International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 2, April 2022, pp. 1477~1487
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i2.pp1477-1487

Robust cepstral feature for bird sound classification

Murugaiya Ramashini[1,2], Pg Emeroylariffion Abas[1], Kusuma Mohanchandra[3], Liyanage C. De Silva[1]
[1]Faculty of Integrated Technologies, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam
[2]Department of Computer Science and Informatics, Uva Wellassa University, Badulla, Sri Lanka
[3]Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Kanakapura, India

Article Info

Article history:

Received Apr 28, 2021
Revised Aug 3, 2021
Accepted Sep 1, 2021

Keywords:

Birds classification
Cepstral features
Gammatone frequency cepstral coefficients
Support vector machine

ABSTRACT

Birds are excellent environmental indicators and may indicate sustainability of the ecosystem; birds may be used to provide provisioning, regulating, and supporting services. Therefore, birdlife conservation-related researches always receive centre stage. Due to the airborne nature of birds and the dense nature of the tropical forest, bird identifications through audio may be a better solution than visual identification. The goal of this study is to find the most appropriate cepstral features that can be used to classify bird sounds more accurately. Fifteen (15) endemic Bornean bird sounds have been selected and segmented using an automated energy-based algorithm. Three (3) types of cepstral features are extracted; linear prediction cepstrum coefficients (LPCC), mel frequency cepstral coefficients (MFCC), gammatone frequency cepstral coefficients (GTCC), and used separately for classification purposes using support vector machine (SVM). Through comparison between their prediction results, it has been demonstrated that model utilising GTCC features, with 93.3% accuracy, outperforms models utilising MFCC and LPCC features. This demonstrates the robustness of GTCC for bird sounds classification. The result is significant for the advancement of bird sound classification research, which has been shown to have many applications such as in eco-tourism and wildlife management.

Corresponding Author:

Murugaiya Ramashini
Faculty of Integrated Technologies, Universiti Brunei Darussalam
Tungku Link Street, BE1410, Brunei Darussalam
Email: ramashini@uwu.ac.lk

1. INTRODUCTION

Birds play a significant role as pollinators or seed dispersers for the stability of the ecosystem, as well as playing a crucial role in maintaining a balanced population of predators and prey in the ecosystem. As such, birdlife conservation and species preservation-related projects are essential for a balanced ecosystem. Implementations of these types of project are demanding, requiring manual work, labour, and physically intensive processes. The emergent of modern and advanced techniques and technologies has somehow made environmental and biodiversity monitoring related researches easier and more feasible. Signal processing and machine learning techniques have been used by many researchers to facilitate demanding and complicated processes. Especially in dense vegetation, bioacoustics signal processing and pattern recognition algorithms have been used for the detection and identification of bird species [1].

To classify bird species according to their acoustic signals, it is necessary to explore the mechanics involved in the production of bird sounds, which utilises the bird's unique vocal organ (the syrinx), an organ not found in any other animal [2]. Syrinx is a vibratory sound-generating organ [3], producing sound through coordinated activities of several muscles that are associated with it, as well as other organs such as the respiratory system. With the exception of vultures, a bird's syrinx is normally located deep within the bird's

chest [4]. This is the junction of two primary bronchi that joins traces of the respiratory system. Generally, a syrinx is composed of a paired set of strong muscles called syringeal, vibrating membranes called labia, several cartilages, and a vocal tract [5] surrounded by an air sac and connected to the air sac of the lungs [4]. The air sac maintains the inhaled air which is passing through the lungs and continuously supplies pressurised expiratory airflow. The syrinx uses all forced expiratory airflow that passes through it to produce vocal sounds. Air sacs, then, act as resonant chambers and play a major role in shaping the spectral composition of the emitted vocal sound [5].

Despite the fact that syrinx is used by every bird for its vocalisation, sounds produced by different species of bird may differ. This is due to the variations in structure, size and location of the syrinx of different bird species as well as its associated organs [4], which play significant roles in determining the frequency of the bird's calls or songs [2], [6]. In the audio signal processing domain, the bird's vocalisation $s(t)$ can be considered as the output of a linear convoluted system. Excitation $e(t)$ is associated with the expiratory airflow, originated from the respiratory system and associated organ activity. The dynamics of the syrinx may be modelled by impulse response $h(t)$, which depends on the bird species. Since the goal is to recover information associated with vocal sound, the analysis requires isolation of $h(t)$. However, as both excitation and impulse response are unknown, recovering $h(t)$ is not as straightforward, with blind deconvolution required. Hence, homomorphic signal processing method is commonly used to retrieve the most vital information of the bird sounds, to allow classification of species. This signal processing method is applied in many signal processing applications such as speech recognition, deconvolution, and pitch detection [7].

