Automatically Distinguishing Adult from Young Giant Pandas Based on Their Call

Yanqiu Zhang[1], Rong Hou[3,4], Longyin Guo[1], Peng Liu[3,4], Shan Zhang[3,4], Peng Chen[3,4](✉), and Qijun Zhao[1,2](✉)

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, Sichuan, People's Republic of China

[2] College of Computer Science, Sichuan University, Chengdu, Sichuan, People's Republic of China
qjzhao@scu.edu.cn

[3] Chengdu Research Base of Giant Panda Breeding, Sichuan Key Laboratory of Conservation Biology for Endangered Wildlife, Chengdu, Sichuan, People's Republic of China

[4] Sichuan Academy of Giant Panda, Chengdu 610086, People's Republic of China

Abstract. Knowing the number of adult and young individuals in the giant panda population is of significant in protecting them. Traditionally, this is done by field investigation, which has high cost and low time efficiency. We approach the problem as a two-class soft-biometric recognition task: given a segment that records the call of a giant panda, we aim to classify the giant panda into adult or young. We construct 1,405 call clips of adult giant panda and 285 call clips of young giant panda and propose a framework for vocal-based giant panda age group classification. Based on the framework, we evaluate the effectiveness of various acoustic features and deep neural networks. The results suggest that (i) adult and young giant pandas show different characteristics in high-frequency acoustic features, (ii) it is feasible to automatically distinguish between adult and young giant pandas based on their call with deeply learned features.

Keywords: Age group classification · Acoustic features · Soft-biometric · Giant panda · Wildlife conservation

1 Introduction

Age, as a soft biometric trait, has important research significance [1]. There are many literatures about the prediction of human age [2–4]. However, only a few studies on the age prediction of animals. For giant pandas, the age structure is related to the sustainability of the population, so it is very important to know the age of them. Sound is one of the main ways of animal communication [5]. It has the advantages of long transmission distance, strong penetration and rich information. Although there are many literatures on vocal-based age prediction [4,6–8], they are all about human beings. It is of great interests to study vocal-based age prediction of giant pandas.

J. Feng et al. (Eds.): CCBR 2021, LNCS 12878, pp. 92–101, 2021.
https://doi.org/10.1007/978-3-030-86608-2_11

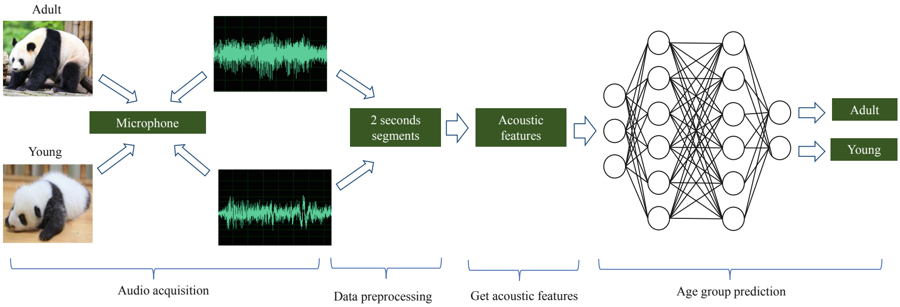


Fig. 1. The call is collected from the Chengdu Research Base of Giant Panda Breeding. We design a framework to predict whether the sound belongs to adult giant panda or young giant panda, so as to analyze the age composition of giant panda.

When using audio signal as the input of network, it is usually necessary to transform the original audio signal into Mel-frequency Cepstral Coefficients (MFCC). MFCC is a feature widely used in automatic speech and speaker recognition. Mel frequency is proposed based on human auditory characteristics. Since human ears are more sensitive to low-frequency signals, MFCC mainly extracts low-frequency information in the signal. In contrast, high-frequency information may play important roles in vocal-based animal communication. Yet, existing literatures almost do not mention how to deal with high-frequency signals of animals. We will for the first time investigate the potential of automatically recognizing the age groups of individual giant pandas (i.e., adult or young) based on their call by using deep neural networks.

We will also study the relative contribution of high-frequency components of the call signal of giant pandas to the age prediction. The rest of this paper is organized as follows. Section 2 briefly introduces related work. Sections 3 and 4, respectively, introduce the dataset and the methods used in this study. Section 5 then presents and analyzes the experimental results. Section 6 finally concludes the paper.

