Mathematical Biosciences and Engineering

MBE, 20(8): 15456–15475.
DOI: 10.3934/mbe.2023690
Received: 06 May 2023
Revised: 19 June 2023
Accepted: 04 July 2023
Published: 24 July 2023

http://www.aimspress.com/journal/MBE

Research article

Automatic recognition of giant panda vocalizations using wide spectrum features and deep neural network

Zhiwu Liao[1,2], Shaoxiang Hu[3], Rong Hou[4], Meiling Liu[5], Ping Xu[5], Zhihe Zhang[5] and Peng Chen[4,*]

[1] Key Laboratory of Land Resources Evaluation and Monitoring in Southwest China, Ministry of Education, Sichuan Normal University, Chengdu, China
[2] Academy of Global Governance and Area Studies, Sichuan Normal University, Chengdu, China
[3] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China
[4] Chengdu Research Base of Giant Panda Breeding, Sichuan Key Laboratory of Conservation Biology for Endangered Wildlife, Chengdu 610081, China
[5] Giant Panda National Park Chengdu Administration, Chengdu 610096, China

* **Correspondence:** Email: capricorncp@163.com.

Abstract: The goal of this study is to present an automatic vocalization recognition system of giant pandas (GPs). Over 12800 vocal samples of GPs were recorded at Chengdu Research Base of Giant Panda Breeding (CRBGPB) and labeled by CRBGPB animal husbandry staff. These vocal samples were divided into 16 categories, each with 800 samples. A novel deep neural network (DNN) named 3Fbank-GRU was proposed to automatically give labels to GP's vocalizations. Unlike existing human vocalization recognition frameworks based on Mel filter bank (Fbank) which used low-frequency features of voice only, we extracted the high, medium and low frequency features by Fbank and two self-deduced filter banks, named Medium Mel Filter bank (MFbank) and Reversed Mel Filter bank (RFbank). The three frequency features were sent into the 3Fbank-GRU to train and test. By training models using datasets labeled by CRBGPB animal husbandry staff and subsequent testing of trained models on recognizing tasks, the proposed method achieved recognition accuracy over 95%, which means that the automatic system can be used to accurately label large data sets of GP vocalizations collected by camera traps or other recording methods.

Keywords: Mel filter bank (Fbank); medium Mel filter bank (MFbank); reversed Mel filter bank (RFbank); gated recurrent unit (GRU); 3Fbank-GRU; vocalization recognition; deep neural

network (DNN)

1. Introduction

Animal vocalization is a fascinating topic and contains valuable evidence about animal behaviors and ecosystems. Giant pandas (Ailuropoda melanoleuca) are monestrous mammals and can make 19 different calls in their first 4–5 weeks of life or during their rather short breeding periods [1–10]. Each call has a specific meaning and the mixture of different calls expresses a certain emotion [1–3].

In the 80s', the most representative system Sphinx [11] was developed by Lee Kaifu et al. of California Miramar University (CMU) based on hidden Markov model (HMM) [12], Gaussian mixture model (GMM) [13] and multivariate grammar model (N-gram) [14]. It is the first speech recognition system which is available to nonspecific person.

One of the earliest works applied a convolution neural network (CNN) to identifying 10 anuran species by bio-acoustic data [15]. In the same year, three of the six teams in the 2016 BirdCLEF challenge submitted CNN systems taking spectrograms as input, including the highest-scoring team [16]. Stowell investigated computational bioacoustics with deep learning and indicated at least 83 of the surveyed articles made use of CNNs until 2021 [17].

Recurrent neural network (RNN) [18,19] brought a new breakthrough for acoustic modeling of speech recognition. RNN created *memory* through the superposition of speech in time. Moreover, long-short term memory (LSTM) solved the problem of gradient vanishing of RNN [20]. The CS-CLDNN (CBAM-Switch-CLDNN) combined the CNN and LSTM models with the convolutional block attention module (CBAM). The recognition accuracy of CS-CLDNN on 20 bird species reached 0.975 [21]. Bergler et al. proposed an orca call types classification based on ResNet18 [22] and Waddell et al. classified six fish call types in the northern Gulf of Mexico Stowell based on ResNet-50 [23].

However, training LSTM was not a trivial task because of its many parameters. Gated recurrent unit (GRU) [24] reduced computation cost through updated gates and reset gates while keeping "memory". Zhang et al. presented a framework composed by convolution module and GRU module, to predict GP's mating success using their vocalizations. However, the recognition accuracy is only about 85% [25].

According to the investigation in [21], there were many efforts on computational Bioacoustics using deep neural networks (DNNs). However, most of works focused on the species classifications based on animal calls, which were easier than the vocalization recognition since the sound differences among species were greater than the differences of the same species vocalizations. Few works devoted to call types classification [22,23,25] with classification accuracy about 85% and still had gaps between existing methods and vocalization recognition.

In summary, although animal vocalization recognition is a very important in animal behaviors analysis and conservation, it is an unsolved open problem even now. Especially, for GP's vocalization recognition, the difficulties are:

1) Sound data collection and labeling is very difficult. The giant panda is a silent animal, only producing calls within four weeks of birth and during a very short mating season. The sound data were all manually labeled by the staff of Chengdu Research Base of Giant Panda Breeding

(CRBGPB). Thus, it required long time to collect GP's calls and very heavy workload to manually label the sound data.
2) There were a lot of ambient noises (such as people talking, opening and closing doors etc.) and calls such as peacocks and other birds.
3) Existing methods cannot be used directly to GP's vocalization recognition because of their low recognition accuracy.

Therefore, we should propose a new framework both in feature extraction and vocalization recognition using DNN to improve the performance of existing methods. After analyzing the GP's sound, we found that it is a broad frequency signal. Thus, two new filters, medium Mel filter bank (MFbank) and reversed Mel filter bank (RFbank), were proposed to extract medium and high frequency features. Combined above two band features with the low-frequency feature extracted by Mel filter bank (Fbank) [17], the three banks' features were sent to a deep network composed by GRUs, named 3Fbank-GRU. 3Fbank indicated the inputs of the DNN were the features extracted by three filter banks and GRUs were the main components of the DNN keeping the 'memory' while reducing computation cost of LSTM.

The main contributions of this paper are:
1) Proposed a new feature-extraction scheme that not only used Fbank to extract low frequency features but also introduced new RFbank and MFbank to extract high and medium frequency features.
2) Proposed a novel DNN composed by GRUs to process the three-bank features of GP's sound data.
3) The proposed 3Fbank-GRU method achieved a high recognition accuracy rate of over 95%, and was suitable for labeling large data sets of GP vocalizations collected by camera traps or other recording methods.

The remainder of this paper is as follows: in the second section, the materials and methods are presented; Experiments and discussion will be given in Sections 3 and 4.

