Spoken Word Recognition Strategy for Tamil Language

AN. Sigappi[1] and S. Palanivel[2]

**[1] Department of Computer Science and Engineering, Annamalai University
Annamalainagar - 608 002, Tamilnadu, India**

**[2] Department of Computer Science and Engineering, Annamalai University
Annamalainagar - 608 002, Tamilnadu, India**

Abstract

This paper outlines a strategy for recognizing a preferred vocabulary of words spoken in Tamil language. The basic philosophy is to extract the features using mel frequency cepstral coefficients (MFCC) from the spoken words that are used as representative features of the speech to create models that aid in recognition. The models chosen for the task are hidden Markov models (HMM) and autoassociative neural networks (AANN). The HMM is used to model the temporal nature of speech and the AANNs to capture the distribution of feature vectors in the feature space. The created models provide a way to investigate an unexplored speech recognition arena for the Tamil language. The performance of the strategy is evaluated for a number of test utterances through HMM and AANN and the results project the reliability of HMM for emerging applications in regional languages.

Keywords: *Speech recognition, Mel frequency cepstral coefficients, Hidden Markov models, Autoassociative neural networks.*

1. Introduction

Speech is the most common form of communication among human beings and it is their intuitive ability to recognize a spoken word, phrase, or sentence effortlessly. The benefits of spoken language interfaces are obvious due to the inherent natural communication approach present in them. However speech recognition systems for regional languages spoken in developing countries with rural background and low literacy rates appears to be still evolving.

Tamil is a classical Dravidian language that is in existence for several hundreds of years. It is classified as agglutinative language and spoken largely by people living in TamilNadu, parts of Srilanka, Singapore, and Malaysia. Tamil language imbibes twelve vowels and eighteen consonants, which combine to form two hundred and sixteen compound characters. In addition to this, there is a special letter called aaytham. Thus, a total of two hundred and forty seven letters constitute the standard Tamil alphabet. The other significant feature present in Tamil language is the presence of unique liquid, which sounds *'zh'*. The information required to perform the basic speech processing tasks is implicitly present in the speech. Owing to the fact that human beings are endowed with both speech production and perception mechanisms, the need for processing the speech signal does not arise. However there is a need to process the speech signal in light of the view that a machine is part of the communication chain.

2. Related Work

A grapheme-based automatic speech recognition system that jointly models phoneme and grapheme information using Kullback-leibler divergence based HMM has been presented and investigated for English language using DARPA Resource Management (RM) corpus [1]. A continuous speech recognizer using a group delay based two level segmentation algorithm has been developed to extract the accurate syllable units from the speech data [2]. Isolated style syllable models have been built for all unique syllables using samples from annotated speech in Tamil language. A formant tracking algorithm has been underlined using the phoneme information in the acoustic speech signal [3]. A robust coupled HMM-based audio video speech recognition (AVSR) system has been developed. The experimental results have been found to record a remarkable increase in the recognition rate compared to the only video based automatic speech recognition systems [4]. A prototype for speech based health information access by low literate community health workers has been developed. The experiences from a pilot study involving the use of community workers in a rural health center has been reported [5]. A privacy

preserving speech recognition model that serves to preserve the privacy between one party with private speech data and one party with private speech recognition models has been realized using HMM [6]. A host of methodologies for effective speech recognition have been articulated and evaluated using SPINE corpus. The use of parallel banks of speech recognizers have been found to improve the performance of recognition [7].

Inspite of the urgent need for automation in all domains, the development of strategies for speech recognition in regional languages is still perceived to be cumbersome due to various issues such as non-availability of speech corpus for training purpose, complexity in the language, lack of phoneme recognizers and difficulty in creating a speech corpus with necessary transcriptions. Though this ordeal is in focus over a period of time, it still remains a challenge and efforts are required to accomplish this with greater precision and reliability. It is proposed to develop a strategy through which a spoken Tamil word can be recognized.

3. Problem Definition

It is an inert requirement to formulate a methodology through which computer systems can assimilate and recognize what a person speaks. A text and speaker dependent medium-sized vocabulary speech recognition mechanism is designed with an ability to recognize one hundred railway station names uttered in Tamil language. It echoes with it a focus to extricate its performance using HMM and AANN models and evolve a platform suitable to acclaim the spoken word from a chosen set of test samples.

