Multi-label classification of frog species via deep learning

Jie Xie
University of Waterloo, Canada
Email: j65xie@uwaterloo.ca

Abstract

Acoustic classification of frogs has received increasing attention for its promising application in ecological studies. Various studies have been proposed for classifying frog species, but most recordings are assumed to have only a single species. In this study, a method to classify multiple frog species in an audio clip is presented. To be specific, continuous frog recordings are first cropped into audio clips (10 seconds). Then, various time-frequency representations are generated for each 10-s recording. Next, instead of using traditional hand-crafted features, a deep learning algorithm is used to find the most important feature. Finally, a binary relevance based multi-label classification approach is proposed to classify simultaneously vocalizing frog species with our proposed features. Experimental results show that our proposed features extracted using deep learning can achieve better classification performance when compared to hand-crafted features for frog call classification.

I. Introduction

As a widely distributed amphibian species, frogs are an integral part of the food web, and are often regarded as a valuable indicator species for environmental health [1]. Although frogs are very important, rapid decline in frog populations has been spotted worldwide. Reasons for this decline can be summarized as habitat loss, invasive species, climate change. To monitor the change of frog population and optimize its protection policy, it is becoming ever more important to gain insights about frogs and the environment. Compared to traditional methods that require ecologists to enter the fields frequently for biodiversity data collection, an acoustic sensor provides a way to collect data over larger spatial and temporal scales [2]. Since large volumes of acoustic data can be generated by an acoustic sensor, enabling automatic methods to study collected data is in high demand.

Many previous studies have developed different methods for classifying frog species by acoustic data [3], [4], [5], [6], [7], [8]. In those studies, various feature vectors and classifiers have been explored for frog call classification. Linear predictive coding (LPC) [9] and Mel-frequency cepstral coefficients (MFCCs) [10], [3] are two well-known features for classifying frog calls. Since LPC and MFCCs describe individual frame within one syllable, all frame-level features of the syllable need to be averaged to characterize the syllable as a whole. Besides LPC and MFCCs, many other acoustic features have been explored for frog call classification, including syllable duration, averaged energy, zero-crossing rate, oscillation rate, Shannon entropy, spectral centroid, spectral flatness, spectral flux, fundamental frequency [4], [5], [6], [7], [8]. Two classifiers, k-nearest neighbor (k-NN) and support vector machine (SVM), are most widely used models for their easy implementation and high accuracy [11], [7], [12]. However, recordings used in those previous studies often have a high signal-to-noise ratio (SNR), and each recording is assumed to include a single species.

In contrast, recordings used in this study have a low SNR and contain many overlapping animal vocal activities, including frogs, birds, crickets. To address those

challenges, multi-label learning is introduced to classify simultaneously vocalizing frog species in low SNR recordings. Various methods have been proposed to classify simultaneously vocalizing birds [13] and frogs [14]. However, hand-crafted features are used in those studies, which are highly affected by the segmentation process. Compared to hand-crafted features, recent use of deep learnings has achieved state-of-the-art accuracy in frog call classification [15], [16], but all recordings used are assumed to have a single species.

In this study, we use a deep learning algorithm to extract features. After splitting continuous recordings into 10-s audio clips, we translate each recording into its time-frequency representation. Then, acoustic features are directly extracted from the time-frequency representation with a pre-trained network. Different from hand-crafted features, we do not need to segment recordings into individual events as previous studies [13], [14], which can increase the robustness of our classification model. To classify simultaneously vocalizing frog species, a binary relevance based multi-label classification approach is used. Eight frog species, which are widely distributed in Queensland, Australia, are selected for the experiment.

The rest of this paper is organized as follows: In section II, we describe the method for frog call classification, which includes data description, feature extraction, and classification. Section III reports experimental results. Section IV presents conclusion and future work.

II. Methods

Our frog call classification method consists of four steps: data description, signal pre-processing, feature extraction, and classification. Detailed information of each step is shown in following sections.