2. Materials and methods

2.1. Sample collection

The data set of GP vocal samples was collected from 176 pandas at the CRBGPB. The subjects include cubs, sub-adult and adult GPs. We used a ShureVP89M directional microphone and a TascamDR-100MK3 handheld recorder (10 Hz~192 0.1/-0.5 dB) to record vocalizations made by GPs. When recording, sampling frequency was set to 48 KHz and 16 bits, since high sampling rate can reduce the distortion of audio and ensure the high frequency vocalizations that were recorded. The audio was saved in WAV format.

At the same time, SONY FDR-AX60 video cameras were used to capture all the video and sound data of the subjects as the basis for classification of vocalization types. The experienced animal husbandry staff of the CRBGPB labeled the vocalizations of the GPs.

20h 20min of data plus the data set previously collected, formed a total of 35h 50min data set. After data collection, the segments containing GP vocalizations were selected from the original datasets manually. Then, the selected segments were classified into 16 types (Table 1). If there was data that could not be labeled accurately because of noises or data missing, it was discarded.

The data was edited with Goldwave audio editing software. The length of each sound clip was 2–4

seconds including at least two peaks of the original signal in order to observe dependency among neighboring vocalizations. Finally, we obtained a total of 12,800 samples which were divided into 16 categories. Each category contains 800 samples.

Table 1. Giant Panda Sound Samples.

Serial number	Age of Giant Pandas	Callout	Behavior or state	Duration	Number of segments
1	Giant Panda Baby	Squeak	Hunger	2–4 s	800
2	Giant Panda Baby	Goo	Full food	2–4 s	800
3	Giant Panda Baby	Mm	Fight and fight	2–4 s	800
4	Young Giant Panda	Haw	Pray for food	2–4 s	800
5	Young Giant Panda	Hum	Playing games	2–4 s	800
6	Adult Giant Panda	Chirping	Not willing	2–4 s	800
7	Adult Giant Panda	Bleat	estrus	2–4 s	800
8	Adult Giant Panda	Bird Scream	Granted	2–4 s	800
9	Adult Giant Panda	Bark	Medium threat	2–4 s	800
10	Adult Giant Panda	Strong Bark	Strong stimulus	2–4 s	800
11	Adult Giant Panda	Ow	Area	2–4 s	800
12	Adult Giant Panda	Howl	Searching	2–4 s	800
13	Adult Giant Panda	Roar	Precursors of direct attack	2–4 s	800
14	Adult Giant Panda	Scream	Strong threat	2–4 s	800
15	Adult Giant Panda	Hiss	Extreme panic	2–4 s	800
16	Adult Giant Panda	Gasp	Gasping	2–4 s	800

2.2. Preprocessing

The sound data contained many systems and environmental noises. Therefore, the noise suppression was carried out to reduce the influence of background noises. We used the minimum mean square error (MMSE) estimation to suppress background noises (see Figure 1 and Table 2).

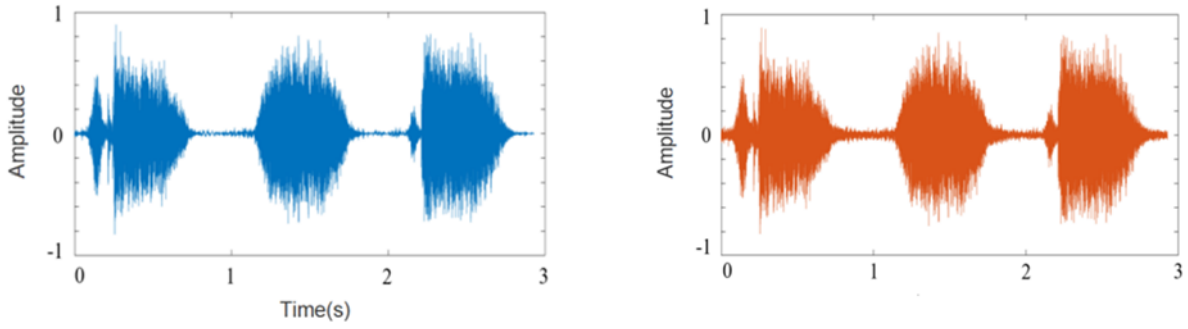

Figure 1. Waveforms of the original signal and denoised signal of a GP's call. From left to right: a waveform of original signal of a GP's call, and a waveform of the denoised signal using MMSE estimation.

Table 2. Signal to noise rate (SNR) of GP sound signals before and after noise reduction.

Signals	1	2	3	4	5
Original Signals	8.8	12.26	15.88	28.24	29.35
Denoised Signals	**19.65**	**26.55**	**38.71**	**54.74**	**56.47**

2.3. Feature extraction

Voiceprint is the acoustic spectrum that carries speech information. The most commonly used voiceprint in sound recognition is Mel filter bank (Fbank) coefficient [21]. However, the classical Fbank cannot get satisfied recognition results for GP vocalization recognition because it only extracts low-frequency information but the frequency spectrum of GP vocalization is very wide. In order to utilize more information from GP's calls, two new filters were proposed to extract other two band coefficients, reverse Mel (RMel) coefficients and medium Mel (Mmel) coefficients, to improve the performance of the whole system.

2.3.1. Fbank feature

A typical framework of Fbank goes through a pre-emphasis filter, is sliced into overlapping frames and a window function is applied to each frame. Afterwards, Fourier transform is performed on each frame (or more specifically a short-time Fourier transform) and then the power spectrum is calculated. Next, convert the linear frequency to Mel frequency and then designs equispaced triangular filters on Mel-scale and coverts it to a linear frequency using Eq (3) to get dense triangles on low frequencies while sparse on other frequencies (Figure 2). Finally, converted triangular filters are applied to the power spectrum to extract the frequency bands [11].

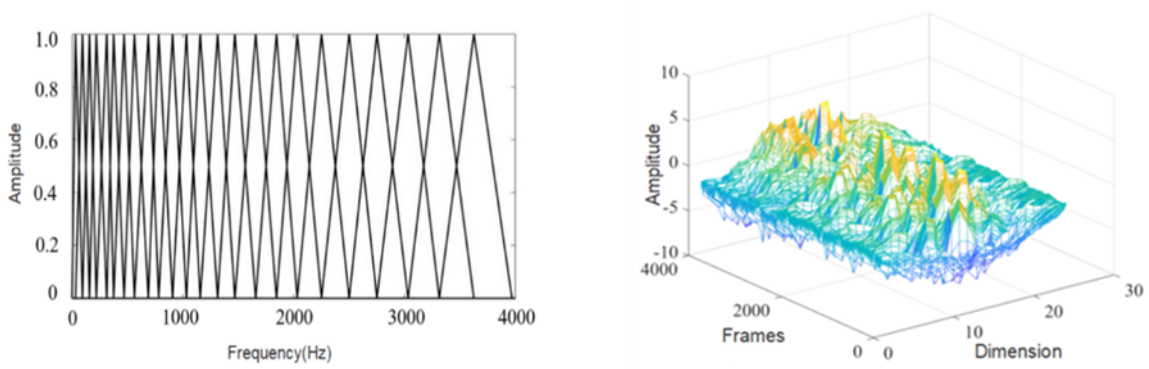

Figure 2. Converted Mel filter banks on linear frequency (left) and the 3D diagram of the Fbank feature (right). The filters are dense on low frequencies and sparse on other frequencies (left). The GP voiceprint, the Fbank feature, was extracted using the method in subsection 2.3.1 (right). The frame length was 256 sampling points, frame shift was 128 sampling points, the length of FFT was 256 and the order of the filter bank was 24.

