Computers and Mathematics with Applications 64 (2012) 1270–1281

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Automatic recognition of frog calls using a multi-stage average spectrum

Wen-Ping Chen [a], Song-Shyong Chen [b], Chun-Cheng Lin [c,*], Ya-Zhung Chen [a], Wen-Chih Lin [d]

[a] *Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan*
[b] *Department of Information Networking Technology, Hsiuping University of Science and Technology, Taichung, Taiwan*
[c] *Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung, Taiwan*
[d] *Liouguei Research Center, Taiwan Forestry Research Institute, Taiwan*

ARTICLE INFO

Keywords:
Frog call recognition
Syllable segmentation
Pre-classification
Multi-stage average spectrum
Template matching

ABSTRACT

The automatic recognition of animal sounds is one of the powerful techniques for replacing the traditional ecological survey method that mainly depends on manpower, which is hence both costly and time consuming. This study developed an automatic frog call recognition system based on the combination of a pre-classification method of the syllable lengths and a multi-stage average spectrum (MSAS) method. In this system, the input frog syllables are first classified into one of the four groups determined by the pre-classification method according to syllable length. Then the proposed MSAS method is used to extract the standard feature template to analyze the time-varying features of each frog species and to recognize the input frog syllable by a template matching method. In all, 960 syllables recorded from 18 frog species are included in this study to evaluate the accuracy of the proposed frog call recognition system. The experimental results demonstrate that the proposed one-level (using the MSAS method only) and two-level (combining the syllable length pre-classification and MSAS methods) recognition methods can provide the best recognition accuracies of 91.9% and 94.3%, respectively, compared with other recognition methods based on dynamic time warping (DTW), spectral ensemble average voice prints (SEAV), *k*-nearest neighbor (*k*NN) and support vector machines (SVMs).

1. Introduction

In the past decades, many ecological habitats have been severely affected by human destruction and natural calamities. Ecologists worldwide are now actively studying and investigating animal habitats to understand the changes in the ecosystems [1,2]. However, traditional ecological survey work mainly depends on manpower, which is both costly and time consuming. It is also problematic to obtain true information because of the difficulty of approaching some sensitive or dangerous study subjects. Hence it is essential to develop an automatic ecological survey system to replace the traditional approach. Recording animal sounds to recognize their species is one of the powerful methods now being used in automatic ecological surveys. This method integrates advanced sensor and digital signal processing technologies, and can automatically process a large amount of ecological data to reduce the manpower costs.

Several animal sound recognition methods have been proposed which extract the features that can characterize the animal sounds to classify their species [3–10]. Taylor et al. [3] developed a recognition method based on spectrogram analysis to identify 22 frog species recorded in north Australia. The peak values in the spectrogram were defined as the features of the frog vocalization. However, spectrogram analysis is time consuming and it is not easy to find accurate

[*] Correspondence to: Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung, No.57, Sec. 2, Jhongshan Rd., Taiping Dist., Taichung 41170, Taiwan. Tel.: +886 4 23924505x7238; fax: +886 4 23924419.
E-mail address: cclin@ncut.edu.tw (C.-C. Lin).

0898-1221/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2012.03.071

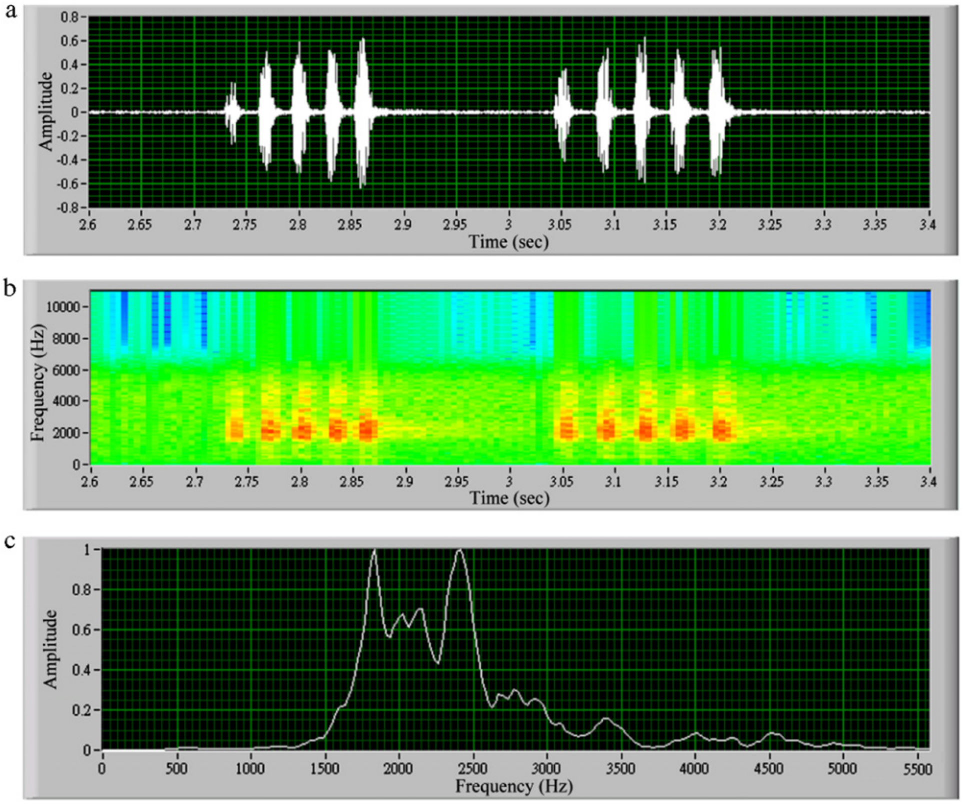

Fig. 1. Illustrations of the sound signals recorded from the Sauteris frog: (a) the time-domain waveform, (b) the time–frequency spectrum, and (c) the frequency spectrum.