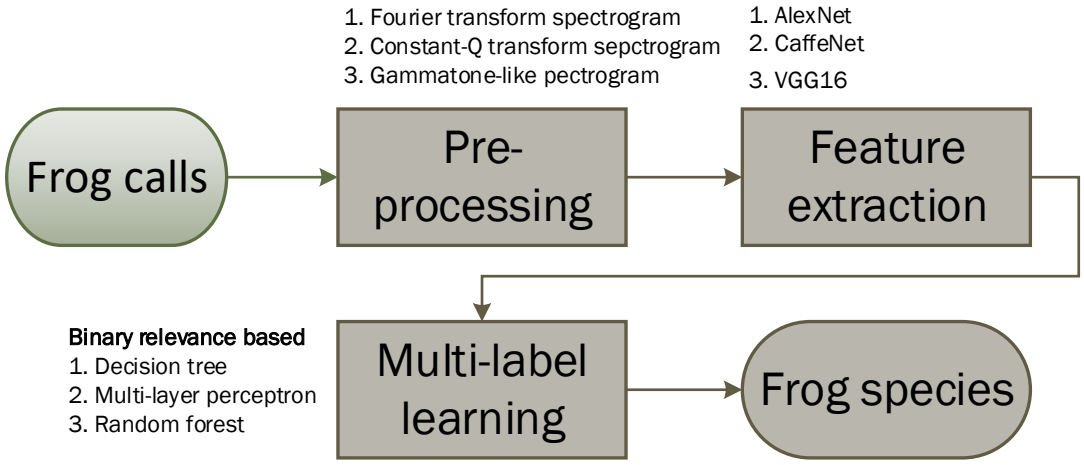

Fig. 1. Flowchart of our frog call classification system using pre-trained deep networks and multi-label learning

A. Data description

Digital recordings in this study were obtained with a battery-powered, weatherproof Song Meter (SM2) box[1]. Recordings were two-channel, sampled at 22.05 kHz and

[1]https://www.wildlifeacoustics.com

saved in WAC4 format. A representative sample of 342 10-s recordings was selected to train and evaluate our proposed algorithm for classifying simultaneously vocalizing frog species in a recording. All those examples were collected between 02/2014 to 03/2014, since it is breeding season for frogs with high calling activity. All the species that are presented in each 10-s recording were manually labeled by an ecologist who is a frog expert. There are totally eight frog species in the recordings: *Rhinella marina* (RMA), *Cyclorana novaehollandiae* (CNE), *Limnodynastes terraereginae* (LTE), *Litoria fallax* (LFX), *Litoria nasuta* (LNA), *Litoria rothii* (LRI), *Litoria rubella* (LRA), and *Uperolela mimula* (UMA). Each recording contains between one and five species.

B. Signal pre-processing

All the recordings were re-sampled at 16 kHz and mixed to mono. Since features are directly calculated by applying deep learning techniques to recordings' time-frequency representations. Three time-frequency representations are tested in this study: fast-Fourier transform spectrogram, constant-Q transform spectrogram, and Gammatone-like spectrogram.

Fast-Fourier transform (FFT) spectrogram is generated by applying short-time Fourier transform (STFT) to each recording. Specifically, each recording was divided into frames of 32 ms with 50 % frame overlap. A fast Fourier transform was then performed on each frame with a Hamming window, which yielded amplitude values for 256 frequency bins, each spanning 31.25 Hz. The final decibel values (S) were generated using

$$S_{tf} = 20 * log_{10} A_{tf} \tag{1}$$

where $t = 0, ..., T-1$, $f = 0, ..., F-1$, t and f represent frequency bin and time index, T and F are 256 frequency bins and 625 frames, A is the amplitude value.

Constant-Q transform spectrogram is generated by applying constant-Q transform to the signal. Compared to STFT, this transform provides a frequency analysis on a log-scale which makes it more adapted to sound with harmonic structures. Here we use 48 filters per octave with lowest and high frequency as 50 Hz and 8000 Hz.

Gammatone-like spectrogram is constructed by first calculating a conventional, fixed-bandwidth spectrogram, then combining the fine frequency resolution of the FFT-based spectra into the coarser, smoother Gammatone responses via a weighting function. Here, each recording was passed through a 64 channel gammatone auditory model filterbank, with lowest and highest frequency as 50 Hz and 8000 Hz. The outputs of each band have their energy integrated over windows of 25 ms with 60% overlap.

