Hidden Markov Model (HMM) Based Murmur Recognition

Sheena Christabel Pravin
Assistant Professor, Department of Electronics and Communication Engineering
Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

Narayani S, Prashanth P
Student, Department of Electronics and Communication Engineering
Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

Abstract- To accommodate the craving need of the vocally handicapped to speak and to provide an added dimension to privacy of communication, we propose Hidden Markov Model (HMM) based Murmur recognition. Data acquisition is done using the Non-Audible Murmur (NAM) Stethoscope–placed on the neck near the glottis; Discrete Wavelet Transform (DWT) is used to de-noise the murmur samples followed by reconstruction. The MFCC features are extracted from the de-noised and reconstructed murmur samples. These feature vectors are used to train the Hidden Markov Model to recognize the murmur. The performance of HMM is evaluated based on its recognition efficiency.

Keywords – De-noising, NAM Stethoscope, DWT, HMM

I. INTRODUCTION

Most of us are gifted with the ability to speak with good speech flow quality but some are deprived of this inherent ability and are vocally handicapped. With various advancements in technology, there is a pressing need that we exterminate this limitation and provide them with a window of opportunity. Also with the growth of the mobile communication, we are able to reach anybody anywhere but seldom do these communications happen with utmost privacy. The peril of losing valuable information in the air to people around, compels us to secure our speech. Murmur recognition is a boon to the vocally deprived; it also helps in secure and private communication. Existing ventures to recognize murmurs are listedbelow: In [1], close talking microphone was used for clean background and NAM microphone for noisy conditions and concluded that NAM microphone was tougher against background noise. Similar proposition by Tomaki Toda, et.al. employed two NAM microphones to detect stereo signals and significantly improved accuracy by reducing non stationary noise[2]. Another work [3] proposed by Tomaki Toda, et.al. discussed various voice conversation techniques for speech signals obtained using NAM microphone and converted it into natural speech using statistical method. Dr. Ganesh S., et.al, proposed HMM for training data and MFCC as features [4]. It was found with the culmination of HMM and MFCC, the recognition accuracy improved in noise based environments. In [5], Campbell N, et.al. used flesh conducted stethoscope for procuring NAM samples and employed Julius Japanese tool kit to test the obtained samples. Also in [6] Panikos Heracleous, et.al showed a relationship between recognition performances and HMM distances; as HMM distance increased, the recognition performance also increased. The refined HMM model by using covariance modeling, parameter estimation and algorithms for training data and compensation of noise was proposed in [7]. Jai Karan Singh, et. al. implemented wavelet transform for de-noising speech signal by thresholding wavelet coefficient, which gave better results even at low noise levels in [8].The new methods to improve degradation of sound quality using blind noise suppression to avoid noise developed during speech is discussed in [9].This paper is organized as: Section II contains murmur acquisition method, Section III explains de-noising of the acquired signal using DWT, Section IV is about HMM training, testing and classification phases, Section V is Results and Discussion and Section VI elaborates conclusion and future applications.

II. EXPERIMENTAL SETUP

Murmur samples were recorded using NAM Stethoscope placed on the neck near the glottis and is connected to the computer via an audio jack. The murmur samples of digits (0-9) were recorded from 20 male speakers and 20 female speakers. The Experimental set up is described in Figure 1.

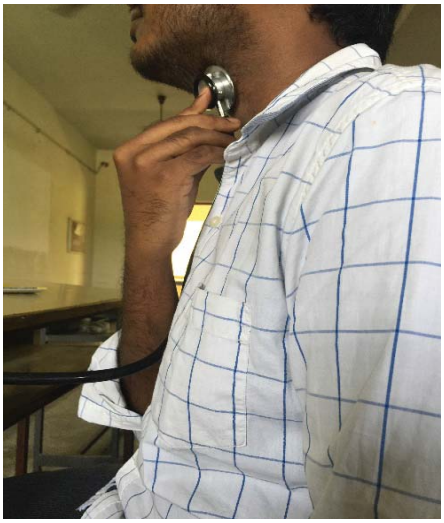


Figure1. Murmur Acquisition using Stethoscope

III. DENOISING OF ACQUIRED SIGNAL USING DWT

The recorded murmur signals are de-noised using Wavelet transform. The Fast Fourier Transform (FFT) gives only the frequency localization of the signals. The partial way to solve the problem of time localization is by using Short Time Fourier Transform (STFT); it has frequency resolution problems due to constant size window. Wavelet Transform, however, provides both time and frequency localization and is hence best suited for de-noising speech signals [10]. The Continuous Wavelet Transform is both time and memory consuming;hence,we chose the Discrete Wavelet Transform. The general method for signal analysis is to decompose the signal to level N using a suitable wavelet. The detailed and approximate co-efficient can be obtained by scaling and dilating the Scaling function and the Wavelet function. The other way to find the wavelet coefficients is to introduce Sub Band Coding.

