Computational Linguistics and Chinese Language Processing
Vol. 19, No. 1, March 2014, pp. 55-68

On the Use of Speech Recognition Techniques to Identify Bird Species

Wei-Ho Tsai* and Yu-Zhi Xue*

Abstract

Wild bird watching has become a popular leisure activity in recent years. Very often, people can see birds or hear their sounds, but have no idea what kind of bird species they are seeing. To help people learn to identify bird species from their sounds, we apply speech recognition techniques to build an automatic bird sound identification system. In this system, two acoustic cues are used for analysis, timbre and pitch. In the timbre-based analysis, Mel-Frequency Cepstral Coefficients (MFCCs) are used to characterize the bird sound. Then, we use Gaussian Mixture Models to represent the MFCCs as a set of parameters. In the pitch-based analysis, we convert bird sounds from their waveform representations into a sequence of MIDI notes. Then, Bigram models are used to capture the dynamic change information of the notes. We chose the top ten common bird species in the Taipei urban area to examine our system. Experiments conducted using audio data collected from commercial CDs and websites show that the timbre-based, pitch-based, and the combination thereof systems achieve 71.1%, 72.1%, and 75.0% accuracy of bird sound identification, respectively.

Keywords: Bird Species Identification, Bigram Model, Gaussian Mixture Model, Pitch, Timbre

1. Introduction

There are more than nine thousand and seven hundred bird species in the world. Although a number of birds are commonly seen, most people cannot recognize any of them. In this study, we attempt to develop automated techniques for identifying bird species from their sounds. Hereafter, this problem is referred to as bird sound identification. It is hoped that the

* Department of Electronic Engineering & Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, No.1, Sec. 3, Chunghsiao E. Rd. Taipei City, 10608, Taiwan, Tel.: +886-2-27712171; Fax: +886-2-27317120
E-mail: whtsai@ntut.edu.tw
The author for correspondence is Wei-Ho Tsai.

techniques can help people learn about such animals by simply recording the bird sounds they hear and sending the recording to our system.

Up to now, there has been very limited published research devoted to bird sound identification. In (Anderson *et al.,* 1996), Anderson *et al*. used dynamic time warping to measure the differences in spectrogram between an unknown bird sound recording and the template bird sound recordings. In (Kogan & Margoliash, 1998), Kogan *et al*. compared the performance of bird sound identification obtained with dynamic time warping and hidden Markov model, in which six acoustic features were used: linear predictive coding coefficients (LPCs), LPC-derived cepstral coefficients, LPC reflection, Mel-Frequency Cepstral Coefficients (MFCCs), log mel-filter bank channel, and linear mel-filter bank channel. In (McIlraith & Card, 1997), McIlraith *et al*. used a backpropagation neural network and multivariate statistics to perform bird sound identification. The acoustic features tested in (McIlraith & Card, 1997) are the number of syllables, average syllable duration, standard deviation of syllable durations, average pause duration, and standard deviation of pause durations. In (Somervuo *et al.,* 2006), Somervuo *et al*. compared three acoustic features on bird sound identification: sinusoidal modeling features, MFCCs, and descriptive features. Nevertheless, it is worth noting that all of the aforementioned studies tackle bird sound identification from the perspective of timbre-based analysis only. They all ignore bird sounds' pitch information, which is an important factor in why a bird sound is often called a bird song.

In this work, we propose a bird sound identification system based on timbre and pitch analyses. In addition to applying the most prevalent speaker-identification method to our system, we devise a method for exploiting the pitch information in bird sounds. Our experiments show that bird sound identification based on pitch information performs slightly better than that based on timbre information. It is further observed that combined use of timbre and pitch information achieves superior performance over the use of the individual information.

The remainder of this paper is organized as follows. Section 2 introduces the configuration of the proposed bird sound system, in which the two major components, timbre-based analysis and pitch-based analysis, are described in Sections 3 and 4, respectively. Section 5 discusses the experiments for examining our system. In Section 6, we present the conclusions and direction of our future works.