reference points in the time axis for the analysis of all frog syllables. Lee et al. [4] also introduced a recognition method based on spectrogram analysis to detect each syllable and calculate the Mel-frequency cepstrum coefficients (MFCCs) of each frame. Their study defined the features by all of the averaged MFCCs in each frame and used linear discriminant analysis for classifying 30 kinds of frog calls and 19 kinds of cricket calls. However the averaged MFCC lost the time-varying features, and consequently it was difficult to accurately identify the species with similar frequency components. Because, in general, the syllables have different lengths, the dynamic time warping (DTW) algorithm which can be used for comparing variable-length sequences has been widely applied in the previous studies [5–7]. Myers and Rabiner [5]applied the DTW algorithms for word recognition, while Kogan and Margoliash [6]and Somervuo et al. [7] used them to recognize bird sounds. Accurate detection of the syllable position is critical for the recognition accuracy of the DTW method. Tyagi et al. [8] further proposed a spectral ensemble average voice print (SEAV) method, and combined the DTW and SEAV methods to improve the recognition accuracy of the bird sounds. The SEAV method extracted the features by calculating the averaged spectrum of each frame, and recognized the bird species by the method of template matching based on the calculation of the Euclidian distance between the SEAV of the reference and the test template [11,12]. However, if the frequency features of the input signal vary with time, the SEAV method cannot provide time-varying features. Recently, Fagerlund [9] applied support vector machines (SVMs) to recognize bird species, while Huang et al. [10] introduced the k-nearest neighbor (kNN) and SVMs to classify frog sounds. Three features, spectral centroid, signal bandwidth and threshold-crossing rate, were extracted to serve as the features for frog sound classification. However, these approaches have their drawbacks. In the kNN method, the k value which varies with the number of syllables needs to be predefined [13], while too many support vectors in the SVM classifier would result in a problem of overfitting, and therefore this approach needs extra processing to reduce the number of vectors [14,15].

Figs. 1(a) and 2(a) demonstrate the time-domain waveforms recorded from the Sauteris and Olive frogs, respectively. All of the frog calls in this paper were digitized at a 22.05 kHz sampling rate and 16-bit resolution. The corresponding time–frequency spectra are shown in Figs. 1(b) and 2(c), respectively, and they have different time-varying features in the frequency components. However, the corresponding frequency spectra as shown in Figs. 1(c) and 2(c) are similar. Hence if the time-varying features are not considered in the design of the frog call recognition system, it is hard to accurately classify the two frog species which have similar frequency spectra. Another important feature of frog calls is the syllable length, as shown in Fig. 3(a) and (b). The syllable length of the bull frog is about 700 ms which is much longer than the 40 ms of the white lipped tree frog. If the syllable length can be included in the frog call recognition system, it would be expected to further enhance recognition accuracy.

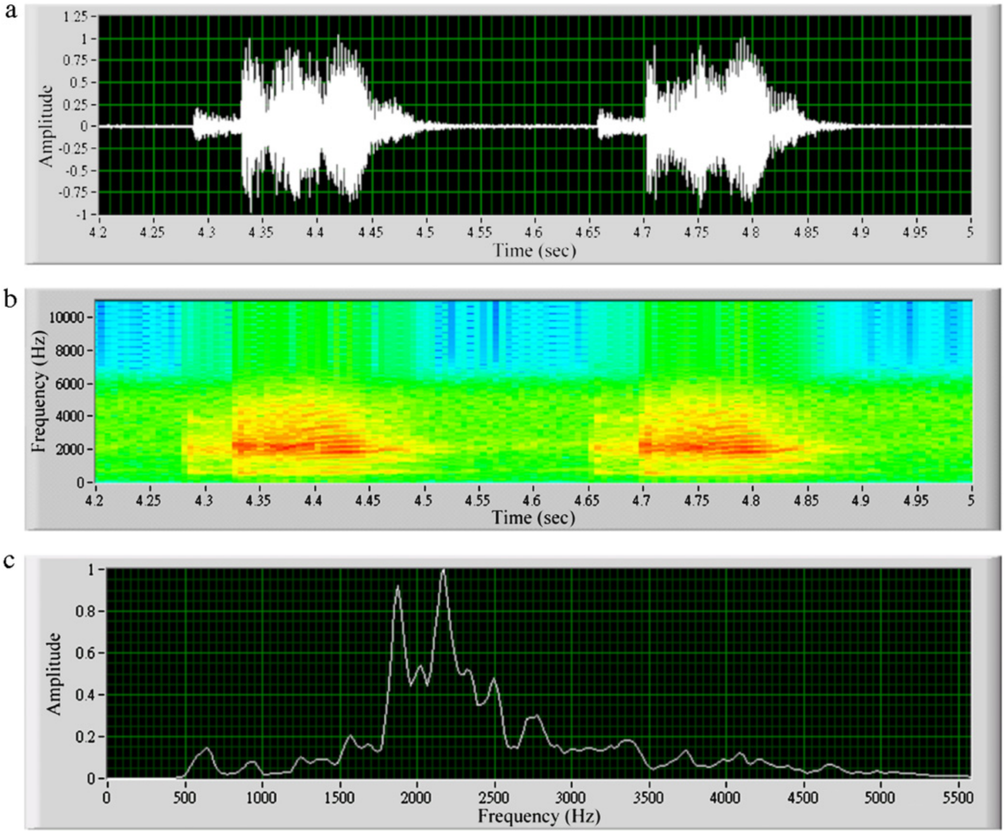

Fig. 2. Illustrations of the sound signals recorded from the Olive frog: (a) the time-domain waveform, (b) the time–frequency spectrum, and (c) the frequency spectrum.

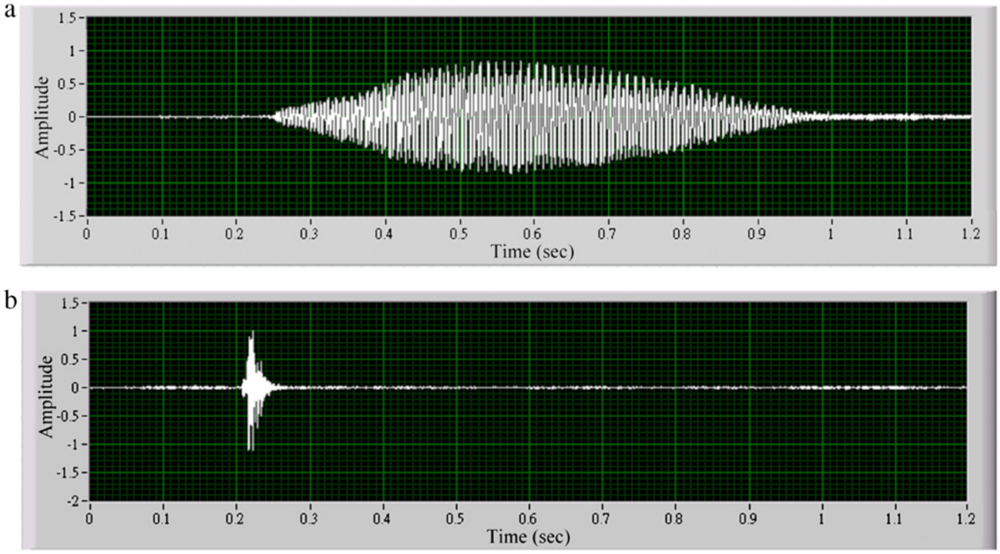

Fig. 3. Examples of the syllables for (a) bull frog, and (b) white lipped tree frog.

