Speaker Independent Speech Recognition Using Maximum Likelihood Approach for Isolated Words

Amaresh P Kandagal#1, V Udayashankara#2

#1 Department of Electronics & Communication Engineering
Sri Siddhartha Institute of Technology, Sri Siddhartha Academy of Higher Education, Tumkur
#2 Department of Electronics & Instrumentation Engineering,
Sri Jayachamarajendra College Of Engineering-Mysore

ABSTRACT

Speech is an intuitive interface for man machine interaction. Minimizing word error rate is a unique challenge to develop Automatic Speech Recognition (ASR) system. Performance of this system is far from perfect. Acoustic model and language models are fundamentals to build robust ASR engine. This paper presents a stochastic procedure for developing phoneme and word level acoustic models. Acoustic features estimated by Mel Frequency Cepstral Coefficients (MFCC) with 35% of overlapping of frames for every 25 milliseconds of a signal. The paper compares and highlights the word and phoneme level acoustic model performances for Kannada language vocabulary. The performance of the system is recorded for different vocabulary sizes, and word error rate (WER) computed for phoneme and word acoustic models. The system presents accuracy of 94.78046% and 97.6% for word and phoneme acoustic models respectively for the vocabulary 90 words. In addition, 98.08% of recognition rate for the vocabulary of 70 words.

Key words: ASR, MFCC, HMM, Acoustic Modeling, Human Computer Interaction

Corresponding Author: Amaresh P Kandagal

1. INTRODUCTION

Automatic speech recognition system automatically translates the natural speech. From six decades, researchers are rigorously trying to achieve robustness but it is not yet up to the mark. Speech is the primary mode to exchange information between entities[1][2]. Speech recognition system provides ease and natural interface for man machine interaction (Human Computer Interaction) in our day-to-day activity to reap maximum benefit from them [3]. Interacting with smart devices in native languages has a great advantage to the society [4]. Most of the people in India are supplemented with one of other languages for reading and writing like English, Telugu, Kannada,etc.. There is a timely need to develop a speech recognition engine for various native languages [5]. There are several applications to build native language speech recognition systems, one important among them is Interactive Voice Response (IVR) systems. The major challenge arises in designing ASR system, is to find the beginning and end of the signal in a word or a sentence [6].
ASR system has many phases such as acoustic analysis and feature extraction module, training phase to construct acoustic models from speech corpus, decision-making module and it is also known as a testing module, which maps unknown speech to the nearest word given in the vocabulary list. Figure 1 illustrates the architecture of ASR system designed using these phases. Based on utterances ASR systems are classified as i) Isolated systems ii) Connected systems iii) Continuous systems and also these are categorized as speaker dependent and speaker independent based on number of speakers system is supported for that is if system is supported only for single user is known speaker dependent system and if it is supported by multiple speakers is known as speaker independent system[7][8]. This paper focuses on design and development of speaker independent isolated speech recognition engine for the Kannada language.
1.1 Isolated Word Recognizers
These systems convert single utterance word to a text at a time identifying the silence between each word[8]. Generally, this kind of systems will have “Listen/Not-Listen” states, where it expects the user to wait between utterances. This sort of speech to text conversion systems are known as isolated speech recognition system[9][10].

1.2 Connected Word Recognizers
These systems work with a minimal silence or pause between isolated words. These allow individual spoken words to ‘run-together’ with a small silence between them[9]. An utterance may be a stand-alone word, or a sequence of multiple words, a sentence or collection of few sentences[7][9][11]. In the word-level acoustic modeling, the connected word-observation sequence is compared against all possible regions of stored word-pattern[12].

1.3 Continuous Speech Recognizers
These systems are potential to recognize continuous speech. To estimate start and end of utterance in a continuous speech is a unique challenge. In these systems, the user can speak naturally, but it is quite difficult to design an algorithm to estimate the content of continues utterance [13] [14].

The remaining part of the paper is organized in different sections. Section 2 describes related work and motivation to carry out the current work, Section 3 explains speech corpus construction. Section 4 describes a methodology for the proposed system, Section 5 presents experimental results and discussions and in Section 6 discusses on conclusions and future enhancements of the proposed work.

2. RELATED WORK AND MOTIVATION
Ankit Kumar et al [5] have implemented a Hindi speech recognition system. MFCC and perceptual linear prediction (PLP) method used for feature extraction. Monophone based acoustic models used and created the vocabulary of 70 words. HTK tool kit used to develop the system. The experimental result shows 95.08% accuracy.

