行政院國家科學委員會專題研究計畫 成果報告

以鳥鳴聲為基礎之鳥種辨識系統架構之研究與開發(I)
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 97-2221-E-216-014-
執 行 期 間 ：97 年 08 月 01 日至 98 年 10 月 31 日
執 行 單 位 ：中華大學資訊工程學系

計 畫 主 持 人：周智勳

計畫參與人員：此計畫無其他參與人員

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

中 華 民 國 98 年 12 月 31 日

行政院國家科學委員會補助專題研究計畫成果報告

以鳥鳴聲為基礎之鳥種辨識系統架構之研究與開發(I)

計畫類別：個別型計畫
計畫編號：NSC97-2221-E-216-014
執行期間：97 年 08 月 01 日至 98 年 10 月 31 日
計畫主持人：周智勳
執行單位：中華大學資訊工程系

Abstract

Feature extraction and classification are two important aspects of an automatic birdsong recognition system. To improve the recognition rate, in this project, a novel feature extraction process based on the Mel-frequency cepstral coefficients and a decision-based neural network classifier with a suitable reinforcement learning rule were developed. The proposed system was applied to two problems, one was the recognition of a set of arbitrary syllables and the other was the recognition of a section of a birdsong. Experimental results, with many comparisons, showed the efficiency of the proposed method.

Keywords: Birdsong recognition, syllable segmentation, Mel-frequency cepstral coefficients, wavelet transformation, decision based neural network.

1. Introduction

Birdsongs are typically divided into four hierarchical levels: note, syllable, phrase, and song [1], of which syllable plays an important role in bird species recognition. To recognize the syllables of two bird species, the DTW algorithm was used [2]. The idea of harmonic spectrum structures was used in [3] to classify four types of syllables. A template-based technique combining time delay neural networks (TDNNs) was proposed in [4] to automatically recognize the syllables of 16 bird species. In [5], syllables were used to solve the problem of the overlapping sound waveforms of multiple birds. A study was done [6] focusing on the histogram of consecutive syllables, called the syllable pair histogram. Similar histograms were used to construct the Gaussian prototypes of a birdsong. In [7]-[9], the number of syllables and the mean and deviation of syllable lengths were combined with other features to form the feature vectors for birdsong recognition.

One feature that has been successfully applied in human voice recognition is the cepstral of the voice waveform, of which the Mel-Frequency Cepstral Coefficients (MFCCs) were obtained based on the fact that the hearing perception of the human being performs better than the Linear Prediction Cepstral Coefficients (LPCCs) [10]-[13]. Using the application of MFCCs, syllables were segmented by using an energy index, and each syllable was partitioned into several frames with which the means of both LPCCs and MFCCs were computed to form the feature vector of the syllable [14]. MFCCs were also combined with the descriptive parameters such as the spectral centroid, the signal bandwidth, the zero crossing rate and the short time energy to form the feature vector [15].

Although MFCCs have been well-applied in bird species recognition, further study on this feature is necessary to increase the recognition rate. Furthermore, the enhancement of the preprocessing as well as the classifier also needs to be studied. In this study, a novel feature extraction process based on the MFCCs and a decision based neural network classifier with suitable reinforcement learning rule were developed to construct an automatic bird species recognition system. The proposed system was applied to two types of birdsong recognition problems with 420 bird species.

The rest of this paper is divided as follows: Section 2 gives a brief review of related studies and problem formulation. The structure of the proposed system is described in section 3. Experimental results are shown in Section 4. Section 5 is the conclusion.

2. Related Studies and Problem Formulation

Although MFCCs have been successfully applied in voice recognition, further study of the computation algorithm is still needed. For example, in [11]-[13], [16], [17] optimal theories were used to obtain the center frequencies and bandwidths of the triangular filters; the discrete cosine transform (DCT) was replaced with the wavelet transform in [18], filter weighting was applied in [19] to assign a weight for each order of MFCCs, and in [20] the MFCCs as well as their first-order and second-order differences were used to form the feature vector.

Usually, the basic unit for computing the MFCCs is the frame in the short time Fourier transform, and the MFCCs of a section of a voice signal containing several frames are combined to compute the feature vector. On the other hand, as described in the Introduction, of the four hierarchical levels of a birdsong, syllable plays the most important role in birdsong recognition. So the objective of this study was to improve the MFCCs and use the results to construct the feature vector of a section of birdsong containing several syllables.

