Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 27 (2006) 93–101

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis

Chang-Hsing Lee [a,*], Chih-Hsun Chou [a], Chin-Chuan Han [b], Ren-Zhuang Huang [a]

[a] Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu 300, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, National United University, Miao-Li 360, Taiwan, ROC

Received 13 August 2004; received in revised form 30 June 2005
Available online 31 August 2005

Communicated by O. Siohan

Abstract

In this paper we propose a method that uses the averaged Mel-frequency cepstral coefficients (MFCCs) and linear discriminant analysis (LDA) to automatically identify animals from their sounds. First, each syllable corresponding to a piece of vocalization is segmented. The averaged MFCCs over all frames in a syllable are calculated as the vocalization features. Linear discriminant analysis (LDA), which finds out a transformation matrix that minimizes the within-class distance and maximizes the between-class distance, is utilized to increase the classification accuracy while to reduce the dimensionality of the feature vectors. In our experiment, the average classification accuracy is 96.8% and 98.1% for 30 kinds of frog calls and 19 kinds of cricket calls, respectively.

MSC: 140.000

Keywords: Linear discriminant analysis; Mel-frequency cepstral coefficients

1. Introduction

Many animals generate sounds either for communication or as a by-product of their living activities such as eating, moving, or flying. Automatic recognition of bioacoustic sounds is valuable for applications such as biological research and environmental monitoring; this is particularly true for detecting and locating animals. In our daily life, we often hear the animal vocalizations rather than see the animals. In general, the animals generate sounds to communicate with members of the same species and thus the animal vocalizations have evolved to be species-specific. Therefore, identifying animal species from their vocalizations is valuable to ecological censusing.

In general, the acoustic signal representing animal vocalizations can be regarded as a sequence of syllables. Thus, a better way to identify animals from their vocalizations is to use a syllable as the acoustic component. It is necessary to segment the syllables of animal vocalizations before the recognition process. Segmentation of speech or audio signals is often based on energy (Lamel et al., 1981; Li et al., 2001; Lu, 2001; Wold et al., 1996; Zhang and Kuo, 2001) and/or zero-crossing rate (Li et al., 2001; Lu, 2001; Tian et al., 2002; Wold et al., 1996; Zhang and Kuo, 2001). A disadvantage of using these segmentation methods to extract syllables from animal vocalizations is that the full syllable cannot be extracted exactly. To overcome this problem, we exploit the frequency information to segment the syllables of animal vocalizations (Harma, 2003).

* Corresponding author. Tel.: +886 3 5186406; fax: +886 3 5186416.
E-mail addresses: chlee@chu.edu.tw (C.-H. Lee), chc@chu.edu.tw (C.-H. Chou), cchan@nuu.edu.tw (C.-C. Han).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.07.004

Once the syllables have been properly segmented, a set of features will be calculated to represent each syllable. The most well-known features for speech/speaker recognition are linear predictive coefficients (LPCs) (Rabiner and Juang, 1993) or Mel-frequency cepstral coefficients (MFCCs) (Picone, 1993; Rabiner and Juang, 1993; Vergin et al., 1999). In this paper, we use the averaged MFCCs in a syllable to identify animals from their sounds due to the fact that MFCCs can represent the spectrum of animal sounds in a compact form. In the next section, we will describe the proposed recognition method for animal vocalizations.

2. The proposed recognition method for animal vocalizations

The recognition system consists of two parts: the training part and the recognition part. The training part is composed of three main modules: syllable segmentation, averaged MFCCs extraction, and linear discriminant analysis (LDA). The recognition part consists of four modules: syllable segmentation, averaged MFCCs extraction, LDA transformation, and classification. A detailed description of each module will be described below.

2.1. Syllable segmentation

The input acoustic signal is first segmented into a set of syllables (Harma, 2003). Each syllable is regarded as the basic acoustic unit for recognition. The syllable segmentation method based on the frequency information is described as follows:

Step 1. Compute the spectrogram of the input bioacoustic signal using short-time Fourier transform (STFT). We denote the spectrogram a matrix $S(f,t)$, where f represents frequency index and t is the frame index.

Step 2. Set $n = 0$.

Step 3. Find f_n and t_n, such that $|S(f_n,t_n)| \geqslant |S(f,t)|$, for every pair of (f,t). Set the position of the nth syllable to be (f_n,t_n).