Hemakumar et al [15] implemented ASR system for speaker dependent continuous Kannada languages by using HMM. Feature extraction is carried out by Linear Predictive Coding (LPC) converts voiced part signal into real cepstrum coefficients. Three state HMM models are created by using Baum-Welch Algorithm. The success rate for individually uttered sentence has 87.76%.

Al-Qatab et al [16] has developed Arabic speech recognition system by using HTK. Feature extraction is performed by MFCC algorithm. It supports continuous and isolated word recognition. It consists of 33 words. Triphone models are developed to estimate the parameters for a Hidden Markov Model. System performance is evaluated for 13 Arabian native speakers. The overall recognition rate for sentence correction is 90.62%, word correction is 98.01 %, and word accuracy is presented as 97.99%.

S.B.Harish et.al [17] presents an application specific Kannada speech recognition system. A database is created from 25 different speakers to support isolated speech recognition system.
MFCC is applied to extract acoustic features. With K-Nearest Neighbor (KNN) classifier 91.5% accuracy is achieved for small vocabulary.

Anusuya et al [18] have developed isolated speech recognition system for the Kannada language. Discrete Wavelet Transform (DWT) and MFCC methods are applied to extract feature vectors. They have created a database for isolated digits from 0 to 10 for the Kannada language. They have noted approximately 90% accuracy for 10 words vocabulary.

However, among these, speaker’s independent Kannada speech recognition systems are less identified. The lack of an efficient Kannada speech recognition system has motivated us to carry out the current research work. Hence in this paper, we have proposed speaker independent isolated word recognition for the Kannada language. The proposed system considers the recognized speech from a variety of contexts like words to phrase and from different users by restricting the contexts of the speech i.e. the words and phrases.

3. SPEECH CORPUS

Generally, the development of speech corpus constructed as application specific or general purpose [19]. In our simulation, we have created application specific corpus by reading the data from a microphone. The text corpus designed according to phonemes histogram values for isolated and connected words. These corpuses are built by taking Indian airport information. All the speech samples collected are speaker independent and these data considered for both word and sentence level in the Kannada language. The corpus contains the following information regarding the airport data namely a) Indian airport names, pronunciations of months, weekdays, am/pm, and digits in Indian local languages. Total 540 speaker’s voice samples collected to build large vocabulary speaker independent ASR system. Speaker’s voice samples captured from a microphone. The data is divided based on age group of the speakers. The signals are included from the age group between 20 to 60 years. Among this 333 are male and 207 female voices. Totally it consists of 2.7 hours of speech data. The recording was conducted at various environments like office, college, and home environments. From each subject, not more than 180 seconds data recorded. During recording, the user is in comfort zone to read the phonetically balanced set of words and sentences. This is sensitive because there should not be emotions in voice recordings. The following settings were set in the Goldwave and Wave Surfer speech capturing and editor tool. Single channel with Pulse code modulation technique selected with the sample rate of 16 kHz [19]. The acoustic signals captured using a microphone.

The smallest utterance of speech sounds that carry useful information known as a phoneme. In linguistics, the utterance of sounds or phonemes is categorized as vowel, diphthongs, semivowel, and consonants. The Kannada language has 49 characters in total; these are broadly divided into 14 vowels, 25 grouped consonants, and 15 miscellaneous consonants. Acoustic models created by training the recorded speech corpus. We have determined 46 monophonic acoustic models. Each model has different states and Gaussian mixers. The number of states for an acoustic model is empirically determined and validated.

For the training phase, 46 acoustic models at phoneme level and 90 acoustic models at words level in the Kannada language are used in the simulation. For the testing purpose, signals that are not trained with the speakers are used. This is evaluated for the recognition accuracy for a speaker independent speech signals.
4. METHODOLOGY

Design and development of speaker independent isolated word recognition have following steps:

1. Perform pre-processing of the acoustic signal
2. Extract the prominent features of the signal using MFCC
3. Training acoustic model using Baum Welch Algorithm
4. Build the Lexicon Models
5. Classify the signals using Viterbi decoder

The system is analyzed and implemented with the following modifications to the conventional system by using Hidden Markov Model Toolkit v3.3, such that the error rate minimized by increasing the recognition accuracy rate.

