JOURNAL OF CRITICAL REVIEWS

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

ENHANCING MOBILITY OF SMART CANE APPLICATION: MALAYALAM DIGIT RECOGNITION SYSTEM

Muhammed Shafi M*

*Research Scholar, Department of Linguistics, University of Kerala

Abstract

Digit recognition is important in many applications such as automatic data entry, PIN entry, voice dialing telephone, automated banking system, etc. The speaker-independent speech recognition system for Malayalam digits is presented in this paper. One of the most important uses of speech recognition technology is creating digit recognition software for the blind. By enabling them to use spoken commands or inputs to engage with digital devices and retrieve information, such a system might empower people who are visually handicapped. The system uses machine learning techniques for recognition and Enhanced Mel Frequency Cepstral Coefficient (EMFCC) as a characteristic for signal processing. On a test set of continuous digit recognition tasks, the system achieved 96.45%-word identification accuracy using 45 male and female voices ranging in age from 15 to 50.

Keywords: Speech Recognition, Digit Recognition, HMM, MFCC, Visually Impaired

INTRODUCTION

Digit recognition systems hold immense importance for visually impaired individuals for several compelling reasons. Firstly, they tackle accessibility challenges by offering an alternative means of inputting and accessing numeric data, surpassing the limitations of traditional methods like keyboard typing or reading printed numbers. This opens up independent access to a plethora of services and information via digital devices. Secondly, they promote autonomy by enabling visually impaired individuals to accomplish tasks such as accessing banking services or entering PINs without relying on external assistance, thus empowering them to manage daily activities independently. Thirdly, these systems improve efficiency by simplifying data entry processes, allowing users to input numeric data efficiently through speech recognition instead of laborious manual methods or assistance from sighted individuals, thereby saving time and energy. Additionally, the implementation of digit recognition systems fosters inclusivity by ensuring accessibility across various applications and services, catering to diverse needs and abilities and facilitating equal participation in society. Moreover, in critical areas like banking or medical services, these systems ensure the accurate and secure input of numeric data, providing a safe and dependable method for visually impaired individuals to access these services without compromising security or privacy. Overall, digit recognition systems play a pivotal role in enhancing accessibility, independence, efficiency, inclusivity, and safety for visually impaired individuals, enabling them to interact more effectively with digital devices and services in their daily lives.

Speech recognition is on the brink of surpassing traditional input methods like keyboards and pointing devices, offering a wide range of valuable applications from telephone directory assistance to office dictation and automatic voice translation. The allure of such applications has fueled research in Automatic Speech Recognition (ASR) since the 1950s, leading to the development of commercial ASR systems like Dragon Naturally Speaking and IBM Viva Voice.

Malayalam, spoken by approximately 38 million people in India, holds a significant position among the country's languages, boasting a rich literary tradition and distinct phonetic structure. With 37 consonants and 16 vowels, Malayalam is written in a syllabic alphabet, maintaining consistency in its literary form across Kerala despite regional dialectical variations. Despite extensive research and development efforts directed towards various Indian languages, including Malayalam, reported work specifically in the domain of Malayalam language has been relatively sparse. Notable among the reported works are a phonetic recognizer and an ASR system based on wavelet techniques.

This paper presents research focused on speaker-independent speech recognition of a continuous set of Malayalam digits without deliberate pauses between each word. Speech recognition is a multifaceted and intricate task, characterized by various challenges. One fundamental issue in speech recognition revolves around managing two primary types of variability: acoustic and temporal. Acoustic variability encompasses diverse accents, pronunciations, pitches, volumes, and other acoustic features, while temporal variability relates to variations in speaking rates.

Contemporary speech recognition systems predominantly rely on Hidden Markov Models (HMM), providing a statistical framework through a Markov process to model speech patterns with a structured hierarchy. HMM excels in capturing temporal dynamics and variability in speech, enabling accurate recognition across diverse scenarios. Artificial Neural

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

Networks (ANN) offer an alternative but face challenges in capturing temporal nuances, designing optimal topologies, slow convergence, and overfitting. Despite these challenges, ongoing research aims to improve ANN performance with techniques like regularization and novel architectures.

