Automated Transcription System for Malayalam Language

Cini Kurian
Department of Computer Applications
Cochin University of Science & Technology, Cochin
Kerala, India

Kannan Balakrishnan
Department of Computer Applications
Cochin University of Science & Technology, Cochin
Kerala, India

ABSTRACT

Malayalam is one of the 22 scheduled languages in India with more than 130 million speakers. This paper presents a report on the development of a speaker independent, continuous transcription system for Malayalam. The system employs Hidden Markov Model (HMM) for acoustic modeling and Mel Frequency Cepstral Coefficient (MFCC) for feature extraction. It is trained with 21 male and female speakers in the age group ranging from 20 to 40 years. The system obtained a word recognition accuracy of 87.4% and a sentence recognition accuracy of 84%, when tested with a set of continuous speech data.

Keywords: HMM, MFCC, Speech Recognition, Transcription systems

1. INTRODUCTION

Speech Recognition technology has tremendous potential as it is an integral part of future intelligent devices, where speech recognition and speech synthesis are used as the basic means of communicating with humans. It will simplify the Herculean task of typing and will eliminate the conventional keyboard [14]. This technology adds a lot in manufacturing and control applications where hands or eyes are otherwise occupied. Disabled, elderly and blind people will no longer need to be away from the internet and Information Technology Revolution. Recently, there has been a large increase in the number of recognition applications for use over telephones, including automated dialing, operator assistance, and remote data access services; such as financial services, for voice dictation systems like medical transcription applications. Such tantalizing applications have initiated research in Automatic Speech Recognition (ASR) since 1950's

Malayalam belongs to one of the four major Dravidian languages of southern India with rich literary tradition. It has official language status in the state of Kerala (one of the southern states of India) and in the Union Territories of Lakshadweep and Mahe. Malayalam is also spoken in the Kanyakumari and Coimbatore districts of Tamil Nadu, southern parts and Kodagu districts of Karnataka State. It is also used by a large population of Indian expatriates living around the globe, including Persian Gulf, United states, Singapore, Australia and Europe. Malayalam language script consists of 53 letters with 37 consonants and 16 vowels. It is a syllable based language written with syllabic alphabet in which all consonants have an inherent vowel /a/. There are different spoken forms in Malayalam even though the literary dialect throughout Kerala is almost uniform. In a multilingual society like India, where there are about 1672 dialect of spoken forms, ASR technology has a wider scope. It would be a vital step in bridging the digital divide between English speaking Indian masses and others. Since there are no input standards for Indian languages, it eliminates the keyboard mapping of different fonts of Indian languages. Semantic integration of speech machine translation and speech synthesis system could facilitate exchange of information between people speaking two different languages

Many research and developments have been taken place in various Indian languages during the recent years [20, 25, 26]. However, Malayalam speech recognition is still in its infancy stage and very less work has been reported in Malayalam. A phonetic recognizer [27], a wavelet based vocabulary word recognizer [16] and number recognition systems [6] are the reported works in Malayalam

ASR is a branch of Artificial Intelligence (AI) and is related with number of fields of knowledge such as acoustics, linguistics, pattern recognition etc [12]. Speech is the most complex signal to deal with. In addition to the inherent physiological complexity of the human vocal tract, physical production system also varies from one person to person. The utterance of a word found to be different, even when produced by the same speaker at different times. Apart from the vast inherent difference across different speakers and different dialects, the speech signal is influenced by the transducers used to capture the signal, channels used to transmit the signal and the environment too can change the signals. The speech also changes with age, sex, and socio economic conditions. In continuous speech, the characteristics of sub-word units are affected by the context and the speaking rate. Sub-word unit selection and the creation of acoustic model are the two most important aspects in continuous speech recognition. In this work, context- dependent triphones [15] are used as the sub-word unit for recognition and Hidden Markov Model is used for acoustic modeling [25].

In most of the current speech recognition systems, the acoustic component of the recognizer is exclusively based on HMM [8, 9, 10]. The temporal evolution of speech is modeled by the Markov process in which each state is connected by transitions, arranged into a strict hierarchy of phones, words and sentences

Artificial neural networks (ANN) [2, 3] and support Vector machines (SVM) [1,7] are other techniques which are being applied to speech recognition problems. In ANN, temporal variation of speech can not be properly represented. SVM, being a binary static classifier, adaptation of the variability of the duration of speech utterance is very complex and confusing. SVM is a binary classifier while ASR, faces multiclass issues.

