# Isolated Digits Recognition in Kannada Language

Gurudath K.P.*
M.Tech Scholar
Dept. of E&C
VVCE Mysore,

D.J. Ravi, PhD
Professor & Head
Dept. of E&C
VVCE Mysore,

## ABSTRACT
In this paper, have implemented the isolated digit recognition in Kannada language using Hidden Markov Model Toolkit (HTK). Hidden Markov models used as pattern recognizer with the help of MFCC as a featured vector of the wave samples. The paper focuses on all isolated digits of Kannada i.e. Sonne to Ommbattu (0 to 9), The system helps in interaction of rural people and the computer or any system. The system data structure is defined at word level and its performance is evaluated.

## Keywords
Automatic Speech Recognition (ASR), Mel frequency Cepstral coefficients (MFCC), Hidden Markov Model (HMM), Isolated Kannada digits, HMM Toolkit (HTK).

## 1. INTRODUCTION
Speech is basic mass media communication for interchange the information from one person to another. Speech is the unique character of human used to communicate and to express the thoughts. In spite of massive developments in several studies of signal processing technology, existing computers and other electronic devices requires assured quantity of physical interfacing with users. The people who belong to physically handicapped and blind, they were not able to operate the computer system. Communicating with computers using speech in the resident language will be excellent key success to the above mentioned crisis. Also, the people who were not able to hear, they can communicate through the speech to text conversion type of communication.

Speech recognition is defined as the process of converting an acoustic signal into some set of text or in general task, taken through an input device called microphone or a telephone. Automatic speech recognition (ASR) is one of the emerging, mounting fields in the construction of speech discipline & technology. Upcoming major improvement in computing technology is man-machine interaction in new generation. Robustness is a term in speech recognition to advance the system. Accuracy is the main concern in speech recognition that can be achieved and maintain by the robustness, still the inferiority of speech or be different in acoustical, articulate or in the phonetic uniqueness of training and testing atmosphere. The Automatic Speech Recognition system offers feasible solution, which gives a good result for the small vocabulary system with the huge data samples.

The speech classifications are of 4 types

1. Isolated Words 2. Connected Words

3. Continuous Speech 4. Spontaneous Speech.

The isolated words are sandwiched between silence, Connected words are run together of two utterances with the small pause or silence between them. The continuous speech is a flow of utterance in which the small silences are in between them. Spontaneous speech is unrelated utterances in between the continuous speech.

The modes of recognition system is of

1. Speaker Dependent 2. Speaker Independent

The speaker dependent system that can operate only with a known trained utterances of peoples. The speaker independent system that can operate with known & unknown utterances of the peoples.

The development of speech recognition system at Bell labs, by [1] for the recognition of isolated numbers for only a speaker and system depends on the spectral resonance of the vowels in the each utterance. The system has been implemented in the European language. Denes et al. [2] Implemented the system with phoneme recognition of 9 consonants and 4 vowels in English, for recognizing phonemes they used pattern matching and spectrum analyzer technique. This was implemented at MIT labs. The work done in the field of speech recognition is more, but work done in the Kannada language is less compared to other sister languages like English, Hindi, Tamil, Telugu, Malayalam, Punjabi [7,8,9,10,11,12,13] etc. In all these languages robustness affects a lot, but some standard data base was already available in English so this is on top in the field of recognition. Communication between human and computer through a Natural language conversational interface acts an extremely key role in civilizing the usage of computers by for the common people. Need more time to bring communication between computer and human as close to human-human interaction seeing that probable.

## 2. AUTOMATIC SPEECH RECOGNITION (ASR)
Speech is a quasi periodic signal lies in the band range between 300 Hz and 3500 Hz. The system which is responsible for recognition of an utterance is called 'Speech Recognition'.

### 2.1 Architecture of ASR
The Fig.1 shows the architecture of speech recognition system. The architecture shows the step by step methods to implement the ASR system. The Automatic speech recognition system includes a different phases:

#### *2.1.1 Pre-Processing*
The speech is an analog signal, which cannot be operated unswervingly by the digital systems, therefore analog signal to be converted into digital form or a structure that can be capable to operate by the recognizer. To realize, the speech signal is to digitize first, after that the digitized or sampled signal is operated by the first-order filter to flatten the signal spectrally.