Support Vector Machines (SVM) are effective classifiers in various domains, but encounter difficulties in handling the complexity and scalability of speech data. SVMs require careful feature representation and parameter tuning, yet efforts persist to enhance their performance, often through integration with other methods like feature engineering. The continual exploration and adaptation of SVMs aim to overcome their limitations and improve their efficacy in speech recognition applications.

In the realm of speech recognition, Support Vector Machines (SVMs) present several notable drawbacks:
i) Static Classification: SVMs are static classifiers, posing challenges in adapting to the variability in speech utterance durations. Unlike dynamic models like Hidden Markov Models (HMMs), which can capture temporal dependencies in speech data, SVMs lack inherent capabilities to handle such variations effectively.

ii) Binary Classification: Originally formulated as binary classifiers, SVMs are inherently designed for two-class classification problems. However, ASR tasks often involve multiclass classification, where distinguishing among multiple speech categories or phonemes is essential. Adapting SVMs for such tasks necessitates strategies like one-vs-all or one-vs-one approaches, potentially introducing complexity and compromising performance.

iii) Scalability Issues: SVM training algorithms may struggle to efficiently handle the large databases typical in ASR tasks. Training SVMs on massive datasets can be computationally intensive and time-consuming, leading to scalability challenges. Consequently, SVMs may not be well-suited for applications requiring rapid training and deployment on extensive speech corpora.

In addressing these challenges, the authors of the discussed work utilize CMU Sphinx, a public domain speech recognition development toolkit, for both training and decoding purposes. Specifically, they employ a phoneme-based trigram model with 5-state Hidden Markov Models (HMMs) and a left-to-right Bakis topology. This approach leverages established techniques in ASR to mitigate the limitations associated with SVMs, providing a framework for developing and evaluating speech recognition systems.

DESIGN AND DEVELOPMENT OF THE SYTEM

The system design, depicted in Figure 1, adopts a structured approach. The input signal undergoes preprocessing before entering the feature extraction module, which extracts relevant features for recognition. During training, parameters for both the acoustic and language modules are estimated using the provided database, resulting in model creation. During testing, the features of the test speech are matched with these trained models for recognition. The trainer component is pivotal, generating a total of 183 models, including 25 base models and 155 triphones. Each state within these models is represented by a single Gaussian. Context-independent tying is employed, reducing the number of states from 1089 to 819 through the use of a decision tree, thereby enhancing efficiency while maintaining representational power. The default frame size and frame shift are set at 25ms and 10ms, respectively. The feature vectors used in the system are 39-dimensional, comprising 13 Mel Frequency Cepstral Coefficients (MFCCs) along with corresponding delta and acceleration coefficients. These coefficients capture essential aspects of the speech signal, facilitating accurate recognition.

To accommodate multiple pronunciations, the system incorporates multiple entries in the pronunciation dictionary. This feature enhances the system's flexibility and robustness in handling variations in speech patterns and pronunciations. Overall, the system's design encompasses preprocessing, feature extraction, model training, and testing phases, utilizing techniques such as context-independent tying and multiple pronunciation handling to achieve effective speech recognition.