In order to preserve the time-varying features of the frog call in the frequency domain, this study proposes an automatic frog call recognition system based on a multi-stage average spectrum (MSAS) method which can first extract the time-varying features of the call, and then recognize the input frog syllable using a template matching method. This study also proposes a pre-classification method of the syllable lengths which can exclude those frog species whose syllable length is not in the same group as that of the input syllable prior to the MSAS analysis. There were 18 frog species and 960 test syllables

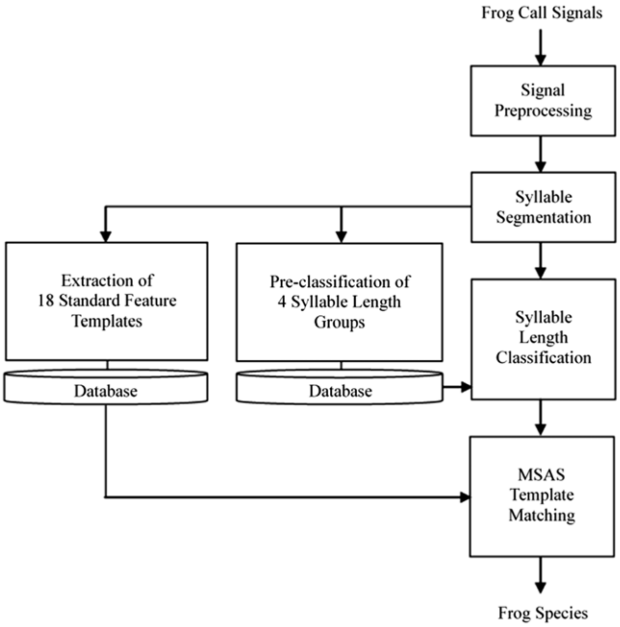
Frog Call Signals
Signal Preprocessing
Syllable Segmentation
Extraction of 18 Standard Feature Templates
Pre-classification of 4 Syllable Length Groups
Syllable Length Classification
Database
Database
MSAS Template Matching
Frog Species

Fig. 4. Block diagram of the automatic frog call recognition system.

included in this study for the evaluation of the performance of the proposed automatic recognition system. The study results demonstrate that the recognition performance of the proposed MSAS method is better than that of other potential methods based on the techniques of DTW, SEAV, kNN and SVM, and the proposed pre-classification method of the syllable lengths can further enhance the recognition accuracy of frog calls.

2. Architecture of the automatic frog call recognition system

The proposed automatic frog call recognition system includes signal preprocessing, syllable segmentation, pre-classification of the syllable lengths, extraction of the standard feature templates using the MSAS method, and recognition of the frog syllables. Fig. 4 illustrates a block diagram of the proposed frog call recognition system. A detailed description of each stage is provided below.

2.1. Signal preprocessing

All of the recorded frog calls were re-sampled at 22.05 kHz, and re-quantified to the range of [−1, 1] by a 16-bit mono format. According to the sampling theorem, the highest available frequency of the input sound signal can be up to 11.025 kHz in this study according to the sampling theorem. Because the low-frequency components with large amplitudes would reduce the contribution of the high-frequency components, this study further introduced a first-order high-pass filter with finite impulse response (FIR) to enhance the high-frequency components by reducing the low-frequency components, defined as follows:

$$y(n) = s(n) - \alpha s\,(n-1) \tag{1}$$

where $s(n)$ is the input frog call, $y(n)$ is the output of the high-pass filter, and the constant α determines the cutoff frequency of the high-pass filter and was set at 0.95 in this study. The filtered signals were then separated into frames with a length of 512 samples. The neighboring frame has an overlap of 256 samples. Each frame was further passed through a Hamming window for the reduction of the edge effects due to the discontinuities at the two sides of each frame. The Hamming window is optimized to minimize the maximum sidelobe in the frequency domain and can get about 40 dB of sidelobe suppression, which can be defined as

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{L-1}\right), \quad 0 \le n \le L-1 \tag{2}$$

where L is the length of the frame and is set at 512 in this study. The signal after the windowing process can be expressed as

$$x(n) = w(n)y(n) \tag{3}$$

where $x(n)$ denotes the signal in a frame after the windowing process. Each frame was then transformed from the time-domain signal $x(n)$ into the frequency-domain signal $X(k)$ by fast Fourier transformation, defined as follows [16]:

$$X(k) = \sum_{n=0}^{L-1} x(n)e^{-j2\pi\, kn/L}, \quad 0 \le k \le L-1 \tag{4}$$

where k is the discrete frequency variable. The spectra of all frames are provided for the following analyses.

2.2. Syllable segmentation

The method of syllable segmentation is mainly based on the combination of the analyses of the frog call energy and zero-crossing rate [17–21]. The energy E of each frame is defined as follows:

$$E = \sum_{k=0}^{L-1} |x(k)| \tag{5}$$

where $x(k)$ denotes the signal in a frame, and L is the length of the frame and is set at 512 in this study. The zero-crossing rate Z is the rate at which the signal changes from positive to negative and back, and is defined by

$$Z = \frac{1}{L}\sum_{k=0}^{L-1} \frac{1}{2}\,|\mathrm{sgn}[x(k+1)] - \mathrm{sgn}[x(k)]| \tag{6}$$

where

$$\mathrm{sgn}[x(k)] = \begin{cases} 1, & x(k) \ge 0 \\ -1, & x(k) < 0. \end{cases} \tag{7}$$

Fig. 5 demonstrates an example of the syllable segmentation for a Swinhoe's frog call. Fig. 5(a) displays the time-domain waveform, and Fig. 5(b) and (c) plot the energy curve and zero-crossing rate, respectively. When only the frog call energy was analyzed and applied to detect the syllable, the detected syllable segment was from points B to C, as shown in Fig. 5(a). It is obvious that the syllable segment cannot be detected accurately by the analysis of the frog call energy due to the low energy levels presented at the regions from points A to B and from points C to D. Hence, the analysis of the zero-crossing rate was further applied to enhance the accuracy of the syllable segmentation. Fig. 5(a) shows that the accurate syllable segments from points A to C can be detected after the adjustment of the zero-crossing rate. The length of the detected syllable segment was calculated for the use of the following analyses.

2.3. Pre-classification of the syllable lengths

Because the lengths of the syllables of most frog species are different, syllable length is one of the useful features for identifying frogs. However it is unavoidable that the calls of some frogs have similar syllable lengths. For this reason this study proposes the pre-classification of the syllable lengths to classify all the syllables into just 4 groups, instead of trying to immediately classify the input frog call into its exact group. At this stage, it is only necessary to compare the features of the input frog syllable with the standard template of each frog species in the same group, as defined in Section 2.4. This pre-classification can first exclude those frog species whose syllable lengths are not in the same group as that of the input syllable, hence increasing both recognition accuracy and speed. The proposed pre-classification of the syllable lengths is based on the binary split method [22], and is described as follows:

Step 1: All of the syllable lengths are labeled as the first group, and the average length is calculated and defined as the centroid of the syllable lengths in that group.