Step 4. Compute the amplitude $A_n(0) = 20\log_{10}|S(f_n,t_n)|$ dB and set the frequency parameter $W_n(0) = f_n$. If $A_n(0) < A_0(0) - 20$ dB, stop the segmentation process. This means that the amplitude of the nth syllable is too small and hence no more syllables need to be extracted.

Step 5. Starting from (f_n,t_n), trace the maximal peak of $|S(f,t)|$ for $t < t_n$ until $A_n(t) < A_n(0) - ß$dB, where ß is the stopping criteria and its default value is 20. Next, trace the maximal peak of $|S(f,t)|$ for $t > t_n$ until $A_n(t) < A_n(0) - ß$dB. The step is to determine the starting time $(t_n - t_s)$ and the ending time $(t_n + t_e)$ of the nth syllable around t_n.

Step 6. Store the amplitude trajectories corresponding to the nth syllable in function $A_n(\tau)$, where $\tau = t_n - t_s, \ldots, t_n + t_e$.

Step 7. Set $S(f,[t_n - t_s, \ldots, t_n + t_e]) = 0$ to delete the area of nth syllable. Set $n = n + 1$ and goto Step 3 to find the next syllable.

Fig. 1 shows the waveform of Olive Frog (*Rana adenopleur*) as well as the segmentation results by using the energy information and the spectrogram frequency information. It is evident that a better result can be obtained. After segmenting each syllable, the averaged MFCCs are extracted to represent the syllable.

2.2. Averaged MFCCs extraction

MFCCs have been the most widely used features for speech recognition (Picone, 1993; Rabiner and Juang, 1993; Vergin et al., 1999), bird song recognition (Kogan and Margoliash, 1998), and audio retrieval (Slaney, 2002) due to their ability to represent the signal spectrum in a com-

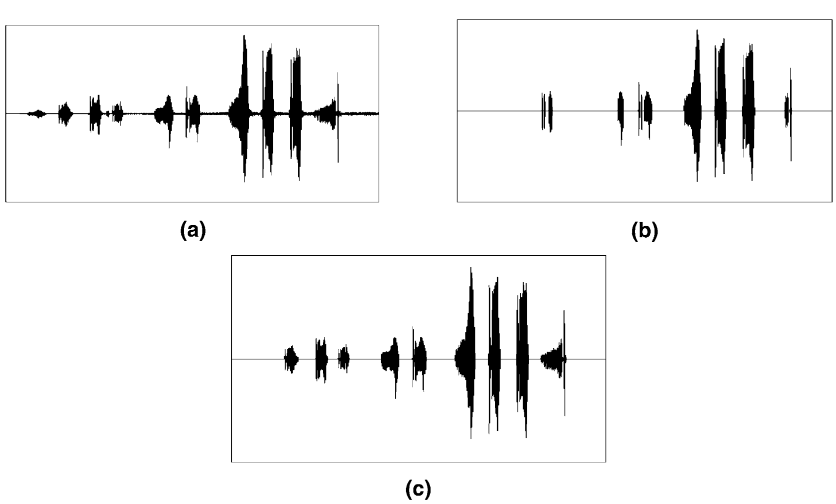

Fig. 1. (a) The waveform of Olive Frog (*Rana adenopleur*). (b) The segmentation result by using the energy information. (c) The segmentation result by using the spectrogram frequency information.

pact form. In fact, the MFCCs have been proven to be very effective in automatic speech recognition or in modeling the subjective frequency content of audio signals. In general, an input signal is first divided into a set of frames. The MFCCs for each frame are then computed and are regarded as the features of this frame. However, the number of frames varies for different syllables. To deal with this problem, the averaged MFCCs of all the frames in a syllable are computed and used as features to represent the syllable. Therefore, the number of features is fixed regardless of the length of the acoustic syllable. A detailed description for deriving the MFCCs of an acoustic signal is given as follows:

Step 1. *Pre-emphasis*.

$$\hat{s}[n] = s[n] - \hat{a}s[n-1], \tag{1}$$

where $s[n]$ is the signal denoting the input syllable, a typical value for $\hat{a}$ is 0.95.

Step 2. *Framing*.Each syllable is divided into a set of overlapped frames with frame size of N samples, and the overlapping size is M samples for each pair of successive frames. Therefore, consecutive frames will never change too much. In our experiments, N is 512 and M is 256.