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

I

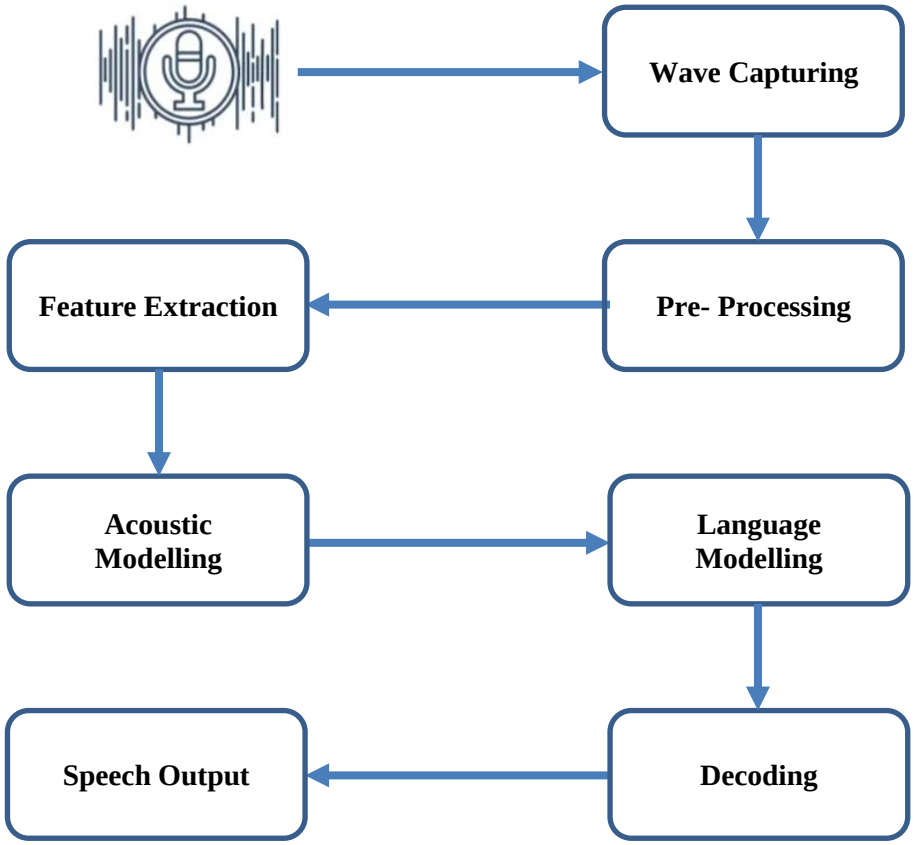

FIGURE 1. Block diagram of speech recognition system

SPEECH DATABASE

The system is designed to recognize sequences of Malayalam digits of any length, resulting in a lexicon size of eleven, which includes silence. Speech recordings were conducted in typical office environments using headsets equipped with microphones featuring a frequency range of 70Hz to 16000 Hz. Recordings were sampled at 16 kHz and quantized by 16 bits using CoolEdit, a Microsoft Wave format tool.

The training database comprises 420 sentences aimed at capturing acoustic variations across word boundaries. To achieve this, the training database consists of a small set of numbers encompassing all possible pairs of digits. Each set includes 20 seven-digit numbers, totaling 20 numbers per speaker. A total of 21 speakers (10 male and 11 female) read 20 continuous strings of digits in a normal manner, ensuring diverse speech samples for training.

Additionally, there is another database comprising utterances from five unknown speakers. These speakers were instructed to articulate any string of digits, resulting in 25 sentences (five speakers each spoke five strings of digits). This database is exclusively utilized for online testing and assessing the speaker independence of the system.

For each utterance, a transcription file is created, along with a language dictionary containing entries for each word in the string. The vocabulary size of the language dictionary is 11, matching the lexicon size. Furthermore, a phonetic dictionary is generated, comprising 27 phoneme-like units, including silence. These transcription and dictionary files are stored separately, facilitating efficient processing and recognition within the system.

FEATURE EXTRACTION

In speech recognition, the representation of signals using specific features is crucial. Various methods exist for parametric representation of sounds, including Linear Prediction Coding (LPC) and Mel-Frequency Cepstral Coefficients (MFCC). Among these, MFCC is widely acknowledged and favored for feature extraction.

MFCCs excel at capturing phonetically significant characteristics of speech by expressing the signal in the Mel-Frequency Scale. This scale features linear frequency spacing below 1000Hz and logarithmic spacing above 1000Hz, aligning well with human hearing's perceptual characteristics. Compared to speech waveforms, MFCCs are less affected by variations in the physical conditions of the speaker's vocal cords.

Figure 2's block diagram illustrates the feature extraction process, where the speech signal undergoes MFCC-based representation. This step is crucial for capturing the speech signal's essential characteristics, facilitating accurate recognition by downstream components of the system.