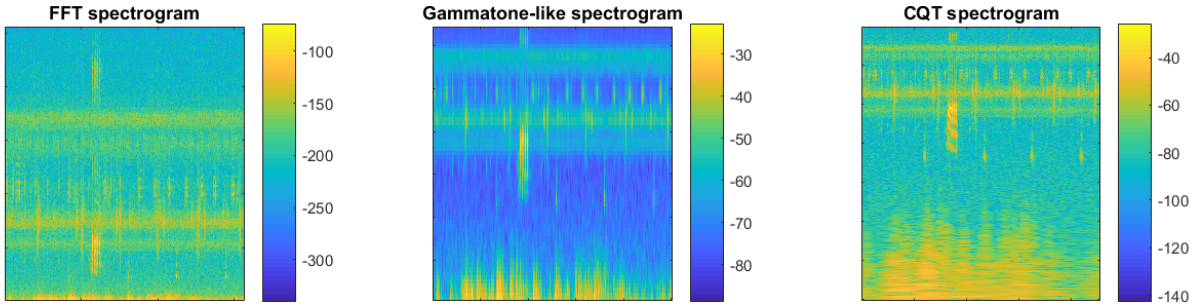

Fig. 2. Spectrogram comparison using three time-frequency representations

C. Feature extraction

Different from [15], we use a deep learning algorithm as a feature extractor. In [16], a pre-trained network is found to achieve higher classification accuracy than training a new network. Also, there are only 342 10-s recordings for the experiment, which are not enough for training. Here, we directly use a pre-trained network to extract features by removing the vectors from one of the final fully connected layers. Multiple different nets, which are trained on ImageNet with different architecture, are used: AlexNet, CaffeNet, and VGG16.

AlexNet [17] was trained on the 1.3 million images in the LSVRC-2010 ImageNet training set and consists of five convolutional layers, two fully connected layers, and a final soft max layer.

CaffeNet [18] has a similar architecture as AlexNet. The difference is that CaffeNet was trained with data that was augmented differently and the pooling and normalization layers were switched.

VGG16 [19] was trained on a subset of the ImageNet database, which was used in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC).

D. Classification

In this study, a binary relevance (BR) based multi-label learning is used for its scalability and flexibility [20]. The principle of the BR method is to solve a multi-label classification problem using multiple binary classifiers. Similar to our previous work [21], three classic single-label learning algorithms are used in this study: decision tree (DT), and k-nearest neighbour (k-NN), and random forest (RF). For each classifier, a grid search is conducted to optimize classification results. To evaluate the multi-label classification model, a tagged representative sample of 342 10-s recordings is used with 5-fold cross-validation.

E. Evaluation metrics

Three evaluation metrics are used: hamming loss, accuracy, and subset accuracy [22]. Hamming loss is defined as the fraction of labels that are incorrectly predicted for an instance and the normalized hamming loss which is normalized over instances is reported. To better interpret these results, a baseline for hamming loss is obtained by considering a non-informative classifier that cannot predict any accuracy labels [13]. Accuracy for a single instance x_i is defined by the Jaccard similarity coefficients between the ground truth y_i and the prediction $h(x_i)$. Subset accuracy is defined as follows:

$$subsetAccuracy = \frac{1}{N}\sum_{i=1}^{N} I(h(x_i) = y_i) \tag{2}$$

where $I(true) = 1$ and $I(false) = 0$. This is a very strict evaluation measure as it requires the predicted set of labels to be an exact match of the true set of labels.

Values for hamming loss, accuracy, and subset accuracy range from zero to one. For hamming loss, zero denotes the perfect result, and one means the wrong prediction of all labels over every instance, whereas for accuracy and subset accuracy, the values have complete opposite meanings.

III. Experimental results and Discussions

In this experiment, three time-frequency representations are first compared using AlexNet. Here, features are extracted from multiple fully connected layers, including layer *fc6*, *fc7*, and *fc8*. As for the classification, a BR based multiple learning is used with three classic single-label learning algorithms as binary classifiers: DT, k-NN, and RF.

