Speaker Independent Isolated Tamil Words Recognition System using Different Classifiers

P.Iswarya
Research Scholar,
Department of Computer Science,
Avinashilingam University for Women,
Coimbatore.
iswaryacbe333@gmail.com

Dr.V.Radha
Professor,
Department of Computer Science,
Avinashilingam University for Women,
radhasrimail@gmail.com

Abstract— The paper is to build a speaker independent isolated speech recognition system for Tamil Language. The system consists of four steps: first step involves recording of Tamil spoken words using high quality studio microphone and stored in the form of wav files. The next step is preprocessing, to filter out noise present in speech signal using pre-emphasis filter. The filtered speech signal features are extracted using Mel Frequency Cepstral Coefficient (MFCC) technique. The extracted features from spoken Tamil words are fed as input to different classifiers such as Probabilistic neural network, Hidden markov model and Support Vector Machine. Each Classifier recognition performance is analyzed, and found that hidden markov model achieves higher recognition rate than other classifiers.

Keywords-Hidden markov model, Mel frequency cepstral coefficients, Probabilistic neural network, Support vector machine.

I. INTRODUCTION

Human have to interact with the computer for task of data processing, which is often done through keyboard, mouse etc as input, and printer, screen, speaker as output. Speech is easy mode of communication for the people to interact with the computer, rather than using keyboard and mouse. Automatic speech recognition is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert into written text [1]. Recognition of speech is difficult task due to speaker variability, noise, environment and microphone used. Tamil is agglutinative language and largely spoken in countries like India, Singapore and Malaysia. Tamil language imbibes twelve vowels, eighteen consonants, and special letter called aaytham. Thus, a total of two hundred and forty seven letters constitute a standard Tamil alphabet [2]. Speech recognizer accepts the spoken words as input in the form of audio format such as wav, raw to take appropriate action for operation. The next stage consists of feature extraction stage and classification stage. The parameters from the feature extraction stage are compared in some form of parameters extracted from signals stored in database or template. Then these parameters for recognition fed in to the classifiers. In the field of speech recognition many research have been carried out using different classifiers. In this paper most widely used classification techniques are chosen that includes Probabilistic Neural Network (PNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM). The extracted MFCC features are given as input to the various classifiers and their recognition performance is analyzed and compared.

The paper is organized as follows. Section 2 describes overview of the system. Section 3 explains about feature extraction technique based on Mel Frequency Cepstral Coefficients (MFCC). Section 4 describes about feed forward probabilistic neural network, Hidden markov model and Support vector machine. Section 5 explores the results and discussion. Finally conclusion is presented in Section 6.

II. SYSTEM OVERVIEW

The speech samples in the form of wav files are recorded by using high quality studio recording microphone at a sampling rate of 16 KHz. A database is created for Tamil language using 10 speakers. Each speaker utters 10 words with 5 repetitions, therefore the total of 500 samples are created. To implement the system, speech signals are pre-processed using pre-emphasis filtering technique. The filtered speech signal is divided into

subsequent 30 ms frames and followed by hamming window function. Then the speech signal undergoes MFCC feature extraction for each frame. The MFCC feature vectors are also called cepstral coefficients which are further given as input to the PNN, HMM, SVM classifiers for recognition. The overall architecture of speech recognition process is shown in Fig. 1.

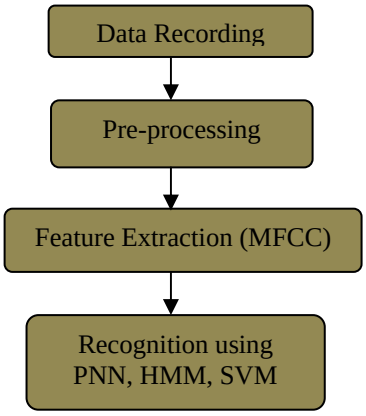

Figure 1.Overall Architecture

III. FEATURE EXTRACTION

The speech signals are generally preproprocessed before it taken for further analysis; this is to improve the recognition accuracy. Pre-processing involves applying filters to the speech signal. Pre-emphasis filter is one of the common filters used in speech enhancement which reduces background noise and resulting in good quality of speech. The original signal and their pre-emphasis of original signal for the word 'amma' is shown in Fig. 2.