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

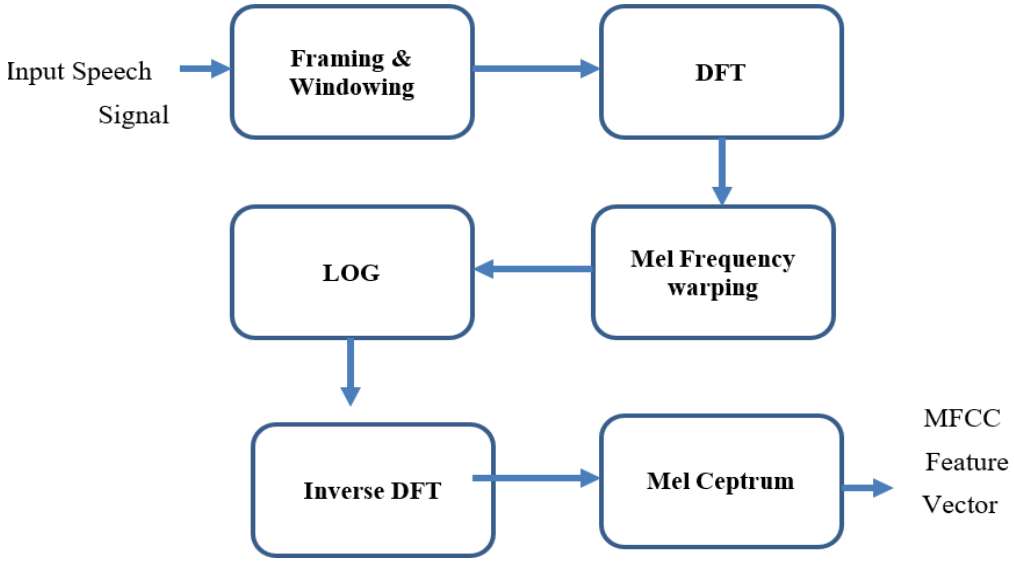

FIGURE 2 . The steps involved in the computation of MFCC

After preprocessing, the speech signal undergoes frame blocking and windowing, segmenting it into short-time segments. Next, the signal in the time domain is transformed into the frequency domain using Fast Fourier Transform (FFT), which breaks down the signal into its constituent frequencies. Subsequently, the spectrum undergoes Mel frequency wrapping. Mel frequency wrapping involves converting the linear frequency scale obtained from the FFT into the Mel frequency scale, which better aligns with human auditory perception. This conversion is achieved by applying a series of triangular filters spaced along the Mel frequency axis. These filters are designed to mimic the non-linear frequency response of the human auditory system. Overall, this sequence of processing steps allows the speech signal to be represented in a format that emphasizes perceptually relevant features, facilitating subsequent analysis and recognition tasks.

The process of Mel-frequency wrapping comprises two main steps: Mel-scale conversion and the creation of filter banks.

Mel-Scale Conversion: In this step, the frequency axis is transformed from the linear scale (measured in Hz) to the Mel scale. The Mel scale is a perceptual scale of pitches that approximates the human auditory system's response to different frequencies. For a given frequency *f* measured in Hz, the corresponding pitch in mels is calculated using the formula:

$$mel(f) = 2595 \times \log_{10}(1 + f / 700)$$

Mel Spectrum coefficients has to be converted to the time domain by applying Discrete cosine Transform (DCT) on it as in (2).

$$Cm = \sum_{k=1}^{N} \cos[m(k - 0.5)\pi / N]Ek, m = 1,2..L \quad (2)$$

Where N is the number filters, L is the number of mel- scale cepstral coefficients, *Ek* is the log energy obtained

This formula maps the given frequency *f* to its equivalent pitch in mels, resulting in a scale that more closely matches human perception.

Filter Banks: Once the frequencies are converted to Mels, the subsequent step involves creating filter banks. Filter banks are a collection of overlapping triangular filters, each centered at a different frequency in the Mel scale. These filters are crafted to mimic the frequency response of the human auditory system. Typically, the number of filters and their widths are determined based on psychoacoustic principles and empirical observations.

Together, these steps convert the linear-frequency spectrum obtained from the FFT into a representation that accentuates perceptually relevant features, rendering it suitable for subsequent analysis and recognition tasks in speech processing.

Following the filter creation process, for each speech frame lasting about 25ms with overlap, a set of Mel-frequency cepstral coefficients is computed. These coefficients, collectively known as acoustic vectors, can be utilized to represent and recognize the voice characteristics of the speaker. Hence, each input utterance is transformed into a sequence of acoustic vectors.