Step 3. *Windowing*.To reduce the discontinuity on both ends of a frame, each frame is multiplied by a Hamming window

$$\tilde{s}[n] = \hat{s}[n]w[n], \quad 0 \leqslant n \leqslant N-1, \tag{2}$$

where $w[n]$ is the Hamming window function

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leqslant n \leqslant N-1. \tag{3}$$

Step 4. *Spectral analysis*.Take the discrete Fourier transform of each frame using FFT

$$X[k] = \sum_{n=0}^{N-1} \tilde{s}[n] e^{-j2\pi \frac{k}{N} n}, \quad 0 \leqslant k \leqslant N-1. \tag{4}$$

Step 5. *Band-pass filtering*.The amplitude spectrum is then filtered using a set of triangular band-pass filters

$$E_j = \sum_{k=0}^{N/2-1} \phi_j(k) A_k, \quad 0 \leqslant j \leqslant J-1, \tag{5}$$

where J is the number of filters, ϕ_j is the jth filter, and A_k is the amplitude of $X[k]$

$$A_k = |X[k]|^2, \quad 0 \leqslant k < N/2. \tag{6}$$

Step 6. *DCT*.The MFCCs for the ith frame are computed by performing DCT on the logarithm of E_j

$$C_m^i = \sum_{j=0}^{J-1} \cos\left(m\frac{\pi}{J}(j+0.5)\right)\log_{10}(E_j), \quad 0 \leqslant m \leqslant L-1, \tag{7}$$

where L is the number of MFCCs.

In the proposed method, the filter bank consists of 25 triangular filters, that is, $J = 25$. The length of MFCCs feature vector for each frame is 15 ($L = 15$). After deriving the MFCCs for each frame, we compute the averaged MFCCs of all frames within the syllable

$$f_m = \frac{\sum_{i=1}^{K} C_m^i}{K}, \quad 0 \leqslant m \leqslant L-1, \tag{8}$$

where f_m is the mth MFCC, K is the number of frames within the syllable, and C_m^i denotes the mth MFCC of the ith frame. In the training phase, the averaging of f_m over all training syllables for the acoustic vocalization of the same species is regarded as the mth feature value, F_m. Since the dynamic ranges of f_m's may be different, we perform a linear normalization process to get the final feature vector F'_m

$$F'_m = \frac{F_m - f_m^{\min}}{f_m^{\max} - f_m^{\min}}, \tag{9}$$

where $f_m^{\max}$ and $f_m^{\min}$ denote the maximum and minimum values of the mth MFCC of all f'_ms for the training syllables, respectively.

From Fig. 1, it seems that each syllable has different sound structure. Fig. 2 shows that the spectrograms of these syllables look similar except that of the last syllable. Table 1 shows the feature vectors extracted from these

Fig. 2. Spectrogram for the example Olive Frog (*Rana adenopleur*).

Table 1
Feature vectors for the extracted syllables (SC denotes the subject code)

SC	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
f_1	0.627	0.667	0.606	0.713	0.727	0.791	0.843	0.829	0.680
f_2	0.680	0.684	0.784	0.768	0.784	0.793	0.783	0.801	0.782
f_3	0.290	0.242	0.342	0.325	0.361	0.310	0.283	0.296	0.511
f_4	0.404	0.456	0.241	0.309	0.305	0.256	0.264	0.270	0.284
f_5	0.368	0.324	0.504	0.463	0.416	0.467	0.417	0.383	0.455
f_6	0.156	0.098	0.418	0.406	0.346	0.462	0.291	0.303	0.439
f_7	0.273	0.250	0.266	0.251	0.162	0.260	0.183	0.154	0.258
f_8	0.457	0.459	0.487	0.536	0.467	0.511	0.510	0.493	0.550
f_9	0.592	0.617	0.351	0.512	0.439	0.562	0.760	0.734	0.478
f_{10}	0.809	0.915	0.641	0.662	0.699	0.698	0.873	0.908	0.676
f_{11}	0.671	0.658	0.420	0.490	0.577	0.456	0.466	0.485	0.475
f_{12}	0.506	0.479	0.653	0.661	0.686	0.609	0.426	0.463	0.727
f_{13}	0.523	0.357	0.559	0.541	0.501	0.511	0.437	0.409	0.525
f_{14}	0.451	0.477	0.390	0.453	0.337	0.453	0.450	0.462	0.396
f_{15}	0.593	0.614	0.410	0.418	0.491	0.425	0.467	0.483	0.482