3. Proposed Birdsong Recognition System

The block diagram of the proposed system, shown in Fig. 3.1, entails preprocessing, feature extraction and

recognition (species decision) as described in detail in the following.

3.1 *Preprocessing*

Preprocessing filters the signal and properly segments the syllables for feature extraction. The procedure for preprocessing in this study is shown in Fig. 3.2 containing four steps: syllable endpoint detection, normalization, pre-emphasis and segmentation.

3.1.1 Syllable endpoint detection

The applied syllable endpoint detection method is described in the following.

1. Compute the short time Fourier transform of $x(t)$ with frame size $N = 512$, and form the spectrogram of the signal. The Hamming window for short time analysis has the form of
$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0 \qquad\qquad\qquad\qquad , \text{otherwise} \end{cases} \quad (3\text{-}1)$$
2. For each frame m, find the frequency Bin bin_m with the greatest magnitude.
3. Initialize the syllable index j, $j = 1$.
4. Compute the frame t at which the maximum magnitude occurs
$$t = \arg\max_{1 \le m \le M}(|X[bin_m]|), \quad (3\text{-}2)$$
and set the amplitude of syllable j as
$$A_j = 20 \cdot \log_{10}|X[bin_t]|(\text{dB}), \quad (3\text{-}3)$$
in which M is the number of frames of $x(t)$, and $X[\cdot]$ denotes the spectrum of $x(t)$.
5. Start from frame t and move backward and forward up to frames h_j and t_j such that both $20 \cdot \log_{10}|X[bin_{h_j}]|$ and $20 \cdot \log_{10}|X[bin_{t_j}]|$ are smaller than $(A_j - 20)$ (dB).
6. Start from frames h_j and t_j, find frames $h_j - \alpha$ and $t_j + \beta$ (α, β > 0) such that both $20 \cdot \log_{10}|X[bin_{h_j-\alpha-1}]|$ and $20 \cdot \log_{10}|X[bin_{t_j+\beta+1}]|$ are greater than $(A_j - 20)$. Then $h_j - \alpha$ and $t_j + \beta$ are called the head frame and tail frame of syllable j.
7. Set $|X[bin_m]| = 0, m = h_j - \alpha, h_j - \alpha + 1, \cdots, t_j + \beta - 1, t_j + \beta$. (3-4)
8. Let $j = j + 1$.
9. Repeat Step 4 to Step 8 until $A_j < A_1 - 20$.

3.1.2 Normalization and pre-emphasis

The normalization process regulates the discrepancy of the voice amplitudes caused by the diversity of recording environments. In this study, the amplitudes were linearly normalized to the region of [-1, 1]. Besides, since the amplitude of the high frequency is usually much smaller than that of the low frequency, pre-emphasis was applied to intensify the signal of the high frequencies. The intensification was accomplished by a finite impulse response (FIR) filter with the following form

$$H(z) = 1 - a \cdot z^{-1}, \quad (3\text{-}5)$$

so that a signal $x[n]$ after the filtering process has the following property

$$\bar{x}[n] = x[n] - ax[n-1], \quad (3\text{-}6)$$

in which α is a scalar usually between 0.9 and 1, which was set as 0.95 in this study.

3.1.3 Segmentation

Focusing on the recognition of a section of birdsong, the segmentation process in this system aimed at segmenting a period of syllables rather than an individual syllable similar to many other researches [2]-[4], [7]-[9]. Extracting the feature vector of a period of syllables is practical for birdsong recognition because the syllables of a birdsong are usually repetitive. After endpoint detection, normalization and pre-emphasis, the segmentation process was done by detecting the repetition of syllables as described in the following

1. Set $i = 1$ the index of the first syllable of the segmentation.
2. Find the nearest syllable j such that the similarity between syllables i and j, sim_{ij}, is smaller than α, j is the last syllable of the segmentation.
3. Set the segmentation length $l = j$.
4. Set $k = j + 1$, the checking index.
5. Set $i = 1$, $l = j$.
6. Compute the similarity, sim_{ki}, between syllable k and syllable i.
7. If $sim_{ki} > \alpha$ (the same type), then
   If $l = k - j$ then Stop. The segmentation is from syllable 1 to syllable l.
   If $i = j$, then $j = j + 1$ go to Step 5.
   set $i = i + 1$ and $k = k + 1$, and go to Step 6.
8. If $i = j$, then $j = j + 1$ go to Step 5.
9. Set $k = k + 1$, $j = j + 1$, $l = l + 1$ and go to Step 6.