Step 2: Split each centroid according to the initial formula defined as follows:

$$\begin{cases} C_g^+ = C_g(1+\varepsilon) \\ C_g^- = C_g(1-\varepsilon) \end{cases} \tag{8}$$

where ε is the split coefficient which has a value between 0 and 1 and was set at 0.05 in this study. C_g is the centroid of the gth group, $1 \le g \le 4$, and C_g^+ and C_g^- are the new centroids split from C_g. The number of groups grows at a rate of 2 for each split.

Step 3: Calculate the differences between each syllable length and the new centroids in the gth group. All of the syllable lengths are further separated into two subgroups in the gth group according to whether they are closer to C_g^+ or C_g^-.

Step 4: Update the centroids C_g^+ and C_g^- which are updated as the average syllable lengths of the two subgroups obtained from step 3.

Step 5: Repeat steps 3 and 4 until the change in the updated centroids is less than a predefined threshold.

Step 6: Repeat steps 2 to 5 until the syllable lengths are separated into 4 groups. The final results are then stored in the database for the use of the following recognition of the frog syllables.

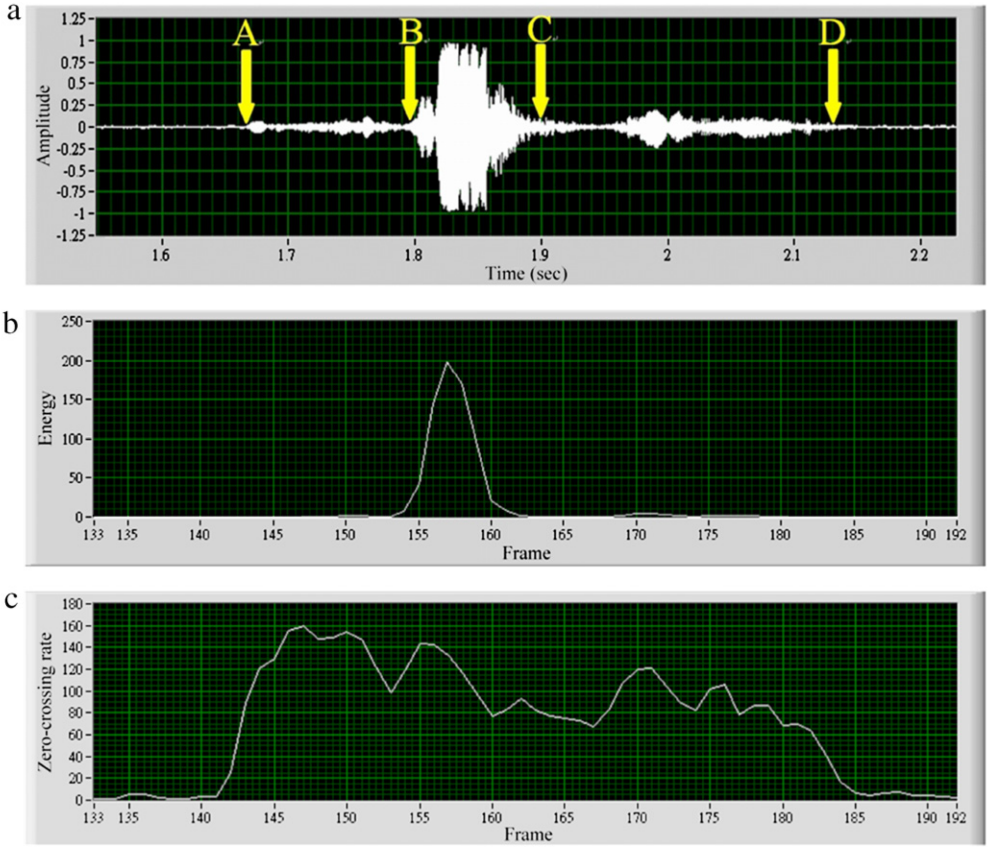

Fig. 5. An example of the syllable segmentation for a Swinhoe's frog call: (a) the time-domain waveform, (b) the energy curve and (c) the zero-crossing rate.

2.4. Extraction of standard feature templates by analysis of the multi-stage average spectrum

Because the frequency feature of the frog calls varies with time, this study proposes a new method based on MSAS analysis to extract the time-varying features within each frog syllable, and to establish a standard feature template for each frog species. The proposed MSAS method can calculate the average spectrum from neighboring frames that have similar spectra, and has the advantage that the energy of the averaged spectrum is more stable than that of a single frame. The extraction of the standard feature template adopted three randomly selected syllables for each frog species. The detailed steps of the MSAS method are described as follows:

Step 1: All of the frames of the input frog syllable were first divided equally into 3 time stages. The total number of frames in the input syllable depends on the syllable length. The following steps are used to reclassify the similar frames into the same stage in accordance with the forward direction of time and to establish the standard feature template. For example, if a frame j is classified into stage i, the frame after frame j can only be classified into stage i or the stage after i.

Step 2: Calculate the average spectrum of each stage defined as follows:

$$S_i(k) = \sum_{j=0}^{N_i-1} \frac{\left|X_j(k)\right|}{N_i}, \quad 0 \le k \le N-1 \tag{9}$$

where $S_i(k)$ denotes the average spectrum of the ith stage, $\left|X_j(k)\right|$ is the amplitude spectrum of the jth frame, N_i is the total number of frames in the ith stage, and k is the discrete frequency variable.

Step 3: Calculate the distance from the spectrum of each frame to the average spectrum of each stage according to the formula of Euclidian distance defined as follows:

$$d(j, i) = \sqrt{\sum_{k=0}^{L-1} [X_j(k) - S_i(k)]^2} \tag{10}$$

where $d(j, i)$ is the distance between the jth frame and the ith stage.

Step 4: Calculate the shortest accumulation distance from the first to the last frame and stage. Fig. 6 illustrates an example of the frame and stage plane for calculating the shortest accumulation distance. The initial point (1, 1) denotes the distance between the first frame and the first stage. Let $\text{Acc}(j, i)$ denote the shortest accumulation distance from the initial point to point (j, i) which is defined as follows:

$$\text{Acc}(j, i) = \min\{\text{Acc}(j-1, i) + d(j, i), \text{Acc}(j-1, i-1) + d(j, i)\} \tag{11}$$

where $\text{Acc}(1, i) = d(1, i)$ for all i, the operation of min {} can obtain the minimum value among the operands. $\text{Acc}(j-1, i) + d(j, i)$ and $\text{Acc}(j-1, i-1) + d(j, i)$ are the two candidate paths, a and b, respectively, as shown in Fig. 6, which are considered in the calculation of the shortest accumulation distance at the point (j, i). Because one frame can only be classified into one stage, the path from point $(j, i-1)$ to (j, i) is forbidden in this study.