4.1 Pre-Processing

In this phase, the signal is removed from the unwanted signal from the background noise, equipment noise and other unwanted signals using spectral subtraction method. From this, the word boundary is detected by short time energy and zero crossing rate of the signal.

4.2 Feature Extraction

The conventional MFCC algorithm is applied to extract the prominent features. The proposed system is developed by making a variation in framing level and overlapping. Parameters are estimated with the 35% of overlapping of frames for every 25 milliseconds of a signal. Feature vector were of thirty-nine, with twelve Mel cepstrum plus log energy and first and second order derivative coefficients i.e. delta and acceleration coefficients [15][20][21].

4.3 Training Acoustic Model Using Baum Welch Algorithm

Speech transcription data (master labelled file) is required to build acoustic models. Labelling data created by using Wave surfer audio editing tool [22] as shown in figure 2. Forty-five minutes of corpus data is manually labelled and remaining data is labelled by automating force alignment algorithm. Tuning acoustic models is a challenging task in the ASR. Table 1 describes few vocabulary words with respective phoneme combinations (HTK format), which were used to build proposed ASR engine. To build the lexicon models, we have considered 3 states, five states, seven states and nine states are considered. Initially, word acoustic models are trained with 7 states with two Gaussian mixers. In the proposed system, acoustic models are developed based on cross validation score of the histograms. The optimal state for a particular word is decided based on the experimental basis.

| Table 1: Vocabulary of words |
|-----------------------------|
| Word | Phoneme combination |
| Name | [Name] | anne |
| Chand | [Chandra] | chandurasa |
| Nina | [Ninja] | Ninna |
| Ombe | [Ombe] | gomber |
| Bebe | [Bebe] | bebe |
| Obatou | [Obatou] | omatou |
| Nirda | [Nirda] | nirdam |
| Ganaka | [Ganaka] | ganaka |

Fig 2: Labelling of acoustic data to build mono phone models
From the simulation, it has been observed that accuracy decreases as the number of states increases, due to over fitting of the feature vector. To obtain the better recognition accuracy for the collected corpus data, the following criteria has considered and applied for fine tuning acoustic models. Bayesian Information Criterion and Akaike Information Criterion used to build the models [23].

For the sentence recognition, connected words lead in the formation of a sentence. Minimum message length criterion was chosen based on the likelihood procedure. Likelihood of the test sequence is computed by forwarding algorithm. For tuning and validation of the acoustic model, numbers of states varied from seven to nine states to calculate the recognition accuracies. Both likelihood and tuning procedures are used to create word acoustic models. For the words, which are, having stop sound utterances in between is implemented with nine states. Some the examples used in the simulation are as follows:

/ Devasthana/ (t is a stop sound),
/Prarambha/ (nasal sound with stop)
/Mandiraa / (d is stop sound)
/Ombathu / (nasal sound with stop)

This is considered because the above words take a long duration of utterance compared with other words. These words are modeled with 8 Gaussian mixtures to have better accuracy rate. The same procedure and the Gaussian mixtures are followed for evaluating and tuning the models with 7 and 5 states.

4.4 Lexicon Models

The lexicon (or dictionary) in general composed on the orthography of words that takes place in the training process and its phonetic transcriptions. At the time of recognition, the phonetic transcriptions in the lexicon module as a constraint. It describes a series of phonomes that are allowed to occur [24]. To build these lexicon models the following phonemes and their combinations are used.

/ a/,/aa/,/e/,/ee/,/i/,/ii/,/o/,/ou/,/l/,/ol/,/l/,/ll/,/p/,/t/,/d/,/k/,/g/,/T/,/D/,/c/,/j/,/h/,/s/,/sh/,/S h/,/m/,/n/,/N/,/L/,/l/,/v/,/v/,/y/,/y/,/ph/,/bh/,/th/,/dh/,/gh/,/Th/,/Dh/,/ch/,/chh/,/nh/.

Table-1 describes the few words sample of vocabulary. All acoustic models trained for five state models with three Gaussian mixers.