The similarity between two syllables was determined by computing the difference between the amplitudes of corresponding frequency Bins. A normalization process was applied for the syllable lengths before the computation. Since the syllable types of a birdsong are usually within 6, in our experiment, α was set as $2 \le l \le 8$. In fact, syllables of the same type in a section of a birdsong usually have small variances in amplitude and length. After segmentation, the segmented syllables were aligned for feature extraction.

3.2 *Feature extraction – the WMFCCs*

After segmenting a period of syllables, the aligned syllables were used to compute the feature vector of the birdsong. The process for obtaining the feature vector WMFCCs is shown in Fig. 3.3 and is described below.

3.2.1 Computing the MFCCs of each frame

The steps for computing the MFCCs of each frame are as follows:

1. Compute the fast Fourier transform (FFT) of each

framed signal.

$$\tilde{x}[k]=\sum_{n=0}^{N-1}x[n]w[n]e^{-j2\pi nk/N},\ 0\le k<N. \quad (3\text{-}7)$$

2. Compute the energy of each triangular filter band

$$E_j=\sum_{k=0}^{N/2-1}\phi_j[k]\left|\tilde{x}[k]\right|^2,\ 0\le j<J, \quad (3\text{-}8)$$

where $\phi_j[k]$ denotes the amplitude(weight) of the j^{th} triangular filter at frequency bin k, E_j denotes the energy of j^{th} filter band, and J is the number of triangular filters.

3. Compute the MFCCs by Cosine transformation

$$c_i(m)=\sum_{j=0}^{J-1}\cos\left(m\frac{\pi}{J}(j+0.5)\right)\log_{10}(E_j), \quad (3\text{-}9)$$

where $c_i(m)$ denotes the m^{th} order MFCC of the i^{th} frame.

Although MFCCs have been well-applied in human voice recognition, the process for birdsong recognition needs to be improved because there is a greater diversity in the vocalizations of different bird species. In the following, an improvement of the MFCCs, called the WMFCCs, is obtained to form the feature vector.

3.2.2 Construct the feature vector by using the WMFCCs

After obtaining the MFCCs of each frame of the aligned birdsong signal, the feature vector of the birdsong was obtained by computing the WMFCCs as described in the following.

1. Collect the first five order MFCCs of all frames of the aligned signal.

$$\{c_1(0),c_1(1),...,c_1(4),c_2(0),...,c_2(4),...,c_i(0),...,c_i(4),...\}. \quad (3\text{-}10)$$

In many studies, the first fifteen order MFCCs have been used to form the feature vector. In this study, for reducing the computation complexity, only the first five orders were used.

2. Align the MFCCs of the same order.

$$s_m=c_1(m),c_2(m),...,c_i(m),...,\ m=0,...,4. \quad (3\text{-}11)$$

3. Compute 3-level wavelet transformation of s_m as shown in Fig. 3.4. The transformation equations are

$$a[n]=\sum_{k=-\infty}^{\infty}h_0[k]s_m[2n-2k], \quad (3\text{-}12a)$$

$$d[n]=\sum_{k=-\infty}^{\infty}h_1[k]s_m[2n-2k], \quad (3\text{-}12b)$$

where $a[n]$ and $d[n]$ denote the low frequency and high frequency components of s_m, and $h_0[n]$ and $h_1[n]$ are the applied low pass and high pass filters in the transformation. The coefficients of these two filters are [21], [22]

$$h_0[n]=[0.3327,0.8086,0.4599,-0.1350,-0.0854,0.0352], \quad (3\text{-}13a)$$

$$h_1[n]=[0.0352,0.0854,-0.1350,-0.4599,0.8086,-0.3327]. \quad (3\text{-}13b)$$

The resulted six sequences of the transformation, called the WMFCCs of s_m, are denoted as s_m^{LLL}, s_m^{LLH}, s_m^{LH}, s_m^{HL}, s_m^{HHL} and s_m^{HHH}.