Step 5: Search for the shortest path from the final point $(N_f, 3)$ back to the initial point (1, 1) along with those paths that have a shorter distance between the two candidate paths after all of the shortest accumulation distances for each frame j and each stage i have been calculated. The points on the shortest path are recorded as a sequence of coordinates $P = \{P(1), P(2), \ldots, P(N_f)\}$ where N_f is the total number of frames in a syllable. For example, if the coordinate sequence of the shortest path is $P = \{(1, 1), (2, 1), (3, 2), (4, 3)\}$, it represents that the spectra of frames 1 and 2 are closest to that of stage 1, and the spectra of frames 3 and 4 are closest to those of stages 2 and 3, respectively. Hence, the frames 1 and 2 will be classified into stage 1, and frames 3 and 4 will be classified into stages 2 and 3, respectively.

Step 6: Reclassify each frame into a new stage according to the coordinate sequence of the shortest path.

Step 7: Repeat steps 2 to 6 until the change in the updated shortest accumulation distance of the final point is lower than the predefined threshold.

Step 8: Calculate the feature template for the input syllable, which is defined as the average spectrum of each stage.

Step 9: Repeat steps 1 to 8 to obtain three feature templates from three randomly selected syllables for each frog species. The standard feature template is defined as the average of the three feature templates, and is stored in the database for the use of the following recognition of the frog syllables.

Fig. 7 demonstrates an example of the training results of the feature template for Swinhoe's frog call. Fig. 7(a) plots the time-domain waveform of the syllable. Fig. 7(b) displays the time–frequency spectrum of the syllable and shows the time-varying features of the frequency components within the frog syllable. Fig. 7(c) shows the relations between frame and stage. The first 12 frames were classified as stage 1, frames 13 to 25 were classified as stage 2, and frames 26–42 were classified as stage 3. Fig. 7(d) illustrates the average spectra of the three stages. It is worth noting that the MSAS method can preserve the time-varying features within the frog syllable.

2.5. Recognition of the frog syllables

After the pre-classification of the syllable lengths and extraction of the standard feature templates, the input frog syllable will be automatically recognized by a one-level or two-level recognition method. In the one-level recognition method, the input syllable was only compared with the standard feature templates of all frog species to find the closest frog species, and was not pre-classified by its syllable length. In the two-level recognition method, the length of the input syllable was first compared with each centroid of the four groups determined from the pre-classification of the syllable lengths, and was classified into the group whose centroid had the minimum difference from the input syllable length. The input syllable was then compared with the standard feature template of each frog species in the same group in order to find the closest species. The template matching method for comparing the input syllable with the standard feature template is described as follows:

Step 1: Calculate the distances from the spectrum of each frame of the input syllable to the average spectrum of each stage of the standard feature template for one frog species according to Eq. (10).

Step 2: Calculate the shortest accumulation distance from the first to the last frame and stage according to Eq. (11). The shortest accumulation distance at the final point $(N_f, 3)$ is applied to analyze the difference between the input syllable and the standard feature template.

Step 3: Repeat steps 1 and 2 to calculate each shortest accumulation distance at the final point $(N_f, 3)$ for all frog species. The input syllable will be recognized as a frog species that has the minimum value among the shortest accumulation distances of all frog species.

3. Experimental results

The proposed frog call recognition system was developed based on the commercial LabVIEW software using a graphical programming language from National Instruments. The frog calls were recorded in a wild field located in the Shan-Ping forest ecological garden in Kaohsiung city, Taiwan. Table 1 lists the family, scientific and common names of the 18 frog species recorded in this study. The input frog calls were re-sampled at 22.05 kHz and digitized to a 16-bit mono format. Three syllables randomly selected from each frog species were used to establish the standard feature templates, and a total

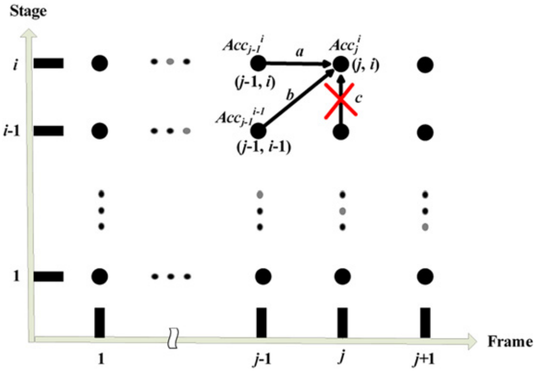

Fig. 6. A frame and stage plane for calculating the shortest accumulation distance.