4.5 Decoding

Given acoustic and languages models, the Viterbi algorithm performs decoding. Recognition module converts unknown observation sequences of acoustic parameters to text. Suppose each word corresponds to a vector of acoustic observation cues A defined as:

\[ A = a_1a_2a_3a_4 \ldots \ldots a_T \]  

Where, \( A_t \) acoustic feature vector observed at time ‘t’ [24], [25]. The decoder module estimates the maximum likelihood for sequence of words \( W \) given the acoustic feature vectorsA, that is,

\[ \hat{W} = \arg\max_W P(W|A) \]  

The probability \( P(W|A) \) estimated using Bayes’ rule given as:

\[ P(W|A) = \frac{P(A|W)P(W)}{P(A)} \]
Given a sequence of acoustic features $A$, decoder estimate the sequence of words $W$ which maximizes the probability $P(A|W)P(W)$ [5][25]. Where $P(W)$ is estimated by the language model it is a prior probability of the word. $P(A|W)$ is the observation likelihood estimated by using HMM acoustic model. The term $P(A)$ is constant and therefore it could be ignored. HVite API generates the result in a transcription file for a given input arguments of observation $A$ that is feature vectors along with HMMs definition, vocabulary list, language model network. The HVite tool processes the observation sequence using Viterbi Algorithm, established by the token passing algorithm [26], which maps with reference to the recognizer's Markov models.

4.6 Performance Analysis

The performance of ASR system measured by two parameters that are Word Error Rate (WER) and Percentage of Accuracy (POA) and these are computed as [5][7][27]:

$$\text{WER} = \frac{S+I+D}{N} \times 100 = 100 - (\text{POA})$$  \hspace{1cm} (4)

$$\text{POA} = \frac{N-D-S-I}{N} \times 100 = \frac{H-I}{N} \times 100$$  \hspace{1cm} (5)

Where $N$ is a total number of words in the validation set, and $D$, $S$, $I$ and $H$ are a number of deletions, number of substitutions, number of insertions and number of correct labels respectively.

5. RESULTS AND DISCUSSIONS

Simulation is carried out at three different levels. At word level, phoneme level and on vocabulary size. The results and discussions are presented for each as discussed below.

5.1 Experiments with Word and Phoneme Acoustic models

These experiments are conducted on the vocabulary size of 90 words. Word and phoneme acoustic models are created with 525 speakers (325 male and 200 female utterances) with the age group between 20-60 years. For testing 8 male and 7 female voice samples are adopted. 3-14 utterances of each speaker are considered with 15 different speakers.

The accuracy obtained at word level is 94.78046% and 97.6% at the phoneme level. The recognition results for phoneme acoustic model is tabulated in Table 2.

5.2 Experiments with Different Vocabulary Sizes

The performance of an ASR system was computed with the different vocabulary size of 10 words – 150 words. Phoneme based acoustic models are used for this experiment. We have achieved higher performance for smaller vocabulary. The WER increases for bigger vocabulary size. The accuracy of the system up to 70 words of vocabulary size is 100%. Above 70 words the performance degrades i.e. increase in vocabulary size. 98.08% accuracy is achieved for 70-word vocabulary size. Testing is performed with 5 speakers, each speaker file contains 7-12 continuous words. Results are plotted is shown in figure 3.
Table 2 Performance Analysis with Phoneme Acoustic Model

| Speaker ID | Number of spoken words (N) | Number of recognized words (H) | Number of Deletions (D) | Number of insertion (I) | Number of substitution (S) | % word correct | % word accuracy | Word error rate (WER) |
|------------|-----------------------------|---------------------------------|-------------------------|-------------------------|-----------------------------|----------------|-----------------|---------------------|
| 1          | 55                          | 55                              | 0                       | 1                       | 0                           | 100            | 98.181818       | 1.818182            |
| 2          | 58                          | 57                              | 0                       | 1                       | 1                           | 98.275862      | 96.551724       | 3.448276            |
| 3          | 65                          | 64                              | 1                       | 1                       | 0                           | 98.461538      | 96.923077       | 3.076923            |
| 4          | 48                          | 47                              | 1                       | 0                       | 0                           | 97.916667      | 97.916667       | 2.083333            |
| 5          | 36                          | 36                              | 0                       | 0                       | 0                           | 100            | 100             | 0                   |
| 6          | 54                          | 53                              | 0                       | 1                       | 1                           | 98.148148      | 96.296296       | 3.703704            |
| 7          | 47                          | 46                              | 1                       | 0                       | 0                           | 97.87234       | 97.87234        | 2.12766             |
| 8          | 45                          | 44                              | 1                       | 0                       | 0                           | 97.777778      | 97.777778       | 2.222222            |
| 9          | 67                          | 65                              | 1                       | 0                       | 1                           | 97.014925      | 97.014925       | 2.985075            |
| 10         | 85                          | 84                              | 1                       | 2                       | 0                           | 98.823529      | 96.470588       | 3.529412            |
| 11         | 65                          | 64                              | 1                       | 0                       | 0                           | 98.461538      | 98.461538       | 1.583462            |
| 12         | 72                          | 71                              | 1                       | 0                       | 0                           | 98.611111      | 98.611111       | 1.388889            |
| 13         | 54                          | 53                              | 1                       | 0                       | 0                           | 98.148148      | 98.148148       | 1.851852            |
| 14         | 43                          | 43                              | 0                       | 1                       | 0                           | 100            | 97.674419       | 2.325581            |
| 15         | 72                          | 71                              | 1                       | 1                       | 0                           | 98.611111      | 97.222222       | 2.777778            |
| Overall result |                        |                                  |                         |                         |                             | 98.54151       | 97.67484        | 2.325156            |