2. System Overview

Figure 1 shows the proposed bird sound identification system. In essence, the system can be divided into two components, namely timbre-based analysis and pitch-based analysis. Both components operate in two phases: training and testing. The purpose of the training phase is to extract the timbre and pitch features in each bird species' sound and to represent the features

as two sets of parametric models. In the testing phase, the system takes as input an unknown sound recording and produces as output two likelihood scores from the timbre-based and pitch-based analyses, respectively. The scores then are combined to serve as the basis of the decision. According to the maximum likelihood decision rule, the system decides an unknown sound recording in favor of bird species B^* when the condition in Eq. (1) is satisfied:

$$B^* = \arg\max_{1 \le i \le N} (\alpha \cdot v_i + \beta \cdot r_i) , \tag{1}$$

where N is the number of bird species; v_i and r_i are the likelihood scores output from the timbre-based and pitch-based analyses with respect to the i-th bird species' models, respectively; and α and β are tunable weights.

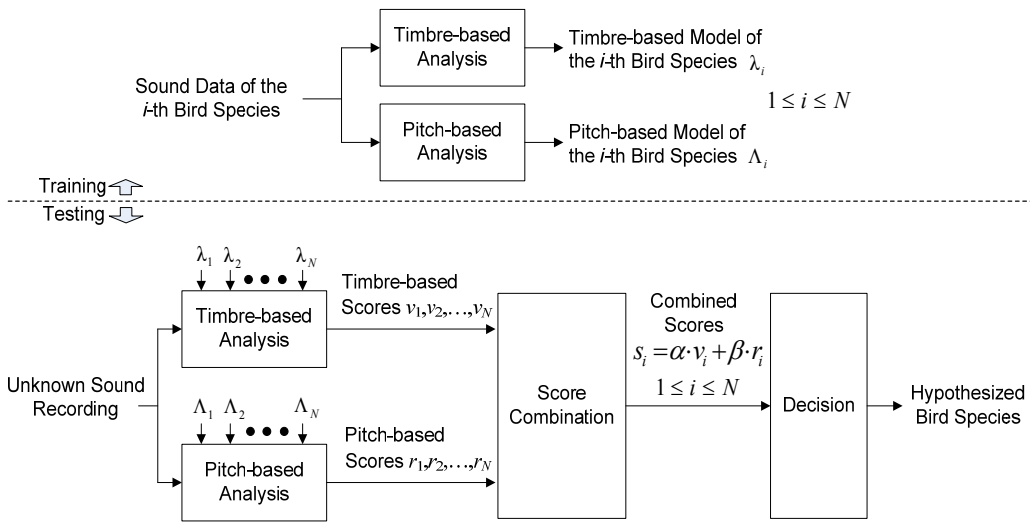

Figure 1. The proposed bird sound identification system.

3. Timbre-based Analysis

Figure 2 shows the procedure of the timbre-based analysis. It consists of feature extraction and Gaussian mixture modeling in the training phase, along with feature extraction and likelihood computation in the testing phase.

3.1 Feature Extraction

Among the timbre-based features investigated in (Kogan & Margoliash, 1998), the Mel-scale Frequency Cepstral Coefficients (MFCCs) feature (Davis & Mermelstein, 1980) has been found to be superior to the others in bird sound identification. To compute MFCCs, a waveform signal first is divided into frames using a P-length sliding Hamming window with $0.5P$-length overlapping between frames. Every frame then undergoes Hamming windowing

and fast Fourier transform (FFT) with size J. Next, each frame is passed through a set of triangular filter banks, equally spaced on a Mel scale. Let $|A_{t,j}|$ denote the signal's magnitude with respect to FFT index j in frame t, where $1 \le j \le J$. Then,

$$X_{t,i} = \frac{1}{B}\sum_{b=1}^{B}\left\{\log\left(\sum_{j=l_b}^{u_b}\left|A_{t,j}\right|^2 T_b(j)\right)\cdot\cos\left(\frac{\pi i}{B}(b-0.5)\right)\right\}, 1 \le i \le B\,, \tag{2}$$

where B is the total number of filter banks, l_b is the lowest frequency index in the b-th bank, u_b is the highest frequency index in the b-th bank, and $T_b(j)$ is the response of the b-th bank. Briefly, MFCCs represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. It is found that the nonlinear mel scale of frequency approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the regular cepstrum.