4. Compute the means of each of the six sequences and denote them as ms_m^{LLL}, ms_m^{LLH}, ms_m^{LH}, ms_m^{HL}, ms_m^{HHL} and ms_m^{HHH}.

5. Form the feature vector by using the six mean values of all the first five order MFCC sequences

$$[ms_0^{LLL}\ ...\ ms_0^{HHH}\ ms_1^{LLL}...ms_1^{HHH}...ms_4^{LLL}...ms_4^{HHH}]^T. \quad (3\text{-}14)$$

An example of the birdsong of the Accipiter nisus is shown in Fig. 3.5, in which part (a) shows the first-order MFCCs of a birdsong segment, and the resulting six WMFCC sequences are shown in part (b). By using the feature vectors obtained in Step 5, the recognition process was achieved by applying the DBNN classifier as described in the next section.

3.3 *Recognition by using the decision based neural network* (*DBNN*)

When applying neural networks (NNs), one of the two learning types, either supervised or unsupervised learning, is adopted. Supervised NNs can also be divided into two types, namely approximation-based formulation and decision-based formulation [23] depending on the arrangement of training data. The function of approximation-based formulation is to approximate the mapping between input and output data so as to minimize the mean square error between the network outputs and the desired outputs. An example of such a type NN is the back-propagation NN with a least mean squares learning rule. The decision-based formulation is usually used to decide which class the input data belongs to meaning that it is more suitable for data classification [23]. The structure of the decision-based neural network (DBNN) applied in this study is shown in Fig. 3.6, in which the weight vectors $\mathbf{w}_j^i$ of the function ϕ_j^i were trained by using reinforcement type learning rules. The training process is described below.

1. Initialize the weight vectors $\mathbf{w}_j^i$ of the function ϕ_j^i, $i=1,2,...,C$, $j=1,2,3$ with random values, where ϕ_j^i is a radius basis function of the form

$$\phi_j^i=\phi(\mathbf{x},\mathbf{w}_j^i)=-\frac{\left\|\mathbf{x}-\mathbf{w}_j^i\right\|^2}{2}, \quad (3\text{-}15)$$

C is the number of classes and **x** is the input feature vector.

2. Input the feature vector, **x**, of the training birdsong whose bird species class is set as the *actual class*.
3. Compute the value of each basis function ϕ_j^i.
4. For each class network i, find the local winner $l_i=\arg\max_j(\phi_j^i)$ and $\phi_{l_i}^i=\max_j(\phi_j^i)$.
5. For all local winners $\phi_{l_i}^i$, $i=1,2,...,C$, find the global winner, $g=\arg\max_i\phi_{l_i}^i$ and $\phi^g=\max_i\phi_{l_i}^i$.
6. If g = *actual class*, then perform the reinforcement

learning

$$\mathbf{w}_{l_g}^{g} = \mathbf{w}_{l_g}^{g} + \eta \nabla \phi(\mathbf{x}, \mathbf{w}_{l_g}^{g}), \quad (3\text{-}16a)$$

else the anti-reinforcement learning

$$\mathbf{w}_{l_g}^{g} = \mathbf{w}_{l_g}^{g} - \eta \nabla \phi(\mathbf{x}, \mathbf{w}_{l_g}^{g}), \quad (3\text{-}16b)$$

where $\mathbf{w}_{l_g}^{g}$ denotes the weight vector of $\phi_{l_g}^{g}$ and η is the learning rate.

In this study, the function ϕ_j^i was defined as a radius basis function as shown in (3-15), so the gradient operation result in (3-16) was

$$\nabla \phi(\mathbf{x}, \mathbf{w}_{l_g}^{g}) = \mathbf{x} - \mathbf{w}_{l_g}^{g}. \quad (3\text{-}17)$$

In the recognition process, the feature vector of a test birdsong was obtained by the same process as the training part. After inputting the feature vector to the DBNN, the global winner of the network (i.e. the network output) indicated the species class the test birdsong belonged to.