Fbank aims to mimic the non-linear human ear perception of sound by being more discriminative at lower frequencies and less discriminative at higher frequencies. However, designing filters directly on linear frequency to mimic human ears is not an in trivial task. The most useful function of Fbank is the equispaced triangular filters designed on Mel-scale. Then, the filters are converted to linear frequency to form dense filters on low-frequencies while sparse filters on medium and high frequencies (Figure 2). We designed Rmel and Mmel according to this method. The formula of converting linear frequency (f) to Mel frequency m is as follows:

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

Then, the equispaced triangular filters are designed on Mel-scale and the m^{th} triangular filter is:

$$H_m(k) = \begin{cases} \frac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ 0, & \text{otherwise} \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1) \end{cases} \tag{2}$$

Generally, a group of M (usually 22–26) triangular filters is used. The m represents the m^{th} filter on Mel-scale, $0 \le m \le M$. $f(m-1), f(m), f(m+1)$ represent the lower, central and upper Mel frequencies respectively.

And then, the Mel-scale on Eq (2) is converted to linear frequency using Eq (3) to get filters $H_f(k), f = 1, \cdots, M$ on linear frequency.

$$f = 700 \times \left(10^{\frac{m}{2595}} - 1\right) \tag{3}$$

The logarithmic energy is calculated as follows:

$$E_m = log_{10}(\sum_{k=0}^{N-1} p(k) H_f(k)),\ 0 \le f \le M \tag{4}$$

where *p(k)* denotes to power spectrum of the sound signal. The obtained value E_m is the Fbank feature.

2.3.2. New voiceprint features (Rfbank, Mfbank features)

1) Rfbank feature

The high-frequencies could be extracted by a group of non equispaced triangular filters which were dense on high frequency and sparse on other frequencies (Figure 3). According to the discussion in subsection 2.3.1, the non equispaced triangular filters could be designed as equispaced on a scale similar to Mel, named reversed Mel-scale first. Then, these equispaced triangular filters were converted to linear frequency to form non equispaced filters.

Here, reversed Mel frequency m_R is:

$$m_R = 4292.2\text{-}2254 \times log_{10}(1 + \frac{8000 - f}{1400}) \tag{5}$$

where f is the linear frequency.

The triangular filter banks designed on reversed Mel-scale is equispaced and the m_R[th] triangular is:

$$H_{m_R}(k) = \begin{cases} \frac{k-f(m_R-1)}{f(m_R)-f(m_R-1)}, & f(m_R-1) \le k \le f(m_R) \\ 0, & \text{otherwise} \\ \frac{f(m_R+1)-k}{f(m_R+1)-f(m_R)}, & f(m_R) \le k \le f(m_R+1) \end{cases} \tag{6}$$

And then, covert the reversed Mel-scale on Eq (6) to linear frequency-scale using following equation to get filters $H_f(k), f = 1, \cdots, M$ on linear frequency.

$$f = 8000 - (10^{\frac{4292.2-m_R}{2254}} - 1) \times 1400 \tag{7}$$

The logarithmic energy calculated using Eq (4) is the Rfbank feature.

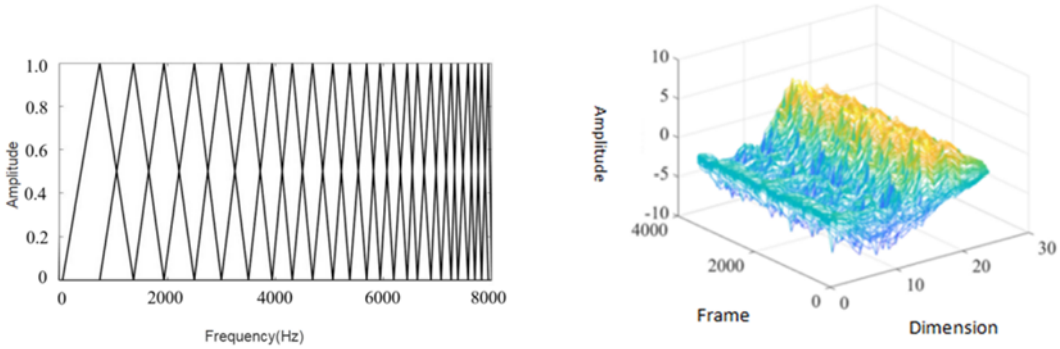

Figure 3. Converted Rmel filter banks on linear frequency (left), the 3D diagram RF of the bank feature (right). The filters are dense on high frequencies and sparse on other frequencies (left). The giant panda voiceprint, the Rfank feature, was extracted using the method in subsection 2.3.2 (right). The frame length was 256 sampling points, frame shift was 128 sampling points, the length of FFT was 256 and the order of the filter bank was 24.

2) Mfbank

According to the discussion in subsection 2.3.1, the medium-frequencies could be extracted by a group of non equispaced triangular filters which were dense on mediun frequency and sparse on other frequencies (Figure 4). The non equispaced triangular filters were designed as equispaced on a scale, named medium Mel-scale first. Then, these equispaced triangular filters will be converted to linear frequency (Hz) to form non equispaced filters.

Here, medium Mel frequency m_m is:

$$m_m = \begin{cases} 1073.05 - 527 \times log_{10}(1 + \frac{2000-f}{300}), & 0 \le f \le 2000 \\ 1073.05 - 527 \times log_{10}(1 + \frac{f-2000}{300}), & 2000 \le f \le 4000 \end{cases} \tag{8}$$

where f is the linear frequency.

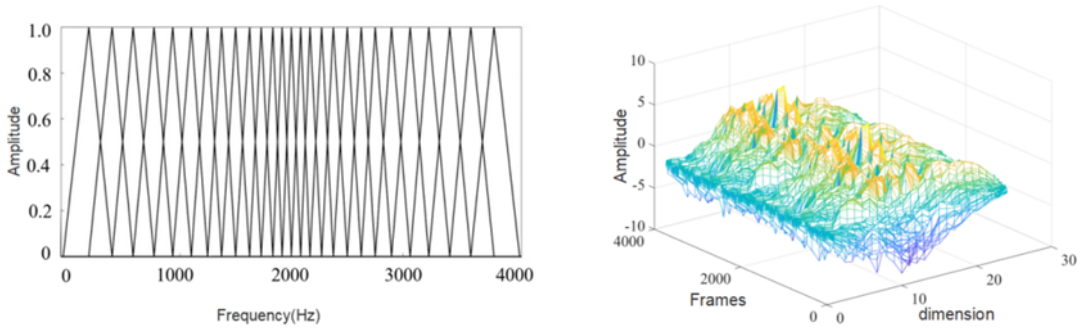

Figure 4. Converted Mmel filter banks on linear frequency (left), the 3D diagram of the Mfbank feature (right). The filters are dense on medium frequencies and sparse on other frequencies (left). The voiceprint feature was extracted using GP the method in subsection 2.3.2, called Mfank feature. The frame length was 256 sampling points, frame shift was 128 sampling points, the length of FFT was 256 and the order of the filter bank was 24.

