2017 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017

Spectrogram Segmentation for Bird Species Classification based on Temporal Continuity

Md. Shamim Towhid
Department of Computer Science and Engineering
Uttara University
Dhaka, Bangladesh
shamim.towhid@gmail.com

Md. Mijanur Rahman
Department of Computer Science and Engineering
Uttara University
Dhaka, Bangladesh
mdmijanur1970@gmail.com

Abstract – **This article presents an enhanced approach for bird species classification from their recorded audio signals. Observing that textures of syllables in audio spectrograms have noticeable discerning capabilities among different bird species, we adopt these texture features for bird species classification. First, we compute spectrogram from recoded audio. We propose an enhanced syllable extraction technique to identify the syllables in the spectrogram. Texture features, based on gray level co-occurrence matrix (GLCM), are computed and used for classification using an ensemble learning method. We obtain satisfactory accuracy when the approach is tested on real audio recordings of 11 different bird species.**

Keywords—Spectrogram Segmentation; GLCM Texture Feature; Pattern Recognition; Ensemble Classifier

I. INTRODUCTION

Many of the ecologically critical bird species are on the verge of extinction. In order to preserve these valuable species, it is imperative to identify and geographically locate them. As their natural habitats are so much diversified and expanded over heterogeneous geography, it is extremely difficult to manually perform these tasks. This, in turn, demands for intelligent automatic systems which can adaptively learn the discerning characteristics of these species.

With advancements in computing technologies, many intelligent systems have emerged with various capabilities of learning and classifications. Many researchers from different disciplines are still actively engaged in further research toward developing systems with more efficiency and accuracy. As a result, a number of research articles have appeared in the literature in the fields of automatic pattern recognition. Many of these articles address the problem of automatic bird species classification. Among them, Chen and Li [1] proposed a time-frequency texture based approach. They segmented the spectrograms of audio signals into syllables and extracted gray level co-occurrence matrix (GLCM) [2] descriptors for each of these syllables. Then they applied Random Forest [3] algorithm to classify bird species with these GLCM descriptors as the required features. A hybrid classification approach was proposed in [4] where the features of interest were static and dynamic cepstral coefficients for each segmented syllable. Adopted classification technique was an optimal mode between Gaussian mixture model and vector quantization. Taking into account the contamination effects of environmental noise, Tjahja et. Al. [5] proposed a supervised method for isolating bird song syllables. They adopted a hierarchical segmentation approach involving both pixel-level and segment level information.

This article extends the proposal of Chen and Li [1] through enhancing the spectrogram segmentation algorithm. The rest of the article is organized as follows. First, an overview of our approach is presented in section II. Then our methodology is presented in section III, which is followed by result analysis in section IV. Finally the article is concluded with some possible future directions in section V.

II. SYSTEM OVERVIEW

The proposed approach consists of four major steps. First, spectrogram of the input audio signal was computed using short-time Fourier transform (STFT). Then, we applied an enhanced segmentation technique, which constitutes our primary contribution in this paper, in order to isolate syllables in the spectrogram. A main problem of the segmentation approach in ref [1] was its limited ability to distinguish signal from noise due to application of a single global threshold. This was overcome in our approach through application of both local and global thresholds and constraining the local maxima to be within a predefined range from the global maxima. To extract dominant texture features, GLCM descriptors were computed for each of the extracted syllables. Then a random forest classifier was used to classify the spectrograms based on the GLCM features. The overall process is pictorially represented in Fig. 1.

III. METHODOLOGY

Accuracy of classification largely depends on the discerning capabilities of the extracted features. As extracted features are often affected by the ambient noise, it is imperative to design robust feature extraction methods. Therefore, this article aims at improving robustness at both

978-1-5386-1150-0/17/$31.00 ©2017 IEEE

feature extraction and classification stages. Adopted approach is step-by-step described in the following sub-sections.

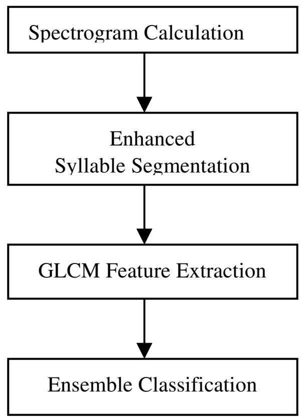


Fig. 1: System overview.

A. Spectrogram Calculation

Usually two methods, namely band-pass filter bank and Fourier transform, are available for spectrogram calculation from time domain signal such as audio signal. Of them, band-pass filter banks are usually used in analog signal processing [6] and Fourier transform is used in digital signal processing. As such we applied Fourier transform, more specifically, short-term Fourier transform (STFT), for spectrogram calculation. An audio signal representing song of a bird belonging to *Cettia Cetti* species and its computed spectrogram are given in the Fig. 2.