Table 2
Distance between each feature vector extracted from the example syllables and the 30 representative feature vectors with the minimum distance highlighted

SC	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
1	0.7415	0.8188	0.6039	0.5936	0.6587	0.5493	0.7214	0.7292	0.6717
2	0.8217	0.9170	0.6944	0.6371	0.7520	0.6797	0.8420	0.8740	0.7215
3	0.7595	0.8941	0.8354	0.8194	0.8182	0.8604	0.9300	0.9503	0.8349
4	0.6782	0.7808	0.7048	0.6665	0.7097	0.6718	0.7571	0.7804	0.7471
5	0.6993	0.7981	0.6162	0.5661	0.6487	0.5693	0.7071	0.7289	0.5668
6	0.8232	0.8872	0.6899	0.6597	0.7372	0.6684	0.8217	0.8341	0.6709
7	0.3099	0.3917	0.4410	0.2813	0.2748	0.2685	0.2748	0.2622	0.3777
8	0.9130	1.0213	0.6975	0.6980	0.7500	0.6806	0.8556	0.8842	0.6811
9	0.8558	0.9303	0.6713	0.6936	0.7061	0.6909	0.8221	0.8495	0.7285
10	0.9542	1.0621	0.7159	0.7853	0.7150	0.7850	0.9780	0.9661	0.7123
11	0.7517	0.8106	0.5296	0.5299	0.6268	0.5263	0.7261	0.7387	0.6011
12	1.0211	1.1240	0.8332	0.8651	0.9174	0.8173	0.9882	1.0073	0.9110
13	0.6558	0.7210	0.6242	0.5835	0.6120	0.5959	0.6438	0.6716	0.6317
14	0.6947	0.8062	0.6483	0.6258	0.6534	0.6830	0.8248	0.8373	0.6077
15	0.5267	0.5939	0.6179	0.5592	0.5456	0.6100	0.7183	0.7158	0.5603
16	0.7915	0.8734	0.7215	0.6223	0.7144	0.6161	0.7501	0.7792	0.6388
17	0.6448	0.7294	0.6343	0.5876	0.6302	0.6157	0.7493	0.7423	0.5883
18	0.6568	0.7204	0.5167	0.5368	0.5833	0.5677	0.7192	0.7357	0.6165
19	1.2702	1.3241	1.3281	1.3180	1.3344	1.3571	1.4431	1.4364	1.2818
20	1.1940	1.2704	1.2935	1.3012	1.3162	1.3569	1.4272	1.4301	1.2355
21	0.6753	0.7434	0.4479	0.4495	0.5182	0.4171	0.5986	0.6129	0.5046
22	0.8649	0.9368	0.7933	0.7902	0.8599	0.8153	0.9374	0.9430	0.7289
23	0.7865	0.8738	0.7021	0.6580	0.6753	0.6612	0.7866	0.7947	0.5952
24	1.0395	1.1656	0.7796	0.8132	0.9295	0.7812	0.9507	0.9850	0.8778
25	0.7724	0.8922	0.7335	0.6541	0.7303	0.6930	0.8256	0.8609	0.6843
26	0.8508	0.9501	0.6928	0.6290	0.7261	0.6534	0.8123	0.8423	0.6436
27	1.0673	1.0742	0.9840	0.8746	1.0122	0.8696	1.0030	0.9894	0.9164
28	0.9436	1.0374	0.7056	0.7272	0.7848	0.7024	0.8491	0.8737	0.7178
29	0.7372	0.8596	0.5467	0.5054	0.6029	0.5086	0.7217	0.7424	0.4600
30	0.9759	1.0893	0.8906	0.8582	0.9431	0.8531	0.9991	1.0190	0.7676

syllables. From this table, we can see that these feature vectors are very close. Table 2 shows the distance between each feature vector extracted from these syllables and the representative feature vectors corresponding to the 30 kinds of frogs. From this table, we can see that the correct frogs (indexed with subject code 7) can be identified using Euclidean distance between two feature vectors.

2.3. Linear discriminant analysis (LDA)

LDA (Baker, 2004; Duda et al., 2000; Slaney, 2002) aims at improving the classification accuracy at a lower-dimensional feature vector space. LDA deals with discrimination between classes. The goal of LDA is to minimize the within-class distance while to maximize the between-class distance.