4. Experimental results

The bird species vocalization database used in this study was obtained from a commercial CD [24] containing both birdcall and birdsong files of 420 bird species recorded in the field in Japan. Each file contained vocalizations of the same bird species. The database of 420 bird species made it much larger than any other used in previous studies. Meanwhile, recordings in the field were usually in a noisy environment, incomplete and interrupted. Because of this, sometimes, only some syllables of a birdsong, rather than a complete birdsong, can be used in the recognition process. So, two recognition problems were applied in the experiments, one was the recognition of a set of arbitrary syllables of a bird species and the other was the recognition of a section of a birdsong. The sampling rate of these vocalization signals was 44.1 kHz with 16-bit resolution and a monotone type PCM format. In the experiment, the frame size was set as 512 samples with three-fourths frame overlapping.

4.1 *Recognition of a set of arbitrary syllables from the song of a bird species*

The purpose of the first experiment was to examine the efficiency of the DBNN. After segmentation, half the syllables of each birdsong file were randomly selected for training and the remaining for testing. The MFCCs of each frame of a training syllable were obtained by using the steps in sec. 3.2.1. After computing the first 15 order MFCCs of each frame, the coefficients of the same order of all frames were averaged and then used to form a 15-dimensional syllable feature vector. Such a feature vector had been successfully used in many studies, so it was used for checking. In the NN training process, the feature vector of the training syllable was used as the input and the corresponding bird species as the desired output. Because back-propagation NN (BPNN) was widely applied in many studies, both BPNN and DBNN were applied in this experiment for comparison.

For recognition, the same feature extraction process was applied to the test syllables. The extracted feature vector of a test syllable was used as the input of the NN whose output indicated the bird species of the test syllable. By comparing the network output and the actual species for each test syllable, the recognition rates (*RR*s) of all test syllables were obtained. Table 4.1 shows the *RR*s of both types of NN. It shows that the DBNN improved the *RR* from about 5% on the BPNN.

4.2 *Recognition of a section of a birdsong*

The second experiment was the recognition of a section of a birdsong containing several consecutive syllables. In this case, both types of feature vectors included the MFCCs of a single syllable (FV1) and the WMFCCs of a section of birdsong (FV2) were applied for comparison. Meanwhile, both types of NN were used. In the experiment, half of each birdsong file was used for training and the remaining for testing. By using the segmentation process, both training and testing data sets contained several birdsong sections of each bird species. When FV1 was applied, all the syllables in a birdsong section (training or testing) were treated individually. In the recognition of a birdsong section, the species was determined by finding the species class with the largest number of NN output. For FV2, the feature vector of a period of syllables in the birdsong section (training or testing) was used as the input of NN. In the recognition process, the NN output indicated the species class of the test birdsong section. In this experiment, the recognition rate *RR* was defined as

$$RR(\%) = \frac{\text{number of test birdsong sections recognized correctly}}{\text{number of test birdsong sections}} \cdot 100\% \quad (4\text{-}1)$$

The *RR*s of all four structures are shown in Table 4.2. It was found that when the segmentation method and the feature vector were fixed, the DBNN improved the *RR*s of 10.68%, 7.35%, 4.48% and 4.72% on BPNN. When the segmentation method and the NN were fixed, the proposed WMFCCs improved the *RR*s of 18.72%, 17.38%, 12.34% and 14.75% on the MFCCs. The above comparisons exhibited the efficiency of the proposed methods, especially the feature extraction. When the proposed two methods FV2 and DBNN were combined, the system achieved an *RR* of 77.89%.

4.3 *Recognition with threatened birdsongs taken into account*

The territorial instinct is innate in some bird species so that they will make threatening noises when there is an intruder. The threatening voice is used to warn other birds in the same group and frighten the invaders away. Such vocalization is usually distinct from the regular birdsong requiring additional categorization. That is, two types of birdsongs are considered if the bird species also makes threatening noises. In this case, the feature extraction process was applied for both regular birdsongs and threatening noises so that some bird species were represented by two types of feature vectors. In the recognition process, the feature vector of the test birdsong section was matched by the NN to the feature vectors of all species to determine the most likely species.

The *RR*s of all four structures used to recognize

birdsongs with the above distinction are shown in Table 4.3, in which titles A and B represent the recognition both without and with the threatening sounds. The results showed that the *RRs* of case B exceeded case A from about 3.56% to 5.22%, and an *RR* of 83.11% was obtained when the proposed two methods were combined.