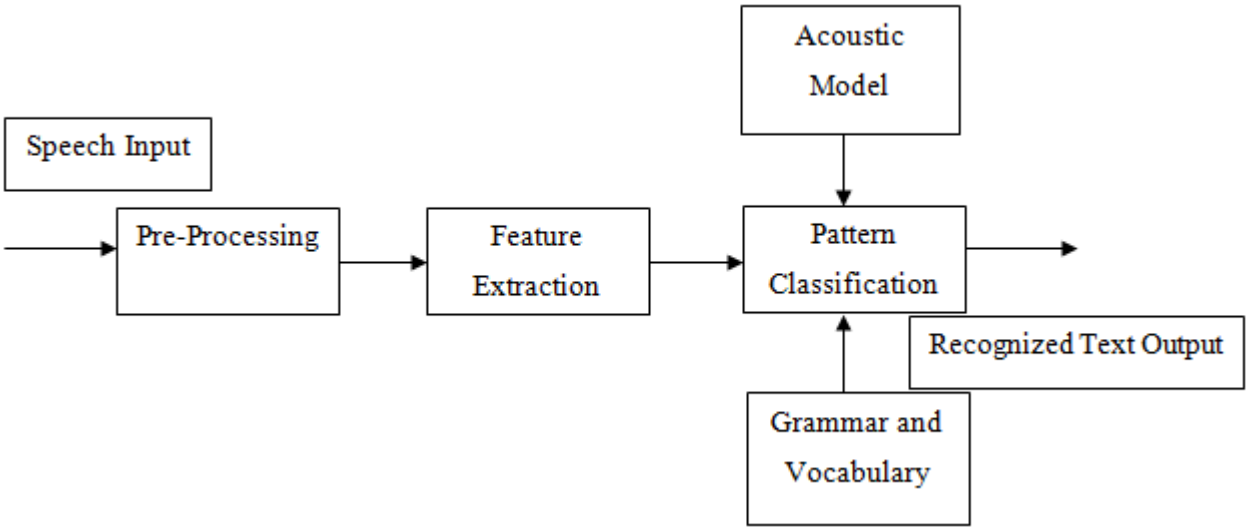


**Fig.1: Structure of ASR.**

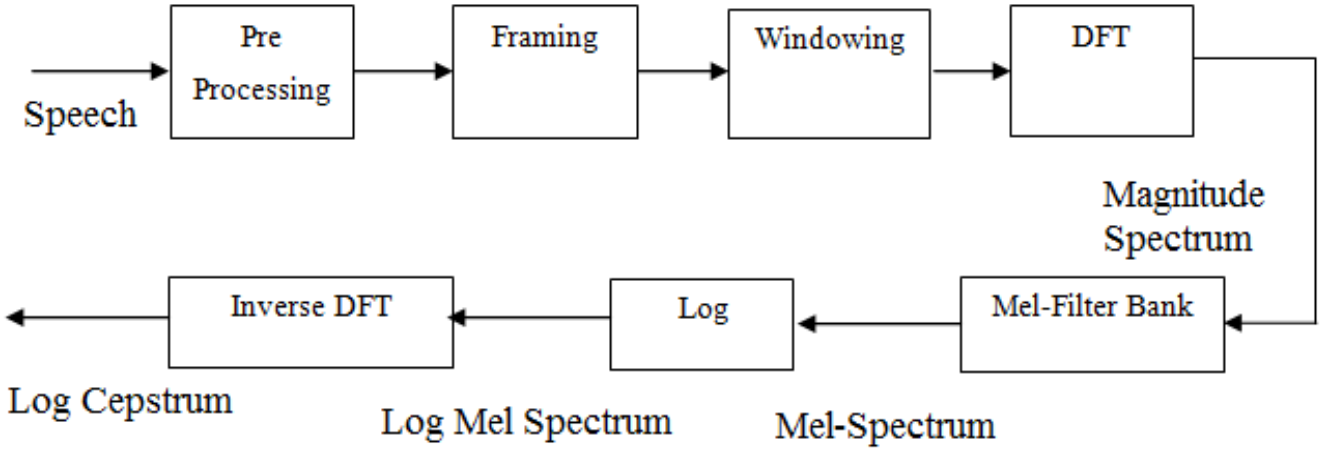


**Fig.2: MFCC Feature Extraction Method.**

The process of increasing the magnitude of higher frequencies with respect to the lower frequencies is called pre-emphasis, the value of pre-emphasis is 0.97.