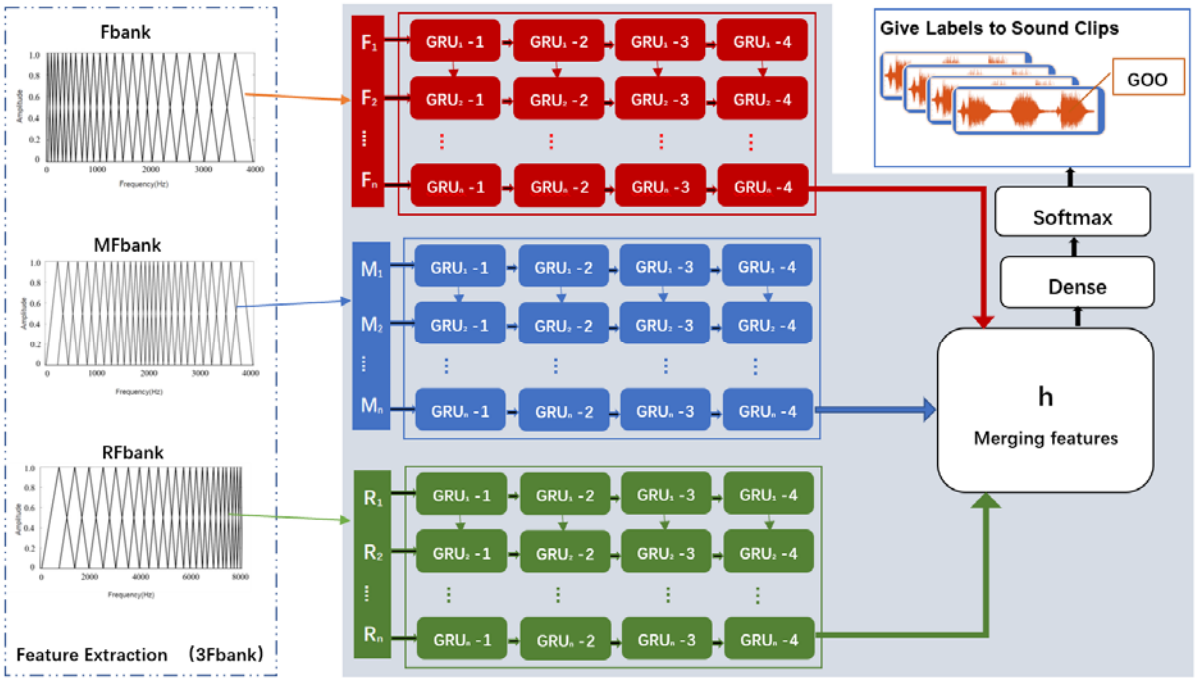

Figure 5. 3Fbank-GRU model structure diagram (structures in gray shades). In order to show the inputs and outputs of the model, inputs extracted by Fbank, Mfbank and Rfbank are shown on the left of the figure while outputs are "labels of GP's sound clips" laid on the top right corner of the figure. Three colors related to three kinds of features: Red is related to the Fbank feature and its 4-layer GRUs; Blue is related to the Mfbank feature and its 4-layer GRUs; Green is related to the Rfbank feature and its 4-layer GRUs. F_i, M_i and R_i, $i = 1...n$ were inputs of Fbank, Mfbank and Rfbank respectively. H received the outputs of the three GRUs and merged them here. Dense composed by the full connected layers. Softmax represented the Softmax layers.

The triangular filter banks designed on reversed Mel-scale is equispaced and the m_m th triangular is:

$$H_{m_m}(k) = \begin{cases} \frac{k - f(m_m - 1)}{f(m_m) - f(m_m - 1)}, & f(m_m - 1) \le k \le f(m_m) \\ 0, & \text{otherwise} \\ \frac{f(m_m + 1) - k}{f(m_m + 1) - f(m_m)}, & f(m_m) \le k \le f(m_m + 1) \end{cases} \tag{9}$$

And then, covert the medium Mel-scale on Eq (9) to linear frequency-scale using following Equation to get filters $H_f(k), f = 1, \cdots, M$ on linear frequency.

$$f = 2000 \pm [1 - 10^{\frac{1073.05 - m_m}{527}}] \times 300, \quad 606.86 \le m_m \le 1073.05 \tag{10}$$

The logarithmic energy calculated using Eq (4) is the Mfbank feature.

2.4. Proposed method: 3Fbank-GRU

A novel DNN model: 3Fbank-GRU model were proposed based on GRUs. That is, Fbank feature, Mfbank feature and Rfbank feature, whose lengths were *n*, were processed by three independent 4-layer GRUs with red, blue and green colors respectively (Figure 5). The final processed results of three 4-layer GRUs were fed into *h* and merged there. Then the merged features were sent to the full connected layers (Dense) and the Softmax layer to get vocalization labels of the sound clips. The labels were encoded by onehot encoder to define the differences of the loss in training.

3Fbank-GRU model was trained using datasets labeled by senior animal husbandry staffs of the CRBGPB. The loss was cross-entropy. During training, the initial value of hyperparameters was: dropout was 0.5, parameter initialization used uniform initialization, batch-size was set to 64 in minibatch training, initial learning rate was set to 0.0001 with variable learning rates and step attenuation (see subsection 3.3). We used 10-fold cross validation to train and test. That is, the samples were randomly divided into 10 equal parts in each category. For each unique group, took the group as a test data set and the remaining 9 groups as a training data set. Fit 3Fbank-GRU on the training set and evaluate it on the test set. The average value of 10 experiments was taken as the final result.

2.5. Method overview

The target of this research is: to provide an automatic recognition system of GP's sound clips to ecological researchers, which can help researchers to find more about the relationship between GP's sound and their behaviors. The recognition can be carried out automatically according to follow procedures (see Figure 6):

1) The collected sound data formed sound database. Clips in the sound database are preprocessing (denoising) firstly. And then, the features were extracted from the preprocessed signals by three banks: Fbank, Mfbank and Rfbank.

2) Given the extracted multi acoustic features: Fbank coefficients, Mfbank coefficients and Rfbank

coefficients, the features were fed into 3Fbank-GRU to learn more discriminative features.

3) The proposed model predicted the labels of GP's vocalizations by using fully connected layers (dense layers) with Softmax activation function based on the acoustic merged features extracted at the 3 independent 4-layer GRUs. Specifically, GRUs generated a probability vector $P \in R^{16\times1}$. The label was assigned as the category with the highest probability.

