Edwin Broomell Newman. "A scale for the measurement of the psychological magnitude pitch." The journal of the acoustical society of america 8.3 (1937): 185-190.

[9] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient and dynamic time warping (DTW) techniques. 2010, *arXiv preprint arXiv*:1003.4083.

[10] Rabiner, Lawrence, and Biing-Hwang Juang. Fundamentals of speech recognition. Prentice-Hall, 1993,Inc.,**542**.90.

[11] Li, Ying, and Zhibin Wu. Animal sound recognition based on double feature of spectrogram in real environment. 2015 *Int Conf Wir Comm Sig Proc* **734**.6

[12] Li, Ying, Hongkeng Huang, and Zhibin Wu. Animal sound recognition based on double feature of spectrogram. 2019 *Chinese J Electron* **28.4**: 667-673.