A. Comparison of three time-frequency representations using three layers and DT

Classification results using AlexNet are shown in Table I. Compared to FFT spectrogram and CQT spectrogram, gamma spectrogram using layer fc7 achieves the best Hamming loss and subset accuracy. According to Fig. 2, gamma spectrogram is the the time-frequency representation with a highest resolution, which is in accordance to the classification result. Therefore, gamma spectrogram is selected for the subsequent analysis.

TABLE I
CLASSIFICATION RESULTS USING ALEXNET. THE BEST VALUE OF EACH EVALUATION METRIC IS IN BOLD. HERE ↓ MEANS THAT A HIGHER VALUE IMPLIES A BETTER PERFORMANCE, BUT ↑ HAS A COMPLETELY OPPOSITE MEANING.

Layer	TF representation	Hamming loss ↓	Accuracy ↑	Subset accuracy ↑
fc6	FFT spectrogram	0.150	0.589	**0.310**
fc6	CQT spectrogram	0.151	0.585	0.301
fc6	Gamma spectrogram	0.147	0.588	**0.310**
fc7	FFT spectrogram	0.156	0.576	0.295
fc7	CQT spectrogram	0.160	0.571	0.289
fc7	Gamma spectrogram	**0.138**	0.591	**0.310**
fc8	FFT spectrogram	0.149	**0.598**	0.301
fc8	CQT spectrogram	0.152	0.584	0.289
fc8	Gamma spectrogram	0.145	0.584	0.284

B. Comparison of three nets using three basic single-label learning algorithms

In this part, gammatone-like spectrogram with fc7 layer is used for various nets and three classifiers due to its best performance in Table I. Table II shows that AlexNet with RF and VGG16 with k-NN are two best classification methods. Among three classifiers, classification performance of DT is the worst. Compared to AlexNet and VGG16, CaffeNet is the worst, which is in consistent with [16].

TABLE II
CLASSIFICATION RESULTS USING THREE NETS AND THREE CLASSIC CLASSIFIERS.

Net	Classifier	Hamming loss ↓	Accuracy ↑	Subset accuracy ↑
AlexNet	DT	0.138	0.591	0.310
AlexNet	k-NN	0.117	0.691	0.444
AlexNet	RF	**0.097**	0.692	0.477
CaffeNet	DT	0.160	0.572	0.298
CaffeNet	k-NN	0.133	0.651	0.418
CaffeNet	RF	0.111	0.647	0.412
VGG16	DT	0.151	0.597	0.319
VGG16	k-NN	0.109	**0.711**	**0.482**
VGG16	RF	0.099	0.679	0.462

C. Comparison with hand-crafted features and baseline

Table III shows the comparison between hand-crafted features and deep learning based features. In our previous studies, the best classification results for Hamming loss are 0.131 and 0.182, which are obtained using multi-label learning and multiple-instance multiple-label learning. The hand-crafted features used for multi-label learning are wavelet-based cesptral coefficients and linear predictive coefficients. However, this kind of global features will cause the information loss in time domain. Features extracted from segmented frog syllable using acoustic event detection are highly affected by the segmentation results. The segmentation process is sensitive to the background noise, and the classification results are not robust.

TABLE III
COMPARISON WITH HAND-CRAFTED FEATURES AND BASELINE

Methods	Hamming loss ↓	Accuracy ↑	Subset accuracy ↑
AlexNet + RF	**0.097**	0.692	0.477
VGG16 + k-NN	0.109	**0.711**	**0.482**
[23]	0.131	—	—
[14]	0.182	—	—
Baseline	0.249	—	—

IV. CONCLUSIONS

This study presents a novel feature extraction method using a deep learning algorithm for classifying simultaneously vocalizing frog calls. Continuous recordings are first segmented into 10-s audio clips. Then, three types of time-frequency representations are used for extracting features with three pre-trained nets. Finally, a binary relevance based multiple-label classification algorithm is used to classify frog species with three single-label learning algorithms: DT, k-NN, and RF. Experimental results on 342 recordings of eight frog species are promising with hamming loss, accuracy and subset accuracy as 0.109, 0.711, and 0.482, respectively. Compared to hand-crafted features, features extracted using deep learning can achieve a better classification performance. Future work will include additional experiments that test a wider variety of audio data from different geographical and environment conditions.

