Insect sound recognition based on MFCC and PNN

ZHU Le-Qing
College of Computer Science and Information Engineering
Zhejiang Gongshang University, Hangzhou,China
e-mail: zhuleqing@zjgsu.edu.cn

***Abstract*—This study aims to provide general technicians who manage pests in production with a convenient way to recognize insects. A viable scheme to identify insect sounds automatically is proposed by using sound parameterization techniques that dominate speaker recognition technology. The acoustic signal is preprocessed, segmented into a series of sound samples. Mel-frequency cepstrum coefficient(MFCC) is extracted from the sound sample as sound features, and probabilistic neural network(PNN) is trained with given features. The testing samples are classified by the PNN finally. The proposed method is evaluated in a database with acoustic samples of 50 different insect sounds. The recognition rate was above 96%. The test results proved the efficiency of the proposed method.**

***Keywords*—insects, sound recognition,MFCC,PNN**

I. INTRODUCTION

The detection and identification of insect pests is often carried out manually using trapping methods, however, recent advances in signal processing and computer technology have introduced the possibility of automatically identifying species by several means including image analysis and acoustics detection. Research into the automated identification of animals by bioacoustics is becoming more widespread mainly due to difficulties in carrying out manual surveys. Although a considerable number of studies have been devoted to the problem of speaker identification, automated acoustic species identification has been considered just a marginal field of pattern recognition and literature on this subject is sparse. Basically, acoustic identification of insects is based on their ability to generate sound either deliberately, as a mean of communication, or as a by-product of eating, flying or locomotion. Provided that the bioacoustic signal produced by insects follows a consistent acoustical pattern that is species specific, it can be exploited for detection and identification purposes. Riede shows that insect sound emissions provide a reliable taxonomic clue and thus can be used to measure biodiversity in his documents[1].

The problem of acoustic insect identification can be divided into two major stages. The first is feature extraction and the second is classification of insects based on the extracted sound features. The features should be capable of separating the insect species from each other in its space, whereas the classifier should be tuned to differentiate the different classes in given feature space. Han [2] used frequency spectrum analysis and BP neural network to recognize stored producted insects, the method was tested in a database including three species(Sitophilus oryzae, Sitophilus zeamais and Tribolium castaneum) and got the recognition rate as high as 81% . Chesmore et al.[3] investigated techniques for automatically identifying Orthoptera (grasshoppers and crickets) with time domain signal processing and artificial neural networks. 25 species of British Orthoptera have been selected as a test set and preliminary results indicate very high classification rates. Pinhas et al. [4] developed a mathematical method to automatically detect acoustic activity of the red palm weevil, which utilize Vector quantization (VQ) and Gaussian mixture modeling (GMM). Ganchev et al.[5] use dominant harmonic, rhythm and duration of pulsations and the 23 linear frequency cepstral coefficients(LFCCs) as feature vector after normalization, and recognize different insect sounds with Probabilistic Neural Network (PNN)-based, Gaussian Mixture Models (GMM)-based, and Hidden Markov Model (HMM)-based classifiers, the approach was evaluated on the singing insects of the North America collection (SINA) and got high accuracy.

This paper use MFCC to extract features from insect sounds, use PNN to classify sounds. We evaluate our approach on stored product insect movement and feeding sounds, movement and feeding sounds of soil invertebrates, defensive stridulation of soil insects, movement and feeding sounds of insects in wood, movement and feeding sounds of insects in plants, wing and abdominal vibration sounds, identification accuracy that exceeds 96 % has been achieved on recognizing specific species.

II. MATERIALS AND PREPROCESSING

A. Sound recordings and data

The material used in this paper is insect sound library[6] established by Richard Mankin's research team from agricultural research service(ARS) of United States department of agriculture (USDA). The time durations of these recorded insect sounds are between 3sec and 60sec. Directly extracting feature from those long sound files not only would be high in computational complexity, but also would affect the recognize accuracy since the background noise are analyzed together with useful signals. An active segmentation as long as 1.2sec is enough to extract useful parameters, which could decrease the computational burden apparently. We segment every sound file into several samples in the preprocessing period according to the sound activity. The relative static segmentations are deleted. A new sound library is constructed from those

This research was supported by the National Hi-Tech Research and Development Program (863) of China (2006AA10Z211).