Feature extraction is important component in speech recognition because the accuracy of recognition mainly depends on the features that are extracted. Different techniques available for feature extraction such as Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC). In this research work the most frequently used MFCC parameters are considered to determine the best feature set for Tamil speech database [3]. The stages involved in extraction of features are pre-emphasis, frame blocking, windowing; filter bank analysis, logarithmic compression and discrete cosine Transformation. The overall process of MFCC feature extraction process is shown in Fig. 3.

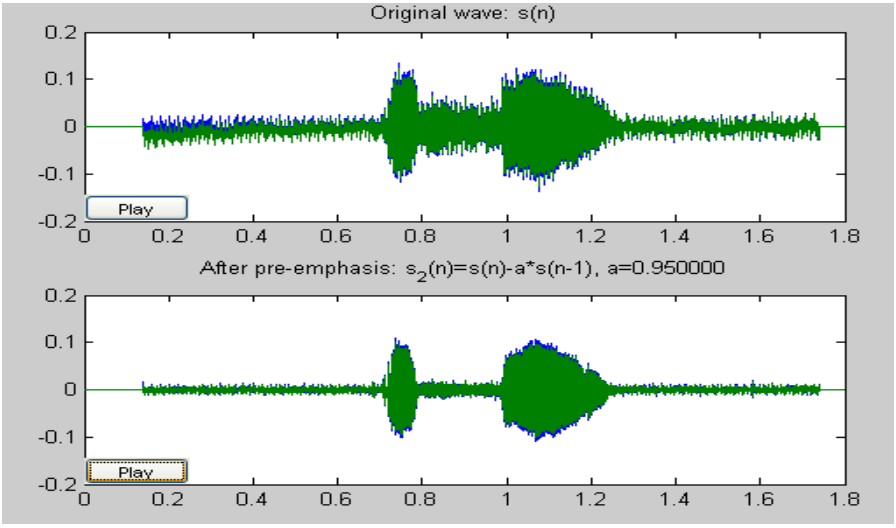

Figure 2.Speech waveform of word "amma" before and after pre-emphasis filter Classification

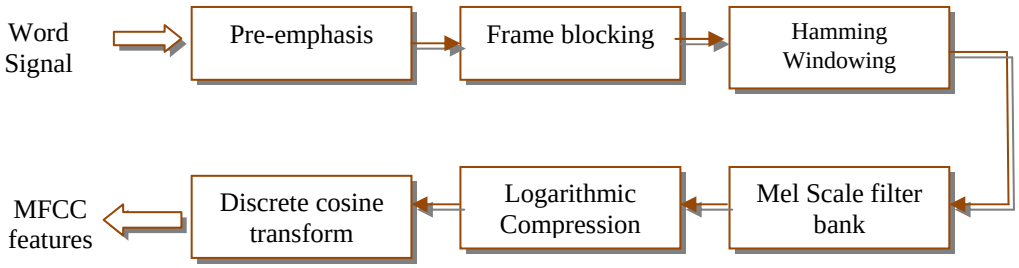

Figure 3. MFCC Feature extraction process

A. Pre-emphasis

Normally Voiced portions of speech have a negative spectral slope this is a psychological characteristic of speech. The pre-emphasis filter spectrally flattens the speech signal to improve efficiency of speech waveform [4]. Typical signal pre-emphasis is given (1)

$$S2(n) = s(n) - a * s(n-1) \quad (1)$$

Where constant 'a' falls between intervals 0.9 to 1.0.

B. Frame Blocking

In speech signal for data processing it is divided into frames of N samples, with adjacent frames being separated by M (M<N). The first frame consists of first N samples and second frame begins after the first frame by M samples. Two consecutive frames have overlapping areas by N-M samples and so on. Typical values of N and M are 256.