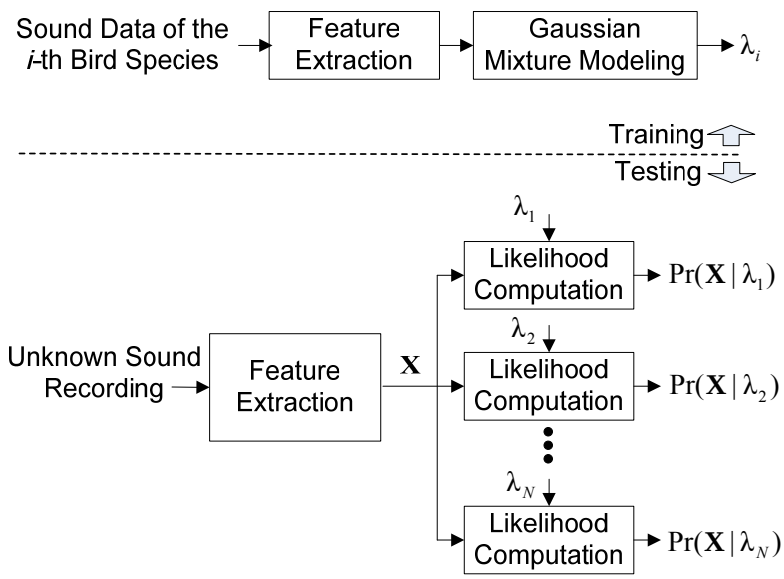

Figure 2. The procedure of the timbre-based analysis.

3.2 Gaussian Mixture Modeling

To capture the collective sound characteristics of each bird species, all of the MFCCs of each bird species are pooled together to form a Gaussian mixture model (GMM) (Reynolds & Rose, 1995). It is assumed that each bird species has its own timbre pattern that reflects in the distribution of MFCCs over a span of time. A GMM approximates the static timbre patterns by a mixture of Gaussian densities. Note that the reason we capture the static timbre patterns rather than dynamic timbre patterns using hidden Markov models (HMMs) (Rabiner, 1989) is to prevent the resulting models from dependence on bird individuals or bird messages.

The parameters of a GMM consist of means, covariances, and mixture weights, which are commonly estimated using the Expectation-Maximization (EM) algorithm (Dempster *et al.,* 1977). Nevertheless, recognizing that the numbers of each bird species' sound samples for training may not be sufficient always, we use the GMM-MAP approach (Reynolds & Quatieri, 2000) to generate each bird species' GMM. Specifically, all of the MFCCs of all of the bird species first are pooled together to form a universal GMM using the EM algorithm. Then, the parameters of the universal GMM are modified with respect to each bird species using the MFCCs of the individual bird species based on maximum *a posteriori* (MAP) estimation. If there are N bird species to be identified, we generate N GMMs, $\lambda_1, \lambda_2, \cdots, \lambda_N$.

3.3 Likelihood Computation

Given an unknown bird sound recording, the system computes its MFCCs $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T\}$ before computing the likelihood probability $\Pr(\mathbf{X}|\lambda_j)$ for each model λ_j:

$$\Pr(\mathbf{X} \mid \lambda_j) = \prod_{t=1}^{T} \sum_{k=1}^{K} w_{j,k} \cdot \frac{1}{\pi^N \left|\mathbf{C}_{j,k}\right|} \exp\left\{-\left(\mathbf{X}_t - \boldsymbol{\mu}_{j,k}\right)' \mathbf{C}_{j,k}^{-1} \left(\mathbf{X}_t - \boldsymbol{\mu}_{j,k}\right)\right\}, \tag{3}$$

where K is the number of mixture Gaussian components; $w_{j,k}$, $\boldsymbol{\mu}_{j,k}$, and $\mathbf{C}_{j,k}$ are the k-th mixture weight, mean, and covariance of model λ_j, respectively; and prime ($'$) denotes the vector transpose.