5. **Conclusion**

The kinds and numbers of birds are good indices in the study of species diversity. So, the investigation of bird species diversity is key in monitoring environment and ecosystem recovery, and automatic bird species recognition has become an invaluable study method in the long-term investigation of bird species. However, recordings in the field are usually in a noisy environment, incomplete or interrupted making birdsong recognition much harder. In order to overcome these problems, both feature extraction and classification were studied. For feature extraction, a novel method for a section of birdsong was proposed. For the study of a classifier, a DBNN with reinforcement learning rules was presented and compared to the BPNN. Two types of recognition problems were considered in the experiments, one was the recognition of a set of arbitrary syllables and the other was the recognition of a section of a birdsong. Experimental results showed that the proposed methods evidently improved the *RRs*, regardless of the presence of threatening noises.

References

[1] C.K. Catchpole and P.J.B. Slater, *Bird Song*: *Biological Themes and Variations*, Cambridge University Press, 1995.

[2] S.E. Anderson, A.S. Dave and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 1209-1219, 1996.

[3] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proceedings of the IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, vol. 5, pp. V-701-4, 2004.

[4] S.A. Selouani, M. Kardouchi, E. Hervet and D. Roy, "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *Proceedings of the ICSC Congress on Computational Intelligence Methods and Applications*, pp. 1-6, 2005.

[5] A. Härmä, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, vol. 5, pp. 545-548, 2003.

[6] P. Somervuo, "A. Härmä, Bird song recognition based on syllable pair histograms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. V-825-8, 2004.

[7] A.L. McIlraith and H.C. Card, "Bird song identification using artificial neural networks and statistical analysis," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, vol. 1, pp. 63-66, 1997.

[8] A.L. McIlraith and H.C. Card, "A comparison of backpropagation and statistical classifiers for bird identification," in *Proceedings of IEEE International Conference on Neural Networks*, vol. 1, pp. 100-104, 1997.

[9] A.L. McIlraith and H.C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2740-2748, 1997.

[10] S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions, Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357-366, Aug 1980.

[11] C.H. Lee, D.H. Hyun, E.S. Choi, J.W. Go and C.Y. Lee, "Optimizing feature extraction for speech recognition," *IEEE Transactions*, *Speech and Audio Processing,* vol. 11, pp. 80-87, Jan. 2003.

[12] S.M Lee, S.H. Fang, J.W. Hung and L.S. Lee, " Improved MFCC feature extraction by PCA-optimized filter-bank for speech recognition," *IEEE Workshop*, *Automatic Speech Recognition and Understanding*, pp. 49-52 Dec. 2001.

[13] M.D. Skowronski, and J.G. Harris, "Increased MFCC filter bandwidth for noise-robust phoneme recognition," *Proceedings of the IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing,* vol. 1, pp. 801-804, 2002.

[14] R.Z. Huang, *Automatic recognition of bioacoustic sounds*, Master Thesis, Department of Computer Science and Information Engineering, Chung Hua University, Taiwan, R.O.C., 2004.

[15] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 38637, 8 pages.

[16] M.D. Skowronski and J.G. Harris, "Improving the filter bank of a classic speech feature extraction algorithm," *Circuits and Systems*, vol. 4, pp. 281-284, May 2003.

[17] S.E. Bou-Ghazale and J.H.L. Hansen., "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 429-442, July 2000.

[18] L.P. Ricotti, "Multitapering and a wavelet variant of MFCC in speech recognition," *IEE Proceedings - Vision*, *Image and Signal Pro*cessing, pp. 29-35, Feb. 2005.

[19] W.W. Hung and H.C. Wang, "On the use of weighted filter bank analysis for the derivation of robust MFCCs," *IEEE Signal Processing Letters,* vol. 8, pp. 70-73, March 2001.

[20] C. Kwan et al., "An automated acoustic system to monitor and classify birds," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 96706, pp. 1-19.

[21] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909-996, Oct. 1988.

[22] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, 1992.

[23] S.Y. Kung, *Digital Neural Networks*, Prentice-Hall, Inc., 1993.

[24] T. Kabaya and M. Matsuda, *The Songs & Calls of 420 Birds in Japan*, SHOGAKUKAN Inc., Tokyo, 2001.