2 Related Work

Due to the increasing awareness of ecological environment protection, researchers began to apply speech recognition algorithms for animals [9]. It is already possible to identify species of birds according to their vocal features [10]. Hendrik et al. constructed a new model to recognize the types of killer whale call, and successfully visualize the feature map [11]. In particular to giant pandas, some researchers proposed a method to recognize individual giant pandas from their call, achieving accuracy of 88.33% on a set of 20 individuals [12]. Others devised a system to automatically predict the mating success of giant pandas based on their call during breeding encounters [13]. However, for both [13] and [12], they do not predict the age of giant pandas, and the subjects are adult giant pandas, and no experiments are conducted on young giant pandas. In [12], the recognition of adult giant panda is not an end-to-end automatic recognition system. The feature of MFCC is extracted in [13], which does not explain the influence of high-frequency information on the speech recognition of giant panda.

The method of determining the age of giant panda is based on chemical signal analysis [14], such as analyzing their excreta. However, it takes a long time for human to collect giant panda's excreta, at which time the chemical signal is to lose effectiveness [15]. So, it is difficult for human to use this information to quickly determine the age groups of the giant panda.

The call of giant pandas contain extremely rich information and giant pandas also have obvious changes in their vocalization [16]. The method of using call for predicting the age has not yet appeared. Based on the above reasons, we use deep learning to predict the age groups of giant pandas by call.

3 Dataset

3.1 Data Acquisition

The data used in this paper are acquired at the Chengdu Research Base of Giant Panda Breeding, by using Shure VP89m, TASCAM DR-100MKII and SONY PCM-D100. The call record is converted from dual channel to single channel, and the sampling rate of single channel call is normalized to 44,100 Hz. There are 1,009 call segments of adult giant pandas, which include the call of both female and male giant pandas. The duration of each call lasts from 0.29 s to 3.25 s. We collect the call from five young giant pandas, and there are 25 calls, each lasting from 5.67 s to 1minute and 5 s.

3.2 Data Preprocessing

In order to ensure the consistency of data dimensions during training, the original call clips of giant pandas are divided into 2 s clips without overlap, and the segments whose duration is shorter than 2 s are expanded to 2 s via zero-padding on the log-mel spectrum. In this study, we mainly focus on investigating whether the call of giant pandas can reveal their age groups, specifically, young and adult. Therefore, we manually exclude the call clips that are contaminated by other sounds (e.g., human voice or collision sounds between objects). Finally, we obtain 285 call clips of young giant pandas, and 1,405 call clips of adult giant pandas.

Table 1. Duration of call data used in this study.

Age group	Adult	Young
Training data	1,560 s	300 s
Validation data	60 s	90 s
Test data	135.2 s	158.4 s

The overall duration of these call clips are 38.4 min. We divide these data into three subsets, i.e., training, validation and test data. Table 1 summarizes the duration of the call data used in this study.

According to the data acquisition conditions in this study, the call data of young giant pandas have relatively weak background noise, while those of adult ones have some strong background noise (e.g., sound of working air-conditioners). To avoid such bias in the data, we first collect some call data of only background sound appearing in the data of adult giant pandas, and then integrate them into the call of young giant pandas via mixing the signal to enforce that the call data of young and adult giant pandas share similar background noise.

Considering the obvious imbalance between the call data of adult and young giant pandas, we augment the data of young giant pandas by adding Gaussian noise and applying SpecAugment [17]. In SpecAugment, we apply frequency masking by adding a mask with value of 0 to the frequency axis of log-mel spectrum, and time masking by adding a mask with value of 0 to the time axis. After the aforementioned preprocessing of background noise synthesis and augment young giant panda data, the training data for young and adult giant pandas both have 1,280 call clips.

4 Method

We compare three acoustic features, MFCC feature, Pre-emphasis feature and IMFCC feature, by six neural networks to analyze their influence on giant panda age group classification.

4.1 Acoustic Feature

MFCC (Mel-frequency Cepstral Coefficients) is a feature widely used in automatic speech and speaker recognition. It assumes that the human ear is more sensitive to the low-frequency part, which is extracted by triangle mel filter.