C. Windowing

To minimize signal discontinuities which occur during beginning and end of each frame, in order to reduce energy at the edges and to prevent abrupt changes at the end points Windowing function is applied. Hamming window is most commonly used window which has least amount of desertion. Hamming equation is given in (2)

$$w(n) = 0.54 - 0.46 \cos(2n\pi/(N-1)) \; 0 \leq n \leq N-1 \quad (2)$$

Where N is the length of Window.

D. Filter Bank analysis

The filters are collectively called as mel scale filter bank and frequency response of filter bank simulate perceptual processing done within the ear. Filter bank analysis is a process of converting time domain speech signals of frame of N samples to frequency domain. A Fast Fourier Transform (FFT) of speech signal is wide it does not follow a linear scale so magnitude is weighted by the series of filter frequency responses. The filter frequency magnitude response is triangular in shape which is equal to auditory critical bandwidth filters that follow a Mel scale [5] represented in (3)

$$Mel(f) = 2595 * \log 10(1 + f/700) \quad (3)$$

E. Logarithmic Compression

The outputs obtained from filter bank analysis are compressed using logarithmic function.

$$X_{m(ln)} = \ln(X_m) \quad 1 \leq m \leq M \quad (4)$$

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

F. Discrete Cosine Transformation

The first few coefficients grouped together as a feature vector of a particular speech frame by applying Discrete Cosine Transformation for filter outputs [2]. The k^{th} MFCC coefficient in the range $l \leq k \leq p$ can be expressed as

$$MFCC_k = \sqrt{2/M} \sum X_{m\,(ln)} \cos(\pi k(m-0.5)M) \qquad (5)$$

Where p is the order of Mel scale spectrum.

In this paper MFCC is implemented with 24 filters and 12 dimensions. For NFFT it uses size of 256. From the sample of Tamil speech signal 12 feature vectors are extracted from MFCC.

IV. RECOGNITION USING PNN, HMM AND SVM

A. Probabilistic Neural Network

Probabilistic Neural Network (PNN) is multilayer feed forward neural network which is similar to back propagation neural network but differs in their learning process. PNN uses supervised learning algorithm and it consists of 3 nodes; input layer, hidden layer and output layer. The input layer consists of N nodes, which represents feature vectors and this layer is fully interconnected with hidden layer. There are no weights in hidden layer and each hidden node represents an example vector, with example acting as the weights to that hidden node. Hidden layer is not fully connected with output layer. The most important function of output layer is determination of class which is done through a winner takes all approach. Fig. 4 represents sample PNN architecture.

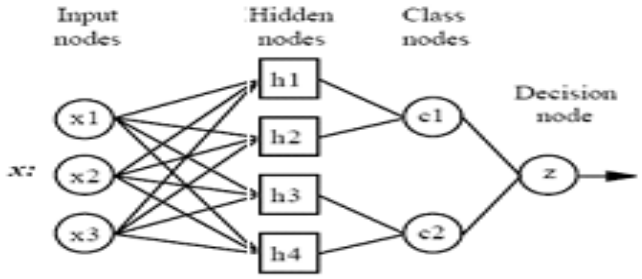

Figure 4.PNN Architecture

PNN is closely related to Parzen window probability density function estimator which is estimated for each of the classes. For each training vector PNN operates on spherical guassian radial basis function centered [6]. The likelihood of an unknown vector belonging to a given class can be expressed in (6)

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^p M_i} \sum_{j=1}^{M_i} \exp\frac{-(\mathbf{x}-\mathbf{x}_{ij})^T(\mathbf{x}-\mathbf{x}_{ij})}{2\sigma^2} \qquad (6)$$

Where i is the class number, j is the pattern number x_{ij} is the j^{th} training vector from class i, **x** is the test vector, Mi is the number of training vectors in class i, p is the dimension of vector **x**, s is the smoothing factor (the standard deviation,) and $f_i(\mathbf{x})$ is the sum of multivariate spherical Gaussians centered at each of the training vectors $\mathbf{x}_{ij}$ for the i^{th} class probability density function (pdf) estimate. Classification decisions are consequently made in accordance with the Bayes' strategy for decision rule, which is $d(\mathbf{x}) = C_i$, if

$f_i(x) > f_k(x)$ for $k \neq i$

where C_i is the class i.

