DISCRIMINATING HUMAN WHISPERS FROM PEST SOUND DURING RECORDINGS IN COCONUT PALM GROOVES USING MFCC AND VECTOR QUANTIZATION

Betty Martin[1] and Vimala Juliet[2]

[1]Research Scholar, Department of Electronics & Control
Sathyabama University, Chennai, India.
[2]Professor and Head, Department of ICE
SRM University, Chennai, India
E:Mail: bettymartin1205@yahoo.co.in

Abstract

The key pest of horticultural and ornamental palm species in Asia are endangering landscape. The pest specially the Red Palm Weevil bores deep into palm crowns ,trunks and offshoots and remains hidden from visual inspection until the palms are nearly dead. Acoustic signals of this boring pest can be recorded from the infested palms using the off-the-shelf recording devices. Whispers from human during recording may result in wrong interpretations confirming the presence of some acoustic activity of the pest. This can be differentiated by using MFCC and VQ. The MFCC is applied on input for sound identification and VQ a quantization method deals with vectors. This paper illustrates how human whispers can be differentiated from pest's acoustic activity during recording.

Keywords: Mel frequency coefficient characteristics (MFCC), Vector Quantization (VQ), Coconut Palm grooves.

I. INTRODUCTION

Sound recognition principle deals with two methods. The first one is sound identification and the other is sound verification. In sound identification, the main task is to use a sound sample to select the identity of the owner that produced the sound from among a population of different sounds. In sound verification, the main task is to use a sound sample to test whether the owner who claims to have produced the sound has in fact done so[1].Sound recognition methods can be divided into text dependent and text independent methods.Fig1 shows block diagram of speaker identification. The MFCC extraction can be well explained with the block diagram of MFCC processor. The block diagram of MFCC processor is shown in Fig 2.

The MfCC extraction starts with the framing of audio input signal. Hamming window is applied to the selected frames. The spectral coefficient of windowed frames are displayed through the use of FFT. Then the FFT spectral coefficients are processed with Mel filter bank to convert them to Mel scale. The logarithm of Mel spectral coefficient are then transformed to frequency domain by discrete cosine transform. Finally the MFCC undergoes Cepstral Mean subtraction to reduce noise present in it.

VQ is a quantization method that deals with vectors rather than individual samples. The training pattern is formed by linking together the MFCC's extracted from audio input training samples. Depending on the size of code book, training patterns are chosen to form code vectors that make a codebook. Here the codebook and training pattern are matrices. The vector is a row of matrix. The codebook is generated by randomly selecting the code vector from training data and calculating centroids that will create codebook.

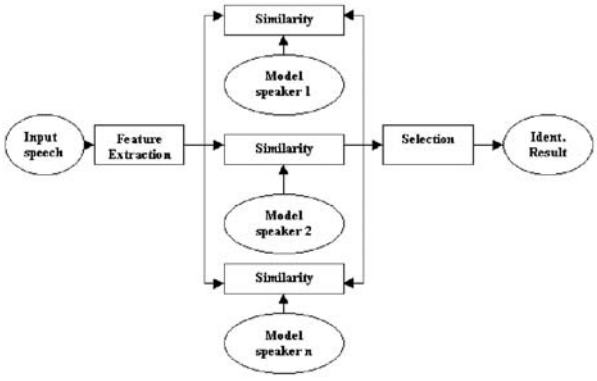

Fig. 1. Block diagram of speaker identification

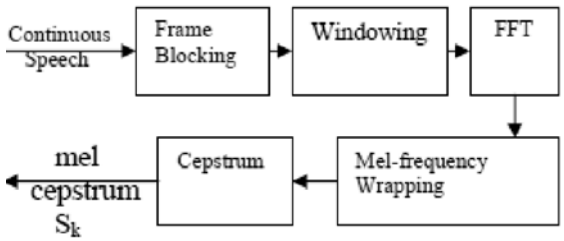

Fig. 2. Block diagram of the MFCC processor

II. MATERIALS AND METHODS

MATLAB is an integrated technical computing environment which combines numeric computation, advanced graphics, visualization and a high level programming language. Using MATLAB, we first form the codebook [3] .The general steps for formation of codebook to identify the sound are given as

- Sound Database
- MFCC
- Training Matrix
- VQ
- Codebook

Fig.3 shows the System overview of creating codebook

A. Sound Database

The first step is to create the sound database containing digitized sound recording of pest and human whispers. The audio input was recorded in the coconut grooves of Tanjore , Pattukottai district, Tamilnadu. The recordings also contained voices of people who helped in recording. The recording was done by a digital voice recorder.