For processing the speech, the signal has to be represented in some parametric form. Wide range of methods exist for parametric representation of speech signals, such as Linear Prediction coding (LPC) [10], and Mel-Frequency Cepstrum Coefficients (MFCC) [8]. MFCCs are less susceptible to the physical conditions of the speaker's vocal tract [21]. In this system, Speech signal is represented by MFCC

2. METHODOLOGIES USED

2.1 Statistical Speech Recognition and HMM

An unknown speech wave form is converted by a front-end signal processor into a sequence of acoustic vectors, $O = o1,o2,o3,\ldots$. The utterance consists of sequence of words $W = w1, w2, w3$ ----*wn*. In ASR it is required to determine the most probable word sequence, W, given the observed acoustic signal *O*. Applying Bays' rule to decompose the required probability, [12]

$$S = \mathrm{arg}w\mathrm{max}P(W/O) = \mathrm{arg}w\mathrm{max}(P(O/W)P(W) / P(O))$$
$$S = \mathrm{arg}w\mathrm{max}P(O/W)P(W)$$

Posterior prior

Hence a speech recognizer should have two components: *P (W)*, the prior probability, is computed by language model, while P *(O/W)*, the observation likelihood, is computed by the acoustic model. Here the acoustic modeling of the recognition unit is done using HMM.

Since HMM is a statistical model in which it is assumed to be in a Markov process with unknown parameters, the challenge is to find all the appropriate hidden parameters from the observable states. Hence it can be considered as the simplest dynamic Bayesian network [8,10]. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. However, in a hidden Markov model, the state is not directly visible (so-called *hidden*), but the variables influenced by the states are visible. Each transition in the state diagram of a HMM has transition probability associated with it [19, 13]. These transition probabilities are denoted by matrix A. Here A is defined as $A = a_{ij}$ where $aij = P\ (t_{t+1} = j \mid j = i\ )$, the probability of being in state j at time $t+1$, given that we were in state i at time t. It is assumed that a_{ij}'s are independent of time. Each state is associated with a set of discrete symbols with an observation probability assigned to each symbol, or is associated with the set of continuous observation with a continuous observation probability density. These observation symbol probabilities are denoted by the parameter B. Here B is defined as $B = b_j(k)$, where $b_j(k) = P(v_k\ \text{at}\ t \mid i_t = j\ )$, the probability of observing the symbol v_k, given that it is in the state j. The initial state probability is denoted by the matrix π, where π is , defined as $\pi = \pi_i$ where $\pi_i = P\ (i_t = 1\ )$, the probability of being in state t at $t = 1$. Using the three parameters A, B, and π a HMM can be compactly denoted as $\lambda = \{ A, B, \pi \}$ An Example of an HMM with five states is shown in Figure 1.

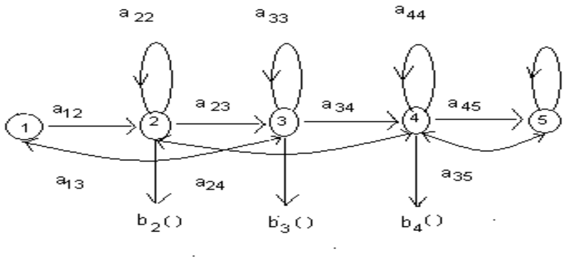

Figure 1 Topology of a five state HMM

There are three fundamental ASR problems that could be addressed with HMM [21]. Problem (i) is Scoring and evaluation problem i.e computing the likelihood of an observation sequence, given a particular HMM. This problem occurs in the recognition phase. Here for the given parameter vector sequence (observation sequence), derived from the test speech utterance, the likelihood value of each HMM is computed using forward algorithm. The symbol associated with the HMM, for which the likelihood is maximum, is identified as the recognized symbol corresponding to the input speech utterance. Problem (ii) is associated with training of the HMM for the given speech unit. Several examples of the same speech segments with different phonetic contexts are taken, and the parameters of the HMMs, λ, have been interactively refined for maximum likelihood estimation, using the Baum- Wetch algorithm [4]. Problem (iii) is associated with decoding or hidden state determination, where the best HMM state sequence is to be determined from the given observation sequence. The Viterbi algorithm [11,28] is employed for solving the problem (iii) as it is computationally efficient.