Cepstral features are the most commonly derived features from the homomorphic signal processing method, and are compact representations of the spectrum which provide a smooth approximation based on logarithmic magnitude [7]. Linear prediction cepstral coefficients (LPCC), mel frequency cepstral coefficients (MFCC), gammatone frequency cepstral coefficients (GTCC), perceptual linear prediction (PLP) cepstral coefficient and greenwood function cepstral coefficients (GFCC) are some types of cepstral features [8]. Cepstral features have been primarily used for speaker identification and speech recognition [9], [10]. However, they have also been employed in applications related to audio retrieval such as singer identification, music classification, environmental sound recognition [11], pitch determination [8] of speech signals [12], and for identification of musical instruments [13].

Within the bird sound classification research, MFCC represents one of the most widely used cepstral feature for bird sound classification [14], [15]. Based on human perception of sound, frequency domain representations of original bird sounds are provided as input to the mel-scale filter bank to produce mel-spectrum, which are then converted to MFCC using cepstral analysis [16]. Each band of the MFCC contains a weighted sum representing the spectral magnitude in the corresponding channel [17]. The calculation of MFCC parameters is efficient and straightforward since it does not involve any tuning parameters. Lee *et al.* [18] use both static and dynamic two-dimensional MFCCs to extract features for their work. MFCCs have also been combined with other methods for feature extraction. For instance, Kogan and Margoliash [19] use both MFCC as well as linear predictive coding (LPC) for feature extraction, whilst a combination of three methods: binned frequency spectrum, MFCC, and LPC, have been used by Leng and Tran [20]. Other cepstral based methods include GFCCs with first and second derivatives [21], and power normalized cepstral coefficients (PNCC) [22].

Audio signal processing researchers have proven that despite the volume of works that have been done using MFCC features for classification purpose, a spectrum of improvement in classification accuracy can still be achieved by considering other cepstral features such as GTCC [23], [24] and LPCC [25], [26]. Especially for non-speech audio classification [11] and noisy environmental audio data, GTCCs have been shown to be more robust [9]. Furthermore, GTCCs have been shown to have a higher resolution at a low frequency than MFCC [27]. Since classification accuracy depends on the features in which the training and testing are done, feature extraction is one of the integral parts of applications related to classification.

Numerous researches have been done in feature extraction and, consequently, classification, however, none of them has, thus far, specifically focused on finding the most suitable feature to classify bird sound. This paper focuses on finding the most robust cepstral features, specifically for bird sound classification purpose. LPCC, MFCC and GTCC features are extracted separately from bird sounds and classified using support vector machine (SVM). Results from the different features are then analysed to determine the most suitable features for the classification of bird sounds.

The following section 2 discusses the research method, composed of data collection, pre-processing, feature extraction, along with the classification method. Section 3 presents classification results using the different features, with the most robust cepstral feature for bird classification subsequently discussed. The final section concludes the paper.

2. RESEARCH METHOD

The general process adopted in this paper is shown in Figure 1. It has two main categories; feature extractions from the collected bird sounds, and classification prediction of unknown bird sounds using a pre-trained model. LPCC, MFCC, and GFCC are the different features extracted from the bird sounds, with SVM used as the classifier. Finally, classification results from the different types of feature are compared to determine the most robust cepstral feature for bird sounds.