B. Mel frequency Coefficient characteristics

The MFCC is a representation of short term power spectrum of a sound based on linear cosine transform of a logarithmic power spectrum on a nonlinear Mel scale of frequency. These recordings underwent MFCC extraction. Training matrices for each sound were performed from MFCC obtained. MFCC extraction in sequence can be explained.

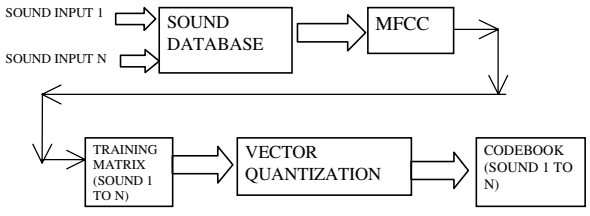

Fig. 3.System overview of creating codebook

III. MFCC Feature Extraction

MFCC feature extraction can be explained as follows.

A. Framing

The audio input signal is segmented into parts ranging from 10-40msec. These divisions are quasi-stationary hence it is framed before extraction occurs. The chosen frame size is 256 samples. If the final frame of audio signal is less than 256 samples the frame is zero padded. Each of the frames is then normalized.

B. Windowing

The frames are the multiplied by a window function. The windowing serves for reduction of spectral distortion that arises due to windowing itself. Hamming window is given by, $W(n), 0 \le n \le N-1$ where

$N =$ number of samples in each frame

$Y[n] =$ Output signal

$X(n) =$ input signal

$W(n) =$ Hamming window, then the result of windowing signal is shown below:

$$Y.n. = Xn * W.n.$$

Figure 4 shows the hamming window applied on audio frame.Then coefficient of a Hamming window is computed from the following equation 1.

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

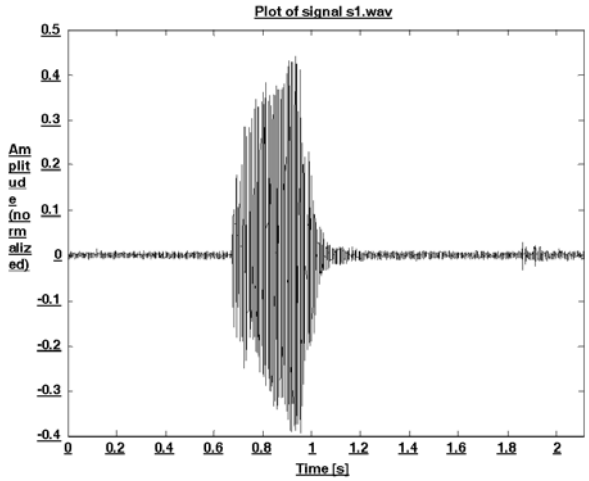

Fig. 4. Hamming window applied on audio frame

C. Fast Fourier Transform - FFT

The next processing step is FFT which converts each frame of n samples from time domain to frequency

domain. The number of samples was selected as 256 since it is in the power of 2 which enables use of FFT. It is a powerful tool and calculates DFT of the input audio frame in an efficient manner. The fast Fourier transform is a fast algorithm to implement Discrete Fourier Transform (DFT) which is defined on the set of N samples.

FFT can be calculated as given in equation 2.

$$X_k = \sum_{n=0}^{N-1} \chi_n e^{-j2\pi kn/N}, \text{ where } k = 0, 1, 2, \ldots, N-1 \quad \ldots (2)$$

D. Mel Frequency Filter Bank

The process of obtaining MFCC involves use of Mel scale filter bank. The spectral coefficient of each frame are converted to Mel scale after applying a filter bank. The Mel scale is logarithmic scale and filter bank is composed of triangular filters that are equally spaced on logarithmic scale. The frequency range in FFT spectrum is very wide and the signal does not follow a linear scale. The bank of filters according to Mel scale is then performed. Fig 5 shows Mel scale filter bank applied on processed frame.