2.2 Mel Frequency Cepstral Coefficient

The mel-based Cepstral Coefficients called Mel-Frequency Cepstral Coefficients (MFCC), which is a popular parameter chosen for speech recognition, extracted from the speech signal. The MFCCs approximate human auditory systems in which frequency scale is wrapped to give high resolution at low frequencies and lower resolution at high frequencies [13]. This mel-scale wrapping improves the performance of speech recognition system compared with linear weight contributions from all frequencies [5,9]

In the training and recognition phase, MFCCs are extracted from speech signal and it is used for further processing. Here speech signal is transformed into a Frequency domain via Fast Fourier Transform (FFT). After mel-warping, the frequency scale is multiplied by a bank of filters. The filter-bank energies are then computed by integrating the energy in each filter. Then by applying Discrete Cosine Transform (DCT), Filter- bank log-energies are converted into Cepstral Coefficients. The flow diagram of generating MFCCs is shown in Figure 2.

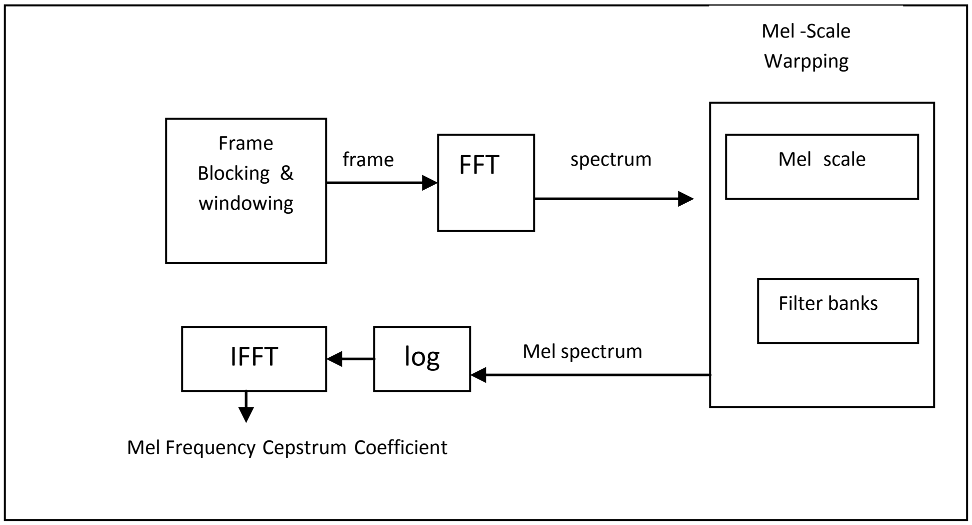

Fig. 2. Feature Extraction using MFCC

3. SYSTEM DEVELOPMENT

The basic architecture of the developed speech recognizer is shown in Figure 3. During training phase the HMM trainer creates acoustic models from three components; feature vectors, language models, and dictionary. Word models which are built from language models and pronunciation dictionary are used for pattern matching in the testing phase along with the syntax and semantics of the language. Figure 4 shows the different steps involved in the development of the proposed system