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

TESTING AND TRAINING

In the training phase of a speech recognition system, the Baum-Welch algorithm is commonly employed to estimate the parameters of the hidden Markov models (HMMs) used in acoustic modeling. This iterative algorithm updates the model parameters based on observed speech data, maximizing the likelihood of the data given the model. Once the models are trained, during testing, the Viterbi algorithm is utilized to decode the input speech signal and determine the most likely sequence of hidden states (phonetic units or words) that generated the observed speech.

Two main components are essential in a speech recognizer:

P(W), the prior probability, computed by the language model, represents the likelihood of encountering a particular sequence of words.

P(O|W), the observation likelihood, computed by the acoustic model, represents the probability of observing the given speech signal given a particular sequence of words.

The acoustic modeling typically employs a machine learning approach, such as HMMs, to estimate the parameters of the model from the training data. This involves finding the appropriate hidden parameters from the observable states, making it akin to a dynamic Bayesian network. In this network, the state transitions and output probabilities are modeled, allowing for the prediction of the hidden states (phonetic units) given the observed speech signal.

In a regular Markov model, the state transitions are the only parameters, and the variables influenced by the states are visible. However, in HMMs, each state has a probability distribution over the possible outputs (observations), allowing for probabilistic modeling of the relationship between the hidden states and the observed speech signal. This capability enables the system to account for the variability and uncertainty inherent in speech recognition tasks.

Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states [6, 10, 13]. Thus, HMM model can be defined as:

$\lambda = (Q, O, A, B, \Pi)$ Where ,Q is $\{q_i\}$ (all possible states),

O is $\{v_i\}$ (all possible observations) ,

A is $\{a_{ij}\}$ where $a_{ij} = P(X_{t+1} = q_j \mid X_t = q_i)$

(Transition probabilities),

B is $\{b_i\}$ where $b_i(k) = P(O_t = v_k \mid X_t = q_{it})$

(Observation probabilities of observation k at state i),

$\Pi = \{\pi_i\}$ where $\pi_i = P(X_0 = q_i)$

(Initial state Probabilities), and database is divided into three equal parts and for each experiments, 2/3 of the data is selected for training and the remaining 1/3 is selected for testing. From the test results word accuracy rate for each set is calculated. Using the above trained model the system has also tested with speech from unknown speakers.

PERFOMANCE EVALUTAION AND DISCUSSION

Digit	Pronounciation	Malayalam Writing	IPA symbol	Pronunciation dictionary
0	puujyam'	പൂജ്യം	pu:jʏm	clp p uu j y a m
1	onnu'	ഒന്ന്	ɪɒn	o n3 u’
2	ran't'u'	രണ്ട്	ɾaɳʈ	r a n: vbd: d:
3	muunnu'	മൂന്ന്	mu:n	m uu n3 u’
4	naalu'	നാല്	nɑ:lə	n aa l u’
5	anjchu'	അഞ്ച്	ɑɲcˈ	a nj clc u’
6	aar’u’	ആറ്	ɑ:ˈ	aa r ‘ u ‘
7	e'zu'	ഏഴ്	e : ʒˈ	ee zh u’
8	et't'u'	എട്ട്	eɖˈ	eclt tu
9	on_patu'	ഒൻപത്	ɒ n p ɑtˈ	o m clp p a clt t u’ o n clp p a clt t u’

TABLE 1. Malayalam digits used in this research

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

Word Error Rate (WER) is the standard evaluation metric for speech recognition. It is computed by SCLITE [8], a scoring and evaluating tool from National Institute of Standards and Technology (NIST). Inputs to SCLITE are the reference text and the output of the decoder is the recognized text (hypothesized sentence). WER align arecognized word string against the correct word string. 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 withat least one word error/ total Number of sentences) * 100

A. Result1: Performance of the system for Trainingdata

For training and testing, the database is divided into three equal parts. For each experiment 2/3rd of the data is taken for training and the remaining 1/3rd for testing. Table 2 gives the digit recognition accuracy and number ('sentence') recognition accuracy obtained. The most confusing pairs obtained was muunu' =>onnu' and the most falsely recognized digit was onnu'.