IMFCC (Invert Mel-frequency Cepstral Coefficients) is a way to extract high-frequency information in audio data. It uses the inverted triangle mel filter, which is more tightly in the high-frequency information part and can extract more high-frequency information.

Pre-emphasis operation can reduce the loss of high-frequency. The so-called Pre-emphasis is to increase the resolution of high-frequency signal in the original audio signal. This way, in the subsequent extraction of MFCC, the damage of high-frequency signal is reduced and more comprehensive information is obtained.

4.2 Input Feature

Three types of input features are considered in this study. 1) MFCC features F_M extracted directly by triangle mel filter M. 2) The raw signal z with Pre-emphasis H, then triangle mel filter M is used to extract acoustic features, and the extracted features are fused with F_M. The final input feature is multi-scale feature F_P. As shown in formula (1). 3) The original signal z is extracted by invert triangle filter I to get IMFCC features, and then fused with F_M to get multi-scale feature F_I. As shown in formula (2).

$$F_P = \{M(H(z)), F_M\}\ . \tag{1}$$

$$F_I = \{I(z), F_M\}\ . \tag{2}$$

4.3 Network Structure

In this paper, six of deep networks are selected for comparative evaluation: AlexNet [18], Xception [19], SENet [20], CRNN [21], Xcep-RNN, and Xcep-RNN-attention. CNN has a very good effect for extracting local features, which helps to extract detailed information from call. However, for voice, the recurrent neural network is more useful because of the characteristics of temporal series data. So, we chose the very typical network in CNN and RNN for experiments.

Xcep-RNN is changing the ordinary convolution of CRNN to depthwise separable convolution, so that the local information extracted from the sample can be realized cross channel interaction and information integration through 1×1 convolution to obtain more comprehensive information. On this basis, the attention [22] is added to obtain more important feature information.

5 Experiments

All the CNN input size is $(180, 32, 1)$, RNN input size is $(180, 32)$. Except for SENet, others batch size are 16, and the SENet is 8. The learning rate is 10^{-3}, and the loss is cross entropy. Each network is trained three times, and 100 epochs each time. The models used in the test are the best three models saved in each network training, and the final evaluation index is the average accuracy of the three models. All experiments are done on Ubuntu, and use the NVIDIA GTX 1070 to help train networks.

5.1 The Effectiveness of Augmentation

We augment the young giant panda call data and experiment the effectiveness of it. The acoustic feature is MFCC. Table 2 shows the result. It obvious to see after augmenting the data, the accuracy of young giant panda has been greatly improved.

Table 2. The effectiveness of augmenting young giant panda data.

Augmentation	Model	AlexNet	Xception	SENet	CRNN	X-RNN	X-RNN-a
No	All	76.88%	78.13%	80.63%	71.25%	71.25%	75.63%
	Adult	91.25%	92.50%	93.75%	81.25%	81.25%	85.00%
	Young	62.50%	63.75%	67.50%	61.25%	61.25%	66.25%
Yes	All	89.38%	90.63%	90.63%	81.88%	81.25%	84.38%
	Adult	97.50%	95.00%	97.50%	83.75%	82.50%	85.00%
	Young	81.25%	86.25%	83.75%	80.00%	80.00%	83.75%

5.2 Compare the Effectiveness of MFCC and Pre-emphasis

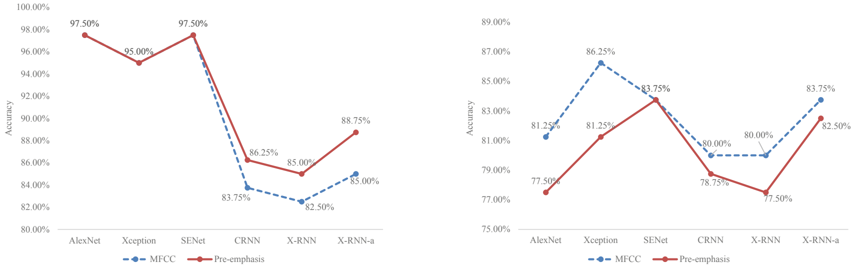
The multi-scale feature F_P obtained by this operation will be put into the network for experiment. The experimental results are shown in Table 3.