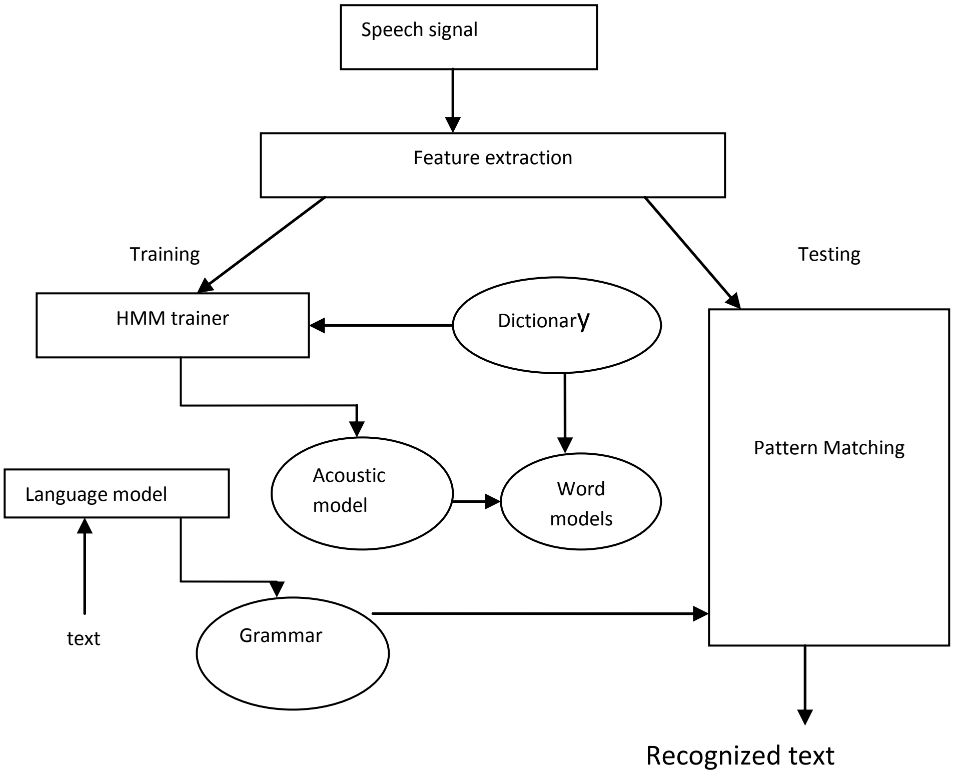

Fig. 3. System Architecture

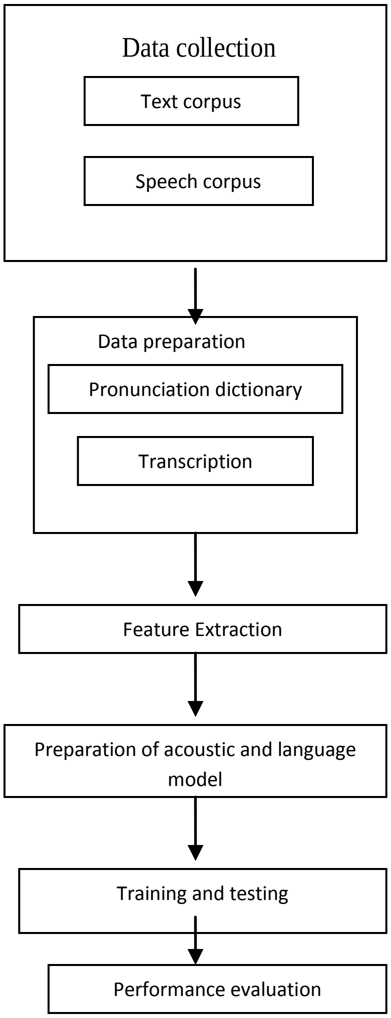

Fig. 4. System Development

3.1 Data Collection

Data collection involves creation of text corpus and speech corpus

3.1.1 Text Corpus

Selecting an optimal set of textual sentences in such a way that it will cover all phonemes, is the first step of system development [6]. The selected sentences should have enough occurrences of each type of co-articulation effects. The quality of text corpus chosen is very important for recognition performance. The text corpus is used as vocabulary for ASR and it is also used as the source for language modeling. Text materials are collected from on-line Malayalam newspapers and a font converter is used to convert the different fonts into desired electronic character code. The corpus balancing tool, CorpusCrt[18] is used to extract a set of phonetically rich sentences, from the text materials. Accordingly, 20 phonetically rich sentences are selected for training.

3.1.2 Speech Corpus

Speech data is used for training and testing an ASR system. The data for testing is collected form 10 male and 11 female speakers. For testing the system, five unknown speakers (whose sounds have not been trained) are selected and speech data of five different sentences collected from them. For training, the speakers were asked to read the 20 sentences. Recording is done in normal office environment using a head set, having microphone with 70Hz to 1600Hz of frequency range. Moreover, it is done with 16 kHz sampling frequency quantized by 16 bit, using a tool named Praat [16]. The speech is saved in Microsoft wave format

3.2 Data Preparation:

Data preparation involves creation of phone list, phonetic dictionary and transcription file.

3.2.1 Phone list

A phone list is a list of all acoustic units in the language that are needed to train models. It should have exactly the same units that are used in dictionaries. Unique phones of the language are also identified and defined. Here the prepared phone list contains 71 phonemes. Eg, clk k , sh, uu,

3.2.2. Phonetic dictionary

A well defined and accurate phonetic dictionary contributes a lot to the accuracy of the recognizing system. Here all words in the training data are mapped into the acoustic unit in the phone list. Here, in addition to the mapping table, some rules for the lexical representation of some sounds are also applied. This is to represent some sounds in Malayalam language in some special contexts. Thus, a phonetic dictionary is created for the 102 words of the vocabulary. Multiple pronunciations are also allowed in this work and it is represented in phonetic dictionary.