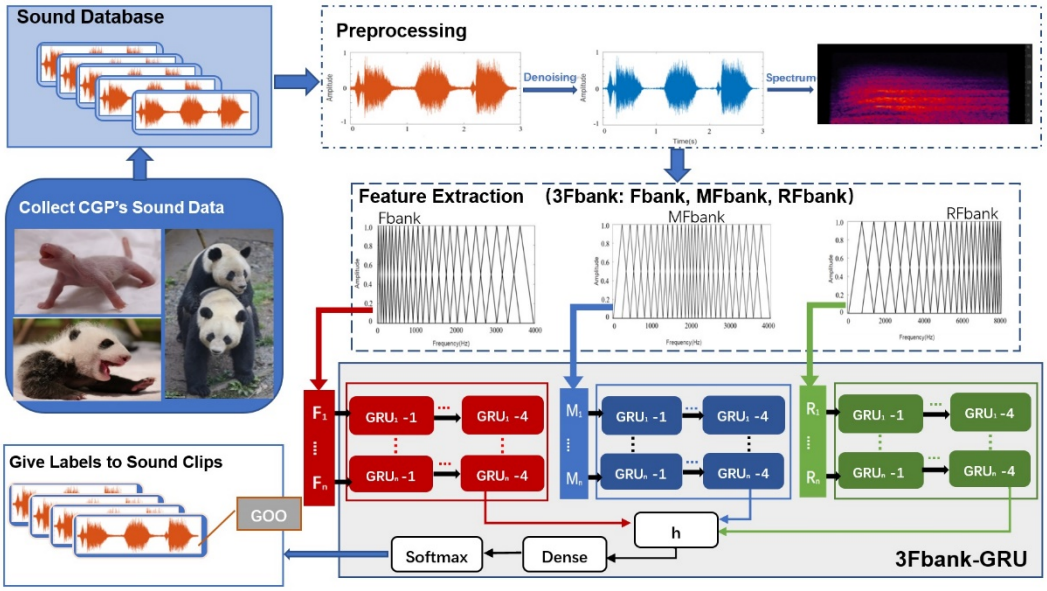

Figure 6. The diagram of GP's vocalization automatic recognition system.

3. Results

3.1. data

There are 12,800 labeled GP vocalization samples, which were divided into 16 categories, each with 800 samples. The test and training data set are chosen according to 10-fold cross validation, which is explained at the end of subsection 2.4.

3.2. Feature selection

MFCC added discrete cosine transform to Fbank filter. In this subsection, we used MFFC features and Fbank features to verify the influence of these two types of voiceprint features on recognition accuracy. The experimental data was set as subsection 3.1 and 2.5. The recognition model was a single-layer GRU network, the hidden layer dimension was 300 and the dropout was 0.5. The initialization was normal initialization, the learning rate was 0.0001, the training process used minibatch, batch size was 64 and the training was 100 epochs.

From Figure 7, the accuracy of the MFCC feature on the training set was 88.54%, the accuracy on the test set was 81.50% and the average accuracy of the 10-fold cross validation on the test set was 82.06%. While the accuracy of Fbank on the training set was 92.65%, the accuracy on the test set was 85.05% and the average accuracy of 10-fold cross validation on the test set was 85.74%.

The experimental results showed that the recognition accuracy of MFCC was 3.68% higher than that of Fbank. Thus, the MFCC feature had better recognition performance than the Fbank feature when using the GRU for giant panda vocalization recognition. Therefore, we used MFCC feature in our recognition framework.

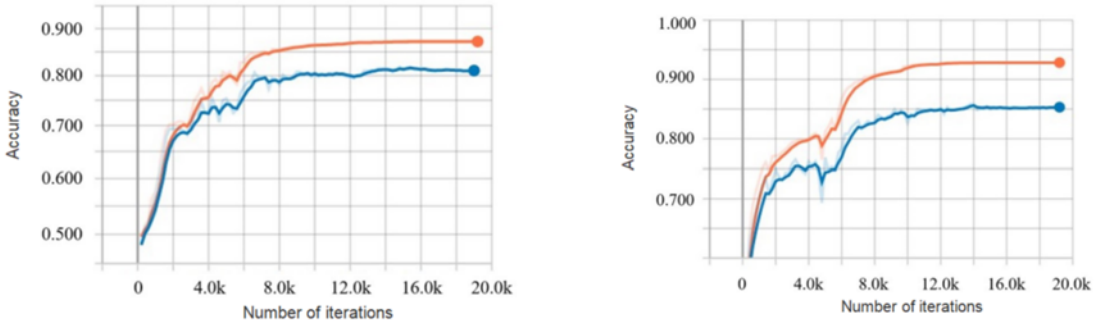

Figure 7. Left: the result of a single experiment of MFCC feature. Right: the result of a single experiment of Fbank feature. The MFCC feature was a 36-dimensional vector while the Fbank feaure was a 24-dimensional vector. The red curve was the accuracy curve of the training set and the blue curve was the accuracy curve of the test set.

3.3. Hyper-parameter selection

3.3.1. Network layer

In order to specify the layer of DNN composed by GRUs, we kept the model parameters except for the layers of DNN constant and observed the recognition accuracies of the different layers (see Figure 8). From Figure 8, we can conclude that the four layers had the best performance in five models and the network layer was set to 4.

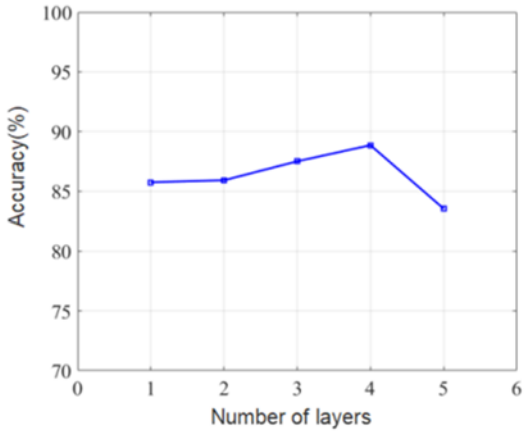

Figure 8. The Fbank feature was selected as the feature for different GRU layers. The number of model layers was taken from 1 to 5 and the rest parameters were kept unchanged. The results were from 10-fold cross validation.

3.3.2. Initialization

From above discussion, when the number of layers of DNN was 4, the recognition accuracy reached 88.85%. Here, we fixed the model parameters and investigated parameter. Three commonly used parameter initialization methods, normal, uniform and orthogonal, were used to conduct experiments, and the experimental results were shown in Table 3.

From Table 3, the uniform Initialization had the highest accuracy of 90.72%. Thus, the uniform was chosen as our initialization method.

Table 3. Experimental results of different initialization methods.

Parameter Initialization method	Normal	Uniform	Orthogonal
(%) of Recognition Accuracy	88.85	**90.72**	84.88

3.3.3. Batch size

Batch size was generally 2 to the power of n, such as 16, 32, 64, etc. Here, the model structure was kept unchanged and uniform initialization was used with batch sizes of 16, 32, 64, 128 respectively (see Table 4). From Table 4, the batch size 64 had the best the recognition accuracy and was selected as the batch size of proposed model.

Table 4. Experimental results of different batch size.