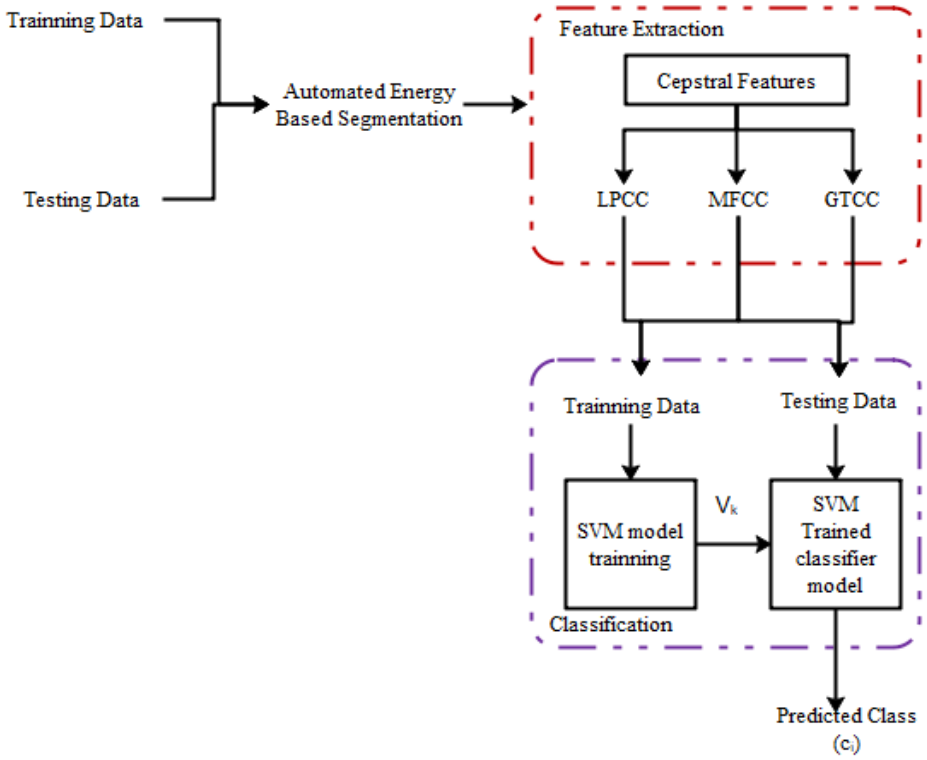

Figure 1. Adopted research method

2.1. Data collection and pre-processing

Birds sounds have been collected from an online database called "Xeno-canto" [28], which contains bird sounds with negligible environmental noise. The data have been properly labelled and verified by experts and are employed for supervised classification. Collected data are divided into two sets: training and testing dataset. Energy-based automated segmentation [29] is performed on both data sets to remove unwanted silent intervals and detect relevant sounds only.

2.2. Feature extraction

Feature extraction is a process of deriving the properties and characteristics of a signal that can be used for further analysis. A suitable feature mimics the properties of a signal in a much efficient way [8], with the ability to describe and represent the birds sound, and also has a significant impact on classification results. Three (3) types of cepstral features are extracted from both training and testing data sets: LPCC, MFCC and GTCC.

2.2.1. Linear prediction cepstral coefficients (LPCC)

The cepstrum possesses many advantages, such as source-filter separation, orthogonality, and compactness. These properties make cepstral coefficients robust and suitable for machine learning. On the other hand, linear prediction coefficients (LPCs) are too sensitive to numerical precision; hence it is desirable to transform LPC into the cepstral domain. The resultant transformed coefficients are referred to as LPCCs [8], which are adopted in this work.

This feature is defined as the inverse Fourier transform (IFT) of the logarithmic magnitude of the linear prediction spectral complete envelope [30], and it provides a more robust and compact representation, which is especially useful for automatic speech recognition and speaker identification [25]. LPCC represents

the human vocal track efficiently, based on linear prediction [31]. Linear predictive analysis is used to estimate the n^{th} sample by using the previous p sample; a linear combination is used as shown in (1).

$$s\hat{}(n) = a_1 s(n-1) + a_2 s(n-2) + \cdots a_p s(n-p) \quad (1)$$

Where, $a_1, a_2, a_3, \ldots, a_p$ are considered to be constants over a sound analysis frame and also known as LPC or predictor coefficients. These coefficients are used to estimate the sound samples. The difference between actual and predicted sound samples is its error and is given in (2) and (3).

$$e(n) = s(n) - s\hat{}(n) \quad (2)$$

$$e(n) = s(n) - \sum_{k=1}^{p} a_k\, s(n-k) \quad (3)$$

Where, s(n) is the sound signal, and e(n) is the error in prediction, s^(n) is a predicted sound signal, and a_ks are the LPCs. To compute a unique set of LPCs, the sum of squared differences between the actual and predicted sound samples could be determined and then minimised. Squared error is given as (4).

$$E_n = \sum_m [s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k)]^2 \quad (4)$$

Where, the number of samples in an analysis frame can be denoted as m. The squared error can be minimised (error minimisation) by differentiating E_n with respect to each and every a_k and then, setting the value to zero as shown in (5).

$$\frac{\partial E_n}{\partial a_k} = 0, for\ k = 1,2,3 \ldots., p \quad (5)$$

After a_k are obtained, cepstral coefficients can be derived from the following recursion,

$$C_0 = log_e p \quad (6)$$

$$C_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k} \quad for\ 1 < m < p \quad (7)$$

Hence, C_m can be denoted as (8),

$$C_m = \sum_{k=m-p}^{m-1} \frac{k}{m} C_k a_{m-k,} \quad for\ m>p \quad (8)$$

2.2.2. Mel frequency cepstral coefficients (MFCC)

MFCCs represent the short-time power spectrum of an audio clip based on the discrete cosine transform of the log power spectrum on a nonlinear mel-scale [32]. In MFCCs, the frequency bands are equally spaced on the mel-scale, which closely mimic the human auditory system, making MFCCs to be common key features in various audio signal processing applications [8]. MFCCs are commonly computed on a warped frequency scale based on known human auditory perception. These frequencies are mapped onto a nonlinear Mel filter bank, transforming them onto the cepstral domain.