ACKNOWLEDGMENT

The authors would like to thank the Eco-acoustic group of Queensland University of Technology and James Cool University for providing the data. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of QUT and JCU.

REFERENCES

[1] S. Böll, B. Schmidt, M. Veith, N. Wagner, D. Rödder, C. Weinmann, T. Kirschey, and S. Loetters, "Amphibians as indicators of changes in aquatic and terrestrial ecosystems following gm crop cultivation: a monitoring guideline," *BioRisk*, vol. 8, p. 39, 2013.

[2] J. Wimmer, M. Towsey, B. Planitz, I. Williamson, and P. Roe, "Analysing environmental acoustic data through collaboration and automation," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 560–568, February 2013.

[3] C.-H. Lee, C.-H. Chou, C.-C. Han, and R.-Z. Huang, "Automatic recognition of animal vocalizations using averaged mfcc and linear discriminant analysis," *Pattern Recognition Letters*, vol. 27, no. 2, pp. 93–101, 2006.

[4] C.-J. Huang, Y.-J. Yang, D.-X. Yang, and Y.-J. Chen, "Frog classification using machine learning techniques," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3737–3743, 2009.

[5] N. C. Han, S. V. Muniandy, and J. Dayou, "Acoustic classification of australian anurans based on hybrid spectral-entropy approach," *Applied Acoustics*, vol. 72, no. 9, pp. 639–645, 2011.

[6] W.-P. Chen, S.-S. Chen, C.-C. Lin, Y.-Z. Chen, and W.-C. Lin, "Automatic recognition of frog calls using a multi-stage average spectrum," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1270–1281, 2012.

[7] J. Xie, M. Towsey, J. Zhang, and P. Roe, "Acoustic classification of australian frogs based on enhanced features and machine learning algorithms," *Applied Acoustics*, vol. 113, pp. 193–201, 2016.

[8] J. B. Alonso, J. Cabrera, R. Shyamnani, C. M. Travieso, F. Bolaños, A. García, A. Villegas, and M. Wainwright, "Automatic anuran identification using noise removal and audio activity detection," *Expert Systems with Applications*, vol. 72, pp. 83–92, 2017.

[9] C. L. T. Yuan and D. A. Ramli, "Frog sound identification system for frog species recognition," in *International Conference on Context-Aware Systems and Applications*. Springer, 2012, pp. 41–50.

[10] C. Bedoya, C. Isaza, J. M. Daza, and J. D. López, "Automatic recognition of anuran species based on syllable identification," *Ecological Informatics*, vol. 24, pp. 200–209, 2014.

[11] C.-J. Huang, Y.-J. Chen, H.-M. Chen, J.-J. Jian, S.-C. Tseng, Y.-J. Yang, and P.-A. Hsu, "Intelligent feature extraction and classification of anuran vocalizations," *Applied Soft Computing*, vol. 19, no. 0, pp. 1 – 7, 2014.

[12] J. Xie, M. Towsey, J. Zhang, and P. Roe, "Adaptive frequency scaled wavelet packet decomposition for frog call classification," *Ecological Informatics*, vol. 32, pp. 134–144, 2016.

[13] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.

[14] J. Xie, M. Towsey, L. Zhang, K. Yasumiba, L. Schwarzkopf, J. Zhang, and P. Roe, "Multiple-instance multiple-label learning for the classification of frog calls with acoustic event detection," in *International Conference on Image and Signal Processing*. Springer, 2016, pp. 222–230.

[15] J. Colonna, T. Peet, C. A. Ferreira, A. M. Jorge, E. F. Gomes, and 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, pp. 73–78.

[16] S. M. M. S. M. B. E. R. Julia Strout, Bryce Rogan, "Anuran call classification with deep learning," *In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, USA,*, March 5-9, 2017.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[20] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[21] L. Zhang, M. Towsey, J. Xie, J. Zhang, and P. Roe, "Using multi-label classification for acoustic pattern detection and assisting bird species surveys," *Applied Acoustics*, vol. 110, pp. 91–98, 2016.

[22] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012.

[23] J. Xie, T. Michael, J. Zhang, and P. Roe, "Detecting frog calling activity based on acoustic event detection and multi-label learning," *Procedia Computer Science*, vol. 80, pp. 627–638, 2016.