Table 3. The result of pre-emphasis.

Model	AlexNet	Xception	SENet	CRNN	X-RNN	X-RNN-a
All	87.50%	88.13%	90.63%	82.50%	81.25%	85.63%
Adult	97.50%	95.00%	97.50%	86.25%	85.00%	88.75%
Young	77.50%	81.25%	83.75%	78.75%	77.50%	82.50%

In the Pre-emphasis experiment, the experimental accuracy of CNN is better than that of RNN. The accuracy of SENet is highest for both adult and young pandas, with adult panda prediction accuracy reaching 97.50% and young panda prediction accuracy 83.75%. In order to better observe the experimental results of MFCC and Pre-emphasis, the adult and young giant panda prediction accuracy in the two experiments is respectively made into a line chart, as shown in Fig. 2.

As Fig. 2 shows, for the prediction accuracy of adult giant panda, the high-frequency information obtained by Pre-emphasis is helpful to the prediction of adult giant panda. For the young giant panda, the high-frequency information obtained by Pre-emphasis is not conducive to the prediction of the young giant panda.



(a) Adult giant panda (b) Young giant panda

Fig. 2. Prediction accuracy of adult and young giant panda by MFCC and pre-emphasis.

5.3 Compare the Effectiveness of MFCC and IMFCC

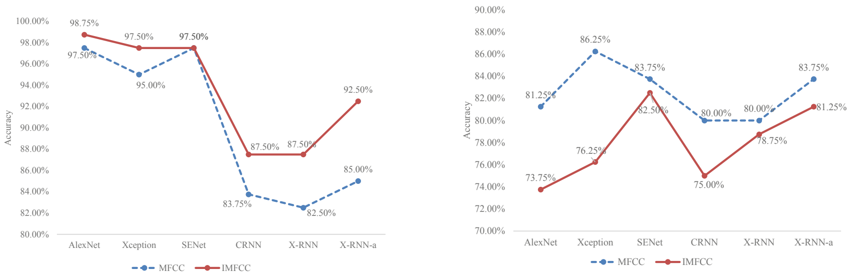
The multi-scale feature F_I is put in six deep networks. The results are shown in Table 4.

Table 4. The result of IMFCC.

Model	AlexNet	Xception	SENet	CRNN	X-RNN	X-RNN-a
All	86.25%	86.88%	90.00%	81.25%	83.13%	86.88%
Adult	98.75%	97.50%	97.50%	87.50%	87.50%	92.50%
Young	73.75%	76.25%	82.50%	75.00%	78.75%	81.25%

In this experiment, SENet still works best. Similarly, the experimental results of IMFCC and MFCC are plotted by line chart, as shown in Fig. 3.

Like the Pre-emphasis features, the prediction accuracy of adult giant panda is improved after using IMFCC. The result shows that the young giant panda's prediction accuracy of IMFCC is lower than that of MFCC.



(a) Adult giant panda (b) Young giant panda

Fig. 3. Prediction accuracy of adult and young giant panda by MFCC and IMFCC.

5.4 Discussion on the Results

Through deep learning to analyze the call of giant pandas, we can effectively recognize the age structure of giant pandas. Pre-emphasis and IMFCC operations are performed on the data to extract the high-frequency information of giant panda call. The experimental results show that both operations can improve the prediction accuracy of adult giant pandas and reduce the prediction accuracy of young giant pandas.

Some literatures can also support our experimental results. These literatures covers not only computer science, but also zoology. The theoretical support of zoological literatures make our results more reliable. The high-frequency information of adult giant panda has certain biological characteristics [12,23], while the high-frequency part of young giant panda is mostly gas noise [24], which is not conducive to the prediction of young giant panda. The research of giant panda age structure prediction based on call is of great application value to the protection and breeding of giant pandas in the wild.

6 Conclusion

We collect a dataset about giant panda call including adult and young, and predict giant panda age groups via their call automatically for the first time. We do experiments on three different acoustic characteristics to analyze the influence of high-frequency signal on giant panda age groups prediction. The results show that it is feasible to automatically distinguish between adult and young giant pandas based on their call with deeply learned features. The age groups prediction of giant pandas based on call is helpful to analyze the age structure of giant pandas in the wild. In the future work, we will try to classify the gender information of giant pandas by their call. We hope to design a set of giant panda attributes detection system based on call.