To balance the spectrum of bird's sounds that generally have steep roll-off, high frequency filtering is commonly used. The transfer function in (9) is the most used pre-emphasis filter, with the value of b, which controls the slope of the filter, usually selected between 0.4 and 1.

$$H(z) = 1 - bz^{-1} \quad (9)$$

The signal is further windowed using hamming window in order to smooth the edges and reduce the edge effect while taking the discrete Fourier transform (DFT). Then, each windowed frame is then converted into a magnitude spectrum DFT, with N taken as the number of points used in this conversion.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \; ; \; 0 \leq k \leq N-1 \tag{10}$$

The magnitude spectrum is wrapped by twenty-six (N=26) overlapping triangular windows with centre frequencies equally distributed on the mel scale. Mel-scale attempts to mimic the nonlinear human ear perception of sound by being less discriminative at higher frequencies and more discriminative at lower frequencies. Based on this, human ears perceived frequency mel unit is measured, and the approximated mel from physical frequency is expressed as in (11).

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{100}) \tag{11}$$

Physical frequency is denoted by f, whilst f_{mel} denotes the perceived frequency. By multiplying the magnitude spectrum with each of the triangular mel weighting filters, X(k) is computed.

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)] \; ; \; 0 \leq m \leq M-1 \tag{12}$$

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k - f(m-1))}{f(m) - f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \dfrac{2(f(m+1) - k)}{f(m+1) - f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{13}$$

$$c(n) = \sum_{m=0}^{M-1} log_{10}(s(m)) \cos(\frac{\pi n(m - 0.5)}{M}); \qquad n = 0,1,2, \ldots . C-1 \tag{14}$$

M denotes the total number of triangular mel weighting filters, and X(k) is the mel spectrum of the magnitude spectrum. As expressed in (13), $H_m(k)$ is the weight given to the k^{th} energy spectrum bin contributing to the m^{th} output band with m ranging between 0 to M-1. The mel spectrum is usually represented in a log scale, and then, to derive cepstral features from log-mel power-spectrum, discrete cosine transform (DCT) is applied.

Since the syrinx has a smooth structure, energy levels in adjacent bands are likely to be correlated; this results in a signal with a frequency peak corresponding to the pitch of the signal and several formats representing low-frequency peaks in the cepstral domain. Generally, the first few MFCCs encompass most of the signal information, and thus, by truncating the higher-order DCT components, the system can be made more robust. Cepstral coefficients c(n) can then be represented by (14), where C is the number of MFCCs. Traditional MFCC systems use only 8-13 cepstral coefficients, with the 0^{th} coefficient often excluded as it represents the average log-energy of the input signal, which only carries little speaker-specific information [33].

2.2.3. Gammatone frequency cepstral coefficients (GTCC)

In recent years, GTCCs have been shown to be more robust to noise in many automatic speech recognition (ASR) systems [23], [24] and noisy environmental sound-related research [9]. GTCCs are based on gammatone (GT) filter banks; these filter banks give cochleagram as the output, which is the frequency-time representation of the sound signal. The extraction process of GTCCs is similar to that of MFCCs, except for the mel-filter bank, which has been replaced by a gammatone filter bank [8].

Gammatone cepstral coefficients (GTCCs) are a biologically inspired modification employing gammatone filters with equivalent rectangular bandwidth bands [23]. There are designed to simulate the process of a human auditory system with frequency resolution feature and filtering characteristics of the cochlear basilar membrane. Gammatone filters are a linear approximation of the filtering function performed by the cochlea in the inner ear, with the ear's frequency analysing sub-bands more delicate at higher frequencies. In fact, GTCC is a modification of the MFCC but uses GT filters, with equivalent rectangular bandwidth (ERB) bands [11]. A Gammatone filter with a centre frequency f_c is defined as (15):

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c + \varphi) \quad (15)$$

where, t refers to the time, ϕ is the phase (usually set to zero), the constant a controls the gain, and the value n defines the order of the filter. The attenuation factor of the i^{th} filter is represented by the factor b, which is defined in (16). b determines the decay rate of the impulse response across the filter bandwidth. In contrast, the bandwidth of each filter is related to the human auditory critical band.

$$b = 25.17(\frac{4.37f_c}{1000} + 1) \quad (16)$$

To obtain a representation similar to spectrograms, a set of gammatone filters, often referred to as channels with different centre frequencies, are used to create a gammatone filter bank. Gammatone filter bank emulates human hearing by simulating impulse response of the auditory nerve fibre, with its shape resembling a tone $\cos(2\pi f_c + \phi)$ modulated with a gamma function $e^{-2\pi bt}$ [27]. All the three (3) cepstral feature coefficients: LPCC, MFCC and GTCC, are extracted from collected training and testing data in this work. To find the most robust cepstral feature to classify bird sounds, all the extracted features are then used for training and testing by the same machine learning (ML) algorithm separately. Classification results from using the different features are then compared to determine the best features for the classification of bird sounds.