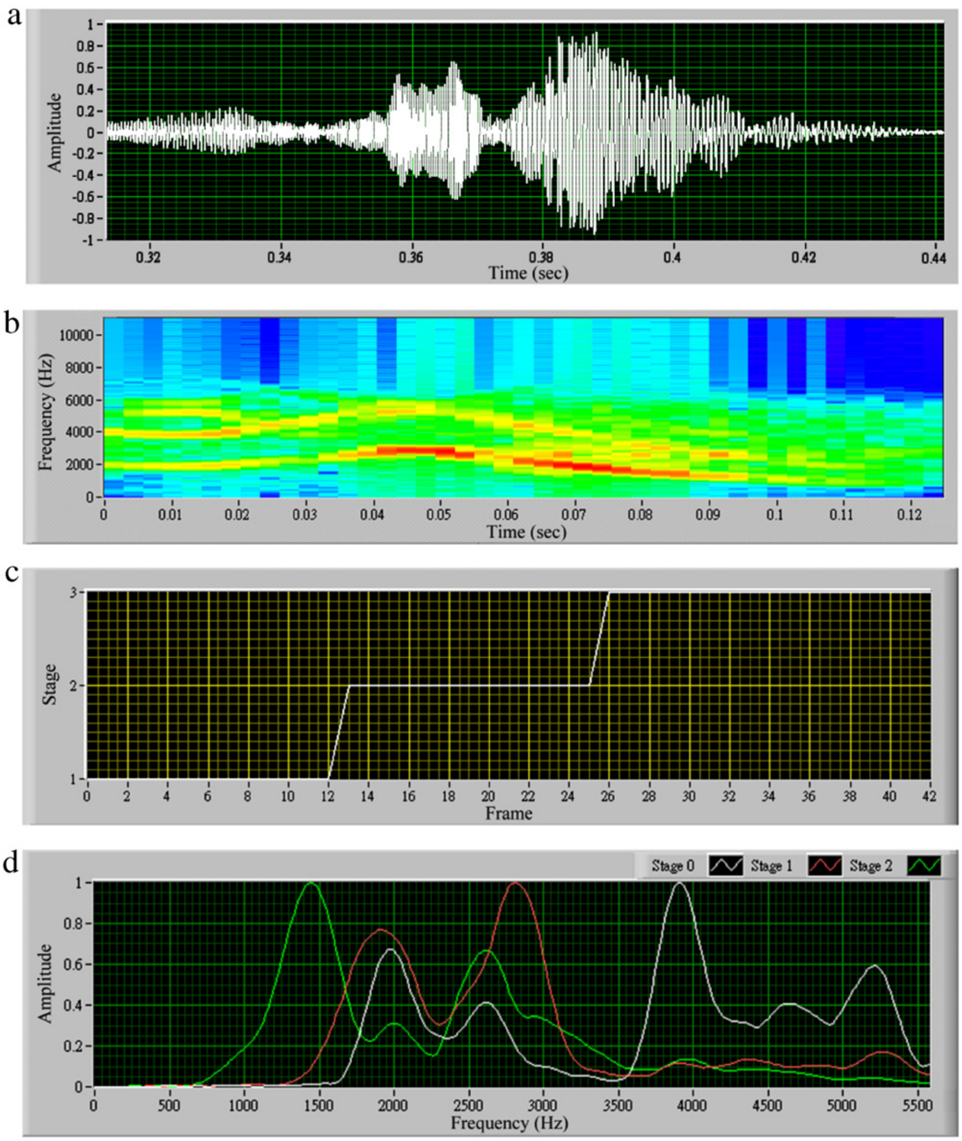

Fig. 7. An example of the training results of the feature template for Swinhoe's frog call: (a) the time-domain waveform of the syllable, (b) the time–frequency spectrum, (c) the relation between frame and stage, and (d) the average spectrum of each stage.

Table 1
Summary of the family, scientific and common names of the 18 frog species recorded in this study.

Family	Scientific name	Common name
Ranidae	Rana adenopleura	Olive frog
	Rana psaltes	Harpist frog
	Rana catesbeiana	Bull frog
	Limnonectes kuhlii	Kuhlis frog
	Rana latouchii	Latouche's frog
	Fejervarya limnocharis	Indian Rice frog
	Pseudoamolops sauteri	Sauteris frog
	Rana swinhoana	Swinhoe's frog
Rhacophoridae	Chirixalus eiffingeri	Eiffinger's tree frog
	Polypedates megacephalus	White lipped tree frog
	Rhacophorus moltrechti	Moltrechtis green tree frog
	Rhacophorus taipeianus	Taipei green tree frog
	Buergeria japonica	Japanese tree frog
Microhylidae	Microhyla ornata	Ornate narrow-mouthed toad
	Kaloula pulchra	Malaysian narrow-mouthed toad
	Microhyla heymonsi	Heymonsis narrow-mouthed toad
	Microhyla steinegeri	Stejneger's paddy frog
Bufonidae	Bufo bankorensis	Common toad

of 960 syllables were applied to test the performance of the proposed frog call recognition system. The recognition accuracy is defined as follows:

$$\text{Accuracy}(\%) = \frac{N_c}{N_s} \times 100\% \tag{12}$$

where N_c is the number of syllables which can be correctly recognized, and N_s is the total number of test syllables.

Table 2 lists the analytical results of the one-level recognition methods for each frog species. The details of the DTW, SEAV, *k*NN, and SVM methods can be found in the previous studies [3–10]. It can be seen from the results in Table 2 that the MSAS recognition method provided the best recognition accuracy of 91.9% compared with 87.4% for *k*NN, 85.3% for DTW, 86.9% for SEAV, and 90.1% for SVM for all frog species. Furthermore, only the MSAS recognition method had an accuracy greater than 70% for all frog species, and in no case was the accuracy of the MSAS recognition method inferior to that of the other methods. The study results also show that, because the spectrum of the Olive frog is similar to that of the Sauteris frog, it is easy to misclassify these two species. The accuracy of the *k*NN, SEAV and SVM recognition methods was only 68.3%, 58.7% and 60.3%, respectively, for the recognition of the Sauteris frog. However the MSAS recognition method can enhance the accuracy for this species to 74.6%. This may be because the MSAS recognition method preserves the time-varying features of the syllables. Although the accuracy of the DTW recognition method can also reach 73% for the recognition of the Sauteris frog, the accuracy was only 62.1% and 66.7% for the white lipped tree frog and the ornate narrow-mouthed toad, respectively. Hence it is obvious that the performance of the MSAS recognition method was better than that of the *k*NN, DTW, SEAV and SVM methods.

This study also compared the two-level recognition methods with the previous study [8]. Table 3 lists the analytical results of the two-level recognition methods for each frog species.

The "SEAV + DTW at Rank Level" method combined the SEAV and DTW methods by summing the top three reference templates, ranked according to the Euclidean distances between the test and the reference template. The "SEAV + DTW at Measurement Level" method combined the SEAV and DTW methods by averaging the first three Euclidean distances. The "SEAV + DTW at Two Levels" method first processed the calls at rank level; then, if two or more frogs were classified into the same species, the measurement level was applied to make the final decision. Detailed descriptions of these methods can be found in the previous study [8]. In the "SEAV + Syllable Length" and the "MSAS + Syllable Length" methods, the length of the input syllable was first compared with each centroid of the four groups determined from the pre-classification of the syllable lengths, and was classified into the group whose centroid had the minimum difference from the input syllable length. The input syllable was then recognized by the SEAV and MSAS methods, respectively. In this way, it was only necessary to compare the input syllables with those in the same group.

The results of the current study show that the "MSAS + Syllable Length" method has the best recognition accuracy of 94.3% compared with 85.2% for "SEAV + DTW at Measurement Level", 88.5% for "SEAV + DTW at Rank Level", 89.6% for "SEAV + DTW at Two Levels", and 90.6% for "MSAS + Syllable Length". Furthermore, there were only three instances where the recognition accuracy of the "MSAS + Syllable Length" method was less than 90%, namely 82.5% for the Sauteris frog, 80% for the Swinhoe's frog, and 89.7% for the white lipped tree frog. The recognition accuracy of the proposed "MSAS + Syllable Length" method for the recognition of the Sauteris frog can reach 82.5%, which is much better than that of the other four methods, namely 66.7%, 66.7%, 66.7% and 68.3%. If the results in Table 3 are compared with those in Table 2, it can be found that the two-level recognition method of "MSAS + Syllable Length" can further enhance the accuracy from 91.9% to 94.3% for all frog species, and from 74.6% to 82.5% for the Sauteris frog compared with the one-level recognition method of the

Table 2
Results of the one-level recognition.