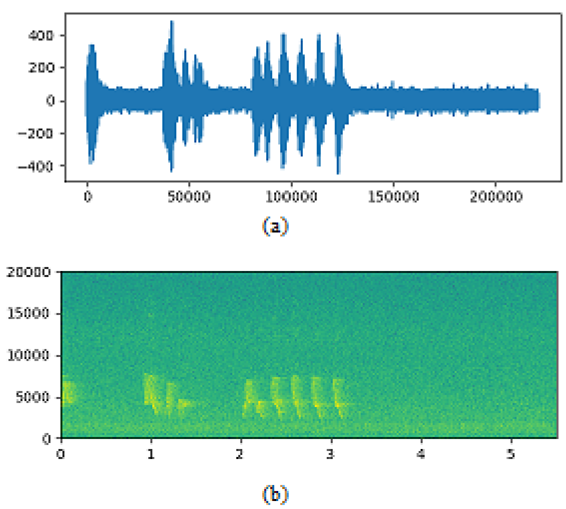


Fig. 2: (a) Input audio signal for *Cettia Cetti* bird song (b) Computed spectrogram.

If we analyze the audio signal and its spectrogram in Fig. 2, it would be clear that for every high amplitude peak in Fig. 2(a), there is a high amplitude region spanning over corresponding time and frequency ranges.

B. Enhanced Syllable Segmentation

Local frequency distribution has been preferred over global distribution by many researchers in bird audio signal classification [1, 5, 6-8]. Therefore, it is necessarily a first step to identify local distributions in the audio signal called syllables. As mentioned, an attractive syllable segmentation technique has been proposed in [1], however, with some potential limitations. Therefore, we adopted their approach and enhanced it to overcome the limitations. Our enhanced approach is described in details in the following. It is worth mentioning that we viewed the spectrograms as 2-D images and addressed spectrogram segmentation problem from the perspective of image segmentation algorithms.

Upon observing bird song audio signals, it became clear to us that there are substantial environmental noises in these captures. Also, we found that the performance of the approach proposed in [1] was greatly affected by these noises. Therefore, we applied Gaussian smoothing as the first step in our segmentation technique. This particularly eliminated noisy spurious peaks from the spectrogram. In addition to Gaussian smoothing, we also use two thresholds. We constrain the intra-syllable amplitude variations within a local threshold and the inter-syllable amplitude variations within a global threshold. Then, we extracted syllable segments as follows.

We find the global amplitude peak in the spectrogram. Let us denote this global peak by a tuple $(S(t_0, f_0), A_0)$ where, $S(t_0, f_0)$ and A_0, respectively, represent position and amplitude of the global peak point, and t_0 and f_0 time and frequency. As the spectrogram was Gaussian smoothed, it is expected that this peak would belong to a syllable extending locally on both sides of the peak. Now our goal is to trace candidate points in the spectrogram on both sides of $S(t_0, f_0)$ that fulfill certain predefined criteria. In order to find all the syllable points on the left of the peak point, first we find a vertical line $S(t_{-1}, f_0 - 10 : f_0 + 10)$, where $f_0 - 10 : f_0 + 10$ represents a frequency range from $f_0 - 10$ to $f_0 + 10$. Then we find the local maximum $(S(t_{-1}, f_{-1}), A_{-1})$ in this range and if $(A_0 - A_{-1}) < Th_l$, where Th_l is the predefined local threshold, then the point $S(t_{-1}, f_{-1})$ is included as a point of the syllable. Then, the local maximum is searched on the left of $S(t_{-1}, f_{-1})$ in the range $S(t_{-2}, f_{-1} - 10 : f_{-1} + 10)$. If the corresponding amplitude A_{-2} is within the threshold T_l, the point $S(t_{-2}, f_{-2})$ is also included as a syllable point and the process continues as long as new points can be found on the left satisfying the threshold. Let us denote the leftmost point within the syllable by t_l. A similar procedure is applied on the right side of global peak point in order to extract the syllable points on the right up to the rightmost point t_r. Therefore, the whole syllable lies within the points t_l and t_r in time. The extracted syllable from the spectrogram in Fig. 2(b) is shown in Fig. 3(b) in Section IV.

In order to find the other syllables, we eliminate all the points of spectrogram between t_l and t_r of the previously extracted syllable. Then the syllable extraction process, as described earlier, is repeated on the remaining points of the spectrogram if the next global peak is within the predefined global threshold Th_g from first global peak. If a detected syllable covers a spectrogram region less than 200ms, it is merged with an adjacent syllable if available, otherwise it is discarded.