4. Components of a Speech Recognition System

The constituents of a typical speech recognition system seen in Fig.1 include a feature extraction component and a template or statistical classifier. The main task of the feature extraction component is to extract features from a speech signal so as to represent the characteristics of the speech signal and yield a few numbers of coefficients that are grouped together to form a feature vector. Subsequent to feature extraction, the sequence of feature vectors is sent to a template or statistical classifier which selects the most likely sequence of word or phonemes.

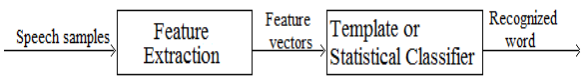


Fig. 1. Components of a speech recognition system

The fundamental issue in speech processing is the manner in which the specific features are extracted in order to perform the desired speech processing tasks. The human ear resolves frequencies non-linearly across the audio spectrum and it is thus desirable to obtain the nonlinear frequency resolution. Mel frequency cepstral coefficients (MFCC) appears to be one of the most successful feature representations in speech recognition related tasks, obtains the coefficients through a filter bank analysis. The mel-cepstrum exploits auditory principles, as well as the decorrelating property of the cepstrum [8]. Fig.2 illustrates the computation of MFCC features for a segment of speech signal [9]. The stages involved in the extraction of features are preemphasis, frame blocking, windowing, filter bank analysis, logarithmic compression, and discrete cosine transformation.

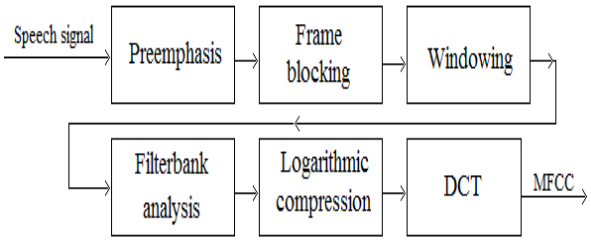


Fig. 2. Computation of MFCC features

(i)Preemphasis: Preemphasis is initiated to spectrally shape the signal so as to obtain similar amplitude for all formants. The speech signal is preemphasised by filtering the speech signal with a first order FIR filter whose transfer function in the *z*-domain is

$$H(z) = 1 - \alpha z^{-1}, 0 \leq \alpha \leq 1 \quad (1)$$

The preemphasised signal is related to the input signal in time domain using the relation. The preemphasis coefficient α lies in the range $0 \leq \alpha < 1$.

$$\hat{s}(n) = s(n) - \alpha s(n-1) \quad (2)$$

(ii)Frame blocking: The statistical characteristics of a speech signal are invariant only within short time intervals of articulatory stability. The preemphasized signal is blocked into frames of N samples (frame size), with adjacent frames being separated by M samples (frame shift). If the l^{th} frame of speech is be denoted by $x_l(n)$ and there are L frames within the entire speech signal, then

$$x_l(n) = \hat{s}(Ml + n), \quad \begin{matrix} 0 \leq n \leq N-1 \\ 0 \leq l \leq N-1 \end{matrix} \quad (3)$$

(iii)Windowing: The next step is to window each frame so as to minimize the signal discontinuities at the beginning and end of the frame. The window is selected to taper the signal at the edges of each frame. If the window is defined as $w(n)$, $0 \le n \le N-1$, then the result of windowing the signal is

$$x_1(n) = x_1(n)w(n), \quad 0 \le n \le N-1 \tag{4}$$

Hamming window is a good choice in speech recognition, considering that the subsequent operation in the feature extraction process integrates all the closest frequency lines. The Hamming window takes the form

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \tag{5}$$

(iv)Filter bank analysis: A Fourier transform is obtained for each frame of the speech signal from which the magnitude is then weighted using a series of filter frequency responses. The center frequencies and bandwidths are chosen to roughly match those of the auditory critical band filters, that follow the mel scale, defined by

$$Mel(f) = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \tag{6}$$

The filters are collectively called as a Mel scale filter bank and the frequency response of the filter banks simulate the perceptual processing performed within the ear.

(v)Logarithmic compression: The filter outputs obtained from filter bank analysis are compressed by a logarithmic function so as to model the perceived loudness of a given signal intensity.

$$X_{m(ln)} = ln(X_m), \quad 1 \le m \le M \tag{7}$$

where $X_{m(ln)}$ is the logarithmically compressed output of the m^{th} filter.