Common name	Total syllable	Correct syllable (accuracy)				
		*k*NN	DTW	SEAV	SVM	MSAS
Olive frog	52	41 (78.9%)	45 (86.5%)	40 (76.9%)	47 (90.4%)	48 (92.3%)
Harpist frog	2	2 (100.0%)	2 (100.0%)	2 (100.0%)	2 (100.0%)	2 (100.0%)
Bull frog	12	12 (100.0%)	12 (100.0%)	12 (100.0%)	12 (100.0%)	12 (100.0%)
Kuhlis frog	64	60 (93.8%)	58 (90.6%)	61 (95.6%)	62 (96.9%)	62 (96.9%)
Latouche's frog	81	73 (90.1%)	74 (91.4%)	72 (88.9%)	73 (90.1%)	74 (91.4%)
Indian Rice frog	74	70 (94.6%)	69 (93.2%)	70 (94.6%)	72 (97.3%)	72 (97.3%)
Sauteris frog	63	43 (68.3%)	46 (73.0%)	37 (58.7%)	38 (60.3%)	47 (74.6%)
Swinhoe's frog	35	26 (74.3%)	28 (80.0%)	25 (71.4%)	26 (74.3%)	28 (80.0%)
Eiffinger's tree frog	84	77 (91.7%)	73 (86.9%)	75 (89.3%)	76 (90.5%)	77 (91.7%)
White lipped tree frog	29	25 (86.2%)	18 (62.1%)	26 (89.7%)	25 (86.2%)	26 (89.7%)
Moltrechtis green tree frog	72	59 (81.9%)	60 (83.3%)	63 (87.5%)	62 (86.1%)	63 (87.5%)
Taipei green tree frog	97	92 (94.9%)	90 (92.8%)	96 (99.0%)	97 (100.0%)	97 (100.0%)
Japanese tree frog	70	58 (82.9%)	54 (77.1%)	60 (85.7%)	61 (87.1%)	62 (88.6%)
Ornate narrow-mouthed toad	15	12 (80.0%)	10 (66.7%)	12 (80.0%)	12 (80.0%)	12 (80.0%)
Kaloula pulchra	7	7 (100.0%)	6 (85.7%)	7 (100.0%)	7 (100.0%)	7 (100.0%)
Microhyla heymonsi	37	30 (81.1%)	33 (89.2%)	30 (81.1%)	34 (91.9%)	34 (91.9%)
Microhyla steinegeri	131	118 (90.1%)	108 (82.4%)	111 (84.7%)	124 (94.7%)	124 (94.7%)
Common toad	35	34 (97.1%)	33 (94.3%)	35 (100.0%)	35 (100.0%)	35 (100.0%)
Total	960	839 (87.4%)	819 (85.3%)	834 (86.9%)	865 (90.1%)	882 (91.9%)

Table 3
Results of the two-level recognition method.

Common name	Total syllable	Correct syllable (accuracy)				
		SEAV + DTW at Measurement Level	SEAV + DTW at Rank Level	SEAV + DTW at Two Levels	SEAV + Syllable Length	MSAS + Syllable Length
Olive frog	52	40 (76.9%)	40 (76.9%)	44 (84.6%)	49 (94.2%)	50 (96.2%)
Harpist frog	2	2 (100.0%)	2 (100.0%)	2 (100.0%)	2 (100.0%)	2 (100.0%)
Bull frog	12	12 (100.0%)	12 (100.0%)	12 (100.0%)	12 (100.0%)	12 (100.0%)
Kuhlis frog	64	56 (87.5%)	61 (95.6%)	62 (96.9%)	61 (95.6%)	62 (96.9%)
Latouche's frog	81	73 (90.1%)	71 (87.7%)	74 (91.36%)	75 (92.6%)	78 (96.3%)
Indian Rice frog	74	69 (93.2%)	72 (97.3%)	72 (97.3%)	71 (96.0%)	72 (97.3%)
Sauteris frog	63	42 (66.7%)	42 (66.7%)	42 (66.7%)	43 (68.3%)	52 (82.5%)
Swinhoe's frog	35	25 (71.4%)	25 (71.4%)	25 (71.4%)	26 (74.3%)	28 (80.0%)
Eiffinger's tree frog	84	73 (86.9%)	76 (90.5%)	76 (90.5%)	77 (91.7%)	78 (92.9%)
White lipped tree frog	29	19 (65.5%)	25 (86.2%)	26 (89.7%)	26 (89.7%)	26 (89.7%)
Moltrechtis green tree frog	72	61 (84.7%)	63 (87.5%)	63 (87.5%)	64 (88.9%)	65 (90.3%)
Taipei green tree frog	97	90 (92.8%)	95 (97.9%)	96 (99.0%)	96 (99.0%)	97 (100.0%)
Japanese tree frog	70	56 (80.0%)	60 (85.7%)	60 (85.7%)	61 (87.1%)	63 (90.0%)
Ornate narrow-mouthed toad	15	11 (73.3%)	13 (86.7%)	13 (86.7%)	14 (93.3%)	14 (93.3%)
Kaloula pulchra	7	7 (100.0%)	7 (100.0%)	7 (100.0%)	7 (100.0%)	7 (100.0%)
Microhyla heymonsi	37	32 (86.5%)	33 (89.2%)	33 (89.2%)	31 (83.8%)	35 (94.6%)
Microhyla steinegeri	131	115 (87.8%)	118 (90.1%)	118 (90.1%)	120 (91.6%)	129 (98.5%)
Common toad	35	35 (100.0%)	35 (100.0%)	35 (100.0%)	35 (100.0%)	35 (100.0%)
Total	960	818 (85.2%)	850 (88.5%)	860 (89.6%)	870 (90.6%)	905 (94.3%)

MSAS. This is because the proposed syllable length pre-classification method can first exclude those frog species whose syllable lengths are not in the same group as that of the input syllable, hence increasing the recognition accuracy and speed.