In LDA, an optimal transformation matrix from an n-dimensional feature space to a d-dimensional space is determined. The transformation matrix is a linear mapping that maximizes the so-called Fisher criterion J_F

$$J_F(A) = \mathrm{tr}((A^T S_W A)^{-1}(A^T S_B A)). \tag{10}$$

Here S_W and S_B are the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix is defined as

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (X_i^j - \mu_j)(X_i^j - \mu_j)^T, \tag{11}$$

where X_i^j is the ith feature vector of class j, μ_j is the mean vector of class j, C is the number of classes, and N_j is the number of feature vectors in class j. The between-class scatter matrix is given by

$$S_B = \sum_{j=1}^{C} (\mu_j - \mu)(\mu_j - \mu)^T, \tag{12}$$

where μ is the mean vector of all classes.

From Eq. (10), we can see that LDA tries to find a transformation matrix that maximizes the ratio of between-class scatter to within-class scatter in a lower-dimensional space. The optimal solution of Eq. (10) is the transformation matrix, $\boldsymbol{A}_{\mathrm{opt}}$, given by

$$\boldsymbol{A}_{\mathrm{opt}} = \arg\max_A \frac{\mathrm{tr}(A^T S_B A)}{\mathrm{tr}(A^T S_W A)}. \tag{13}$$

The transformation matrix $\boldsymbol{A}_{\mathrm{opt}}$ can be determined by finding the eigenvectors of $S_W^{-1} S_B$.

2.4. Automatic recognition of animal vocalizations

At the recognition part, each input acoustic signal is first segmented into a set of syllables. The averaged MFCCs for each syllable are calculated. The same linear normalization process is applied to each MFCC. The normalized MFCCs are transformed to a lower-dimensional feature vector by the transformation matrix $\boldsymbol{A}_{\mathrm{opt}}$ derived by LDA. For every species, the distance between the feature vector of the test syllable and the feature vector representing this species is calculated. The one with minimum distance is regarded as the identified species. In this paper, the distance between two feature vectors is the Euclidean distance. That is, the subject code that represents the identified species is determined by

$$r = \arg\min_{1 \leqslant k \leqslant C} \sum_{m=1}^{d} (F'_m - F_m^k)^2, \tag{14}$$

where C is the number of classes, d is the dimension of the feature vector, F'_m is the mth feature of the testing syllable,

Table 3
Frog call database (SC denotes the subject code)

SC	Scientific name (popular name)	Ns (Experiment 1)	Ns (Experiment 2)
1	*Bufo bankorensis* (Central Formosan Toad)	38	38
2	*Bufo melanosticus* (Spectacled Toad)	47	48
3	*Hyla chinensis* (Chinese Tree Frog)	45	46
4	*Microhyla butleri* (Butler's Narrow-Mouthed Toad)	15	15
5	*Microhyla heymonsi* (Heymonsi's Narrow-Mouthed Toad)	234	235
6	*Microhyla ornata* (Ornate Narrow-Mouthed Toad)	192	193
7	*Rana adenopleura* (Olive Frog)	35	36
8	*Rana catesbiana* (American Bull Frog)	12	13
9	*Rana guentheri* (Guenther's Amoy Frog)	15	15
10	*Rana kuhlii* (Kuhli's Wart Frog)	52	52
11	*Rana latouchii* (Brown Wood Frog)	95	95
12	*Rana limmocharis* (Indian Rice Frog)	38	38
13	*Rana rugulosa* (Chinese Bull Frog)	97	98
14	*Rana swinhoana* (Swinhoe's Frog)	13	13
15	*Rana sauteri* (Sauter's Frog)	131	132
16	*Rana taipehensis* (Taipei Grass Frog)	27	27
17	*Buergeria japonica* (Japanese Tree Frog)	60	60
18	*Buergeri robusta* (Brown Tree Frog)	66	67
19	*Chirixalus eiffingeri* (Eiffinger's Tree Frog)	10	10
20	*Chirixalus idiootocus* (Meintein Tree Frog)	112	112
21	*Polypedates megacephalus* (White lipped Tree Frog)	22	23
22	*Rhacophorus arvalis* (Farmland Tree Frog)	46	46
23	*Rhacophorus aurantiventris* (Orange-Belly Tree Frog)	38	38
24	*Rhacophorus moltrechti* (Moltrecht's Tree Frog)	190	191
25	*Rhacophorus prasinatus* (Emerald Tree Frog)	52	52
26	*Rhacophorus taipeianus* (Taipei Green Tree Frog)	49	49
27	*Microhyla steinegeri* (Steinger's Narrow-Mouthed Toad)	2	3
28	*Kaloula pulchra* (Malaysian Narrow-Mouthed Toad)	6	6
29	*Rana longicrus* (Long-Legged Frog)	9	9
30	*Rana psaltes* (Harpist Frog)	65	65