2.3. Classification

It has been proven that solutions to many existing issues can be solved by integrating audio signal processing techniques with ML algorithms [8]. In this work, a well-established supervised classifier called SVM [34], is used to train the model as well as to predict unknown bird sounds. The three different types of extracted cepstral features are separately used to train the SVM model, and consequently, the trained models are then used for testing. SVM is chosen due to its proven high accuracy results as well as statistical learning theory and structural risk minimisation properties [29]. SVM finds the best hyperplane, which can be described as the most significant margin between classes. In other words, it finds the data points which separate different classes accurately [35]. The polynomial kernel function (K_f) can be denoted as (17) for order n,

$$K_f(x_i, x_j) = (x_i^T, x_j + c)^n \quad (17)$$

where, x_i and x_j are the vectors of two input space, and c is the constant that allows trade-off to influence the higher-order and lower order. The cubic kernel function is used i.e. n is set to 3, in this paper.

3. RESULTS AND DISCUSSION

Based on the above discussion, fifteen (i.e. C=15) endemic Bornean bird sounds have been collected from an online database, "Xeno-Canto" [28]. These sounds have been divided into a training dataset (with six hundred (600) samples) and a testing dataset (with hundred (100) samples). Forty (40) samples from each bird species are used for training, and ten (10) samples from each bird species for testing. Table 1 lists the different birds and their corresponding abbreviations.

From both training and testing dataset, LPCC, MFCC, and GTCC features have been extracted. Thirteen coefficients (i.e. N=13) from each feature are extracted, with the individual extracted features used for training and then, testing; using SVM as the classification method. The predicted results are compared to find the most robust cepstral feature for bird sound classification.

K-fold cross-validation is performed whilst training the model to prevent overfitting. The dataset is split into several folds, with accuracy of each fold estimated, to ensure that every observation from the original dataset has the chance of appearing during training. In this work, five (05) fold cross-validation has been used for training the model using GTCC, MFCC, and LPCC features separately. Figure 2 shows the classifier's performance per class whilst training using GTCC features with five-fold cross-validation. The row and column represent the output and targeted classes, respectively. The green percentage values on the far-right column and the bottom row represent the percentage of correctly classified entries in that row and column, respectively. On the other hand, red percentages give the percentage of correctly classified entries in that row and column, respectively. Overall, the classifier provides 89.3% training accuracy, with 10.7% wrong classifications.

Table 1. Names of the birds with the corresponding abbreviations used

Bird No	Bird Name	Abbreviation
1	Bornean Blue Flycatcher	BBF
2	Bushy Crested Hornbill	BCH
3	Black Copped White-eye	BCW
4	Bornean Spider Hunter	BSH
5	Bornean Tree Pie	BTP
6	Bornean Whistler	BW
7	Collared Kingfish	CK
8	Green Pitta	GP
9	Golden Whiskered Barbet	GWB
10	Hotted Pitta	HP
11	Malaysian Banded Pitta	MBP
12	Malaysian Pied Fantail	MPF
13	Rhinocerous Hornbill	RH
14	Savanna Nightjar	SN
15	White-Crowned Forktail	WCF

Figure 2. The performance of the classifier per class while training GTCC features with five-fold cross-validation.

Figure 3 illustrates classification prediction accuracies for models separately trained and tested using the three (3) cepstral features. Using GTCC features results in the highest accuracy of 93.33% for all fifteen (15) birds, whilst using MFCC and LPCC features give 87.33% and 86.67% accuracies, respectively. Table 2 lists the class-wise SVM prediction accuracies based on LPCC, MFCC and GTCC features. As a whole, the SVM model trained using GTCC predicts eight (8) bird species with 100% accuracy whilst models trained with MFCC and LPCC features predict only five (5) and one (1) bird species, respectively, with 100%

accuracy. Also, with the exception of MBP bird, the model utilising GTCC features predict all other birds more than 80% correctly. Further analysis has also shown that four (4) samples out of the hundred and fifty (150) test samples, have been predicted wrongly by all three (3) feature-based classification process.

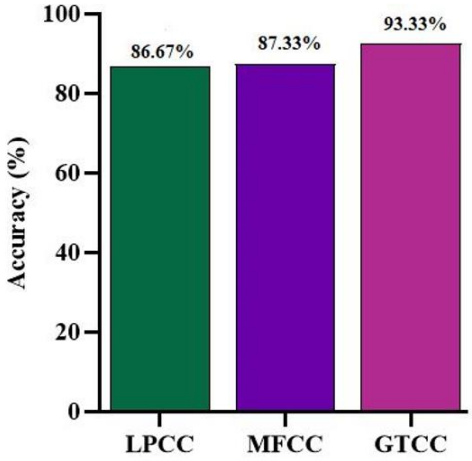

Figure 3. SVM prediction accuracy based on cepstral features

Table 2. Class wise SVM prediction accuracy for cepstral features

	Prediction Accuracy (%)		
Bird Name	LPCC	MFCC	GTCC
BBF	70	90	100
BCH	80	90	80
BCW	80	90	100
BSH	90	90	100
BTP	90	90	90
BW	90	70	90
CK	90	90	100
GP	90	90	90
GWB	80	70	90
HP	90	90	100
MBP	90	70	70
MPF	80	90	90
RH	90	90	100
SN	100	100	100
WCF	90	100	100
Average (%)	86.67	87.33	93.33