Fig 3: Result for Various Vocabulary Sizes
5.3 Comparative Analysis

In this section, we have compared current work with existing similar work as explored and discussed in section 2. Table 3 illustrates the tabular comparison of the existing system with the reported works.

| Authors/Year          | Language Supported | Speaker Independence | Vocabulary size | Acoustic model | % accuracy | Word Error Rate |
|-----------------------|--------------------|----------------------|-----------------|----------------|------------|-----------------|
| Al-Qataab et al 2010  | Arabic             | Yes                  | 33              | Phoneme        | 97.99%     | 2.01%          |
| Anusuya et al 2010    | Kannada            | No                   | 10              | -              | ~90%       | ~10%           |
| Hemakumar et al 2014  | Kannada            | No                   | 20 Sentences    | Phoneme        | 87.76%     | 12.24%         |
| Ankit Kumar et al 2014| Hindi              | Yes                  | 70              | Phoneme        | 95.08%     | 4.92%          |
| S.B. Harish et al 2017| Kannada            | Yes                  | 12              | -              | 91.5%      | 8.5%           |
| Our work 2017        | Kannada            | Yes                  | 90              | Phoneme        | 97.6%      | 2.4%           |
|                       |                    |                      |                 | Word           | 94.7%      | 5.3%           |

5.4 Discussions

Word acoustic models require more memory. Any new addition of words to vocabulary increases the risk to build word acoustic models. More the words, more (storage area) memory it utilizes simultaneously increasing the tuning of word acoustic models (Number of states and Gaussian mixers) in the dictionary. For large vocabulary ASR system, phoneme acoustic models are feasible compared to word acoustic models. For application specific and limited vocabulary, ASR systems can be designed efficiently with word acoustic models.

6. CONCLUSIONS AND FUTURE ENHANCEMENT

This paper discusses speaker independent Kannada speech recognition system performances at phoneme and word levels. It is observed that framing and overlapping concepts in feature extraction procedure are prominent in reducing the computational time of the system. The unigram-mono phone models with 8 Gaussian mixtures have resulted in better recognition rate at both word and phoneme level. As the vocabulary size increases, acoustic modeling and tagging of states becomes complex and reduces recognition accuracy for speaker independent systems.

As a future enhancement, the proposed system can be extended to construct the acoustic models using bi-gram and trigram models to evaluate the speech recognition system for large vocabulary sizes varied with a number of speakers.

ACKNOWLEDGEMENTS

We thank Dr. Anusuya M A, Professor, Department of Computer Science and Engineering, JSS Science and Technology University, Mysuru, for her significant support in understanding and implementing this presented paper.
REFERENCES

[1] Cantor, A. Speech recognition: An accommodation planning perspective. Proceedings of CSUN 2001 Conference, Los Angeles. Northridge: California State University. 2001

[2] Paul, D.B, “An efficient A stack decoder algorithm for continuous speech recognition with a stochastic language model,” in Proc. ICASSP '92, San Francisco, CA, pp. 25-28, Mar. 1992.

[3] Pantic, M., Pentland, A., Nijholt, A., & Huang, T. Machine Understanding of Human Behavior. In Proceedings AI for Human Computing (AI4HC’07), workshop at IJCAI 2007 (pp. 13-24). IJCAI. 2007

[4] Muskan, Naveen Aggarwal, "Punjabi Speech Recognition: A Survey", Proc. of the Intl. Conf. on Advances in Engineering and Technology - ICAET-2014.