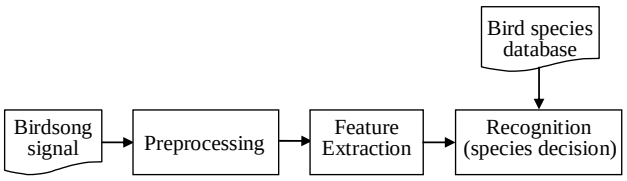


Figure 3.1 Block diagram of the proposed system

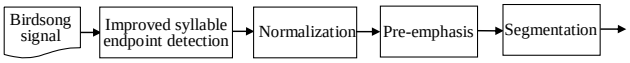


Fig. 3.2 Block diagram of the preprocessing

Figure 3.3 Flow chart of computing the feature vector WMFCCs

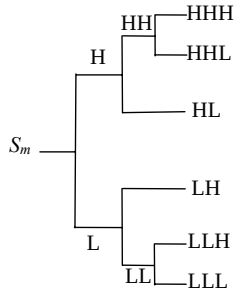
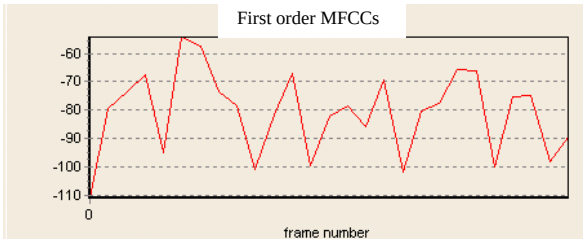
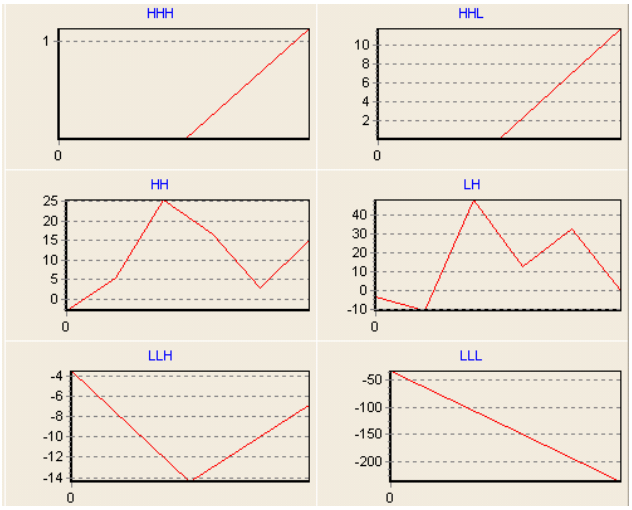


Figure 3.4 Applied 3-level wavelet transformation.



(a) First-order MFCCs of the birdsong segment.



(b) Six WMFCC sequences obtained from (a)

Figure 3.5 WMFCC sequences obtained from a birdsong segment of the Accipiter nisus.

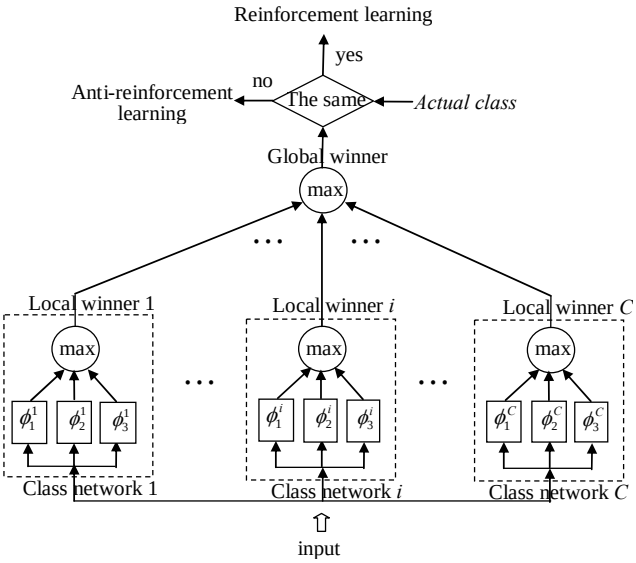


Figure 3.6 Structure of the applied DBNN.

Table 4.1 *RR*s of using different types of NN

MFCC+BP	MFCC+DBNN
59.76%	64.93%

Table 4.2 *RR*s of using various structures

Structure	***RR***
FV1+BPNN	54.69%
FV1+DBNN	65.37%
FV2+BPNN	73.41%
FV2+DBNN	77.89%

Table 4.3 *RR*s of using various structures

Structure	**A**	**B**
RS+FV1+BPNN	54.69%	59.16%
RS+FV1+DBNN	65.37%	69.24%
RS+FV2+BPNN	73.41%	76.97%
RS+FV2+DBNN	77.89%	83.11%

A: Recognition without the distinction of threatened voices.
B: Recognition with the distinction of threatened voices.