Table 4
Cricket call database (SC denotes the subject code)

SC	Scientific name (popular name)	Ns (Experiment 1)	Ns (Experiment 2)
1	*Gryllotalpa fossor* (Mole Cricket)	79	80
2	*Teleogryllus occipitalis* (Oil guord)	24	24
3	*Teleogryllus mitratus* (Oil guord)	3	3
4	*Teleogryllus emma* (Oil guord)	9	10
5	*Gryllus bimaculatus* (Painted Mirror)	5	6
6	*Brachytrupes portentosus* (Formosan Giant Crickets)	66	67
7	*Loxoblemmus equestris* (Coffin-headed cricket)	9	9
8	*Dianemobius flavoantennalis* (Flowered Bell)	22	22
9	*Homoeogryllus japonicus* (Horse bell)	2	3
10	*Scleropterus punctatus* (Rocky bell)	6	7
11	*Oecanthus longicaudus* (Bamboo bell)	6	7
12	*Xenogryllus marmoratus* (Pagoda Bell)	5	6
13	*Anaxipha pallidula* (Yellow Bell)	57	58
14	*Svistella bifasciatata* (Golden Bell)	14	15
15	*Homoeoxipha lycoides* (Inky bell)	4	4
16	*Mecopoda elongta* L. (Weaving Lady)	74	74
17	*Gryllotalpa fossor* (Mole Cricket)	135	135
18	*Teleogryllus occipitalis* (Oil guord)	1	2
19	*Teleogryllus mitratus* (Oil guord)	5	5

F_m^k is the mth feature of the kth species, and r is the subject code for the rth species as shown in Tables 3 and 4.

3. Experimental results

Two audio databases of 30 frog calls and 19 cricket calls derived from compact disk are used for the experiments (see Tables 3 and 4). The sampling frequency is 44,100 Hz and each sample is digitized in 16 bits. Most of the calls are field recordings with additional sounds in the background. Some of the calls are generated by multiple individuals vocalizing simultaneously. Each acoustic signal is first segmented into a set of syllables, in which half is used for training and half for testing. Two scenarios for dividing the syllables equally into training and testing sets are conducted: (1) the syllables are alternately divided into training and testing sets, that is, the odd-numbered syllables are used for training whereas the even-numbered syllables for testing; (2) the first-half of the syllables are used for training and the second half of the syllables are used for testing. In the second scenario, the training and testing syllables may be extracted from the calls of different frogs/crickets. The classification accuracy (CA) is defined as

$$\mathrm{CA} = \frac{N_{\mathrm{CA}}}{N_{\mathrm{S}}} \times 100, \tag{15}$$

where N_{CA} is the number of syllables which were recognized correctly and N_{S} is the total number of test syllables (shown in Tables 3 and 4).

3.1. Experiment 1—alternate training and testing

In this experiment, the syllables are alternately divided into training and testing sets. The odd-numbered syllables are used for training whereas the even-numbered syllables for testing. Tables 5 and 6 compare the recognition results of the proposed averaged MFCC (AMFCC) method with HMM and averaged LPC (ALPC) for 30 frog calls and 19 cricket calls, respectively. In addition, multiple feature vector templates were compared in these two tables, where LPC-i and MFCC-i denote that i ($i > 1$) vector templates are used to model the syllables derived from the same species. From these two tables, we can see that AMFCC greatly outperforms HMM and ALPC. HMM, which exploits the temporal information among the frames within a syllable, does not provide better performance than ALPC or AMFCC. Since variations between consecutive frames within a syllable are not regular and thus temporal information extracted for identification purpose is not very essential. In addition, most of the sounds are recorded in the field with additional sounds/noise in the background. Thus, the feature vector extracted from each syllable is not so stable. On the other hand, the averaged LPC/MFCC can attenuate the effect of background noise by averaging the feature vectors of all frames within the syllable. Therefore, the proposed AMFCC is adequate for the identification of frogs and crickets. From Tables 5 and 6, we can also see that if each species is modeled by more than one vector template, the classification accuracy will increase as well.