[5] Ankit Kumar, Mohit Dua and Tripti Choudhary, “Continuous Hindi Speech Recognition Using Monophone Based Acoustic Modeling”, International Journal of Computer Applications (IJCA) (0975 – 8887). 2014.

[6] Amarendra P Kandagal and V. Udayashankara, “Automatic Bimodal Audiovisual Speech Recognition: A Review, IEEE Conference on Contemporary Computing and Informatics (IC3I), Page(s):940-945, Nov 2014.

[7] M.A. Anusuya, S.K. Katti, Speech Recognition by Machine: A Review, International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009

[8] U. G. Patil, S. D. Shirbahadurkar, A. N. Paithane, Automatic Speech Recognition of isolated words in Hindi language using MFCC, International Conference on Computing, Analytics and Security Trends (CAST), 19-21 Dec. 2016.

[9] D. Raj Reddy, "Speech Recognition by Machine: A Review", PROCEEDINGS OF THE IEEE, VOL. 64, NO. 4, APRIL 1976

[10] Preeti Saini, Parneet Kaur, Automatic Speech Recognition: A Review, International Journal of Engineering Trends and Technology- Volume4, Issue2- 2013

[11] Rajan Mehta, Mamta, R.K. Aggarwal, Automatic Speech Recognition: A Survey, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), Volume 3, January 2014

[12] R. Makhijani, and R. Gupta, “Isolated word speech recognition system using dynamic time warping,” International Journal of Engineering sciences and emerging Technologies, vol. 6, no. 3, pp. 352-367, 2013.
[13] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition from features to supervectors,” Speech Communication, vol. 52, no. 1, pp. 12-40, 2010.

[14] H. Ali, N. Ahmad, X. Zhou, K. Iqbal, and S. M. Ali, “DWT features performance analysis for automatic speech recognition of Urdu,” Springerplus, vol. 3, pp. 1-10, 2014.

[15] Hemakumar G and Punitha P, “Speaker Dependent Continuous Kannada Speech Recognition Using HMM”, IEEE, International Conference on Intelligent Computing Applications, 2014.

[16] Al-Qatab, B.A.Q. and Ainon, R.N. "Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)", International Symposium in Information Technology (ITSim), Vol. 2, pp.557–562, June 2010.

[17] S.B.Harish, et.al, "Kannada Speech Recognition Using MFCC and KNN Classifier for Banking Application", International Journal Research of Innovative research in computer and communication Engineering, ISSN (Online): 2320, Vol.5, Issue 1, January 2017.

[18] Anusuya MA and Katti SK “Mel Frequency Discrete Wavelet Coefficients For Kannada Speech Recognition Using PCA” In Proceedings of international conference on advances in computer science ACEEE, pp. 225 – 227, 2010.

[19] Amaresh P Kandagal, V Udayashankara,” Speech Corpus Development for Speaker Independent Speech Recognition for Indian Languages ",Grenze International Journal of Computer Theory and Engineering,Vol 3, Issue 4,pp-81-87,2017

[20] F. Guojiang, “A novel isolated speech recognition method based on neural network,” International Conference on Networking and Information Technology, vol. 17, pp. 264-269, 2011.

[21] N. S. Nehe and R. S. Holambe, “DWT and LPC based feature extraction methods for isolated word recognition,” EURASIP Journal of Audio, Speech, Music Processing, vol. 2012, no. 1, pp. 1-7, 2012.

[22] K. Sjolander and J. Beskow. Wavesurfer-an open source speech tool. In Proceedings of International Conference on Spoken Language Processing, pages 464–467, 2000.

[23] Bowen Zhou and John H. L. Hansen,"Efficient Audio Stream Segmentation via the Combined T2 Statistic and Bayesian Information Criterion",IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 13, NO. 4, JULY 2005

[24] D. Jurafsky and J. H. Martin, —Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, (2009).
[25] Lei Wang, “Research on speech recognition algorithm based on HTK toolbox”, 3rd International Conference on Materials Engineering, Manufacturing Technology and Control (ICMEMTC) 2016.

[26] Young, S.J., et al., “The HTK Book”, Microsoft Corporation and Cambridge University Engineering Department, Cambridge. 2009.

[27] Kuldeep Kumar and R.K. Aggarwal, “A Hindi speech recognition system for connected words using HTK”, International Journal Computational Systems Engineering, Vol. 1, No. 1, 2012