4. Pitch-based Analysis

As bird sound is often regarded as a type of music, it is reasonable to assume that each bird species has its own pitch pattern that can be exploited to distinguish from other species. Pitch is the reciprocal of fundamental frequency; hence, a bird sound recording can be viewed as a sequence of fundamental frequencies. We then can model the variations of the fundamental frequencies to characterize each bird species' sounds. Nevertheless, considering that the estimation of fundamental frequency is prone to numerical errors, we use MIDI note numbers instead of fundamental frequencies to explore the pitch information in bird sounds. The MIDI note numbers can be treated as the non-linear quantization of fundamental frequencies and can absorb the numerical errors during the estimation of fundamental frequencies. Figure 3 shows the procedure of pitch-based analysis. It consists of MIDI note extraction for converting sound recordings from waveform representations into MIDI note sequences and bigram modeling for characterizing the underlying pitch information in the note sequences.

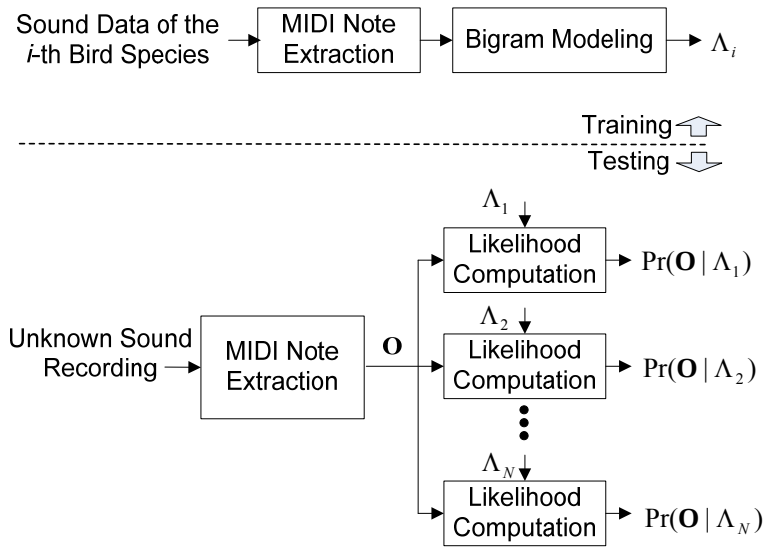

Figure 3. The procedure of pitch-based analysis.

4.1 MIDI Note Extraction

Let e_m, $1 \le m \le M$, be the inventory of possible notes produced by a bird. Our aim is to determine which among the M possible notes is most likely produced at each instant in a bird sound recording. We apply the strategy in (Yu *et al.,* 2008) to solve this problem. First, the bird sound is divided into frames using a P-length sliding Hamming window, with $0.5P$-length overlapping between frames. Every frame then undergoes a Fast Fourier Transform (FFT) with size J. Let $x_{t,j}$ denote the signal's energy with respect to FFT index j in frame t, where $1 \le j \le J$, and $x_{t,j}$ has been normalized to the range between 0 and 1. Then, the signal's energy on the m-th note in frame t can be estimated by:

$$\hat{x}_{t,m} = \max_{\forall j, U(j)=e_m} x_{t,j} \quad , \tag{4}$$

and

$$U(j) = \left\lfloor 12 \cdot \log_2 \left(\frac{F(j)}{440} \right) + 69.5 \right\rfloor, \tag{5}$$

where $\lfloor \ \rfloor$ is a floor operator, $F(j)$ is the corresponding frequency of FFT index j, and $U(\cdot)$ represents a conversion between the FFT indices and the MIDI note numbers.