3.2. Experiment 2—progressive training and testing

In this experiment, the first-half of the syllables are used for training and the second half of the syllables are used for testing. Tables 7 and 8 compare the recognition results of AMFCC with HMM and ALPC for 30 frog calls and 19 cricket calls, respectively. In addition, multiple feature vector templates were compared in these two tables. From these two tables, we can see that AMFCC greatly outperforms HMM and ALPC. However, if each species is

Table 5
Classification accuracy of 30 frog calls for alternate training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	96	81	97	97	92	92	92
2	75	93	98	100	100	100	100
3	67	100	100	100	100	100	100
4	89	100	100	100	100	100	100
5	80	88	87	87	96	99	98
6	85	89	88	85	93	91	91
7	78	82	94	100	100	100	100
8	77	100	100	100	100	100	100
9	100	100	100	100	100	100	100
10	100	84	98	100	100	100	100
11	83	68	81	77	96	99	96
12	60	89	100	100	100	100	100
13	93	100	97	97	100	100	100
14	100	53	46	62	61	77	77
15	100	83	85	90	96	95	97
16	95	85	89	89	88	89	85
17	99	100	100	100	100	100	100
18	100	98	100	100	98	100	100
19	40	100	100	100	100	100	100
20	91	100	100	100	100	100	100
21	96	63	95	95	95	100	100
22	75	82	85	80	91	93	96
23	67	76	97	97	97	100	100
24	89	95	97	98	100	100	100
25	80	96	98	96	80	90	88
26	85	59	76	80	100	100	100
27	78	100	100	100	100	100	100
28	77	66	67	67	83	100	100
29	100	100	89	100	100	100	100
30	100	86	89	91	100	100	100
Average	81.9	88.9	91.9	92.2	96.8	97.6	97.4

Table 6
Classification accuracy of 19 cricket calls for alternate training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	65	97	97	97	100	100	100
2	92	100	96	100	100	100	100
3	100	100	100	100	100	100	100
4	33	22	89	89	100	100	100
5	100	100	100	100	100	100	100
6	86	84	85	88	100	98	98
7	89	100	89	89	88	89	89
8	100	90	91	91	100	100	100
9	0	0	50	50	0	50	50
10	17	83	100	100	33	100	100
11	83	100	100	100	100	100	100
12	100	100	100	100	100	100	100
13	79	98	100	98	100	100	100
14	64	100	100	100	100	100	100
15	100	100	100	100	100	100	100
16	92	91	84	88	95	96	96
17	99	91	95	94	100	100	100
18	100	100	100	0	100	100	100
19	20	80	80	80	100	100	100
Average	91.4	91.6	92.8	93.3	98.1	98.9	98.9

model by more than one vector template, the classification accuracy does not increase accordingly, especially for the recognition of cricket calls. The reason to explain the phenomenon is that in the case of progressive training

Table 7
Classification accuracy of 30 frog calls for progressive training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	66	89	97	95	97	100	100
2	98	91	96	94	100	100	100
3	89	97	100	100	100	100	100
4	40	100	100	100	100	100	100
5	76	71	70	87	95	89	97
6	82	80	84	86	97	95	93
7	53	58	53	94	100	100	100
8	54	100	100	92	100	100	100
9	93	100	100	100	100	100	100
10	50	76	98	100	100	100	100
11	80	66	74	75	100	100	100
12	68	100	100	100	100	100	100
13	81	97	94	100	100	100	100
14	62	46	54	46	38	46	54
15	77	87	87	91	88	92	75
16	63	85	85	85	92	96	93
17	75	100	100	100	100	100	100
18	96	100	99	100	97	97	97
19	100	100	100	100	100	100	100
20	99	100	100	100	100	100	100
21	43	60	74	78	56	48	74
22	15	30	48	50	73	72	61
23	79	68	87	82	100	100	100
24	94	96	98	98	99	99	100
25	67	96	96	96	96	100	100
26	94	85	92	90	100	100	100
27	0	100	100	33	100	100	0
28	0	66	100	50	100	100	83
29	89	88	89	56	100	100	33
30	98	69	80	69	100	100	98
Average	78.9	83.9	86.8	89.8	96.2	95.5	94.6

Table 8
Classification accuracy of 19 cricket calls for progressive training and testing