Eg. Yaatra y aa t r a
nikuti n i k u t i

3.2.3 Transcription file

The transcription file contains the sequence of words transcribed orthographically, written exactly as they occurred during the training speech. Each speech file in the database is transcribed into its corresponding orthographic representation. Hence here the transcription file is created for the 400 training sentences.

Eg:- <s> innu aaluva shivaratri </s>

4. FEATURE EXTRACTION

Mel-Frequency Cepstral Features are extracted from the speech signal using a window size of 25msec and a window shift of 10msec. From each frame of speech, 12 cepstral coefficients are obtained. The delta and acceleration coefficients are appended to the derived cepstral coefficients to obtain a 39 dimensional vector coefficient. These acoustic vectors are used for representing the voice characteristic of the speaker [14]. Therefore, each input utterance is transformed into a sequence of acoustic vectors.

5. ACOUSTIC MODELING

Characteristics of the basic recognition units [10] are represented by Acoustic models. Here, the Semi-Continuous Hidden Markov models (SCHMMs) [141] are chosen to represent context dependent phones (triphones). The phone likelihood is computed from the HMMs. The likelihood of the word is computed from the combined likelihood of all the phonemes. The acoustic model thus built is a 5 state SCHMMs, with states clustered using a decision tree.

6. LANGUAGE MODELING

The Language Model attempts to convey behavior of the language and is used to support the recognition process. The Language Model module provides word-level language structure, which can be represented by any number of pluggable implementations. These implementations typically fall into one of the two categories, viz; graph-driven grammars and stochastic N-Gram models. The stochastic N-Gram models provide probabilities for words given the observation of the previous N-1 words. The trigram model [N=3], based on the previous two words is powerful, as most words have a strong dependence on the previous two words and it can be estimated reasonably well with an attainable corpus. The trigram based language model with back-off is used for recognition. The language model is created using the CMU statistical LM toolkit [16].

7. TRAINING AND TESTING

For training and testing the system, the data base is divided into three equal parts- 1, 2, 3 and the experiment is conducted in a round robin fashion. For each experiment, 2/3rd of the data is taken for training and 1/3rd of the data is used for testing. In the experiment I, part 1 and part 2 of data is given for training. Then the same trained system is taken for testing the system with part 3 of the database. In experiment II, part 1 and part 3 of the data base is taken for training and part II of the database is used for testing. In experiment III, part 2 and part 3 of the database is taken for training and tested with part 1 of the database. The result obtained from each training and testing experiment in terms of Word Accuracy, Sentence Accuracy, Number of words deleted, inserted, substituted are detailed in table 1.

Table 1: Performance Evaluation of the System with Training and Testing Data

	TRAINING					TESTING				
	%					%				
Experiment	Word Accuracy	Deletions	Substitutions	Insertions	Sentence Accuracy	Word Accuracy	Deletions	Substitutions	Insertions	Sentence Accuracy
1	93.32	1.76	4.97	0.19	88.57	79.82	9.43	10.7	0	66.43
2	94.24	1.89	3.86	0	88.21	86.89	2.8	10.2	0	84.29
3	94.04	1.83	4.12	0	89.29	88.37	1.69	9.93	1	83.57

8. PERFORMANCE EVALUATION AND DISCUSSION

Word Error Rate (WER) is the standard evaluation metric for speech recognition. It is computed by SCLITE [13], a scoring and evaluating tool from National Institute of Standards and Technology (NIST). The inputs to the SCLITE are the reference text and the recognized text (output of the decoder). After each training and testing experiment, recognized text and the reference text are converted into the sclite compatible form and fed as input in sclite. Then detailed results are obtained in terms of WER, SER, number of word deletions; insertions and substitutions. If N is the number of words in the correct transcript; S, the number of substitutions; and D, the number of Deletions, then,

$$WER = ((S + D + I)N)/100$$

Sentence Error Rate (S.E.R) = (Number of sentences with at least one word error/ total Number of sentences) * 100

The recognition performance evaluation of ASR systems is to be measured on a corpus of data different from the training data. For this purpose, initially, the system is trained with the complete data base. Then, a separate test corpus, is created in such a way that it is different from the trained text corpus however contains only words from the pronunciation dictionary. In order to test the speaker independency of the system, speakers involved in the creation of training corpus are avoided. Hence a set of five sentences are created and five unknown speakers are selected. Speakers read the text corpus and recognition performance of the system is reported in terms of WER and SER.The WER obtained is 85% and SER is 82%. These results prove that the developed system is speaker independent with less error. This could be further reduced by training the system on a larger training data and also adding recordings from speakers of different dialects.