Ideally, if note n_m is sung in frame t, the resulting energy, $\hat{x}_{t,m}$, should be the maximum among $\hat{x}_{t,1}, \hat{x}_{t,2}, \ldots, \hat{x}_{t,M}$. Nevertheless, it is sometimes the case that the energy of a true note is smaller than that of its harmonic note. To avoid the interference of harmonics in the estimation of true notes, we use the strategy of Sub-Harmonic Summation (SHS) (Piszczalski & Galler, 1979), which computes a value for the "strength" of each possible note by summing

the signal's energy on a note and its harmonic note numbers. Specifically, the strength of note n_m in frame t is computed using

$$y_{t,m} = \sum_{c=0}^{C} h^c \hat{x}_{t,m+12c} \ , \tag{6}$$

where C is the number of harmonics considered, and h is a positive value less than 1 that discounts the contribution of higher harmonics. The result of this summation is that the true note usually receives the largest amount of energy from its harmonic notes. Thus, the true note in frame t can be determined by choosing the note number associated with the largest value of the strength. Nevertheless, recognizing that a note usually lasts several frames, the decision could be made by including the information from neighboring frames. Specifically, we determine the sung note in frame t by choosing the note number associated with the largest value of the strength accumulated for adjacent frames, *i.e.*,

$$o_t = \arg\max_{1 \le m \le M} \sum_{b=-W}^{W} y_{t+b,m} \ , \tag{7}$$

Further, the resulting note sequence is refined by taking into account the continuity between frames. This is done with median filtering, which replaces each note with the local median of notes of its neighboring $\pm W$ frames to remove jitters between adjacent frames. In the implementation, the range of e_m is set to be $60 \le e_m \le 120$, corresponding to fundamental frequency from to 261.6 to 8591 Hz.

4.2 Bigram Modeling

After converting bird sounds into sequences of MIDI notes, we use a bigram model (Huang *et al.,* 2001) to capture the dynamic information in the note sequences. The bigram model consists of a set of bigram probabilities and unigram probabilities. The bigram probabilities $\Pr(e_j|e_i)$, $1 \le i, j \le M$, account for the frequency of a certain note e_i followed by another note e_j, while the unigram probabilities $\Pr(e_i)$ account for the frequency of occurring a certain note e_i. It is assumed that each bird species has its own pitch pattern that reflects in the frequency of occurrence of one or a pair of notes. For N bird species to be identified, we generate N bigram models $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$.

4.3 Likelihood Computation

In the testing phase, an unknown bird sound recording first is converted into a sequence of notes $O = o_1, o_2, \cdots, o_T$, then tested against each bigram model $\Lambda_i, 1 \le i \le N$. The results of testing are likelihood probabilities:

$$\Pr(\mathbf{O} \mid \Lambda) = \Pr(o_1) \cdot \prod_{t=2}^{T} \Pr(o_t \mid o_{t-1}) \ . \tag{8}$$

5. Experiments

5.1 Bird Sound Data

The bird sound data used in this study stem from the commercial CDs and websites listed in Table 1. To facilitate the experiments, all of the sound data were converted into PCM WAV with 22.05-kHz sampling rate and 16-bit quantization resolution. We chose ten bird species commonly seen in the Taipei urban area, including *Dicrurus aeneus*, *Dendrocopos canicapillus*, *Pomatorhinus ruficollis*, *Stachyris ruficeps*, *Megalaima oorti*, *Heterophasia auricularis*, *Hypsipetes madagascariensis*, *Myiophonus insularis*, *Otus spilocephalus*, and *Dendrocitta formosae*. The data were divided into two subsets, training and testing. The amount of sound data with respect to each bird species is listed in Table 2.