(vi)DCT: Discrete cosine transform (DCT) is thereafter applied to the filter outputs and the first few coefficients are grouped together as a feature vector of a particular speech frame. If p is the order of the mel scale cepstrum, the feature vector is obtained by considering the first p DCT coefficients. The k^{th} MFCC coefficient in the range $1 \le k \le p$ can be expressed as

$$MFCC_k = \sqrt{\frac{2}{M}} \sum X_{m(ln)} \cos(\pi k(m - 0.5)/M) \tag{8}$$

5. Hidden Markov Models (HMM)

Hidden Markov models (HMMs) are widely used in automatic speech recognition applications because of their accuracy in recognition. A Hidden Markov model is characterized by the following [10]:

(i) N, the number of hidden states in the model. The individual states are indicated as $S = S_1, S_2, ..., S_N$, and the state at time t as q_t.
(ii) M, the number of distinct observation symbols per state. The observation symbols are denoted as $V = v_1, v_2, ..., v_M$.
(iii) The state transition probability distribution $A=a_{ij}$, where

$$a_{ij} = p[q_{t+1} = S_j | q_t = S_i], \quad \begin{matrix} 1 \le i \\ j \le N \end{matrix} \tag{9}$$

(iv) The observation probability distribution in state j, $B = b_j(k)$, where

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j], 1 \le j \le N, 1 \le k \le M \tag{10}$$

(v) The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = P[q_1 = S_i], \quad 1 \le i \le N \tag{11}$$

A complete specification of a HMM is as follows:

$$\lambda = \{a_{ij}, b_j(k), \pi\}, \quad \begin{matrix} \sum_j a_{ij} = 1, \ \forall i \\ \sum_k b_j(k) = 1, \ \forall j \end{matrix} \tag{12}$$

An isolated word recognizer using HMM is shown in Fig. 3. A training set of K occurrences and the features extracted from each occurrence of the word constitutes an observation sequence for every word in the vocabulary. The HMM constructed for each word estimates the model parameters (A, B, π) that optimises the likelihood of the training set observation vectors. The observation sequence for the test utterance is determined from the speech signal from where the most likelihood calculation is made for all possible models, using which the word whose model likelihood appears to be the highest is selected.

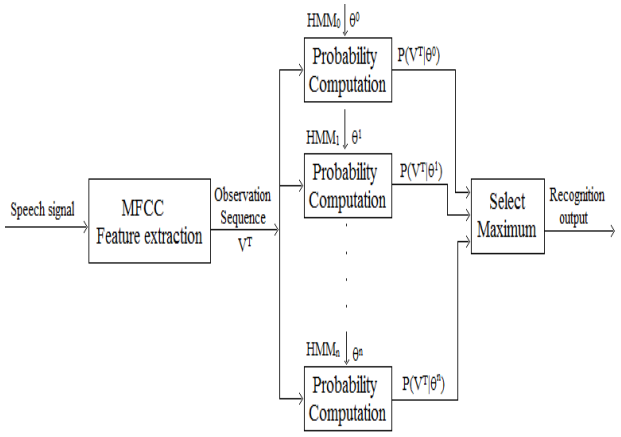


Fig. 3. Isolated word Recognizer using HMM

6. Autoassociative Neural Networks (AANN)

Autoassociative neural network models (AANNs) are feed forward neural networks bestowed with an ability to perform identity mapping of the input space [11] and are increasingly used in speech processing applications. The distribution capturing capability of the AANN model is described using the five layer AANN model as shown in Fig. 4. This model comprises of three hidden layers in which the processing units in the first and third hidden layer are nonlinear whereas the units in the second hidden layer can be either linear or nonlinear. The first layer in the network is the input layer, the third is the compression layer and the last is the output layer. The second and fourth layers contain more units than the input layer, while the third layer is with fewer units than the first or fifth layer. In any case, the number of units in the input and output layers are the same.