Batch size	16	32	64	128
Recognition Accuracy (%)	87.88	89.06	92.65	90.28

3.3.4. Dropout

The dropout values were 0.2, 0.5, 0.7 and "without dropout". The dropout 0.5 had the highest recognition accuracy and was the dropout of proposed model (see Table 5).

Table 5. Experimental results of different dropout.

Dropout	0.2	0.5	0.7	Not used
Recognition Accuracy (%)	91.33	92.65	90.47	86.92

3.3.5. Summary

In summary, the feature of the ablation experiment was the Fbank feature and hyper-parameters were: four layers of DNN, 64 batch size, 0.5 dropout and uniform initialization.

3.4. Model comparison

Proposed 3Fbank-GRU were compared with Gaussian mixture model-hidden Markov model (GMM-HMM) [12,13], Fbank-GRU, Fbank-LSTM [19], 3Fbank-LSTM and KD-CLDNN [21]. All

three models used 16 categories (800 samples each) of GP labeled vocal samples in experiments and used 10-fold cross validation to train and test the system.

3Fbank-GRU/3Fbank-LSTM (proposed method) extracted three kinds of features: Fbank, Rfbank, Mfbank and automatically gave labels of GP test sound clips (Figure 5). It had three independent GRUs whose parameters were not shared with each other. Each GRU's layer was four and its hidden layer dimension was 300.

Fbank-GRU/Fbank-LSTM's feature was its Fbank coefficients and the recognition model was a one-layer GRU/LSTM with 300 hidden layer dimension. The other hyperparameters of training initialization both of Fbank-GRU, Fbank-LSTM, 3Fbank-LSTM and 3Fbank-GRU were set as described in subsection 3.3.

We built an HMM for each category of samples. The number of states of an HMM was three, the number of Gaussian mixture elements for each state was three, the covariance matrix was the diagonal matrix and the number of iterations was 2000 times.

CS-CLDNN proposed in [21] was a CNN-LSTM-DNN model. The low-frequency features of bird voices were extracted by Mel FBank and sent to two CNN modules. In each module, the convolutional layer was activated by Swish function, followed by a convolutional block attention module (CBAM) and MaxPooling. Then, the low frequency features and features extracted by two CNN modules were fed to LSTM and full-connected DNN to classification. CS-CLDNN was considered as a state-of-art method in bird voice classification because of its high accuracy.

Table 6 showed average recognition accuracies of GP's 16 vocalization categories using 6 models. Not surprisingly, GMM-HMM had the lowest recognition accuracy 87.1%. The performance of GRU and LSTM were very similar both using features extracted by Fbank and 3Fbank. That is, the recognition accuracy of Fbank-GRU was 92.6% while Fbank-LSTM was 92.8%. The difference between two models was 0.2%. Moreover, the recognition accuracy of 3Fbank-GRU (96.9%) was 0.4% higher than the 3Fbank-LSTM (96.5%). Considering the training LSTM was more complex than the GRU, GRU was the more cost-effective model of the two.

Table 6. Average recognition accuracy of 16 types of GP vocalizations with different models. The bold number are the best result.

Model	GMM-HMM	Fbank-GRU	Fbank-LSTM	KD-CLDNN [21]	3Fbank-LSTM	Proposed Model 3Fbank-GRU
Average recognition accuracy	87.1%	92.6%	92.8%	93.9%	96.5%	**96.9%**

Although the average classification accuracy of 20 kinds of bird sounds using CS-CLDNN achieved 97.5%, the average recognition accuracy of GP's 16 vocalization categories was only 93.9%. However, CS-CLDNN was with the highest recognition accuracy in models whose features were extracted only by Fbank. The other two models, Fbank-GRU and Fbank-LSTM were with recognition accuracies 92.6% and 92.8% respectively. The recognition accuracy of CS-CLDNN was higher 1.3% than Fbank-GRU while 1.1% than Fbank-LSTM. Thus, CS-CLDNN was the best model in models using Fbank features.

When 3Fbank features were introduced, the recognition accuracies were improved by nearly 3% even compared to CS-CLDNN with the highest accuracy using Fbank features. That is, the accuracy of 3Fbank-LSTM was 2.6% higher than CS-CLDNN while 3Fbank-GRU was 3% higher than CS-CLDNN. It was a big improvement.

Just as above discussion, LSTM and GRU have very similar recognition performance and models with 3Fbank features are big improvements over models with Fbank features. Thus, in order to observe the performance of different kinds of models in each of GP's 16 vocalizations recognition, GMM-HMM, Fbank-GRU and 3Fbank-GRU are chosen to observe the recognition performance.

The vocalizations with high recognition accuracies using the Fbank-GRU model included the cub's "goo" and "chirping" and "gasp" of the adult GPs (see Table 7). The recognition accuracies of the above-mentioned three calls were higher than 96%. There were eight calls with low recognition accuracies: "squeak" of GPs' cub, "mm-hmm" of sub-adult GPs and "hum", "bird scream", "scream", "hiss" of adult GPs. Among them, the recognition accuracies of the six calls were lower than 91% while the classification of "bark" and "strong bark" was vague with more cases of miscommunication. The recognition accuracy rates for the remain six vocalizations were between 92% and 95% respectively.

Table 7. Recognition accuracy of 16 types of GP vocalizations with different models. The bold numbers are the best results.

Serial number	Age of Giant Pandas	Callout	Recognition accuracy (GMM-HMM)	Recognition accuracy (Fbank-GRU)	Recognition accuracy (proposed 3Fbank-GRU)
1	Giant Panda Baby	Squeak	85.35%	89.07%	**97.38%**
2	Giant Panda Baby	Goo	91.23%	96.50%	**98.74%**
3	Giant Panda Baby	Mm-Hmm	87.01%	90.62%	**96.56%**
4	Young Giant Panda	Haw	86.90%	92.77%	**96.88%**
5	Young Giant Panda	Hum	85.60%	90.54%	**96.62%**
6	Adult Giant Panda	Chirping	92.52%	97.06%	**98.27%**
7	Adult Giant Panda	Bleat	89.14%	94.21%	**97.95%**
8	Adult Giant Panda	Bird scream	82.61%	89.64%	**96.35%**
9	Adult Giant Panda	Bark	84.58%	91.93%	**95.29%**
10	Adult Giant Panda	Strong bark	85.17%	91.43%	**94.67%**
11	Adult Giant Panda	Ow	85.20%	94.14%	**97.21%**
12	Adult Giant Panda	Howl	87.67%	92.72%	**96.53%**
13	Adult Giant Panda	Roar	89.56%	94.34%	**96.73%**
14	Adult Giant Panda	Scream	84.78%	90.61%	**96.25%**
15	Adult Giant Panda	Hiss	84.51%	89.97%	**96.39%**
16	Adult Giant Panda	Gasp	92.27%	96.02%	**98.52%**

It is obviously that the recognition accuracies using Fbank-GRU for all categories were improved comparing with the accuracies using HMM-GMM. The average improvement rate was 5.47%. This implies that DNNs composed by GRUs are promising methods in GP's automatic vocalization recognition.