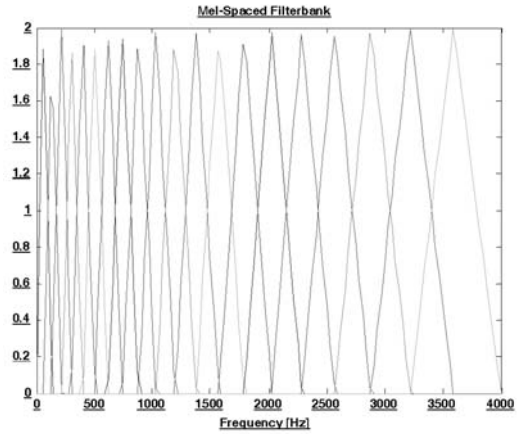

Fig. 5. Mel scale filter bank applied on processed frame

The above figure shows a set of triangular filters used to compute weighted sum of filter spectral components so that output process approximates to Mel scale. Each filters magnitude frequency representation is triangular in shape and is equal to unity at the centre frequency and decreases linearly to zero at centre frequency of 2 adjustant filters[4].Each filter output is sum of its filtered spectral components. The equation used to compute Mel for given frequency f in HZ is given by the equation (3).

Mel scale is calculated as

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right) = 1127 \log_e\left(\frac{f}{700} + 1\right) \quad \ldots\ldots\ldots (3)$$

E. Discrete Cosine Transform – DCT

The Discrete Cosine Transform is applied to the log of the Mel-spectral coefficients to obtain the Mel-Frequency Cepstral Coefficients. The DCT is defined as in equation 4.

$$y(k) = \omega'(k) \sum_{n=1}^{N} \chi(n) \cos\frac{\pi(2n-1)(k-1)}{2N}, k = 1, \ldots, N \quad \ldots (4)$$

Where

$$\omega'(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \le k \le N \end{cases}$$

The result of conversion is called MFCC. The set of coefficient is called acoustic vectors. Each input is transformed into a sequence of acoustic vectors .

F. Cepstral Mean Subtractions – CMS

We convert logarithmic Mel spectrum back to time and this is called MFCC. The cepstral representation of signal spectrum provides a good representation of local spectral properties of signal for given frame analysis. The Mel spectrum coefficients are real numbers which can be converted to time domain using DCT. The MFCC's may be calculated using equation(5)

$$\tilde{C}_n = \sum_{k-1}^{k} (\log \tilde{S}_k)\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right],$$

where $n = 1, 2, \ldots K$ ... (5)

- **Training matrices**

Training matrices for each of the sounds were later formed from available MFCC matrices as obtained in the sequence. The training matrices were then utilized to obtain codebooks which serve as references for each sound after VQ is applied.

- **Vector Quantization - VQ**

VQ is a method that deals with vectors rather than samples. A training pattern is formed by linking

MFCC extracted from available training samples. Depending on the size determined for codebook, training patterns are chosen to form code vectors that make up codebook. Codebook and training pattern are matrices. The codebook can be generated by either randomly selecting code vectors from training data or calculating centroids that will create codebook.

- **Codebook Formation**

Lloyd's Algorithm

For design of codebook we use Lloyd's Algorithm . The description can be done in few steps.[3]

Initialization

Each of the training samples underwent MFCC calculation and were stored as rows of training matrix which serve as input to give rise to codebook representing each input.

Vector Coding

Each vector in the training matrix is categorized with respect to codebook. This is done by calculating Euclidean distance between given training vectors and code vector in codebook. Once the code vector is less, the criteria is achieved.

Codebook Updating

Once training vectors have been labeled, linking them to proper code vectors, the codebook is updated and as a cluster. The centroids of all given clusters are calculated and each centroid replaces code vector indicated by vectors in a cluster.

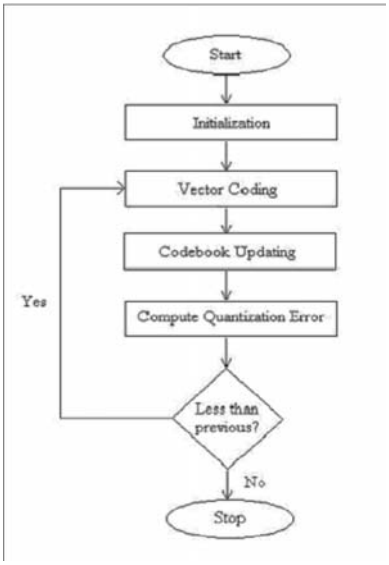

Fig. 6. Lloyd's Algorithm

Quantization Error calculation

This is done by calculating the Euclidean distance between each training vector and each code vector and then adding these distance together. The summation represents total quantization error of codebook.