計畫成果自評：

1. 研究內容與原計畫相符程度、達成預期目標情況：

   本計劃完成新類型特徵向量的攫取(應用wavelet transformation於MFCCs中)，並應用新型式分類器，對鳥鳴聲辨識率有所貢獻。

2. 研究成果之學術或應用價值、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等。

   本計劃研究成果之一部份發表於下列國際研討會中：

   Chih-Hsun Chou and Pang-Hsin Liu, "Bird Species Recognition by Wavelet Transformation of a Section of birdsong," *Proceedings of the First International Symposium on Cyber-Physical Intelligence* (CPI-09), pp.189-193, July 7-10, 2009, Brisbane, Australia. (EI)

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

98 年 9 月 30 日

附件三

報告人姓名	周智勳	服務機構及職稱	中華大學資訊工程系
時間 會議 地點	2009/9/12~2009/9/14 日本-京都	本會核定 補助文號	計劃編號： NSC 97-2221-E-216-014
會議名稱	(英文) The Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2009)		
發表論文題目	(英文) **1.** A New Feature Integration Approach and its Application to 3D Model Retrieval 2. Modulation Spectral Analysis of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification		

報告內容應包括下列各項：

一、參加會議經過

這次會議的地點在京都車站前的一棟會議廳，交通相當方便。我和同事在會議前一天晚上抵達，就住在車站附近。

報告時間被排在 session 議程第一天的早上，會議流程同一時間只有三、四個 sessions 同時進行，因此每個 session 聆聽的人數也比較多。會議室場地還不錯，然而會議室外的空間較小，不是很方便做其他的交流，也沒看到什麼參展的廠商，也許是因為這樣，所以沒有 poster 型式，交流機會也較少。

午餐發的餐卷可以直接拿去附近的餐廳用餐，主要是蕎麵類的定食。第一天晚上的迎賓宴採半雞尾酒式，大家站著聊天，台灣來的學者相當多，大陸的也不少，會場到處充斥著中文。

第二天晚上正式的晚宴也是自助式，菜色不錯，涵蓋了不少日本食物的特色。晚宴中也頒發了一些獎狀及謝函，並穿插了茶藝的表演，只是表演擺的位置不是很好，很多人看不到。

二、與會心得

1. 會場的佈置似乎不像國內辦研討會的熱鬧，國內辦研討會，會花不少心思在會場佈置上，感覺比較有氣氛。
2. 國內研討會都會提供甜點，國外好像大都不來這一套，較為陽春?
3. 此次研討會沒有參展的攤位，較特殊。
4. 會議路徑指示不甚明確，需加強。
5. 此次之與會人員，相當專業，參與態度非常積極，值得效法。

三、考察參觀活動(無是項活動者省略)

無。

四、建議

1. 大會網頁提供的相關資訊，包括會場地圖，搭車方式，飯店位置等，算是充足的，值得學習。
2. 應該可以建一份資料庫提供查詢，這資料庫是有關每位學者參加研討會後，所帶回來的論文集資料名稱。如此，當其他學者恰好需要此論文級上的某篇論文時，可以直接與該位學者聯繫，達到資訊互通的效果。
3. 國內也常舉辦國際性研討會，可以於議程中安排半日遊，讓外國學者增加認識台灣的機會，或於晚宴時安排有代表性的表演，如此對推展觀光也許有一些幫助。

五、攜回資料名稱及內容

Proceedings of the5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing

六、其他

1. 日本人蠻親切也蠻有禮貌的，英文程度平均而言好像弱了點，這可能跟他們的翻譯制度完善有關。
2. 晚宴氣氛不錯，交談熱烈。
3. 日本的 JR 火車真準時，什麼時候台鐵也能這樣…。
4. 除了火車準時，人民有禮貌，街道更是乾淨。想要發展服務業，提昇觀光產業的台灣，這些是值得學習的。不過，台灣的物價比日本便宜，這是相對有利點之一。
5. 感謝國科會工程處的補助。