978-0-7695-4356-7/11 $26.00 © 2011 IEEE
DOI 10.1109/CMSP.2011.100

resulting samples and our experiments are carried on with those preprocessed samples.

B. *Preprocessing*

Assuming that all the input insect sounds are digital signals that have been sampled and quantified, the preprocessing would include magnitude normalization and segmentation.

1) Normalization

The sound signal is normalized by dividing the maximum value of the magnitude, i.e.:

$$\tilde{x}(i) = x(i) / \max_{0 \le i < n-1} x(i) \tag{1}$$

where $x(i)$ is the original signal， $\tilde{x}(i)$ is the normalized signal，n is the length of the signal.

2) Pre-emphasis

Since sound signal degraded in power with the increasing of frequency, most of the energy concentrates in lower frequency bands, and the signal-to-noise ratio of high frequency components would degraded to an unacceptable level. Pre-emphasis is a way to boost only the signal's high-frequency components, while leaving the low-frequency components in their original state. The pre-emphasis factor α is computed as

$$\alpha = \exp(-2\pi F \Delta t) \tag{2}$$

where Δt is the sampling period of the sound. The new sound y is then computed as:

$$H(z) = 1 - \alpha z^{-1} \tag{3}$$

3) Segmentation

The sound signal is first enframed with certain width, the most expressive parts are extracted from them. The segmentation is based on the detector of acoustic activity, which estimates the pre-emphasis energy for a frame of K successive samples as:

$$E(k) = \sum_{i=1}^{K} (x(kL+i) - ax(kL+i-1))^2 \qquad k = 0,\ldots,M-1 \tag{4}$$

where x is the input signal，k is the frame number，L is a predefined step size which defines the degree of overlapping between two successive frames, and

$$W = \lfloor (K - N + L)/L \rfloor \tag{5}$$

is the number of frames in a recording with length of K sample points. The operator $\lfloor\ \rfloor$ stands for rounding towards the smaller integer value. N is the amount of successive sample points in a frame. Since the subsequent estimates of the energy are for overlapping groups, the precision of border detection depends on the step size L. In this paper, we consider L=80 (equivalent to time resolution 3.2 milliseconds at 25000 Hz sampling frequency), which provides a good trade-off between temporal resolution and computational demands. For obtaining a smooth estimation of $E(k)$ we used a group size N=256 samples, which corresponds to frame size of 10.24 milliseconds. We use the 10% maximum value in short-term energy sequence as threshold th, if the short-term energy is less than th, we consider the signal is background sound. The data between two neighboring background sound is segmented out as one sample. If this sample is too short, we deleted it as noise. The maximum length of a sample is 1.2 seconds, if the segmented sample is longer than 1.2s, just shortened it to 1.2 seconds with the peak value in the middle of the window. As shown in Fig.1, a sound signal with length of 6 seconds is broken into three samples that are less than 1.2s after normalization.

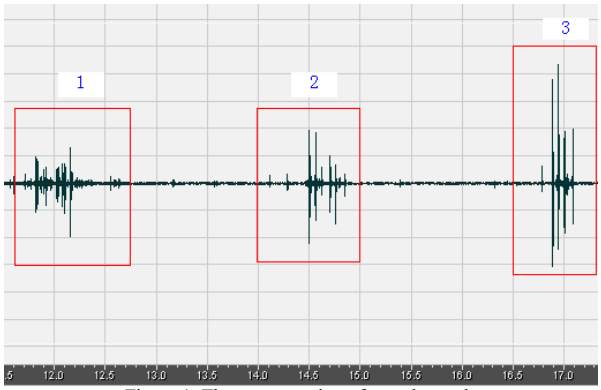

Figure 1. The segmentation of sound sample

III. MFCC FEATURE EXTRACTION

The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC[7, 8]. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. The relationship between Mel-frequency and linear frequency is shown as (Fig.2)：

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700}) \tag{6}$$

MFCCs are commonly derived as follows (fig.3):

1) Pre-amphasis: refer to section II.