B. Hidden Markov Model

Hidden markov model is rich in mathematical structure and it has three components such as acoustic model, pronunciation dictionary and language model. The block diagram of speech recognition using HMM is shown in Figure 5.

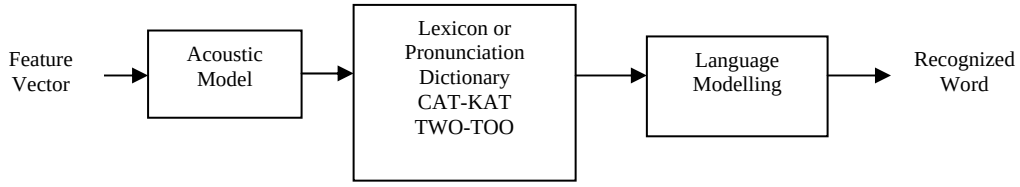

Figure.5.Block diagram of ASR using HMM

The first and foremost step after feature extraction process is to build acoustic model based on statistical representations. An unknown sample is scored against the acoustic model and the model with maximum score wins, resultant as recognized word. The context-independent based HMM model is used in this research work [7]. Each phoneme in acoustic component has unique HMM model and it is in the form of three states. The paper implements Vector Quantization technique [8] for mapping the feature vector to HMM state. Pronunciation dictionary or lexicon contains detail about the words and their corresponding monophone or phoneme sequence. The lexicon helps in constructing unique HMM model. Sometimes many words with same phoneme sequence will occur, during that case language model helps. It uses context information to narrow down the recognized word to resemble the given grammar construct. Thus in HMM to recognize a given sentence or word, all these three components will work together.

C. Support Vector Machine

Support Vector Machine (SVM) offers a discriminative solution to the pattern recognition problem involved in ASR [9]. Initially SVM is designed to classify linear applications, and later it is extended to non-linear type of data by the use of kernel-trick. The SVM constructs the hyper plane in a high dimensional space for classification tasks. The paper uses polynomial kernel with third degree functionality to classify this non-linear data. ASR system implements multiclass SVM involving one vs one type is chosen to test against all other classes individually. SVM can achieve good generalization ability with minimum structural risk.

V. RESULTS AND DISCUSSION

The performance of speech recognition system is generally measured in terms of recognition rate and Word error rate. Recognition rate is ratio between the number of words correctly recognized * 100 and total number of words. Word error rate is given as 100 – Recognition rate. To develop real time speaker independent Automatic Speech Recognition (ASR) focus on minimizing the word error rate to zero and recognition accuracy to 100%.

For Implementation working platform of MATLAB version 7.11 is used. In our experiment ten Tamil words (அம்மா, விலங்கு, மலர், சிரிப்பு, கண்ணாடி, கவிதை, தண்ணீர், முகம், தாமரை, விடை) were taken as input. The Tamil speech sample database is created with 10 speakers utters 10 words each with 5 repetitions, total of 500 samples created. In our experiment ten Tamil words were taken as input. Each speech signal is divided into subsequent 30ms frames and from each frame 12 feature vector MFCC coefficients are extracted. These MFCC features are fed in classifiers for evaluation and corresponding results are presented in Table.1.

TABLE I. RESULTS OF TAMIL SPEECH RECOGNITION SYSTEM USING DIFFERENT CLASSIFIERS

Tamil Words	Recognition rate of PNN	Recognition rate of HMM	Recognition rate of SVM
அம்மா	99%	99%	99%
விலங்கு	98%	99%	99%
மலர்	99%	99%	99%
சிரிப்பு	83.33%	95.00%	91%
கண்ணாடி	93.33%	98.00%	95%
கவிதை	86.66%	90.00%	85%
தண்ணீர்	90%	98%	92%
முகம்	86.66%	94.00%	94%
தாமரை	86.66%	94.00%	92%
விடை	90%	99%	99%
Total	91.26%	**96.50%**	95%

The PNN, HMM, SVM obtains the average recognition accuracy of 91.26%, 96.50%, 95% and Word Error Rate is about 8.74%, 3.50%, 5% respectively. The representation of WER in graphical form is shown in Fig.6.