Experiment Number	Word Recognition Accuracy		Digit Recognition Accuracy	
	In Percent		In Percent	
	Train	Test	Train	Test
1	99.8	98.5	99.29	96.43
2	99.8	98.9	97.86	97.29
3	99.3	95.9	98.57	97.86
Average	99.63	97.76	98.57	97.19

Table:1performance of the system for training data

B. Result 2: Performance of the system for test(unseendata)

In order to test the system for live application, five speakers whose voice was unknown to the system were selected.They were asked to utter any sequence of digits of any length. The test result gave an accuracy of 96.45%.

CONCLUSIONS

The paper introduces a recognition system tailored for Malayalam language numbers, leveraging Hidden Markov Models (HMMs). This system excels in recognizing sequences of Malayalam digits pronounced without pauses between them. By offering a user-friendly interface for inputting numeric data into computers through spoken numbers, the system aims to enhance accessibility and convenience for users. The reported accuracy of the system is deemed satisfactory, indicating its effectiveness in accurately recognizing spoken Malayalam digits.However, the paper suggests that further accuracy improvements are attainable by expanding the training data. This expansion could entail incorporating utterances from a larger and more diverse pool of speakers, encompassing variations in age, accent, and other demographic factors. By enriching the diversity and volume of training data, the system can better adapt to the myriad speech patterns and styles encountered in real-world scenarios. This enhancement would bolster its robustness and generalization capability, ultimately leading to improved accuracy and performance in recognizing spoken Malayalam numbers.

REFERENCES

[1] A. Sperduti and A. Starita, "Supervised Neural Networks for Classification of Structures", IEEE Transactions on Neural Networks, 8(3): pp.714-735, May 1997.
[2] C. J.C. Burges, "A tutorial on support vector machines for pattern recognition," Knowledge Discovery Data Mining, vol. 2, no. 2, pp. 121–167, 1998.
[3] Davis S and Mermelstei P, "Comparison of parametric representations for Monosyllabic word Recognition in continuously spoken sentences", IEEE Trans. On ASSP,vol. 28, pp.357 – 366.
[4] Dimov, D., and Azamanov, I. (2005)." Experimental specifics of using HMM in isolated word Speech recognition". International Conference on Computer Systems and Technologies – CompSysTech '2005'.
[5] E. 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.
[6] F.Felinek, "Statistical Methods for Speech recognition" MIT Press, Cambridge, Massachusetts, USA,1997.
[7] Huang, X., Alex, A., and Hon, H. W. (2001). "Spoken Language Processing; A Guide to Theory, Algorithm and System Development", Prentice Hall, Upper Saddle River, New Jersey.
[8] Jurafsky, D., and Martin, J.H (2007). "Speech and Language processing: An introduction to natural Language processing, computational linguistics, and Speech recognition ", 2nd edition, http://www.cs.colorado

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

.edu/~martin/slp2.html.
[9] Krishnan, V.R.V. Jayakumar A, Anto P B (2008), "Speech Recognition of isolated Malayalam Words Using Wavlet features and Artificial Neural Network". DELTA2008, 4th IEEE International Symposium on Electronic Design, Test and Applications, 2008.Volume,Issue, 23-25 Jan.2008 Page(s):240–243.
[10] M Kumar., et al. "A Large Vocabulary Continuous Speech recognition system for Hindi", IBM Research and Development Journal, September 2004.
[11] Rabiner L R, "A Tutorial on Hidden Markov Models and selected Application s in Speech Recognition" Proc. IEEE, vol. 77, 1989, pp. 257 – 286.
[12] S.S Stevens and J. Volkman (1940), " The relation of pitch to frequency", American Journal of Psychology, vol. 53(3), pp 329-353.
[13] Saumudravijaya K, "Hindi Speech Recognition" (2001), J. Acoustic Society India, 29(1), pp 385-395.
[14] Singh, S. P., et al. "Building Large Vocabulary Speech Recognition Systems for Indian Languages" International Conference on Natural Language Processing, 1:245-254, 2004.
[15] Syama R, Suma Mary Idikkula (2008) " HMM Based Speech Recognition System For Malayalam", ICAI'08 – The 2008 International Conference on Artificial Intelligence, Monte Carlo Resort, Las Vegas, Nevada, USA (July 14- 17, 2008).
[16] http://www.icsi.berkeley.edu/Speech/docs/sctk- 1.2/sclite.htm.