Figure 4 depicts the confusion matrix for SVM classification, using GTCC features only. Specifically, for MBP bird, it can be seen that only 7 out of the 10 MBP bird sounds are predicted correctly, i.e. 70% accuracy, with three wrongly predicted bird sounds classified as BBF birds. BCH bird has a reported accuracy of 80%, with two BCH birds wrongly predicted as BW and MPF. Other birds report accuracies of 90% and above. Also, from Figure 4, it can be seen that a total of 5 birds are wrongly predicted as BBF birds: 3 MBP birds, 1 BTP bird and GWB bird.

To understand more on the impacts of these features, GTCC, LPCC and MFCC have also been combined; and collectively used for training of SVM classification model. Figure 5 also shows the prediction accuracy of combined features, clearly showing that even though using combination of features provide significant improvements in accuracies when compared to MFCC and LPCC separately, these improvements are actually less than the model using GTCC features only. Using the LPCC feature alone produces 86.67%, but when combined with GTCC or MFCC separately, it produces 92.67% accuracy. However, when it combined with both GTCC and MFCC together, accuracy dropped slightly to 91.33%. Similarly, MFCC shows an approximately 5% increase in prediction accuracy when combined with GTCC and LPCC separately but reduces when all features are combined. However, the accuracy of GTCC is dropped whenever it is combined with other features.