The recognition accuracies using 3Fbak-GRU were improved in all GP's vocalization compared with both GMM-HMM and Fbank-GRU. As discussed in the previous paragraph, the performance of Fbank-GRU was better than GMM-HMM. Thus, we will only compare the proposed method (3Fbank-GRU) with Fbank-GRU.

Six vocalizations with low recognition rate using Fbank-GRU: "squeak", "mm-hmm", "hum", "bird scream", "scream" and "hiss" whose recognition accuracies were about 90% are significantly

improved with more than 6% improvement rates and their recognition accuracies were all over 96%. The improvement of recognition accuracies for "barking" and "strong bark" were relatively small. There were still misclassifications, but recognition accuracies were increased more than 3%. Three calls: "goo", "chirping" and "gasp" which had high recognition accuracies also had a slight improvement and their recognition accuracies exceed 98%. The recognition accuracies of the other kinds of vocalizations were also improved.

In summary, all recognition accuracies using the proposed model were over 95% and the proposed model was best among the three recognition methods. The proposed DNN improved average accuracy by 4.3% compared with Fbank-GRU and by 9.77% compared with GMM-HMM.

4. Discussion

Both for the infant and adult GPs, vocalizations convey important information of breeding or needs to the mother. Considering captive GP's, their low success rate of natural mating and low birth rate of newborns, vocalization recognition is very important to the management of captive GPs.

4.1. Spectrogram characters of GP's voice

When Fbank features were used in speech recognition, the low frequency components of speech signals usually reflect the essential information of speech well. Nevertheless, for the GPs, the recognition accuracy of only using the Fbank feature and the GRUs was unsatisfied because of broad spectrum of animal calls.

Table 8. Voice frequencies of giant pandas.

Serial number	Age of Giant Pandas	Callout	Frequency range of concentrated energy (Hz)	Average frequency (Hz)
1	Giant Panda Baby	Squeak	260–10,200	6800
2	Giant Panda Baby	Goo	570–2200	1550
3	Giant Panda Baby	Mm-Hmm	875–7900	2950
4	Young Giant Panda	Haw	490–4100	1980
5	Young Giant Panda	Hum	910–5000	2830
6	Adult Giant Panda	Chirping	410–1900	1155
7	Adult Giant Panda	Bleat	280–4500	1890
8	Adult Giant Panda	Bird scream	410–6000	3160
9	Adult Giant Panda	Bark	270–3220	1745
10	Adult Giant Panda	Strong bark	220–3420	1820
11	Adult Giant Panda	Ow	150–2940	1545
12	Adult Giant Panda	Howl	180–3530	1855
13	Adult Giant Panda	Roar	310–4260	1985
14	Adult Giant Panda	Scream	560–6380	3370
15	Adult Giant Panda	Hiss	600–8450	3850
16	Adult Giant Panda	Gasp	350–1600	860

From Table 8, we can see that the frequency distribution of GP's vocalizations is wide. Baby panda "cuckoo" sound, adult pandas "chirping" and "gasp" are the lowest frequency over all calls and their frequencies of concentrated energy are below 2000 Hz. The baby giant panda's "goo", the young giant panda's "mm", the adult giant panda's "hum", "bird scream", "scream" and "hiss" have high frequency, whose average frequency is above 2000 Hz. The frequency of "squeak" is the highest, and can reach about 10.2 kHz. The average frequency of the vocalizations of the rest types is between 1500 and 2000 Hz.

In order to observe the change trends between the recognition accuracy of Fbank-GRU and the average frequency of GP's vocalizations, the GP's vocalizations were coded to their serial numbers defined in Table 1. Moreover, the recognition accuracy of Fbank-GRU and 3Fbank-GRU was magnified by 40,000 times and subtracted by 35,000 to show it with the same order of the average accuracy's values (see Figure 9).

Figure 9. The average frequency of GP's vocalizations (the blue curve), the magnification and translation versions of recognition accuracy using Fbank-GRU (the red curve) and using 3Fbank-GRU (the orange curve). In order to observe the change trends of recognition accuracy using Fbank-GRU and 3Fbank-GRU and the relation between each of model and the average frequency of GP's vocalizations, the recognition accuracy of two models was magnified by 40,000 times and subtracted by 35,000 to let them be of the same order of magnitude as the average frequency. The x axis shows the serial number of GP's vocalizations defined in Table 1.

Observing Figure 9, we can conclude:

1) The recognition accuracy of Fbank-GRU (the red curve) and the average frequency (the blue curve) had opposite trends. That is, the minimum of the recognition accuracy curve was the maximum of the average frequency curve, which mean the GP's vocalizations with high average frequency were with low recognition accuracy while the low average frequency were with high recognition accuracy. In other words, Fbank-GRU had good performance only in low frequency signal.

2) The orange curve of 3Fbank-GRU floated above the red curve of Fbank-GRU, which mean the recognition accuracy of 3Fbank-GRU was higher than the Fbank-GRU. In addition, the orange curve was smoother than the red curve. Thus, the 3Fbank-GRU can handle wide average frequency signals better than the Fbank-GRU.

As mentioned above, the GP's vocalization frequency is very wide. Therefore, extracting high frequency and medium frequency features can help us increase the recognition accuracy. Two new features, Mfbank and Rfbank, were proposed to improve the performance of Fbank features and the GRUs. In order to handle three features, three GRUs with not shared parameters were designed in our framework. The output of the three GRUs were spliced into one output finally (Figure 6). Experimental results were obviously improved compared with the Fbank-GRU model and the GMM-HMM. All accuracies of 16 vocalizations using the proposed model were over 95%, which mean that the system based on the proposed model can be used directly for the automatic recognition of GP's vocalizations.

4.2. Robustness in noise

There are environmental and device noises in collected sound dataset. Fortunately, the highest energy of these noises concentrated on the low frequency. Thus, the most of noises can be suppressed by traditional high-passed filters. The MMSE was used to reduced noises (see subsection 2.2 and Table 2).

4.3. Weakness

However, there are some weaknesses that need further study.

1) There are 16 categories of sample data in our paper, some kinds of calls are not included because of the small amount of data, such as "single call" and "low ow".

2) Although the recognition results of 16 kinds of panda calls are satisfactory, but the differences between each type of calls are not analyzed.

3) The number of layers, Initialization and other parameters of the identification model have been discussed in this paper. However, the limited number of panda calls samples cannot fully reflect the statistical characteristics of all kinds of parameters. More data are needed to optimize the model.

4.4. Future works

1) Collect more types and numbers of giant panda calls to train and design new DNN model.

2) Analyze the differences between the frequency features of different kinds of panda calls and study the voiceprint features more suitable for panda voice signals.

3) Focus on both automatically recognize, locate and segment the PD's vocalizations.