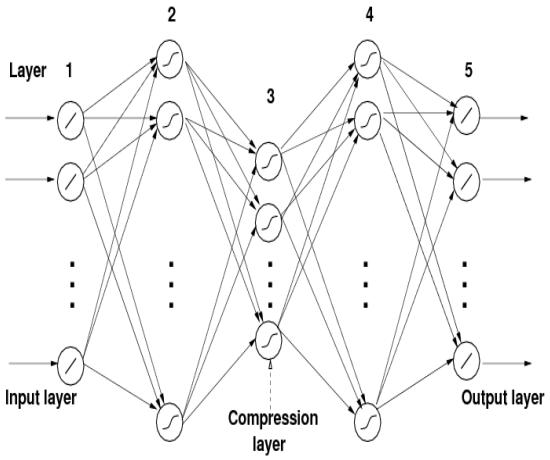
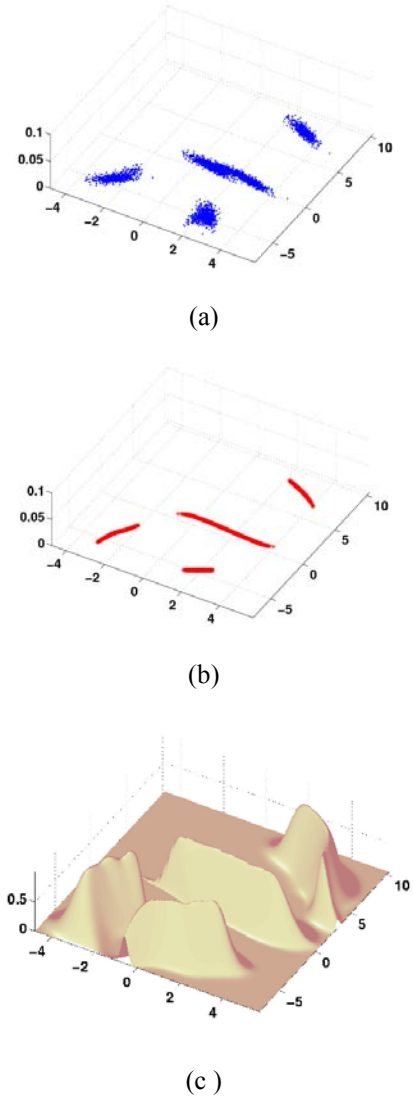


Fig. 4. A five layer autoassociative neural network model

The cluster of points in the input space determines the shape of the hypersurface obtained by the projection onto the lower dimension space, as the error between the actual and the desired output vectors is minimized. The space spanned by the one dimensional compression layer in Fig.5(b) corresponds to the two dimensional data shown in Fig.5(a) and the network structure *2L 10N 1N 10N 2L*, where *L* denotes a linear unit and *N* denotes a non linear unit. The integer values indicate the number of units present in that layer. The nonlinear output function for each unit is *tanh(s)*, where *s* is the activation value of the unit. The network is trained using the backpropagation algorithm. The AANN thus captures the distribution of the input data depending on the constraints imposed by the structure of the network.



(a)

(b)

(c)

Fig. 5 Distribution capturing capability of AANN. (a) Artificial two dimensional data. (b) Two dimensional output of the AANN model. (c) Probability surfaces realized by the AANN.

The error for each input data point is plotted in the form of some probability surface as given in Fig. 5(c). The error e_i for the data point i in the input space is plotted as $p_i = exp(-e_i/\alpha)$, where α is a constant. Though p_i is not strictly a probability density function, the resulting surface is termed as probability surface. The plot of the probability surface shows a larger amplitude for a smaller error e_i, indicating a better match of the network for that data point. The constraints imposed by the network is seen by the shape the error surface takes in both cases. An ideal expectation pertaining to the distribution of data is oriented to achieve the best probability surface, best defined in terms of some measure corresponding to a low average error.

7. Experimental Results

7.1 Speech Corpus Creation

The strategy is evolved for a speech recognition task, with a view to identify the spoken utterances of specific words in Tamil language. The vocabulary includes one hundred railway station names in TamilNadu. The procedure necessitates the creation of an in-house speech corpus that contains the audio recordings of the words uttered by speakers. An omnidirectional microphone is used for recording the voices and the recording is carried out in a laboratory environment. Wavesurfer software is used for the recording purpose and the recorded files are stored in .wav format. The sampling rate is chosen to be 16 KHz. The speakers are chosen to be in different ages ranging from twenty one to sixty and of both genders in order to ensure good training and replicate the actual scenario. Twenty speakers are chosen and a speaker is made to utter each station name five times and hence a total of ten thousand samples constitute the speech corpus. The experiments are conducted using the database created, from which eight thousand samples are used for training purpose and the remaining two thousand samples for testing the performance of the methodology.