C. GLCM Feature Extraction

Among the available statistical approaches to texture feature extraction, GLCM has drawn much attraction of the researchers in the relevant community. GLCM basically represents the joint probability of simultaneous appearance of two pixels with certain inter-distance, d and relative orientation θ. The basic metric in GLCM statistics is the statistical frequency $n(i,j\,|\,d,\theta)$ of occurrences when two pixels with distance d and orientation θ have intensity levels i and j respectively [1]. In this article, N different intensity levels have been considered.

$$n(i,j\,|\,d,\theta)=card\{(x,y)\,|\,f(x,y)=i,\,f(x+\Delta x,y+\Delta y)=j\,\}$$

Here, $\Delta x = d\cos\theta$ and $\Delta y = d\sin\theta$. This metric is normalized over all the combinations of (i,j) to obtain corresponding probability,

$$p(i,j\,|\,d,\theta)=\frac{n(i,j\,|\,d,\theta)}{\sum_i\sum_j n(i,j\,|\,d,\theta)}$$

As many as 14 different GLCM texture features have been proposed [1, 9-11]. Out of them, following five measures are considered as texture features in this article.

$$Angular\ Second\ Moment=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}p(i,j)^2 \tag{1}$$

$$Correlation=\frac{1}{\sigma}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(i-\mu)(j-\mu)p(i,j) \tag{2}$$

Where, $\sigma=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(i-\mu)^2 p(i,j)$

$$Contrast=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(i-j)^2 p(i,j) \tag{3}$$

$$Entropy=-\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}p(i,j)\log|p(i,j)| \tag{4}$$

$$Homogeneity=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\frac{p(i,j)}{1+(i-j)^2} \tag{5}$$

While the first feature, angular second moment, also known as energy, basically reflects the uniformity of the texture, correlation, which constitutes the second feature, can be interpreted as the principal direction of the texture. Further, entropy and contrast represent the complexity of texture and the clarity of texture, respectively. Finally, homogeneity measures local changes in image texture number. We considered $d=1$ and $\theta=0^\circ, 45^\circ, 90^\circ, 135^\circ$ for the extraction of all these features.

D. Ensemble Classification

We adopted Random Forest [3], a robust ensemble classifier, for the task of classification. The features vector consisted of 20 GLCM texture descriptors computed in the previous step. We considered 100 trees in the construction of the random forest with a maximum depth of 10.

Random Forest, which incorporates a collection of decision trees for the classification task, is an important step forward in the strive for finding an optimal and robust classifier. RF is widely well reputed for its fast performance in in training as well as evaluation, which strongly motivated us to adopt it as the classifier.

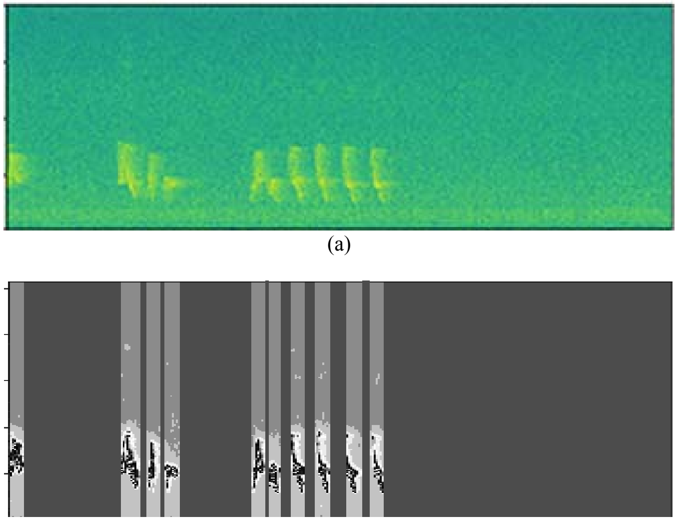
IV. RESULTS

We experimented our system on recoded audio signals of 10 different bird species. Statistics of the recordings for these audio signals are given in Table 1. Recorded audio signals are divided into 10 classes based on bird species and vocalization types. For example, calling audio of *Galerida theklae* bird has been assigned class index 3, while song audio of the same bird has been assigned class index 4. Number of extracted syllables for each class is given in the last column of the table.