9. CONCLUSION

The objective of this study is to build a transcription system for Malayalam language. It is achieved while working with medium size vocabulary. The system is trained with 21 speakers, and it gives sentence recognition accuracy 84%. The accuracy results are highly encouraging while considering the training data and vocabulary size. This proves that the goal of creating a large vocabulary continuous transcription system is an easily achievable task in the near future.

10. REFERENCES

[1] A.Ganapatiraju , J. Hamaker and J. Picones, "Support vector machines for Speech Recogntion " Proceedings of the International Conference on Spoken Language Processing , pp 292-296, Sydney, Australia , November, 1999.

[2] A. Sperduti and Starita , " Supervised Neural Networks for Classification of structures "IEEE Transactions on Neural Networks, 8(3) , pp 714-735, May 1997.

[3] Behrman, L. Nash,J . Steck, V. Chandrashekar and S.Skinner, Simulations of Quantum Neural Networks", Information Sciences, 128(3-4): pp 257-269, October 2000.

[4] Baum, L.E, T. Petrie , G. Soules and N. Weiss, (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains, Ann. Math , Statist, vol 41, no, 1, pp 164-171.

[5] Chegalvarayan, R. and L. Deng , (1997), " HMM based speech recognition using state-dependent discriminatively derived transforms on mel-warped DFT features", IEEE Trans. Speech, Audio Processing, vol.5.pp 243-256.

[6] Cini Kurian , Kannan BalaKrishnan, (2009), " Speech Recognition of Malayalam Numbers", IEEE Transaction of Nature and Biologically Inspired Computing (NABIC-2009) pp 1475-1479.

[7] C.J.C. Burges, A tutorial on Support Vector Machines on Pattern "knowledge Discovery Data Mining, vol 2, no, 2 , pp. 121-167 , 1998.

[8] Davis S and Mermelstein P, "Comparative parametric representations of monosyllabic word recognition in continuously spoken sentences" IEEE Trans. ASSP vol 28 pp 57-336.

[9] Dimov, D., and Azamonov , I (2005). "Experimental specifics using HMM in isolated word speech recognition" International conference on Computer Systems and Technologies – CompSysTech , 2005.

[10] F. Felinek, "Statistical Methods for Speech Recognition" MIT Press , Cambridge, Massachusetts, USA, 1997.

[11] Forney, G.D., (1973), The Viterbi Algorithm, Proc. IEEE, vol . 61, pp. 268-277.

[12] Huang, X., Alex, A., and Hon, H.W (2001). "Spoken Language Processing; A Guide to Theory, Algorithm and System Development", Pentice Hall, Upper Saddle River, New Jersey .

[13] Jankowski , C.H , D.V and Lippman, (1995), A comparison of signal Processing front ends for Automatic word recognition , IEEE Trans. Speech , Audio, Processing, vol, 2, pp. 286-293.

[14] Jurasky, D., and Martin, J.H (2007). "Speech and Language Processing: An introduction to Natural Language Processing, Computational linguistics, and speech recognition" 2nd Edition .

[15] Kai-Fu Lee " Context-Dependent phonetic Hidden Markov Models for speaker Independent Continuous speech recognition, IEEE Transaction on Acoustics, Speech and Signal Processing vol 38, No. 4 , April 1990.

[16] Krishnan, V.R ; V. Jayakumar A, Anto P.B (2008) , "Speech Recognition of isolated Malayalam words using wavelet features and Artificial Neural Networks " DELTA 2008. 4th IEEE International symposium on Electronic Design, Test and Applications, 2008 volume Issue 23-25 Jan, 2008. Page(s) 240 – 243

[17] Mosur K, Ravishankar , Kevin A. Lenzo , Sphinx II User Guide CMU, 2001.

[18] Pallett et al., D, 1990. Tools for the analysis of bench mark speech recognition tests in ICASSP, volume 1

[19] P.Boersma, "Praat a system for doing phonetics by computer", Glot International, vol 5, 9/10, pp 341-345, 2005