A biorthogonal wavelet is a wavelet where the associated wavelet transform is invertible but not necessarily orthogonal. Designing biorthogonal wavelets allows more degrees of freedom than orthogonal wavelets. One additional degree of freedom is the possibility to construct symmetric wavelet functions.In the biorthogonal case, there are two scaling functions$\phi, \tilde{\phi}$, which may generate different multiresolution analyses, and accordingly two different wavelet functions$\psi, \tilde{\psi}$. Therefore, the numbers M and N coefficients in the scaling sequences$a, \tilde{a}$may differ. The scaling sequences must satisfy the following biorthogonality condition [11]

$$\sum_{n\in\mathbb{Z}} a_n \tilde{a}_{n+2m} = 2 \cdot \delta \qquad (3.1)$$

Then the wavelet sequences can be determined as

$$b_n = (-1)^n \tilde{a}_{N-1-n} \quad (n = 0, \cdots, N- \qquad (3.2)$$

$$\tilde{b}_n = (-1)^n a_{N-1-n} \quad (n = 0, \cdots, N- \qquad (3.3)$$

The Discrete Wavelet Transform is implemented using the filter banks to perform multi-resolution analysis. The sampled signal is high-pass and low-pass filtered to get the detailed and approximate co-efficient respectively. These detailed co-efficient correspond to details in the data set. The approximate co-efficient obtained in stage I of the filter is fed as input to a stage II and so on until stage III (see Figure 2).

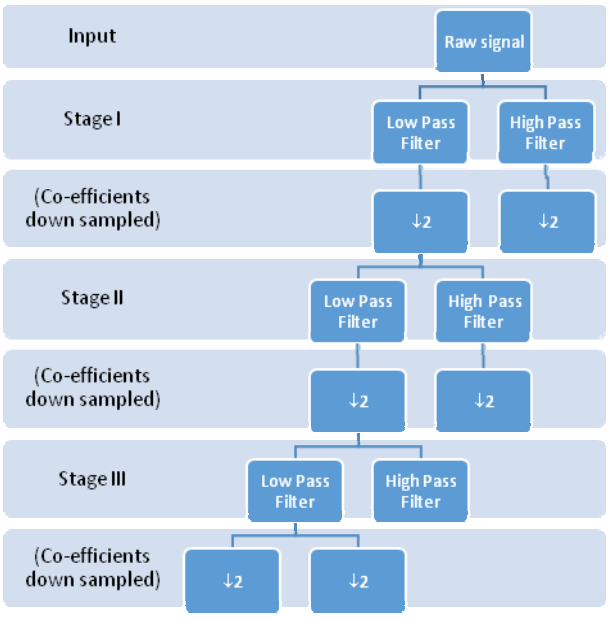


Figure 2. DWT Flow-graph [11]

IV. FEATURE EXTRACTION

Speech feature extraction is a fundamental requirement of any speech recognition system. In a human speech recognition system, the goal is to classify the source files using a reliable representation that reflects the difference between utterances. Cepstrum is the Fourier Transformer of the log with unwrapped phase of the Fourier Transformer. MFCC is very similar in principle with the human ear perception, and are especially good for speech recognition and speech synthesis. MFCC uses human hearing perceptions, which cannot perceive frequencies over one KHz. It consists of six block. The first step enhances the spoken word signal at high frequencies. Then framing facilitates use of FFT. To get energy distribution over frequency domain, we perform FFT. Before applying FFT, we will multiply each frame with window function, to keep the continuity between first and last point. N set of triangular bandpassfilters are multiplied to calculate energy in each band pass filter. Here we use Mel frequency triangular band pass filters. To compress the dynamic range of values we take log Then Discrete Cosine Transform is applied to get the MFCC. They represent the acoustic features of speech. [12, 13, 14]

Fig. 3. MFCC Overall Process

At the acoustic feature extraction stage, input speech is converted into MFCCs of 39 dimensions (12MFCC, 12-ΔMFCC, 12-ΔΔMFCC, P, ΔP and ΔΔP, where P stands for raw energy of the input speech signal).