7.2 Feature Representation

The input speech data is preemphasised using a first order digital filter, with a preemphasis coefficient of 0.97. It is then segmented into ten millisecond frames with an overlap of fifty percent between adjacent frames and windowed using Hamming window. The twelve MFCC coefficients are thereafter extracted from the filter bank output through the use of DCT. In addition to the log energy, the 0'th cepstral parameter C_0 is appended to give thirteen MFCC coefficients. The time derivatives are added to the basic static parameters thus obtained, in order to further enhance the performance of the speech recognition strategy. The thirteen first order regression coefficients, known as delta coefficients and the thirteen second order regression coefficients, referred to as acceleration coefficients are appended to the static coefficients to yield thirty nine mel frequency cepstral coefficients. The problem of separating the background silence from the input speech is another factor that accords a significant impact on the implementation of a speech recognition system [10]. It is typically done on the basis of signal energy and signal durations. A frame is judged to be nonsilent if its total energy is less than 40 dB below the maximum total energy computed across all the frames in the utterance. The other frames are considered to be silent frames and are not taken into account for purpose of model creation.

7.3 Model Construction

The 39 dimensional MFCC feature vectors extracted from the nonsilence frames of the speech signal corresponding to each word are given as input to estimate the parameters of HMM and is implemented using HTK toolkit. The HTK tools HInit and HRest provide isolated word style training using a flat start mechanism and a HMM for each station name is generated individually. Once an initial set of models are created, the tool HRest performs a Baum-Welch reestimation of the entire training set. Each of the models are reestimated until no change occurs in the state transition probabilities and finally the required models are made available to represent each station name. The structure of the AANN model used in the experiments is chosen by systematically varying the number of units in the second and third layers and by varying the number of epochs required for training the network.

7.4 Performance

The performance of the strategy presented in this work is evaluated through the voice samples recorded from the speakers that are set aside for testing. The recognition rate is used as the performance measure and it is defined as the number of words correctly recognized. It is given by the equation

$$r = \frac{c}{t} \times 100 \quad (13)$$

where r represents the recognition rate, c the number of words correctly recognized during testing, and t the total number of words in the vocabulary.

The experiment using HMM is conducted by varying the

number of states in the model and also the number of mixtures in each state. The results shown in Fig.6 indicate that the HMM with 5 states and 4 mixtures in each state yields a recognition rate of 95.0%. Similarly the performance of the strategy using AANN is evaluated by varying the number of units in the second (N_s) and third (N_t) layers. These results seen in Fig.7 explain that the network structure *39L 80N 30N 80N 39L* trained for 600 epochs offers the best results with a recognition rate of 90.0%.

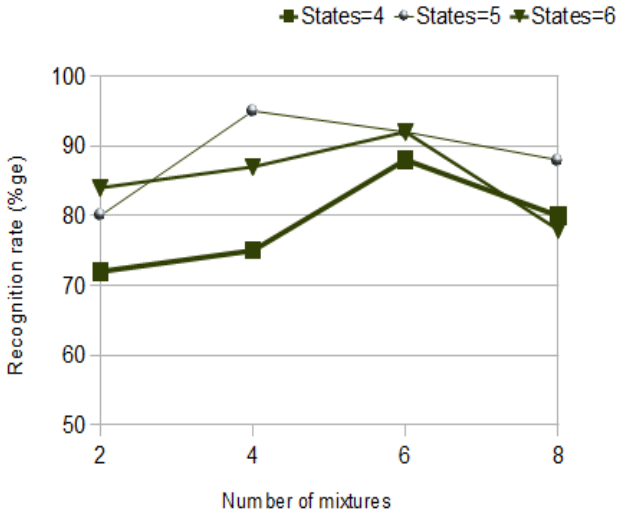


Fig. 6. Recognition results for various states and mixtures in the HMM

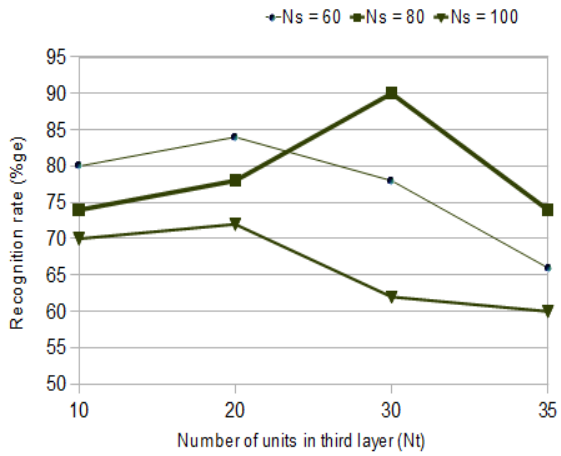


Fig. 7. Performance of different AANN network structures

The results summarized in Fig.8 indicate that HMM is a reliable model promising for speech recognition applications in comparison to AANN. It also follows that the responses are closely related to the strength of the samples used. It precisely points to the need for a rich corpus to accomplish the worthy use of the structure.