Table 1: Audio recording statistics

Class Index	Bird Species	Vocal. Type	# of Audio files	# of Syll.
1	*Aegithalos caudatus*	Call	24	88
2	*Linaria cannabina*	Song	20	64
3	*Galerida theklae*	Call	16	84
4	*Galerida theklae*	Song	28	108
5	*Luscinia megarhynchos*	Call	16	80
6	*Sylvia cantillans*	Call	24	124
7	*Sylvia melanocephala*	Call	32	96
8	*Sylvia undata*	Call	32	152
9	*Sylvia undata*	Song	36	156
10	*Turdus philomelos*	Call	16	80
11	*Cettia cetti*	Song	17	96

The results of the syllable extraction algorithm are shown in Figure 3. In order to keep brevity of representation, we have shown spectrograms and extracted syllables for a single audio file. While Figure 3(a) shows spectrogram computed from this audio file, Figure 3(b) shows the extracted syllables. As seen in this figure, the algorithm has successfully extracted most of the syllables.



(a)

(b)

Fig. 3: (a) Computed spectrogram of class 11 bird species
(b)Results of syllable extraction

Classification accuracy for each bird species is given in Table 2. For proper training and testing of the classification algorithm, the syllable sets were split into training sets containing 70% of the syllables and test sets with the rest. Cardinalities of these training and test sets are also shown in the table. As seen in this table, the worst classification accuracy was obtained for class 4, where out of 40 test syllables, 30 were accurately classified thus giving an accuracy of 75%. The best case was obtained for class 2, where out of 18 test syllables, 17 were accurately classified thus giving an accuracy of 94.44%. On the total, 297 syllables out of 339 were accurately classified giving 87.61% average accuracy which may be considered quite satisfactory.

Table 2: Classification accuracy

Index Number	# of training syllable	# of test syllables	Accurate Prediction	Accuracy (%)
1	58	30	26	86.67
2	46	18	17	94.44
3	65	19	16	84.21
4	68	40	30	75
5	56	24	21	87.5
6	81	43	40	93.02
7	66	30	28	93.33
8	108	44	35	79.54
9	115	41	38	92.68
10	56	24	22	91.67
11	70	26	24	92.31
Total	789	339	297	87.61

V. Conclusion

A bird species classification technique based on recorded audio signal was presented. The main contribution of this article was to enhance the technique of syllable extraction from spectrogram presented in [1]. This enhanced algorithm improved the ability to distinguish signal from noise through application of separate thresholds in the local and global scopes. Also an enhanced pixel tracing technique was introduced to find segment boundaries of the syllables. The classification results showed the efficacy of the presented technique.

For future direction, there are plenty of rooms for the improvement of the syllable segmentation technique. The present technique still involves thresholds, which is always undesirable in signal processing techniques. Fuzzy logic approaches may be used to avoid these thresholds. Other classification techniques including connectionist approaches may be explored. However, training is a time consuming factor there which should be taken into consideration.

References

[1] S. S. Chen, and Y. Li, “Automatic Recognition of Bird Songs Using Time-frequency Texture,” *5th IEEE International Conference on Computational Intelligence and Communication Networks*, pp. 262–266, 2013.

[2] C. H. Lin, R. T. Chen, and Y. K. Chan, “A smart content-based image retrieval system based on color and texture feature*,” Image and Vision Computing*, vol. 27, no. 6, pp. 658–665, 2009..

[3] L. Breiman, "Random Forests," *J. Machine Learning*, vol. 45, no. 1, pp. 5- 32, October 2001.

[4] C. H. Lee and C. C. Han, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp 1541-1550, November 2008.

[5] T.V. Tjahja, X. Z. Fern, R. Raich,and A.T. Pham, "Supervised hierarchical segmentation for bird song recording", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 763-767, April 2015.

[6] E. Cakır, S. Adavanne, G. Parascandolo, K. Drossos, T. Virtanen, "Convolutional recurrent neural networks for bird audio detection", *arXiv:1703.02317* [cs.SD].

[7] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, DOI: 10.1109/ICASSP.2011.5946906, pp 2012-2015, May 2011.

[8] R. Sun, Y. W. Marye, H. Zhao, “Wavelet transform digital sound processing to identify wild bird species”, *IEEE International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR),* pp 306-309, July, 2013.

[9] R. Asery, R. K. Sunkaria, L. D. Sharma, A. Kumar, "Fog Detection using GLCM based Features and SVM", *IEEE Conference on Advances in Signal Processing (CASP)*, pp 72-76, June, 2016.

[10] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural features for image classification", *IEEE Transactions on Systems, Man and Cybernetics*, vol SMC-3, no 6, pp 610-621, 1973.

[11] F. Albregtsen, "Image Processing Laboratory", *Department of Informatics, University of Oslo*, November 5, 2008