Classification of singing insect sounds with convolutional neural network

Sho Hibino, Chifumi Suzuki* and Takanori Nishino†

Faculty of Urban Science, Meijo University, 4-102-9 Yadamnam, Higashi-ku, Nagoya, 461-8534 Japan

(Received 25 June 2021, Accepted for publication 5 August 2021)

Keywords: Singing insect sound, Classification, Convolutional neural network, Data augmentation

1. Introduction

An insect chirp is a familiar sound that reminds us of the changing the seasons [1], and some insects have been the subject of natural phenomenon observations in Japan [2]. There are over a million species of insects in the world [3]. Even if they are the same species, insects have different patterns and colors in different regions. Therefore, we have difficulty in discriminating insects by their appearance. Furthermore, catching a glimpse of them is not always easy. However, in the case of insects that make sounds, their sounds can be observed. In addition, because each species makes different sounds, we believe it is possible to distinguish between insect species.

In this study, we used machine learning to classify insects so that people without specialized knowledge could classify the species [4]. We believe that applications that teach non-experts insect species based on singing insect sounds are important for education. Several previous researches used insect sounds to classify singing insects [5,6]. Previous research achieved 90% accuracy for 307 species with a Gaussian mixture model [5]. Another previous research using a support vector machine to discriminate insect sounds yielded up to 87.7% for distinguishing four species of crickets and up to 96.5% for distinguishing four species of cicadas [6]. Our experiment used a convolutional neural network to perform multi-class classification and investigated a method for determining insect species. Additionally, we investigated whether some data augmentation methods are effective for insect sounds.

2. Experimental conditions

2.1. Insect sound data

Seven insect species were chosen for discrimination, namely, three cicadas species (Aburazemi (*Graptopsaltria nigrofuscata*), Minminzemi (*Hyalessa maculaticollis*), and Tsukutsukuboushi (*Meimuna opalifera*)), three crickets species (Enma-koorogi (*Teleogryllus emma*), Mitsukado-koorogi (*Loxoblemmus doenitzi*), and Tsuduresase-koorogi (*Velarifictorus mikado*)), and one pine cricket specie (Matsumushi (*Xenogryllus marmoratus*)). We recorded insect sounds in Inuyama City, Aichi Prefecture, and used Minminzemi sounds from YouTube [7,8]. The sound data was recorded monaurally with a recorder (TASCAM, DR-07X) at 16-bit quantization and 44.1-kHz sampling frequency. The sound data

obtained from YouTube was converted to monaural using Audacity [9] with 16-bit quantization and 44.1-kHz sampling frequency.

Each insect's sound data was divided into 5-second segments, and 100 pieces of sound data were prepared for each type of insect.

2.2. Network structure

Figure 1 shows the network structure [10] used in the experiment. This network's input data is a spectrogram of insect sounds. Four convolutional layers with varying filter sizes were used in the first convolutional layer. The lengths are (1×8) , (1×16) , (1×32) , and (1×64) , respectively, and the output of the convolution in the frequency axis is fed into the next convolution layer. The filters with sizes of (8×1) , (16×1) , (32×1) , and (64×1) in the time axis are convolved in the second convolutional layer, respectively. The outputs of these four convolution layers are concatenated, and in the third convolution layer, a filter of size (1×16) is convolved in the frequency axis. A filter of size (16×1) is convolved in the time axis in the fourth convolution layer. The Global Average Pooling is performed in the next layer. Finally, all nodes are combined into seven in the fully connected layer, and classification is conducted using the softmax function.

2.3. Data augmentation

The experiment's data was divided using the hold-out method. There were 100 pieces of sound data for each type of insect, and they were divided into 50 for training and 50 for evaluation, respectively. Therefore, the number of training and test data was 350 apiece. We used three types of data augmentation methods: time shift, time stretch, and white noise [10,11]. The time shift method corresponds to the way of trimming the waveform because it performs a cyclic shift of the waveform. The white noise method is equivalent to varying the signal-to-noise ratio. Some insects change their chirp style depending on the temperature [12], and the time stretch method is thought to simulate such a change in chirp style.

The time shift method shifted a waveform with a shift length determined using a random number that follows a uniform distribution. The shift length is selected from 0.83 s (1/6 segment), 1 s (1/5 segment), 1.25 s (1/4 segment), 1.67 s (1/3 segment), and 2.5 s (1/2 segment). The time stretch method randomly stretched the original waveform. A stretched factor was used a random number that follows a uniform distribution from 0.8 to 1.2 in increments of 0.01. The white noise method added a random value to each sample of the

*e-mail: csuzuki@meijo-u.ac.jp

†e-mail: nishino@meijo-u.ac.jp

[doi:10.1250/ast.42.354]

Fig. 1 Network structure.

original waveform. This random value was a product of a random number that follows a standard normal distribution and a random integer that follows a uniform distribution from 1 to 50, and finally divided by 1,000. The original waveform amplitude values within ± 0.3 account for 97.9% of the total. Each data augmentation method generated 350 data, respectively. In addition, after using the white noise method, the time shift or time stretch method was randomly applied. Because we prepared four types of augmented data, the total number of data obtained by the data augmentation method was 1,400.