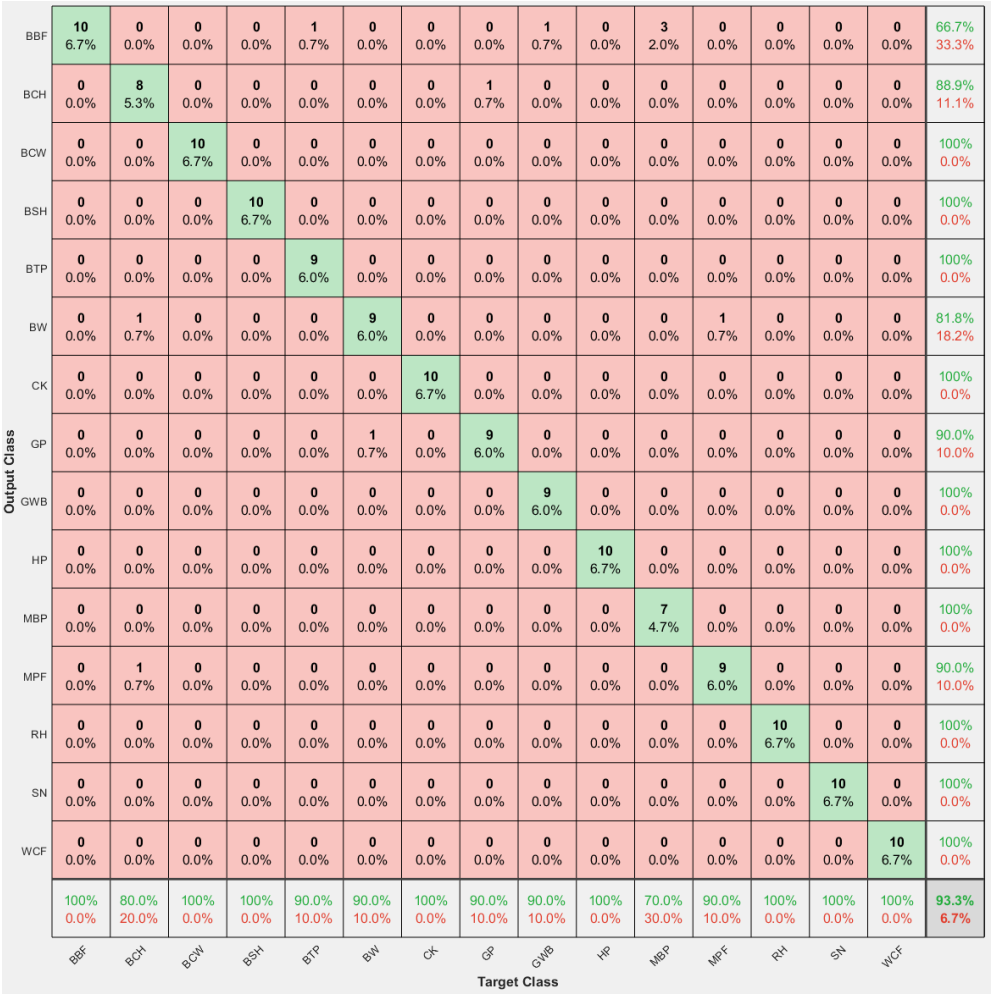

Figure 4. Confusion matrix of GTCC based SVM prediction

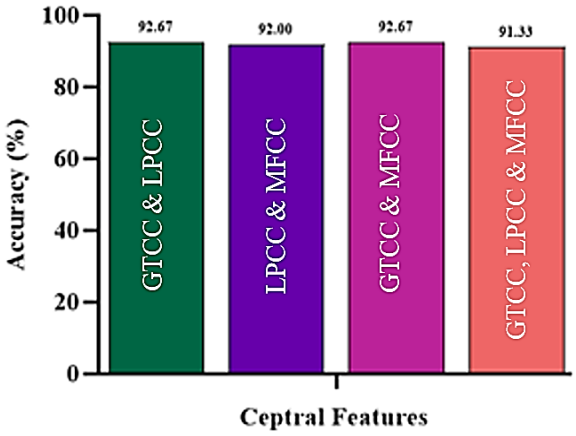

Figure 5. SVM prediction accuracy based on combined cepstral features

4. CONCLUSION

Regardless of applications, a classifier requires robust and discriminatory features to give good classification accuracy. The challenge is to extract the most appropriate and suitable features for a particular purpose. Feature selection is an important process, requiring the extracted feature to be compact but, at the same time, capable of highlighting important characteristics of the signal. Ideally, the extracted features

should not require a lot of processing by reducing the size of the signal significantly whilst still able to describe and represent the signal entirely and accurately. Therefore, this work aims to find the most robust feature for bird sound classification. While investigating more on bird's vocal track (syrinx), there are several common properties that can be found between human vocal track and speech processing, with that of bird sounds. Also, the literature has indicated that homomorphic signal processing method and its derivatives called cepstral features, particularly MFCC, is most suitable for the bird's sounds analysis. As such, focus has been made to explore different cepstral features in this paper, in order to obtain the most robust cepstral feature for birds sound.

The basic structure of any typical audio classification has several stages after data collection; the first stage is pre-processing, which is done on the audio signal for noise cancellation, silence reduction, and normalisation. In this work, fifteen (15) endemic Bornean birds' sounds have been collected and segmented using automatic energy-based segmentation in order to remove silence and noise in the recording. Then, in the feature extraction stage, three (3) cepstral features, namely LPCC, MFCC, and GTCC, have been extracted from the segmented audio signal. Finally, six hundred (600) samples have been used for training, and a hundred and fifty (150) samples used for testing. SVM classifier is used in this work separately for each features types, both training and testing. It can be seen that GTCC feature-based classification outperforms the other two LPCC and MFCC based classifications, despite the fact that MFCC has been more widely used by many researchers over the years. Combining the three cepstral features does not increase the accuracy over using GTCC features only. The result is significant as it shows that using GTCC alone would give reasonably high accuracy (93.3%) for bird sound classification. However, there is still room for further improvement through the investigation of different properties of birds sound, combining GTCC with other signal features, and implementing the technique in real-time on portable multimedia devices, which may also give new directions to this work.

REFERENCES

[1] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, vol. 80, pp. 1–9, Jun. 2014, doi: 10.1016/j.apacoust.2014.01.001.

[2] S. Nowicki and P. Marler, "How do birds sing?," *Music Perception*, vol. 5, no. 4, pp. 391–426, Jul. 1988, doi: 10.2307/40285408.

[3] O. N. Larsen and F. Gollerf, "Role of syringeal vibrations in bird vocalizations," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 266, no. 1429, pp. 1609–1615, Aug. 1999, doi: 10.1098/rspb.1999.0822.

[4] R. A. Suthers, "How birds sing and why it matters," in *Nature's Music*, Elsevier, pp. 272–295, 2004.

[5] L. R. Hernandez-Miranda and C. Birchmeier, "Mechanisms and neuronal control of vocalization in vertebrates," *Opera Medica et Physiologica*, vol. 4, no. 2, pp. 50–62, Dec. 2018, doi: 10.20388/omp2018.001.0059.

[6] M. J. Ryan and E. A. Brenowitz, "The role of body size, phylogeny, and ambient noise in the evolution of bird song," *The American Naturalist*, vol. 126, no. 1, pp. 87–100, Jul. 1985, doi: 10.1086/284398.

[7] L. Su, "Between homomorphic signal processing and deep neural networks: Constructing deep algorithms for polyphonic music transcription," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2017, pp. 884–891, doi: 10.1109/APSIPA.2017.8282170.