Acknowledgments. This work was supported by the Chengdu Research Base of Giant Panda Breeding [No. 2021CPB-C01; No. CPB2018-02; No. 2020CPB-C09; No. 2021CPB-B06].

References

1. Anda, F., Becker, B.A., Lillis, D., Le-Khac, N.A., Scanlon, M.: Assessing the influencing factors on the accuracy of underage facial age estimation. In: The 6th IEEE International Conference on Cyber Security and Protection of Digital Services (Cyber Security). (2020)
2. Guo, G., Yun, F., Huang, T.S., Dyer, C.R.: Locally adjusted robust regression for human age estimation. In: IEEE Workshop on Applications of Computer Vision (2008)

3. Hernandez-Ortega, J., Morales, A., Fierrez, J., Acien, A.: Detecting age groups using touch interaction based on neuromotor characteristics. Electr. Lett. **53**(20), 1349–1350 (2017)
4. Li, X., Malebary, S., Qu, X., Ji, X., Xu, W.: iCare: automatic and user-friendly child identification on smartphones. In: the 19th International Workshop (2018)
5. Liuni, M., Ardaillon, L., Bonal, L., Seropian, L., Aucouturier, J.-J.: ANGUS: real-time manipulation of vocal roughness for emotional speech transformations. arXiv preprint arXiv:2008.11241 (2020)
6. Ming, L., Han, K.J., Narayanan, S.: Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. Comput. Speech Lang. **27**(1), 151–167 (2013)
7. Meinedo, H., Trancoso, I.: Age and gender classification using fusion of acoustic and prosodic features. In: Interspeech Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 2010
8. Kaushik, M., Pham, V.T., Chng, E.S.: End-to-end speaker height and age estimation using attention mechanism with LSTM-RNN. arXiv preprint arXiv:2101.05056 (2021)
9. Oikarinen, T., Srinivasan, K., Meisner, O., Hyman, J.B., Parmar, S., Fanucci-Kiss, A., Desimone, R., Landman, R., Feng, G.: Deep convolutional network for animal sound classification and source attribution using dual audio recordings. J. Acoust. Soc. Am. **145**(2), 654–662 (2019)
10. Solomes, A.M., Dan, S.: Efficient bird sound detection on the bela embedded system. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)
11. Schröter, H., Nöth, E., Maier, A., Cheng, R., Barth, V., Bergler, C.: Segmentation, classification, and visualization of orca calls using deep learning. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8231–8235 IEEE (2019)
12. Lu, H.: Analysis and research of giant panda individual identification system based on voiceprint (2019)
13. Yan, W., et al.: Automatically predicting giant panda mating success based on acoustic features. Glob. Ecol. Conserv. **24**, e01301 (2020)
14. White, A.M., Zhang, H.: Chemical communication in the giant panda (ailuropoda melanoleuca): the role of age in the signaller and assessor. J. Zool. **259**(2), 171–178 (2003)
15. Yuan, H., Liu, D., Wei, R., Sun, L., Zhang, S.: Anogenital gland secretions code for sex and age in the giant panda, ailuropoda melanoleuca. Can. J. Zool. **82**(10), 1596–1604(9) (2004)
16. Charlton, B.D., Owen, M.A., Keating, J.L., Martin-Wintle, M.S., Swaisgood, R.R.: Sound transmission in a bamboo forest and its implications for information transfer in giant panda (ailuropoda melanoleuca) bleats. Sci. Rep. **8**(1) (2018)
17. Park, D.S., et al.: Specaugment: a simple data augmentation method for automatic speech recognition. In: Interspeech 2019 (2019)
18. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**(2) (2012)
19. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

21. Shi, B., Xiang, B., Cong, Y.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Patt. Anal. Mach. Intell. **39**(11), 2298–2304 (2016)
22. Vaswani, A., et al.: Attention is all you need. arXiv (2017)
23. Jing Zhu, Z.M.: Estrus calls of giant pandas and their behavioral significance. Curr. Zool. (1987)
24. Cannan Zhao, P.W.: The sound spectrum analysis of calls in the bay giant panda. Discov. Nat. (1988)