3. Results

3.1. Performances based on input features

We calculated the mel-spectrogram and the power spectrogram for each data and compared the correct rate of using each spectrogram. Both spectrograms were generated using the librosa module [13] in Google Colaboratory [14]. We used a Hanning window with a length of 1,024 points and a frame shift of 128 points. The training and discrimination were performed three times, and the average correct rate was obtained. Each time, the number of training epochs was 10. According to the results of the experiment, the average correct rate of the mel-spectrogram was 80.2%, and that of the power spectrogram was 68.2%. Because the mel-spectrogram's correct rate was higher than that of the power spectrogram, it is appropriate to use the mel-spectrogram as the input feature.

3.2. Effects of data augmentation

We conducted experiments with various combinations of original data and augmented data to determine which data augmentation method is more effective. The mel-spectrogram was used as input feature. The number of training epochs was 10, and the number of test data was 350. When four types of augmented data were individually added to the original data, the number of training data was 700, respectively. When we used all of the augmented data obtained by the time shift, the time stretch, and the white noise, we had 1,400 training data. Moreover, when all of the augmented data was used, the total number of training data was 1,750. In addition, we trained the network both without and with the mixup method [15] in the input layer.

Under each training data condition, the training and discrimination were performed three times. The average correct rate is shown in Table 1. When using a single data

Table 1 Correct rate [%] (the mean of three times training).

Types of data augmentation method	Without mixup	With mixup
(Base line)	(80.2)	(90.7)
Time shift	90.3	94.6
Time stretch	92.4	95.3
White noise	84.0	90.9
White noise and time shift/time stretch	78.5	88.7
Three types of data augmentation	93.0	95.2
All augmentation	95.9	97.8

augmentation method, these results show that the mixup method is effective. The combination of the original data and the time stretch is highly accurate. Furthermore, the accuracy was maximized by utilizing all of the augmented data. As a result, even in the insect sound classification, increasing the number of training data by using the data augmentation method helps to improve the correct rate.

3.3. Overall evaluation

Classification experiments were conducted using a network trained with all of the augmented data and the mel-spectrogram as an input feature. The number of training epochs was 59 without the mixup and 50 with the mixup. As a result, the correct rate was 97.7% (342/350) for without the mixup and 99.7% (349/350) with the mixup. The sound data of the Minminzemi was misidentified as the sound of the Tsukutsukuboshi in the case of with the mixup. The mixture of several insect sounds was one of the possible causes for the misidentification.

4. Conclusion

In this study, we used a convolutional neural network to distinguish between insect species based on their chirps. The experiment proved that it was possible to distinguish insects from their chirps with 99.7% accuracy. In the future, it will be necessary to thoroughly evaluate performance using the cross-validation method. Furthermore, the recording conditions must be considered so that the sounds of insects are not mixed, and discrimination experiments with a larger number of target insects must be conducted.

Acknowledgment

A part of this work was supported by JSPS KAKENHI Grant Number 19K12193. The authors would like to thank Enago (www.enago.jp) for the English language review.

References

- [1] S. Hozumi, T. Inagaki and K. Fukuda, “Effects of sounds of crickets on human’s KANSEI,” *J. Jpn. Soc. Kansei Eng.*, **8**, 1137–1144 (2009) (in Japanese).
- [2] Japan Meteorological Agency, <https://www.data.jma.go.jp/sakura/data/> (in Japanese) (accessed 25 June 2021).
- [3] N. E. Stork, “How many species of insects and other terrestrial arthropods are there on earth?,” *Annu. Rev. Entomol.*, **63**, 31–45 (2018).
- [4] S. Hibino, C. Suzuki and T. Nishino, “Insect sound classification with convolutional neural network,” *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 189–190 (2021) (in Japanese).
- [5] T. Ganchev and I. Potamitis, “Automatic acoustic identification of singing insects,” *Bioacoustics*, **16**, 281–328 (2007).
- [6] T. Wada, J. Akita and A. Kitagawa, “Environmental monitoring system in cooperation with smartphones by automatic identification of insect sounds —Chu-lingual,” *IPSJ J.*, **53**, 1017–1021 (2012) (in Japanese).
- [7] alpsdake, “Hyalessa maculaticollis singing,” <https://youtu.be/sHpHvkbUIM0> (accessed 25 June 2021).
- [8] miyusaluin, “Minminzemi,” https://youtu.be/_n73qJ-8DPQ (accessed 25 June 2021).
- [9] Audacity, <https://www.audacityteam.org/> (accessed 25 June 2021).
- [10] Y. Shibui, <https://qiita.com/cvusk/items/61cdbce80785eaf28349> (in Japanese) (accessed 25 June 2021).
- [11] L. Nanni, G. Maguolo and M. Paci, “Data augmentation approaches for improving animal audio classification,” *Ecol. Inform.*, **57**, Article 101084 (2020).
- [12] J. VanCleave, “Cricket thermometer,” in *Janice VanCleave’s Biology for Every Kid: 101 Easy Experiments That Really Work* (John Wiley & Sons, New York, 1990), pp. 104–105.
- [13] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg and O. Nieto, “librosa: Audio and music signal analysis in python,” *Proc. 14th Python Science Conf.*, pp. 18–25 (2015). <https://librosa.org/> (accessed 25 June 2021).
- [14] Google Colaboratory, <https://colab.research.google.com/notebooks/intro.ipynb> (accessed 25 June 2021).
- [15] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv:1710.09412* (2017).