2) Hamming windowed：Sound signal is quasi stationary, only stationary in short term. In order to analyze it with methods used in stationary signals, the quasi stationary signal should be segmented into short pieces that we call frames. A window function should be added on to the signal to segment out a piece of wave that contains N sampling points. Rectangular window function would

bring about Gibbs phenomenon at end points. In order to minimize the signal discontinuities at the boundaries of each frame, we multiply each frame with a raised cosine windowing function—Hamming window:

$$\omega_H = 0.54 - 0.46\cos(\frac{2\pi n}{n-1}) \qquad (7)$$
$$n = 0,1,...,N-1$$

where N is the length of a frame which equals to the width of Hamming window. In order to capture information that may occur at the window boundaries, frames are somewhat overlapped.

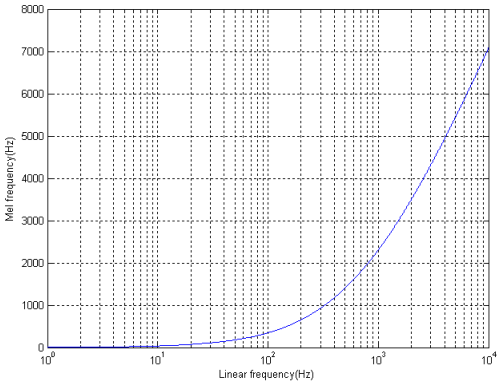

Figure 2. The relation between Mel pitch and frequency

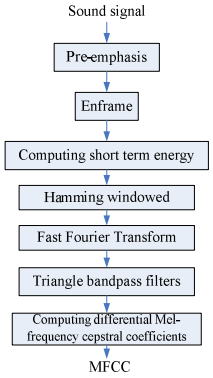

Figure 3. The flowchart of MFCC extraction

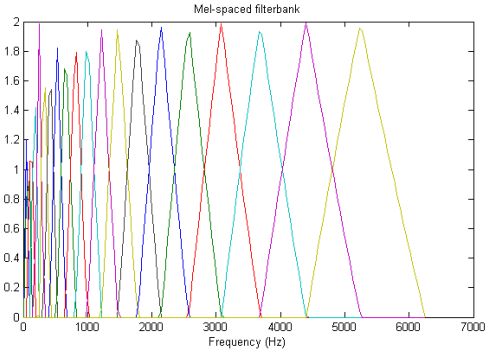

Figure 4. Bank of Mel filters

3) We set frame length N to 256, take Fast Fourier Transform(FFT) of each frame, then the spectrum of the m^{th} frame is:

$$S(k,m) = \sum_{n=0}^{255} s(n,m)\exp(-f\frac{2\pi mk}{256}) \qquad (8)$$

where $\{s(n,m)|n=0,1,\ldots,255\}$ are 256 sampling points of the m^{th} frame. The square modulus of the above spectrum makes the power spectrum.

4) Map the powers of above obtained spectrum to mel scale, filtered with M Mel bandpass filters, resulting in a group of coefficients $m1,m2,\ldots$. The Mel filters are actually triangular overlapping windows in mel scale(fig.4).

5) Take the logs of the powers at each of the mel frequencies.

6) Take the discrete cosine transform of the list of mel log powers, as if it were a signal.

The MFCCs are the amplitudes of the resulting spectrum:

$$C_n = \sum_{k=1}^{M} \ln x'(k)\cos[\pi(k-0.5)n/M] \qquad (9)$$
$$n = 1,2,...,L$$

where $x'(k)$ is input power spectrum of the k^{th} filter. M is the number of Mel filters, L is the number of frames.