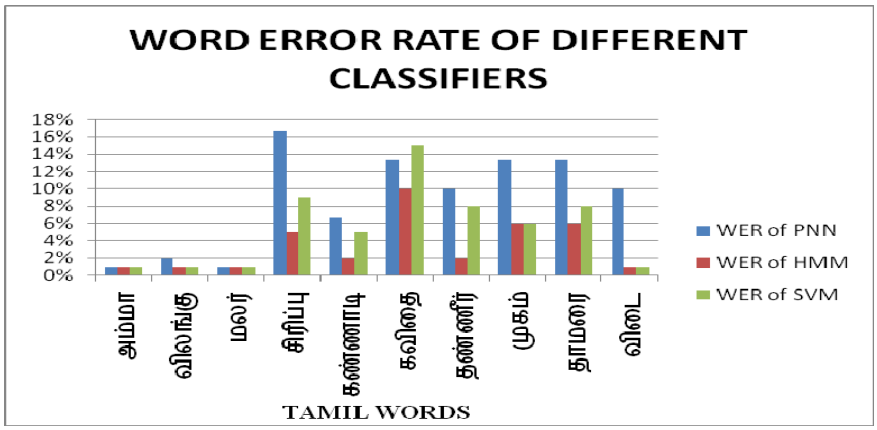

Figure. 6. Graphical representation of Word Error rate

The experimental results show that the hidden markov model achieves maximum recognition rate than Probabilistic neural network and Support vector machine. Also SVM performs better than PNN. Recognition accuracy decreases and Word error rate increases when vocabulary size grows.

VI. CONCLUSION

In this paper isolated speaker independent Tamil word recognition system is developed with MFCC feature vector and three different classifiers. The Mel frequency coefficients are fed in to the each classifier, to find the best classifier for speech recognition system. Compare with different recognition results, the HMM shows greater recognition accuracy than PNN and SVM. As the vocabulary is small the recognition system gives minimum word error rate, and in future, medium or large vocabulary for continuous or isolated speech signal can be developed with different feature extraction techniques like Wavelet packet or Wavelet analysis or Linear Predictive Cepstral coefficients (LPCC). Ensemble model performance is analyzed and implemented.

REFERENCES

[1] Ms.Vimala.C and Dr.V.Radha, “Speaker Independent Isolated Speech recognition system for Tamil language using HMM” , International conference on communication technology and System design 2011.
[2] AN.Sigappi and S.Palanivel, “Spoken word recognition strategy for Tamil language” , IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.
[3] Dr.V.Radha et al, “Isolated word recognition system using Back propagation network for Tamil Spoken Language”, First International Conference on Computer Science, Engineering and Information Technology, CCSEIT 2011, Tirunelveli, Tamil Nadu, India, September 23-25, 2011. Proceedings.
[4] Milan Sigmund, “Voice recognition by computer”, Tectum Verlag DE, 2003
[5] S.rojathai et.al,” A Novel Speech Recognition System for Tamil Word Recognition based on MFCC and FFBNN”, European Journal of Scientific Research ISSN 1450-216X Vol. 85 No 4 September, 2012.
[6] Raymond Low and Roberto Togneri,” Speech Recognition Using the Probabilistic Neural Network”, Proceedings of ICSLP 1998.
[7] Nirav S. Uchat, “Hidden Markov Model and Speech Recognition” Seminar report, Department of Computer Science and Engineering Indian Institute of Technology, Mumbai.
[8] Dan Jurafsky. CS 224S / LINGUIST 181 Speech Recognition and Synthesis. World WideWeb, http://www.stanford.edu/class/cs224s/.
[9] R. Solera-Urena et.al, “SVMs for Automatic Speech Recognition: A Survey” Progress in Nonlinear Speech Processing, Springer Berlin Heidelberg, pp 190-216, 2007.