Table 1. Source of our bird sound data

CDs or Websites	Audio Types
"Songbirds" CDs published by WIND RECORDS CO., LTD.	44.1kHz Sampling Rate 16-bit Quantization Resolution CDs
"Birdwatcher's guide to the Taipei region" CDs published by Department of Information and Tourism,Taipei City Government	44.1kHz Sampling Rate 16-bit Quantization Resolution CDs
http://archive.zo.ntu.edu.tw/	44.1kHz Sampling Rate 16-bit Quantization Resolution WAV Files
http://macaulaylibrary.org/index.do	Streaming Audio

Table 2. The amount of sound data with respect to each bird species

Bird Species	**Training Data:** Total Duration (sec)	**Testing Data:** No. of Sound Samples
Dicrurus aeneus	130	77
Dendrocopos canicapillus	35	102
Pomatorhinus ruficollis	125	155
Stachyris ruficeps	66	81
Megalaima oorti	93	219
Heterophasia auricularis	91	86
Hypsipetes madagascariensis	42	157
Myiophonus insularis	27	25
Otus spilocephalus	27	111
Dendrocitta formosae	39	73

5.2 Experiment Results

Our experiments were conducted to examine the timbre-based component and pitch-based component separately before evaluating if the performance of bird sound identification could be further improved by combining the two components. The performance was characterized with the accuracy:

$$\text{Accuracy (in\%)} = \frac{\text{Tonal number of correctly - identified recordings}}{\text{Tonal number of testing recordings}} \times 100\%$$

5.2.1 Accuracies Obtained with the Timbre-based Analysis

In the timbre-based analysis, the MFCC feature vectors, each consisting of 20 coefficients, were extracted from the bird sound data, using a 30-ms Hamming-windowed frame with 15-ms frame shifts. The FFT size was set to be 2048. Table 3 shows the identification accuracies obtained with various numbers of mixture Gaussian densities used in GMM. The best accuracy in Table 3 is 71.1%, achieved with 64 mixtures. Table 4 shows the confusion matrix of the identification for the case of 64 mixtures. We can see from Table 4 that the timbre-based analysis performs best in identifying *Pomatorhinus ruficollis*, whereas it performs worst in identifying *Dendrocitta formosae*.

Table 3. Identification accuracies (in %) obtained with various numbers of mixture Gaussian densities used in GMM.

# mixtures	**4**	**8**	**16**	**32**	**64**	**128**
Dicrurus aeneus	55.8	58.4	58.4	59.7	64.9	62.3
Dendrocopos canicapillus	59.8	59.8	60.8	62.7	62.7	62.7
Pomatorhinus ruficollis	80	81.9	83.2	83.2	82.6	81.9
Stachyris ruficeps	69.1	69.1	70.4	71.6	70.4	69.1
Megalaima oorti	72.1	73.1	74	74.4	74.9	74.9
Heterophasia auricularis	74.4	75.6	78	77.9	76.7	76.7
Hypsipetes madagascariensis	62.4	63.7	65	66.2	68.8	67.5
Myiophonus insularis	68	72	76	76	76	76
Otus spilocephalus	64	64	64	64	67.6	65.8
Dendrocitta formosae	50.7	52.1	56.2	57.5	56.2	54.8
Average Accuracy	**67.1**	**68.2**	**69.5**	**70.3**	**71.1**	**70.3**

Table 4. Confusion matrix of the identification for the case of 64 mixtures.

True \ Identified	*Dicrurus aeneus*	*Dendrocopos canicapillus*	*Pomatorhinus ruficollis*	*Stachyris ruficeps*	*Megalaima oorti*	*Heterophasia auricularis*	*Hypsipetes madagascariensis*	*Myiophonus insularis*	*Otus spilocephalus*	*Dendrocitta formosae*
Dicrurus aeneus	64.9	10.4	0	9.1	0	6.5	6.5	0	2.6	0
Dendrocopos canicapillus	9.8	62.7	14.7	2.9	3.9	2.9	2.9	0	0	0
Pomatorhinus ruficollis	0	6.5	82.6	0	0	0	0	6.5	0	4.5
Stachyris ruficeps	7.4	3.7	6.2	70.4	6.2	3.7	0	2.5	0	0
Megalaima oorti	0	0.9	0	0	74.9	0	13.7	0	6.8	3.7
Heterophasia auricularis	3.9	0	8.1	0	0	76.7	5.8	5.8	0	0
Hypsipetes madagascariensis	0	6.4	0	4.5	8.3	0	68.8	2.5	15.9	0
Myiophonus insularis	0	0	16	0	0	8	0	76	0	0
Otus spilocephalus	2.7	13.5	0	9	0	7.2	0	0	67.6	0
Dendrocitta formosae	2.7	0	11	0	0	0	21.9	0	8.2	56.2