5. Conclusions

This paper is mainly based on using the voiceprint features and deep learning to identify GP vocalizations. Two new voiceprint features, Mfbank features and Rfbank features, were extracted by designing two new filters. These two types of voiceprint features were combined with the

Fbank voiceprint features to improve GP vocalization recognition. We designed a DNN model named 3Fbank-GRU to realize the automatic recognition of GP vocalizations based on the features of Fbank, Mfbank and Rfbank. Finally, our experiments showed that the recognition accuracies of 16 vocalizations were over 95% and improved average accuracy by 4.3% compared with the Fbank-GRU and by 9.77% compared with GMM-HMM. As the first effort to design a system for the automatic recognition of GP vocalizations, the proposed model will help researchers to better understand the role that vocalizations play in giant panda behavior. Moreover, since the system based on 3Fbank-GRU can label sound clips automatically, it will greatly reduce the work burden of manually analyzing vocalizations.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research is supported by the Program of Natural Science Foundation of Sichuan Province (2022NSFSC0020), National and Regional Studies Archival Center for Higher Education, Ministry of Education, Japan and South Korea Institute, Sichuan Normal University (2022RHZD003), Chengdu Science and Technology Program (2022-YF09-00019-SN) and the Program of Chengdu Research Base of Giant Panda Breeding (NO. 2020CPB-C09, NO. 2021CPB-B06).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. G. Peters, A note on the vocal behaviour of the giant panda, Ailuropoda melanoleuca (David, 1869), *Z. Saeugetierkd.*, **47** (1982), 236–246.
2. D. G. Kleiman, Ethology and reproduction of captive giant pandas (Ailuropoda melanoleuca), *Z. Tierpsychol.*, **62** (1983), 1–46.
3. G. B. Schaller, J. Hu, W. Pan, J. Zhu, *The Giant Pandas of Wolong*, University of Chicago Press in Chicago, 1985.
4. B. Charlton, Z. H. Zhang, R. Snyder, The information content of giant panda, Ailuropoda melanoleuca, bleats: acoustic cues to sex, age and size, *Anim. Behav.*, **78** (2009), 893–898. https://doi.org/10.1016/j.anbehav.2009.06.029
5. B. Charlton, Y. Huang, R. Swaisgood, Vocal discrimination of potential mates by female giant pandas (Ailuropoda melanoleuca), *Biol. Lett.*, **5** (2009), 597–599. https://doi.org/10.1098/rsbl.2009.0331
6. M. Xu, Z. P. Wang, D. Z. Liu, Cross-modal signaling in giant pandas, *Chin. Sci. Bull.*, **57** (2012), 344–348. https://doi.org/10.1007/s11434-011-4843-y

7. A. S. Stoeger, A. Baotic, D. Li, B. D. Charlton, Acoustic features indicate arousal in infant giant panda vocalisations, *Ethology*, **118** (2012), 896–905. https://doi.org/10.1111/j.1439-0310.2012.02080.x
8. B. Anton, A. S. Stoeger, D. S. Li, C. X. Tang, B. D. Charlton, The vocal repertoire of infant giant pandas (Ailuropoda melanoleuca), *Bioacoustics*, **23** (2014), 15–28, http://doi.org/10.1080/09524622.2013.798744
9. B. D. Charlton, M. S. Martin-Wintle, M. A. Owen, H. Zhang, R. R. Swaisgood, Vocal behaviour predicts mating success in giant pandas, *R. Soc. Open Sci.*, **10** (2018), 181323. https://doi.org/10.1098/rsos.181323
10. B. D. Charlton, M. A. Owen, X. Zhou, H. Zhang, R. R. Swaisgood, Influence of season and social context on male giant panda (Ailuropoda melanoleuca) vocal behaviour, *PloS One*, **14** (2019), e0225772. https://doi.org/10.1371/journal.pone.0225772
11. K. F. Lee, H. W. Hon, R. Reddy, An overview of the SPHINX speech recognition system, *IEEE Trans. Acoust. Speech Signal Process.*, **38** (1990), 35–45. http://doi.org/10.1109/29.45616
12. L. R. Bahl, P. F. Brown, P. V. D. Souza, R. L. Mercer, Maximum mutual information estimation of hidden Markov model parameters for speech recognition, in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, **11** (1986), 49–52. http://doi.org/10.1109/ICASSP.1986.1169179
13. D. A. Reynolds, R. C. Rose, Robust text-independent identification using Gaussian mixture speaker models, *IEEE Trans. Speech Audio Process.*, **3** (1995), 72–83. http://doi.org/10.1109/89.365379
14. W. B. Cavnar, J. M. Trenkle, N-gram-based text categorization, in *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, (1994), 14. http://doi.org/161175.10.1.1.21.3248&rep=rep1&type=pdf
15. J. Colonna, T. Peet, C. A. Ferreira, A. M. Jorge, E. F. Gomes, J. Gama, Automatic classification of anuran sounds using convolutional neural networks, in *Proceedings of the Ninth International c* Conference on Computer Science & Software Engineering*, ACM, (2016), 73–78. http://doi.org/10.1145/2948992.2949016
16. H. Goëau, H. Glotin, W. P. Vellinga, R. Planqué, A. Joly, LifeCLEF bird identification task 2016: the arrival of deep learning, in *CLEF: Conference and Labs of the Evaluation Forum*, Évora, Portugal, (2016), 440–449.
17. D. Stowell, Computational bioacoustics with deep learning: a review and roadmap, *PeerJ*, **10** (2021), e13152. http://doi.org/10.7717/peerj.13152
18. A. Graves, A. R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, (2013), 6645–6649. http://doi.org/10.1109/ICASSP.2013.6638947
19. F. A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: Continual prediction with LSTM, *Neural Comput.*, **12** (2000), 2451–2471. http://doi.org/10.1049/cp:19991218
20. F. A. Gers, N. N. Schraudolph, J. Schmidhuber, Learning precise timing with LSTM recurrent networks, *J. Mach. Learn. Res.*, **3** (2002), 115–143. http://doi.org/10.1162/153244303768966139
21. J. Xie, S. Zhao, X. Li, D. Ni, J. Zhang, KD-CLDNN: Lightweight automatic recognition model based on bird vocalization, *Appl. Acoust.*, **188** (2022), 108550. http://doi.org/10.1016/j.apacoust.2021.108550

22. C. Bergler, M. Schmitt, R. X. Cheng, H. Schröter, A. Maier, V. Barth, et al., Deep representation learning for orca call type classification, in *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22*, Springer, **11697** (2019), 274–286. http://doi.org/10.1007/978-3-030-27947-9_23
23. E. E. Waddell, J. H. Rasmussen, A. Širović, Applying artificial intelligence methods to detect and classify fish calls from the northern gulf of Mexico, *J. Mar. Sci. Eng.*, **9** (2021), 1128. http://doi.org/10.3390/jmse9101128.
24. J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, preprint, arXiv:1412.3555.
25. W. Yan, M. Tang, Z. Chen, P. Chen, Q. Zhao, P. Que, et al., Automatically predicting giant panda mating success based on acoustic features, *Global Ecol. Conserv.*, **24** (2020), e01301. https://doi.org/10.1016/j.gecco.2020.e01301

AIMS Press