4. Discussion and conclusions

This study has demonstrated that the proposed MSAS method can extract the time-varying features of the frog syllable, and the proposed pre-classification method of the syllable lengths can first exclude those frog species whose lengths are not in the same group as that of the input syllable for the automatic recognition of frog calls. An FIR high-frequency filter is also proposed to enhance the high-frequency components by decreasing the low-frequency components. This can avoid the problem of the features of the frog calls being dominated by the low-frequency components with large amplitude spectra. This study also introduced the combination of the analyses of the frog call energy and zero-crossing rate to detect the syllables, which can increase the accuracy of the endpoint detection of the syllables compared with only analyzing the energy or zero-crossing rate. The proposed MSAS method separates each syllable into three time stages, and the number of frames included in each time stage is not fixed or dependent on the time-varying features of the frog syllable. The example illustrated in Fig. 7 demonstrates that the averaged spectra of the three time stages differed significantly for the Swinhoe's frog call, and hence it was possible to preserve the time-varying features of the frequency components. The recognition performance of the proposed one-level MSAS method was compared with other potential one-level methods based on the DTW, SEAV, *k*NN and SVM techniques. The SEAV method is mainly based on the analysis of the average spectrum, so it is difficult to capture the spectral variation of the frog syllables. Because the DTW method compares the spectra of the standard and test templates for each frame, it can retain the time-varying information. However, each frame only includes 512 sample points since spectrum variation among frames that is too great may affect recognition accuracy. Although the SVM method also provided a good recognition accuracy of 90.1%, the proposed MSAS method can provide the best accuracy of 91.9% among the five methods.

The proposed pre-classification method of the syllable lengths is based on the binary split method which can separate all syllable lengths in the same group into two groups for each split according to the closeness of the syllable length. All of the syllable lengths are pre-classified into four groups. The input test syllable can first be compared with the centroids of the four groups to determine the group that the test syllable belongs to, and the following template matches only include the standard templates in the same group. The proposed two-level method, combining the pre-classification method of the syllable lengths and the MSAS method, was further compared with other two-level methods including"SEAV + DTW at Rank Level", "SEAV + DTW at Measurement Level", "SEAV + DTW at Two Levels", and "SEAV + Syllable Length". The previous study results of Tyagi et al. [8] have demonstrated that the two-level methods of "SEAV + DTW at Rank Level" and "SEAV + DTW at Two Levels" have improvements of 5.7% and 11.4%, respectively, in comparison with the one-level method of SEAV for the recognition performance of bird calls, but "SEAV + DTW at Measurement Level" has a degradation of 5.7%. The experimental results of this study also showed that both "SEAV + DTW at Rank Level" and "SEAV + DTW at Two Levels" have better recognition accuracy of frog calls compared with the SEAV method, but that the accuracy of "SEAV + DTW at Measurement Level" is worse. However, the proposed two-level method presents the best recognition performance among the five two-level methods, and can further increase the recognition accuracy from 91.9% to 94.3% compared with the one-level MSAS method. In conclusion, the proposed pre-classification method of the syllable lengths and the MSAS method are promising techniques for the design of an automatic recognition frog call system. Future works can focus on the reduction of background noises and the recognition of hybrid frog calls.

Conflict of interest statement

No author had a financial or personal conflict of interest related to this research or its source of funding.

Acknowledgments

The authors would like to thank the National Science Council (NSC) of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 100-2221-E-151-011.

References

[1] N. Keyghobadi, The genetic implications of habitat fragmentation for animals, Canadian Journal of Zoology 85 (2007) 1049–1064.
[2] S.M.P. Sullivan, K.T. Vierling, Experimental and ecological implications of evening bird surveys in stream-riparian ecosystems, Environmental Management 44 (2009) 789–799.
[3] A. Taylor, G. Grigg, G. Watson, H. McCallum, Monitoring frog communities: an application of machine learning, in: Proceedings of the Conference on Innovative Applications of Artificial Intelligence, 1996, pp. 1564–1569.
[4] C.H. Lee, C.H. Chou, C.C. Han, R.Z. Hunag, Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis, Pattern Recognition Letters 27 (2006) 93–101.
[5] C. Myers, L.R. Rabiner, Performance tradeoffs in dynamic time warping algorithms for isolated word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (1980) 623–635.
[6] J.A. Kogan, D. Margoliash, Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study, Journal of the Acoustical Society of America 103 (4) (1998) 2185–2196.
[7] P. Somervuo, A. Harma, S. Fagerlund, Parametric representations of bird sounds for automatic species recognition, IEEE Transactions on Audio, Speech, and Language Processing 14 (6) (2006) 2252–2263.
[8] H. Tyagi, R.M. Hegde, H.A. Murthy, A. Prabhaka, Automatic identification of bird calls using spectral ensemble average voice prints, in: Proceedings of the Thirteenth European Signal Processing Conference, 2006.
[9] S. Fagerlund, Bird species recognition using support vector machines, EURASIP Journal on Applied Signal Processing (2007) 1–8. http://dx.doi.org/10.1155/2007/38637.
[10] C.J. Huang, Y.J. Yangb, D.X. Yang, Y.J. Chen, Frog classification using machine learning techniques, Expert Systems with Applications 36 (2009) 3737–3743.
[11] P. Danielsson, Euclidean distance mapping, Computer Graphics and Image Processing 14 (1980) 227–248.
[12] C.L. Li, K.C. Hui, Feature recognition by template matching, Computers and Graphics 24 (4) (2000) 569–582.
[13] N. Razmjooy, B.S. Mousavi, F. Soleymani, A real-time mathematical computer method for potato inspection using machine vision, Computers & Mathematics with Applications 63 (2012) 268–279.
[14] W. Chen, T. Liu, B. Wang, Ultrasonic image classification based on support vector machine with two independent component features, Computers & Mathematics with Applications 62 (7) (2011) 2696–2703.
[15] L. Gonzalez-Abril, F. Velasco, J.A. Ortega, L. Franco, Support vector machines for classification of input vectors with different metrics, Computers & Mathematics with Applications 61 (9) (2011) 2874–2878.
[16] A. Antoniou, Digital Signal Processing: Signals, Systems, and Filters, Mcgraw-Hill, United State of America, 2006.
[17] L. Lamel, L. Labiner, A. Rosenberg, J. Wilpon, An improved endpoint detector for isolated word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 29 (1981) 777–785.

[18] S. Ishimitsu, M. Nakayama, T. Yoshirni, Construction of the noise-robust body-conducted speech recognition system, in: Proceedings of the Second International Conference on Innovative Computing, Information and Control, 2007, pp. 123–126.
[19] X. Zhao, D. O'Shaughnessy, A new hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation, in: Proceedings of the Canadian Conference on Electrical and Computer Engineering, 2008, pp. 145–148.
[20] Y. Tian, Z. Wang, D. Lu, Nonspeech segment rejection based on prosodic information for robust speech recognition, IEEE Signal Processing Letters 9 (11) (2002) 364–367.
[21] Lie Lu, Hong-Jiang Zhang, Stan Z. Li, Content-based audio classification and segmentation by using support vector machines, Multimedia Systems 8 (2003) 482–492.
[22] S. Chatterje, T.V. Sreenivas, Optimum transform domain split VQ, IEEE Signal Processing Letters 15 (2008) 285–288.