[8] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, Jan. 2020, Art. no. 107020, doi: 10.1016/j.apacoust.2019.107020.

[9] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7204–7208, doi: 10.1109/ICASSP.2013.6639061.

[10] B. Ayoub, K. Jamal, and Z. Arsalane, "Gammatone frequency cepstral coefficients for speaker identification over VoIP networks," in *2016 International Conference on Information Technology for Organizations Development (IT4OD)*, Mar. 2016, pp. 1–5, doi: 10.1109/IT4OD.2016.7479293.

[11] X. Valero and F. Alías, "Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

[12] A. Camacho, "Comment on 'Cepstrum pitch determination' [J. Acoust. Soc. Am. 41, 293–309 (1967)]," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2706–2707, Nov. 2008, doi: 10.1121/1.2988293.

[13] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1933–1941, Mar. 1999, doi: 10.1121/1.426728.

[14] M. Ramashini, P. E. Abas, U. Grafe, and L. C. De Silva, "Bird sounds classification using linear discriminant analysis," in *2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, Nov. 2019, pp. 1–6, doi: 10.1109/ICRAIE47735.2019.9037645.

[15] M. A. Acevedo, C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide, "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," *Ecological Informatics*, vol. 4, no. 4, pp. 206–214, Sep. 2009, doi: 10.1016/j.ecoinf.2009.06.005.

[16] A. Franzen and I. Y. H. Gu, "Classification of bird species by using key song searching: A comparative study," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, 2003, vol. 1, pp. 880–887, doi: 10.1109/ICSMC.2003.1243926.

[17] V. M. Trifa, A. N. G. Kirschel, C. E. Taylor, and E. E. Vallejo, "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2424–2431, Apr. 2008, doi: 10.1121/1.2839017.

[18] C.-H. Lee, C.-C. Han, and C.-C. Chuang, "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, Nov. 2008,

doi: 10.1109/TASL.2008.2005345.
[19] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2185–2196, Apr. 1998, doi: 10.1121/1.421364.
[20] Y. R. Leng and H. Dat Tran, "Multi-label bird classification using an ensemble classifier with simple features," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec. 2014, pp. 1–5, doi: 10.1109/APSIPA.2014.7041649.
[21] K. Adi, M. T. Johnson, and T. S. Osiejuk, "Acoustic censusing using automatic vocalization classification and identity recognition," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 874–883, Feb. 2010, doi: 10.1121/1.3273887.
[22] A. Badi, K. Ko, and H. Ko, "Bird sounds classification by combining PNCC and robust Mel-log filter bank features," *Journal of the Acoustical Society of Korea*, vol. 38, no. 1, pp. 39–46, 2019, doi: 10.7776/ASK.2019.38.1.039.
[23] R. Fathima and P. E. Raseena, "Gammatone cepstral coefficient for speaker identification," *International Journal of Scientific and Engineering Research*, vol. 4, no. 10, pp. 795–798, 2013.
[24] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 4625–4628, doi: 10.1109/ICASSP.2009.4960661.
[25] S. Misra, T. Das, P. Saha, U. Baruah, and R. H. Laskar, "Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis," in *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, Mar. 2015, pp. 1–4, doi: 10.1109/ICCPCT.2015.7159307.
[26] P. Chen, "A comparative study of LPCC and MFCC features for the recognition of assamese phonemes," *International Journal of Engineering and Technology*, vol. 2, no. 3, pp. 1–10, 2013.
[27] H. Wang and C. Zhang, "The application of Gammatone frequency cepstral coefficients for forensic voice comparison under noisy conditions," *Australian Journal of Forensic Sciences*, vol. 52, no. 5, pp. 553–568, Sep. 2020, doi: 10.1080/00450618.2019.1584830.
[28] F. Briggs *et al.*, "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, Jun. 2012, doi: 10.1121/1.4707424.
[29] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, Dec. 2007, Art. no. 038637, doi: 10.1155/2007/38637.
[30] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974, doi: 10.1121/1.1914702.
[31] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 498–502, doi: 10.1109/CONFLUENCE.2016.7508171.
[32] M. G. M. Milani, P. E. Abas, L. C. De Silva, and N. D. Nanayakkara, "Abnormal heart sound classification using phonocardiography signals," *Smart Health*, vol. 21, Jul. 2021, Art. no. 100194, doi: 10.1016/j.smhl.2021.100194.
[33] V. R. Reddy, *Language identification using spectral and prosodic features*. Springer, 2015.
[34] D. Chakraborty, P. Mukker, P. Rajan, and A. D. Dileep, "Bird call identification using dynamic kernel based support vector machines and deep neural networks," in *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 2017, pp. 280–285, doi: 10.1109/ICMLA.2016.60.
[35] I. The MathWorks, "Predictive maintenance toolbox ™ reference R 2020 a," 2020.