### *2.1.2 Feature Extraction*

The extraction of a feature is a process of finding set of features of the utterance that have correlated with the acoustic signal, shown in Fig.2 Features are the parameters that can estimate by signal processing. Rejecting the unwanted information while maintaining the required part of the signal is the process in feature extraction. Maintaining required part of the signal includes measurement of the delta, energy, frequency response, acceleration and some other perceptual measurements and statistical conditions to form the observation vectors [14]. The Majority of the speech recognition work uses the feature called Mel-Frequency Cepstral Coefficient (MFCC) in now a days and this work is also using the same. In MFCC, the main benefit is that it uses Mel-frequency scaling which is very fairly accurate to the human auditory system [10]. The human auditory perception is resulted from the Mel-scale. The Mel-scale is a logarithmic scale used to extract the features and also for humanizing the recognition performance. The MFCC is given by [10]

$$Mel(f) = 2595 \log_{10}[1 + f/700] \quad \ldots(1)$$

The preprocessed samples first divide into a set of frames, which is further operated with the hamming window with 50% of overlapping, the outcome of hamming window is applied with Fourier transform and pass it via Mel-filter bank of 24 channels, taking logarithm for the Mel-filtered samples and apply Inverse Fourier transform to get MFCC as a feature vector of 12.

### *2.1.3 Pattern Classification*

The acoustic parameters or features of a word for the test samples is recognized or classified by the Pattern classifier. The classification problem is solved by judgment of the most probable sequence of words W given the acoustic sample L the below equation is from Baye's theorem is given by Equation (2). L is an acoustic observation sequence, W is the words that find by the classifier which maximizes the

$$P(w_i|L) = P\left[\frac{L}{wi}\right] P[wi]/P[L] \quad \ldots(2)$$

probability of [L/Wi]P[Wi], P[Wi] is the prior probability predictable by language model. P[L/Wi] is scrutiny of likelihood, Known as acoustic model.

### *2.1.4 Acoustic Models*

Acoustic modeling, acting a significant part to get better the accuracy of the system. Establishing the statistical representation of feature vector sequences estimated from the acoustic modeling of the speech signal. HMM is one of the most familiar acoustic models.

The Pronunciation modeling is integrated with the acoustic modeling, which is the pronunciation model, includes the multi-sequence speech units such as phonemes which forms the large speech units called vocabulary or phrases are the objective of the recognition system. In sort to achieve the robustness in speech recognition, and to reshape the features of speech, acoustic modeling has also used the feedback from the recognizer.

### *2.1.5 Language Models*

The production of accurate value of likelihood of the word can be made by the language modeling. In order to generate the probability, a language model has the structural constraints available in the language. After occurring of each utterance sequence it determines the probability of the word.

## 2.2 Pattern Recognition

A pattern is an object, as the name indicates it is process or event. A character of the pattern is a set of patterns have average attributes and usually originating from the source. During the process of the recognition character may assign the object. In machine learning system label is assigned to input value for pattern recognition. Classification is the example for pattern recognition, which assigns the all input value to one of the given class (ex: conclude whether a particular input voice is "Male" or "Female").

All though the pattern recognition encompasses the general problem of another type of outputs. Regression assigns the genuine output from the each input, Sequence labeling which assigns each sequence of values to class (for example, part of speech assigning to each input sentence. The aim of pattern recognition algorithm is by considering the statistical variations of input, provides the realistic answer by matching "most likely" to the all possible input [14].

### *2.2.1 Statistical Pattern Recognition*

The natural construction to originate solutions to problems of pattern recognition is statistical method, which clearly recognizes the nature of the probabilistic information to be processed and form to which the result should express [14].