5.2.2 Accuracies Obtained with the Pitch-based Analysis

We then tested the pitch-based analysis component. The length of frame and FFT size were the same as the settings in computing MFCCs. Table 5 shows the resulting confusion matrix of the identification. We obtained an average identification accuracy of 72.0%, which is slightly higher than that obtained with the timbre-based analysis. Comparing Tables 4 and 5, we can see that the misidentified cases for timbre-based analysis and pitch-based analysis are different. This indicates that combined use of the two components would achieve higher identification accuracy than the use of an individual component.

5.2.3 Combined Use of the Timbre-based and Pitch-based Analyses

Finally, we examined the proposed system based on the combination of timbre-based analysis and pitch-based analysis. Table 6 shows the identification accuracies obtained with different settings in the value of α and β. We can see from Table 6 that the combined use of timbre-based analysis and pitch-based analysis does perform better than both timbre-based analysis and pitch-based analysis used solely. It also can be seen that the resulting accuracies are not sensitive to the values of α and β, as long as they are set to a certain range. Table 7 shows the confusion matrix of the identification for the case of $\alpha = 0.4$ and $\beta = 0.6$, which

achieves an average accuracy of 75.0%. We can see from Table 7 that the overall system improves the accuracies of identifying almost every bird species, compared to Tables 4 and 5. This result confirms the validity of the proposed system.

Table 5. Confusion matrix of the identification using pitch-based analysis.

True \ Identified	*Dicrurus aeneus*	*Dendrocopos canicapillus*	*Pomatorhimus ruficollis*	*Stachyris ruficeps*	*Megalaima oorti*	*Heterophasia auricularis*	*Hypsipetes madagascariensis*	*Myiophomus insularis*	*Otus spilocephalus*	*Dendrocitta formosae*
Dicrurus aeneus	61	19.9	0	10.4	0	5.2	3.9	0	0	0
Dendrocopos canicapillus	2.9	71.6	12.7	0	2	0	3.9	0	2.9	3.9
Pomatorhinus ruficollis	0	7.7	82.9	0	0	0	0	7.7	0	1.9
Stachyris ruficeps	1.2	0	7.4	75.3	2.5	1.2	0	0	0	0
Megalaima oorti	0	1.4	0	1.8	82.2	0	11.4	0	2.7	0.5
Heterophasia auricularis	0	0	11.6	0	0	76.7	5.8	5.8	0	0
Hypsipetes madagascariensis	0	6.4	0	2.5	9.6	0	63.1	2.5	15.9	0
Myiophonus insularis	0	0	4	0	12	28	0	56	0	0
Otus spilocephalus	7.2	21.6	5.4	0	0	0.1	0	0	64.9	0
Dendrocitta formosae	5.4	0	8.2	0	0	0	24.7	0	2.7	58.9

Table 6. Accuracies (in %) obtained with different settings in the value of α and β.

α \ β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.9	73.3	-	-	-	-	-	-	-	-
0.8	-	74.0	-	-	-	-	-	-	-
0.7	-	-	74.1	-	-	-	-	-	-
0.6	-	-	-	74.6	-	-	-	-	-
0.5	-	-	-	-	74.9	-	-	-	-
0.4	-	-	-	-	-	75	-	-	-
0.3	-	-	-	-	-	-	74.8	-	-
0.2	-	-	-	-	-	-	-	74.7	-
0.1	-	-	-	-	-	-	-	-	73.1

Table 7. Confusion matrix of the identification for the case of α = 0.4 and β = 0.6