V. HMM MODELING

In HMM based speech recognition, a Markov model is used for generating the sequence of observed speech vectors corresponding to each word. It is shown in Fig.5. A Markov model is a finite state machine which changes state once every time unit. A speech vector $\boldsymbol{O}_t$ is generated from the probability density $\mathbf{b}_j\mathbf{X}_t$ every time state j is entered at time t. Furthermore, the transition from state *i* to state *j* is also probabilistic and is governed by the

discrete probability a_{ij}. The six state model of this process is shown in the Fig.3. It moves through the state sequence *X*=1,2,2,3,4,4,5,6 in order to generate the sequence $\mathbf{O_1}$to $\mathbf{O}_2$. Notice that, the entry and exit states of a HMM are non-emitting [15].

The joint probability that the model M moving through the state sequence X generates O is calculated simply as the product of the transition probabilities and the output probabilities. So for the state sequence *X* in Fig. 5.

$$P(O, X|M) = a_{12} b_2(O_1) a_{22} b_2(O_2) a_{23} b_3(O_3 \quad (5.1)$$

However, in practice, the underlying state sequence *X* is hidden and only the observation sequence O is known. This is why it is named as *Hidden Markov Model.* The required likelihood is evaluated for an unknown X by combining over all the possible state sequences that is X= x (1) x (2) x (3)….x (t) [16]

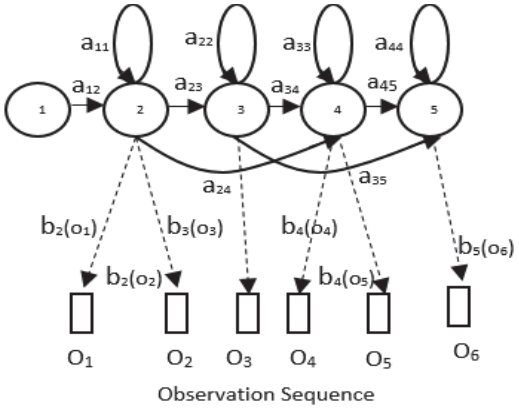


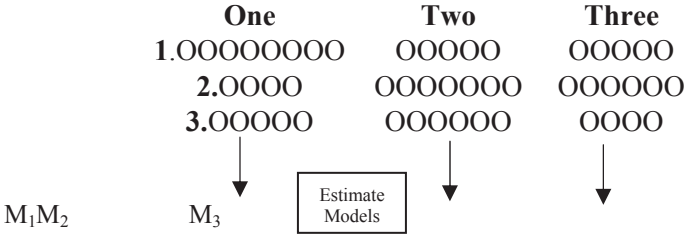
Figure 4. Hidden Markov Model

$$P(O|M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o)_t a_{x(t)x(t+1)} \quad (5.2)$$

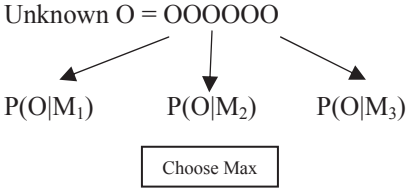
Where$x(0)$ is constrained to be the model entry state and $x_{(t+1)}$ is constrained to be the model exit state [17]. We have a vocabulary of ten words (0-9 digits) to be recognized. For each distinct word we had developed a distinct HMM model. For each word in the vocabulary, we have n training set of samples spoken by distinct persons. For each unknown word to be recognized we measure the observational sequence by feature analysis of the speech and it is followed by the calculation of the maximum likelihood path for all the distinct models. Then we select the word whose model has the highest likelihood.

A. Training

Training examples



B. Recognition



C. *Classification*

Let $\lambda_i$denote the parameter set for word i. When presented with an observation $O_1,\ldots,O_T$, the selection is done as follows.

$$Predictedword = \arg\max_i f(O_1, \ldots, O_T) \quad (5.2)$$

And we recognize that exactly what the forward algorithm computes. [18]

VI. Results And Discussion

The below figure shows the time domain and FFT of digit 5.