SC	HMM (%)	ALPC (%)	LPC-2 (%)	LPC-3 (%)	AMFCC (%)	MFCC-2 (%)	MFCC-3 (%)
1	90	92	94	96	98	100	100
2	83	100	83	83	100	100	100
3	100	100	100	100	100	100	100
4	50	20	90	90	100	100	100
5	50	100	100	100	100	100	100
6	99	97	94	99	100	100	100
7	89	100	89	89	88	89	89
8	68	86	86	95	100	100	100
9	33	100	33	67	100	33	67
10	86	85	86	29	85	86	0
11	71	100	100	0	100	100	0
12	100	100	100	67	100	100	17
13	93	94	98	36	100	100	45
14	87	93	87	7	93	93	7
15	100	100	100	0	100	100	0
16	99	98	93	92	98	99	77
17	100	97	99	84	100	100	46
18	0	100	100	100	100	100	100
19	20	80	60	80	80	60	0
Average	91.2	94.6	94.0	79.5	98.9	98.5	69.1

and testing, the diversity of the training syllables is not large enough to learn all the possible calls, especially for the experiment on cricket calls.

Comparing Tables 3 and 5 (respectively, Tables 4 and 6), we can see that alternate training and testing performs better than progressive training and testing. This is due to the fact that for progressive training and testing not all frog/cricket calls are well trained since the training and testing syllables may be extracted from the calls of different frogs/crickets.

4. Conclusions

In this paper we propose a method capable of identifying frogs/crickets automatically from the sounds they generate. Each syllable corresponding to a piece of vocalization is first segmented. The averaged MFCCs (AMFCC) over all frames within a syllable are used as vocalization features such that the effect of background noise can be attenuated. Linear discriminant analysis (LDA) is used to reduce the feature dimension and increase the classification accuracy. Experimental results have shown that AMFCC greatly outperforms HMM and ALPC. In the experiments, the average classification accuracy is up to 96.8% and 98.1% for 30 kinds of frog calls and 19 cricket calls, respectively. From the experimental results, we can see that the proposed simple approach is adequate for the identification of frogs and crickets since the sounds generated by them are simpler than other sounds. If the animal sounds are more complex like bird songs in which the types of sounds and the syntactical arrangements of the sounds change significantly, a more complicated system may be required to do the recognition process.

Acknowledgments

The authors would like to thank the anonymous referees for their valuable comments that improved the representation and quality of this paper. This research was supported in part by Chung Hua University under contract CHU-94-TR-02 and the National Science Council of ROC under contract NSC-92-2213-E-216-020.

References

Baker, M.C., 2004. The chorus song of cooperatively breeding laughing kookaburras: characterization and comparison among groups. Ethology 110 (1), 21–35.

Duda, R., Hart, P., Stork, D., 2000. Pattern Classification. Wiley, New York.

Harma, A., 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. Internat. Conf. on Acoust. Speech Signal Process. 5, 545–548.

Kogan, J.A., Margoliash, D., 1998. Automated recognition of bird song elements from continuous recordings using DTW and HMMs. Journal of the Acoustical Society of America 103 (4), 2185–2196.

Lamel, L.F., Rabiner, L.R., Rosenberg, A.E., 1981. An improved endpoint detector for isolated word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 29, 777–785.

Li, D., Sethi, I.K., Dimitrova, N., McGee, T., 2001. Classification of general audio data for content-based retrieval. Pattern Recognition Letters 22 (5), 533–544.

Lu, G.J., 2001. Indexing and retrieval of audio: A survey. Multimedia Tools and Applications 15, 269–290.

Picone, J.W., 1993. Signal modeling techniques in speech recognition. Proceedings of the IEEE 81, 1215–1247.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Slaney, M., 2002. Mixture of probability experts for audio retrieval and indexing. IEEE Internat. Conf. on Multimedia Expo. 1, 345–348.

Tian, Y., Wang, Z., Lu, D., 2002. Nonspeech segment rejection based on prosodic information for robust speech recognition. IEEE Signal Processing Letters 9 (11), 364–367.

Vergin, R., O'Shaughnessy, D., Farhat, A., 1999. Generalized Mel-frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. IEEE Transactions on Speech and Audio Processing 7 (5), 525–532.

Wold, E., Blum, T., Keislar, D., Wheaten, J., 1996. Content-based classification, search, and retrieval of audio. IEEE Multimedia Magazine 3 (3), 27–36.

Zhang, T., Kuo, C.-C.J., 2001. Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on Speech and Audio Processing 9 (4), 441–457.