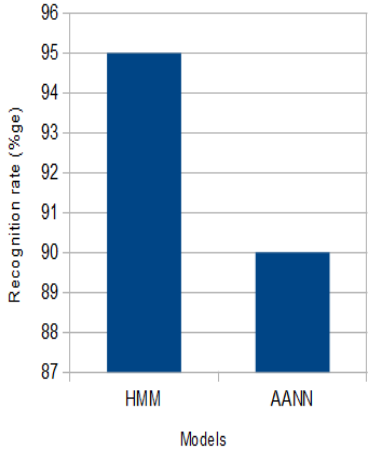


Fig. 8. Comparison of HMM and AANN results

4. Conclusion and Future Work

A strategy for recognizing spoken words in Tamil language has been developed. The strategy has been evolved using HMM and AANN models constructed using the mel frequency cepstral coefficient features obtained from the speech signal. The HMM with 5 states and 4 mixtures has been found to yield better recognition results in comparison to the AANN model with a structure *39L 80N 30N 80N 39L*. The performance of the strategy has been found to explicitly portray the suitability of HMM over AANN for a speech recognition task in Tamil language. The consistency and robustness of the approach have been proved to be its highlights and allow its use in unforeseen environments. Besides if new features can be explored to characterize the speech signal more accurately it will go a long way in arriving at higher recognition rates.

References

[1] Mathew Magimai.-Doss, Ramya Rasipuram, Guillermo Aradilla, and Herve Bourlard, “Grapheme-based automatic speech recognition using KL-HMM”, in Proceedings of Interspeech, Aug 2011.

[2] A. Lakshmi, and Hema A. Murthy, "A syllable based continuous speech recognizer for tamil", in Intl. Conf. on Spoken Language Processing, Sept 2006.

[3] M. Lee, J. van Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 5, Sept 2005, pp. 741-751.

[4] Yashwanth .H., Harish Mahendrakar, and Suman David, "Automatic speech recognition using audio visual cues", in IEEE India Annual Conferenc, Dec 2004, pp. 166-169.

[5] Jahanzeb Sherwani, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld, "HealthLine: Speech-based access to health information by low-literate users", in IEEE/ACM Intl. Conference on Information and Communication Technologies and Development, Dec 2007.
[6] Paris Smaragdis and Madhusudana Shasashanka, "A framework for secure speech recognition", IEEE transactions on Audio, Speech, and Language Processing, Vol.15, No.4, May 2007, pp.1404-1413.
[7] John H. L. Hansen, Ruhi Sarikaya, Umit Yapanel, and Bryan Pellom, "Robust speech recognition in noise: An evaluation using SPINE corpus", in EUROSPEECH 2001, Sept 2001, pp. 4148-4153.
[8] S. B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol.28, No.4, Aug. 1980, pp.357-366.
[9] "The HTK Book", Cambridge University Engineering Department, 2002.
[10] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol.77, N0.2, Feb 1989, pp.257-285.
[11] B. Yegnanarayana, and S. P. Kishore, "AANN: an alternative to GMM for pattern recognition", Neural Networks, Vol.15, No.3, April 2002, pp.459-469.

AN. Sigappi received her Bachelors degree in Computer Science and Engineering from Annamalai University in 1993 and her Masters degree in the same discipline from Anna University in 2001. She is currently pursuing her doctoral research work at Annamalai University. Her career includes both administrative and academic experience spanning over 15 years and she is presently working as Associate Professor in the Department of Computer Science and Engineering at Annamalai University. Her research interest includes speech and image processing, software engineering, management information systems and intelligent systems.

S. Palanivel received the B.E(Hons) degree in Computer Science and Engineering from Bharathidasan University in 1989 and followed it up with Masters degree in the same discipline from Bharathiar University in 1994. He completed his Ph.D in Computer Science and Engineering from the Indian Institute of Technology Madras in the year 2005. He is currently serving as Associate Professor in Computer Science and Engineering at Annamalai University. He carries with him 17 years of teaching experience and over 20 publications in international conferences and journals. His research interests include speech processing, image and video processing, pattern classification and neural networks.