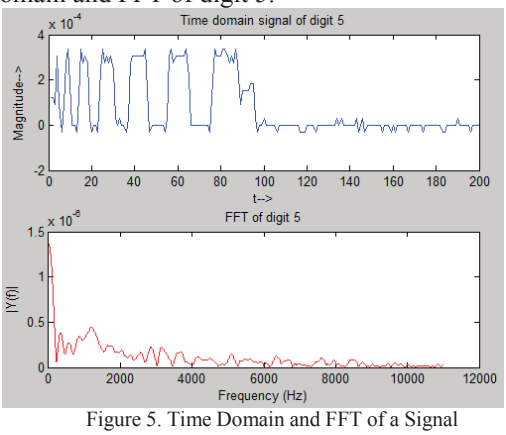


Figure 5. Time Domain and FFT of a Signal

The Three Stage DWT waveform for digit 5 is shown in Figure 6. The detail coefficients extracted to the third stage and the approximate coefficeint are plotted. The de-noised signals are reconstructed and then features are extracted.

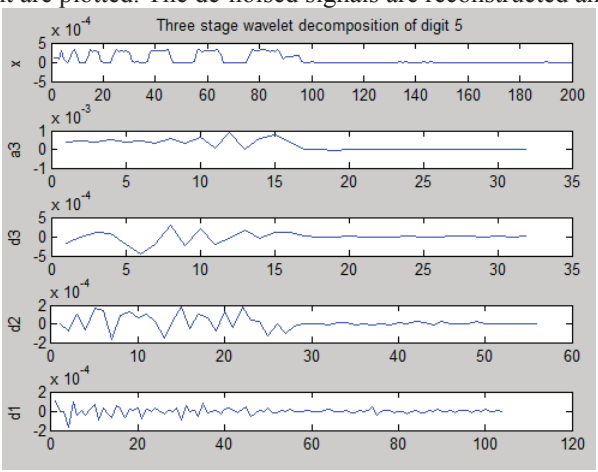


Figure 6. Three Stage DWT Waveform

We could seek to find the parameters A that maximize the log-likelihood of sequence of observations. This corresponds to finding the likelihoods of transitioning from one state to anotherwhich makes a set of observations most likely. Let us define the log-likelihood a Markov model [22].

$$l(A) = \log P(\vec{Z}; A) = \sum_{t=1}^{T} \log A_{z_t} \quad (5.3)$$

In the last line, we use an indicator function whose value is one when the condition holds and zero otherwise to select the observed transition at each time step. When solving this optimization problem, it is important to ensure that solved parameters A still make a valid transition matrix. In particular, we need to enforce that the outgoing probability distribution from state 'i' to 'j' always sums to one and all elements of A are non-negative. The HMM Training for digit 8 with the log maximum likelihood Plot is shown in Figure 7.

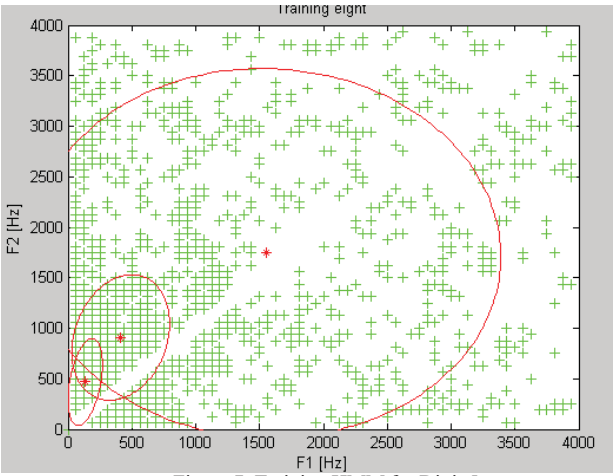


Figure 7. Training HMM for Digit 5

The validation and classification of data with recognition efficiency is shown in Figure 9.

```
Recognized word as seven!
Recognized word as seven!
Recognized word as eight!
Recognized word as eight!
Recognized word as seven!
Recognized word as eight!
Recognized word as eight!
Recognized word as six!
Recognized word as six!

ans =

    0.9130
```

Figure 9. Snapshot of HMM based Classificationand Efficiency

VII. Conclusion And Future Work

Murmur recognition is the goal of this work. Murmur samples were recorded from male and female speakers and de-noised using DWT employing the Bior-4.4 wavelet. The de-noised signal was reconstructed from the DWT coefficients. MFCC features were extracted from this de-noised signal and fed to the HMM for training followed by testing and classification. The performance of HMM was evaluated and found to give an efficiency of 91.30%. Further, this work would be extended to continuous digit recognition using HMM.