The algorithm is an iterative process and quantization error determines how many times algorithm can be repeated. If algorithm runs at least two times, quantization error of previous time is compared to newly computed quantization error. If present error is less than previous one, speakers codebook will be modified. If not the algorithm is repeated starting from second process i.e. Vector coding. If previous one is less than present one the algorithm terminates execution.

IV. RESULTS AND DISCUSSION

We take vectors in two dimensional case without loss of information. Figure 7 shows vectors in space for each signal. Associated with each cluster of vectors is a representative code word. The representative codeword is determined to be the closest in Euclidean distance from input vector. The Euclidean distance is defined by:

$$d(x_1 y_1) = \sqrt{\sum_{j=1}^{k} (x_j - y_y)^2}$$

where x_j is the jth component of the input vector, and y_{ij} is the jth component of the codeword y_i. The vector Quantization deals with acoustic vectors and output index of codeword that offers lowest distortion. The lowest distortion is found by evaluating Euclidean distance between input vector and each codeword in codebook. Once the closest codeword is found, index of that codeword is sent out. At the output side, when

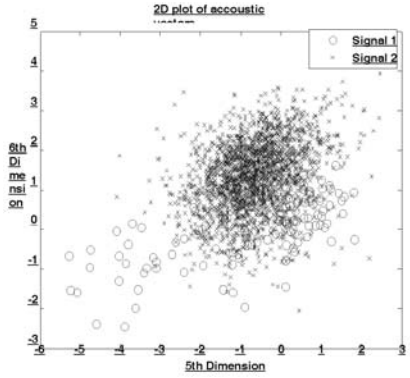

Fig. 7. shows 2D plot of acoustic vectors. The input vectors are marked with ° in red for human

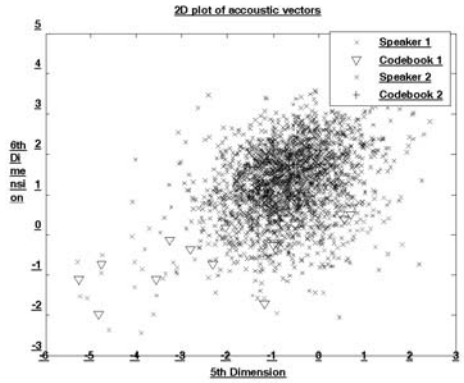

Fig. 8. shows codewords in 2D space. ∇ and + symbol in black represent codeword for the human and pest signal respectively.

it receives the index of codeword, it replaces the index with associated codeword.

In the matching phase the test sample that is to be identified is taken and similarly processed as in training phase to form feature vector. The stored feature vector which gives minimum Euclidean distance with the input sample feature is identified. Figure 8 shows codewords in 2Dspace. ∇ and + symbol in black represent codeword for the human and pest signal respectively. In this result, since the feature vectors do not overlap the resultant, output results in maximum Euclidean distance. This confirms the audio input of two different species.

V. CONCLUSION

This paper shows results of how human whispers can be discriminated from pest sound during recording. The technique explained here uses VQ and MFCC. The result confirms these techniques for discriminating and recognizing sound. Mel frequency and Hamming window combines to give best performance. This work can further be tried with LPC.

REFERENCES

[1] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman "Speaker Identification Using Mel Frequencycepstral Coefficients." *3rd* International Conference On Electrical & Computer Engineering.ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh

[2] Patricia Melin, Jerica Urias, Daniel Solano, Miguel Soto, Miguel Lopez, and Oscar Castillo' Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms'. Engineering Letters, 13:2, EL_13_2_9 (Advance online publication: 4 August 2006).

[3] Jose Boris Sanchez "Speaker Indentification Based On An Integrated SystemCombining Cepstral Feature Extraction And Vector Quantization

[4] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi'Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617

HTTPS://SITES.GOOGLE.COM/SITE/JOURNALOFCOMPUTING/138

ACKNOWLEDGEMENT

We thank Sathyabama University for supporting this work. We thank Dr. I Henry Louis, M.D. Hi Tech Coconut Corporation for his assistance in visual as well as acoustic inspection of the weevil. We would also like to thank Palanidurai Devar for providing his coconut plantations for observation. We thank Mr. Pughalendhi, Sound Engineer for signal recording and analysis. We thank Dr. Xavier Suresh, Head of the department of Bioinformatics for reviewing an early version of the manuscript .We thank Janani Jayaraj Architect for her continuous support through out the project.

Betty Martin is an Assistant Professor in Electronics and Control from Sathyabama University, She is now pursuing her research work in Acoustics. She has published 3 papers in National Conference, 1 International Conference,1 National Journal.