Standard MFCC can only reflect the static characteristic of sound. First order differentials of MFCC (△MFCC), a kind of dynamic parameter, can reflect dynamic characteristic of sound and has better robusticity. △MFCC can be calculated as:

$$d(n) = \frac{1}{\sqrt{\sum_{i=-k}^{k} i^2}} \sum_{i=-k}^{k} i \times c(n+i) \qquad (10)$$

where c is MFCC, d is first order differentials of MFCC, k is a constant, generally k=2.

IV. PNN-BASED IDENTIFICATION

The probabilistic neural network was developed by Donald Specht[9, 10]. This network provides a general solution to pattern classification problems by following an approach developed in statistics, called Bayesian classifiers. The probabilistic neural network uses a supervised training set to develop distribution functions within a pattern layer. These functions, in the recall mode, are used to estimate the likelihood of an input feature vector being part of a learned category, or class. The learned patterns can also be combined, or weighted, with the a priori probability, also called the relative frequency, of each category to determine the most likely class for a given input vector. If the relative frequency of the categories is unknown, then all categories can be assumed to be equally likely and the determination of category is solely based on the closeness of the input feature vector to the distribution function of a class.

Probabilistic neural networks can be used for classification problems. When an input is presented, the first layer computes distances from the input vector to the training input vectors and produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce a vector of probabilities as its net output. Finally, a competed transfer function on the output of the second layer picks the maximum of these probabilities, and produces a 1 for that class and a 0 for the other classes. The architecture for this system is shown as fig.5.

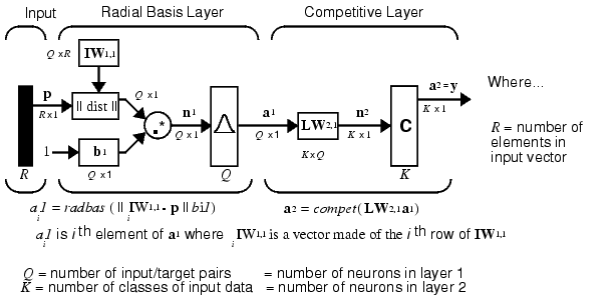

Figure 5. Structure of neural network

It is assumed that there are Q input vector/target vector pairs. Each target vector has K elements. One of these elements is 1 and the rest are 0. Thus, each input vector is associated with one of K classes.

The first-layer input weights, IW1, 1 (net.IW{1,1}), are set to the transpose of the matrix formed from the Q training pairs, P'. When an input is presented, the || dist || box produces a vector whose elements indicate how close the input is to the vectors of the training set. These elements are multiplied, element by element, by the bias and sent to the radbas transfer function. An input vector close to a training vector is represented by a number close to 1 in the output vector a1. If an input is close to several training vectors of a single class, it is represented by several elements of a1 that are close to 1.

The second-layer weights, LW1, 2 (net.LW{2,1}), are set to the matrix $\boldsymbol{T}$ of target vectors. Each vector has a 1 only in the row associated with that particular class of input, and 0's elsewhere. The multiplication Ta1 sums the elements of a1 due to each of the K input classes. Finally, the second-layer transfer function, compete, produces a 1 corresponding to the largest element of n2, and 0's elsewhere. Thus, the network classifies the input vector into a specific K class because that class has the maximum probability of being correct.

The PNN introduced by Specht is essentially based on the well-known Bayesian classifier technique commonly used in many classical pattern-recognition problems.

Consider a pattern vector 'x' with 'm' dimensions that belongs to one of two categories K_1 and K_2. Let $F_1(x)$ and $F_2(x)$ be the probability density functions (PDF) for the classification categories K_1 and K_2, respectively[11].