References

[1] Noise- Robust Whispered Speech Recognition Using A Non Audible Murmur Microphone With Vts Compensation proposed *by Chen-Yu Yang , Georgina Brown , Liang Lu , Junichi Yamagishi , Simon King*

[2] Blind Noise Suppression For Non - Audible Murmur Recognition With Sterio Signal Processing proposed by *ShuntaIshii, Tomoki Toda, Hiroshi Saruwatari, Sakriani Sakti, and Satoshi Nakamura*

[3] Speech Recognition Using Hidden Markov Model And Vitterbi Algorithm proposed by *Mr. Sanjay Bhardwaj, Mr. Sunil Pathania, Mr. Rajesh Akela*

[4] Implementation Of Text Dependant Speaker Independent Isolated Word Speech Recognition Using Hmm by *Ms. Rupali S Chavan, Dr. Ganesh S. Sable*

[5] BlindNoiseSuppressionFor Non-Audible Murmur Recognition With Stereo Signal Processing by *Shunta Ishii, Tomoki Toda, Hiroshi Saruwatari, Sakriani Sakti, Satoshi Nakamura*

[6] Analysis And Recognition Of Nam Speech Using Hmm Distances And Visual Information by *Panikos Heracleous, Viet-Anh Tran, Takayuki Nagai, and KiyohiroShikano*

[7] The Application Of Hidden Markov Models In Speech Recognition by *Mark Gales and Steve Young*

[8] Noise Reduction Of Speech Signal Using Wavelet Transform With Modified Universal Threshold by *Rajeev Aggarwal, Jai Karan Singh ,Mukesh Tiwari ,Dr. AnubhutiKhare*

[9] Blind Speech Extraction For Non-Audible Murmur Speech With Speaker's Movement Noise by *Miyuki Itoi, t Ryoichi Miyazaki, t Tomoki Toda, t Hiroshi Saruwatari, t KiyohiroShikano*

[10] Sheena Christabel Pravin ,et.al.,"Feature Extraction from Non-Audible Murmur (NAM) for the Vocally Handicapped using Wavelet Transform", International Journal of Computer Applications (0975 – 8887) Volume 135 – No.6, February 2016

[11] Local Feature or Mel Frequency Cepstral Coefficients - Which One is Better for MLN-Based Bangla Speech Recognition? by Foyzul Hassan, Mohammed Rokibul Alam Kotwal, Md. Mostafizur Rahman, Mohammad Nasiruddin, Md. Abdul Latif and Mohammad Nurul Huda

[12] F.Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska Delacreetaz, and D.A. Reynolds, "*A tutorial on text-independent speaker verification*," EURASIP Journal on Applied Signal Processing, Hindawi Publishing Cor-poration, vol. 4, pp. 430–451, 2004.

[13] D.K. Kim and N.S. Kim, "*Maximum a posteriori adaptation of HMM parameters based on speaker space projection*," Speech Communication, vol. 42, no. 1, pp. 59–73, Jan. 2004.

[14] WuChou,"Discriminant-Function-Based Minimum Recognition Error Rate Pattern Recognition Approach to Speech Recognition," PROCEEDINGS OF THE IEEE, VOL. 88, NO. 8, AUGUST 2000.

[15] Dimov, D., and Azmanov "Experimental specifics of using HMM in isolated word speech recognition." International Conference on Computer Systems and Technologies Comp Sys Tech'2005.

[16] N. Najkar, F. Razzazi, and H. Sameti, "*A novel approach to HMM-based speech recognition system using particle swarm optimization*," in BIC-TA 2009 - Proceedings, 2009 4th International Conference on Bio-Inspired Computing: Theories and Applications, 2009, pp. 296-301.

[17] Steve Young, Gunnar Evermann, Mark Gales "The HTK Book" Microsoft Corporation http://www.ee.columbia.edu/~dpwe/LabROSA/doc/HTKBook21/node5.html

[18] Isolated-word speech recognition using hidden Markov models, Hakon Sandsmark, December 18, 2010