True \ Identified	*Dicrurus aeneus*	*Dendrocopos canicapillus*	*Pomatorhinus ruficollis*	*Stachyris ruficeps*	*Megalaima oorti*	*Heterophasia auricularis*	*Hypsipetes madagascariensis*	*Myiophonus insularis*	*Otus spilocephalus*	*Dendrocitta formosae*
Dicrurus aeneus	67.5	13	0	9.1	0	5.2	2.6	0	2.6	0
Dendrocopos canicapillus	2.9	75.5	9.8	0	0.1	0	3.9	0	2.9	3.9
Pomatorhinus ruficollis	0	5.8	85.2	0	0	0	0	5.8	0	3.2
Stachyris ruficeps	1.2	0	6.2	75.3	2.5	1.2	0	1.2	0	0
Megalaima oorti	0	1.4	0	1.8	83.1	0	9.1	0	3.2	1.4
Heterophasia auricularis	0	0	10.5	0	0	80.2	4.7	3.5	0	0
Hypsipetes madagascariensis	0	6.4	0	1.3	9.6	0	65.6	1.3	15.9	0
Myiophonus insularis	0	0	4	0	4	12	0	80	0	0
Otus spilocephalus	7.2	21.6	5.4	0	0	1.8	0	0	64	0
Dendrocitta formosae	4.1	0	5.5	0	0	0	21.9	0	2.7	65.8

6. Conclusion

This work has developed an automatic bird sound identification system, with the motivation of helping people learn to identify bird species from their sounds. The system is built on speech recognition techniques, along with specific tailoring to handle the bird sound characteristics. Two acoustic cues were investigated for analysis, timbre and pitch. In the timbre-based analysis, we used MFCCs to characterize the bird sound. Then, GMMs were used to represent the MFCCs as a set of parameters. In the pitch-based analysis, we converted bird sounds from their waveform representations into a sequence of MIDI notes. Then, Bigram models were used to capture the dynamic change information of the notes. Our experiments, conducted using audio data of the ten most common bird species in the Taipei urban area, show that the timbre-based, pitch-based, and the combined system achieves 71.1%, 72.1%, and 75.0% accuracy of bird sound identification, respectively.

Despite the potential, the performance of the proposed bird sound identification system still leaves considerable room for improvement. In the future, we will try to include more characteristics of bird sounds, such as the concept of bird calls and bird songs, into our system design. In addition, we have to scale up our sound database to hundreds or thousands of bird

species to validate the proposed identification system.

Acknowledgement

This work was supported in part by the National Science Council, Taiwan, under Grant No. NSC 99-2628-E-027-005.

References

Anderson, S. E., Dave, A. S., & Margoliash, D. (1996) .Template-based automatic recognition of birdsong syllables from continuous recordings. *J. Acoust. Soc. Amer.*, *100*(2), 1209-1219.

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. on Acoustic, Speech and Signal Processing.*, *28*(4), 357-366.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.*, *39*, 1-38.

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken Language Processing*, Prentice Hall.

Kogan, J., & Margoliash, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. *J. Acoust. Soc. Amer.*, *103*(4), 2187-2196.

McIlraith, A. L., & Card, H. C. (1997). Birdsong recognition using backpropagation and multivariate statistics. *IEEE Trans. Signal Process.*, *45*(11), 2740-2748.

Piszczalski, M., & Galler, B. A. (1979). Predicting musical pitch from component frequency ratios. *Journal of the Acoustical Society of America*, *66*(3), 710-720.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.*, *3*(1), 72-83.

Reynolds, D., & Quatieri, T. (2000). Speaker Verification Using Adapted Gaussian ixture Models. *Digital Signal Processing, 10*, 19-41.

Somervuo, P., Härmä, A., & Fagerlund, S. (2006). Parametric representations of bird sounds for automatic species recognition. *IEEE Trans. Audio, Speech, Language Process.*, *14*(6), 2252-2263.

Yu, H. M., Tsai, W. H., & Wang, H. M. (2008). A query-by-Singing system for retrieving karaoke music. *IEEE Trans. Multimedia*, *10*(8), 1626-1637.