From Bayes' discriminant decision rule, 'x': belongs to K_1 if:

$$\frac{F_1(x)}{F_2(x)} > \frac{L_1}{L_2}\cdot\frac{P_2}{P_1} \tag{11}$$

Conversely, 'x' belongs to K_2 if

$$\frac{F_1(x)}{F_2(x)} < \frac{L_1}{L_2}\cdot\frac{P_2}{P_1} \tag{12}$$

where L_1 is the loss or cost function associated with misclassifying the vector as belonging to category K_1 while it belongs to category K_2, L_2 is the loss function associated with misclassifying the vector as belonging to category K_2 while it belongs to category K_1, P_1 is the prior probability of occurrence of category K_1, and P_2 is the prior probability of occurrence of category K_2 In many situations, the loss functions and the prior probabilities can be considered equal. Hence the key to using the decision rules given by equations (11) and (12) is to estimate the probability density functions from the training patterns. In the PNN, a nonparametric estimation technique known as Parzen windows is used to construct the class dependent probability density functions (pdf) for each classification category required by Bayes' theory. This allows determination of the chance a given vector pattern lies within a given category. Combining this with the relative frequency of each category, the PNN selects the most likely category for the given pattern vector. Both Bayes' theory and Parzen windows are theoretically well established, have been in use for decades in many engineering applications, and are treated at length in a variety of statistical textbooks. If the j^{th} training pattern for category K_1 is x_j, then the Parzen estimate of the pdf for category K_1 is given by equation (13) as

$$F(x) = \frac{1}{(2\pi)^{m/2}\sigma^m n}\sum \exp\left[\frac{(x-x_j)^T(x-x_j)}{2\sigma^2}\right] \tag{13}$$

where, n is the number of training patterns, m is the input space dimension, j is the pattern number, and σ is an adjustable smoothing parameter. However, the choice of σ in general has been found to be not too sensitive to variations in its value.

Probabilistic neural networks (PNN) can be used for classification problems as these networks generalize well. A PNN is guaranteed to converge to a Bayesian classifier providing it is given enough training data. The only factor that needs to be selected for training is the smoothing factor, which is the deviation of the Gaussian functions -too small deviations cause a very spiky approximation that cannot generalize well; too large deviations smooth out details.

V. EXPERIMENTAL RESULTS

Proposed approach is implemented with Matlab on Intel Core2 2.16GHz,1G RAM PC. There are total 50 different insect sounds in the experiment. Every class of sound has several different samples, 1 sample is used to establish the PNN, the rest are used as testing.

Table I is the performance evaluation of proposed method. The spread value of PNN was chosen as 0.5. The identification rate with less than 1.2 seconds of testing data on 50 insect sounds for MFCC exceeds 96% and average identification time

is less than 10 seconds. Table 2 summarizes the final results for different types of insect sounds. As we can see, except that the recognition rate for the movement and feeding sounds of insects in wood sounds is relatively lower, 92.44%, the identification rate of the rest are all above 93%.

TABLE I. PERFORMANCE OF AUTOMATIC ACOUSTIC INSECT IDENTIFICATION ALGORITHM BASED ON MFCC AND PNN

Preprocessing time(ms)	364
Feature extraction time（ms）	244
Training time（s）	625
Identification time(s)	9.547
Identification rate(%)	96.17

TABLE II. A COMPARISON ON RECOGNITION ACCURACY FOR DIFFERENT TYPES OF INSECT SOUNDS

insect sound types	MFCC-PNN
Stored Product Insect movement and feeding sounds (7 classes)	0.9375
Movement and feeding sounds of soil invertebrates[12-15] (15 classes)	0.9931
Defensive stridulation of soil insects[16](2 classes)	1.000
Movement and feeding sounds of insects in wood[17-19] (13 classes)	0.9244
Movement and feeding sounds of insects in plants(1 class)	1.000
Wing and abdominal vibration sounds[20](12 classes)	0.9658

VI. CONCLUSION

In this paper we accomplish the task of the automatic acoustical identification of insects by employing signal parameterization methods and state-of-the-art pattern matching techniques in a manner that resembles the methodology of speaker recognition. The proposed automatic identification method employed MFCC as sound feature and PNN as classifier which demonstrated good performance in recognizing 50 specific sounds of insects. However, much work has still to be carried out, most of the sound files in this paper were selected from noise-free sections of recorded signal, in the future, we'll try to detect and separate insect sounds from noises mixed background before recognition.

ACKNOWLEDGMENT

The authors would like to acknowledge Richard Mankin's research team in agricultural research service(ARS) of United States department of agriculture (USDA) for sharing their insect sound library.

REFERENCES

[1] Riede, K., *Acoustic monitoring of Orthoptera and its potential for conservation.* Journal of Insect Conserv., 1998. **2**: p. 217-223.

[2] Han P. Voice-pattern recognition of storedproducted insects. *Computer engineering*. **29**(22): p. 151-154, 2003.

[3] Chesmore E.D., Nellenbach C., *Acoustic methods for the automated detection and identification of insects.* Acta Horticulturae 2001(562): p. 223-231.

[4] Pinhas J. , Soroker V., Hetzroni A., Mizrach A., Teicher M. , Goldberger J., *Automatic acoustic detection of the red palm weevil.* Computers and Electronics in Agriculture 2008. **63**(2): p. 131-139

[5] Ganchev, T., Potamitis, I., Fakotakis, N. *Acoustic monitoring of singing insects*. in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2007. Honolulu.

[6] Mankin, Richard, *Sound Library* 2009.

[7] Mermelstein, P., *Distance measures for speech recognition, psychological and instrumental.* Pattern Recognition and Artificial Intelligence, 1976: p. 374–388.

[8] Davis S.B., Mermelstein P., *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980. **28**(4): p. 357–366.

[9] F., Specht D. *Probabilistic neural networks for classification, mapping or associative memory*. in *Proceedings of IEEE international conference neural networks* 1988.

[10] D.F., Specht, *Probabilistic neural networks.* Neural Networks, 1990. **3**(1): p. 109-118.

[11] Gill, G.S., Sohal,J. S., *Battlefield Decision Making: A Neural Network Approach.* Journal of Theoretical and Applied Information Technology, 2008. **4**: p. 697-699

[12] Arbogast, T.R., Kendra E.P. , Mankin,W.R., Mcgovern,E.J., *Monitoring insect pests in retail stores by trapping and spatial analysis.* Journal of economic entomology, 2000. **93**(5): p. 1531-1542

[13] Brandhorst-Hubbard JL, Flanders KL, Mankin RW, Guertal EA, Crocker RL., *Mapping of soil insect infestations sampled by excavation and acoustic methods.* Journal of economic entomology, 2001. **94**(6): p. 1452-14588.

[14] Zhang ML., Crocker RL., Mankin RW.,Flanders KL.,Brandhorst-Hubbard JL. , *Acoustic identification and measurement of activity patterns of white grubs in soil.* Journal of Economic Entomology 2003. **96**(6): p. 1704-1710.

[15] Mankin, RW, Mizrach, A., Hetzroni, A., Levsky, S., Nakache, Y., Soroker, V. , *Temporal and spectral features of sounds of wood-boring beetle larvae: identifiable patterns of activity enable improved discrimination from background noise.* Florida Entomologist 2008. **91**(2): p. 241-247.

[16] Vulinec, Kevina, *Dung beetles (Coleoptera: Scarabaeidae), monkeys, and conservation in Amazonia* The Florida Entomologist, 2000. **83**(3): p. 229-241

[17] Thoms, EM, *Use of an acoustic emissions detector and intragallery injection of spinosad by pest control operators for remedial control of drywood termites (Isoptera: Kalotermitidae).* Florida Entomologist 2000. **83**(1): p. 64-72.

[18] Mankin, RW, Smith,MT,Tropp,JM,Atkinson,EB,Jong,DY, *Detection of Anoplophora glabripennis (Coleoptera : cerambycidae) larvae in different host trees and tissues by automated analyses of sound-impulse frequency and temporal patterns.* Journal of Economic Entomology 2008. **101**(3): p. 838-849.

[19] Mankin RW., Osbrink WL., Oi FM.,Anderson JB. , *Acoustic detection of termite infestations in urban trees.* Journal of Economic Entomology 2002. **95**(5): p. 981-988.

[20] Hay-Roe MM, Mankin RW *Wing-click sounds of Heliconius cydno alithea (Nymphalidae : Heliconiinae) butterflies.* Journal of insect behavior 2004. **17**(3): p. 664-674.