Machine intelligence is based on the patterns and pattern classes of statistical modeling. Along with a statistical model applies probability theory or decision theory to get algorithm. Training patterns set the learning algorithm.

Pattern recognition includes a sequence of problems of huge practical importance, from speech recognition and classification of handwritten characters, to reveal the error in machinery and medical diagnosis.

# 3. HIDDEN MARKOV MODEL (HMM)

The Hidden Markov Model is statistical model, extracts the feature vectors and also supportive of the recognition. HMM is an arithmetical model for prearranged series of symbols, performing as a stochastic finite state machine, assumes that which is built up from a restricted set of possible states, each of these states were associated with the probability distribution function (pdf).

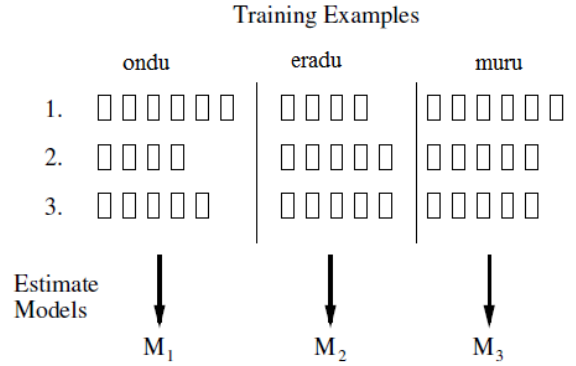


**Fig.3: Digits training using HMM.**

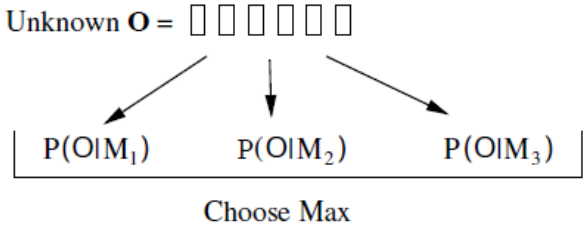


**Fig.4: Unknown utterance recognition using HMM.**

The wave of training samples can be done by using hidden Markov model which is supported by Hmm Toolkit (HTK). The training process is done by initializing the proto model for each word and re-estimating of each models number of times, for re-estimation Baum-Welch re-estimation method is employed. The completion of training process the unknown utterance is recognized or verified by employing Viterbi decoding method. The HTK supports all these features in one package with different function in it. Internally the HTK has a number of features which are used for a huge number of applications, along with it can do file organizing, and fault reporting and memory supervision. The HTK is used in the field of signal processing, but the major work was done by using HTK in the field of Speech signal processing.

## 3.1 Training and Recognition

### *3.1.1 Data Preparation*

The entire speech samples are recorded with silent environment in the background. The recorded data are divided into training data set and testing data set. The system development is done with the training data set along with supporting labeling for them is given by the task grammar. The wave samples are recorded at a sampling frequency rate of 16K Hz in a mono channel as a .wav format, from a speaker with 10 Kannada language Digits i.e.0 to 9 (270*10= 2700). Out of 2700 wave samples, 2500 wave samples are used for training the system and remaining 200 wave trials are used for testing the system. Recording and labeling is done by Pratt and Wave Surfer respectively.

### *3.1.2 Training Process*

The training the wave samples by extracting the MFCC (Hcopy in HTK) feature and generate proto model and train with huge number samples and creating the estimated model the testing the system. Recording and labeling is done by Pratt and Wave Surfer. Create proper grammar and vocabulary to support the models created using HTK, creating the Proto

model for each word and then it is initialized and re-estimated using Balm-Welch algorithm (Hint & HRest of HTK).

### *3.1.3 Testing Process*

The Testing of utterances is by employing the Viterbi search decoding algorithm, which finds result by most likelihood ratio. The Hvite of HTK will do the decoding of input speech. The input speech sample is preprocessed and it breaks into frames and extract the mfcc feature from it using the distance formula (minimum distance).

**Table 1: Isolated digits & their recognition rate**

| Number | Uttered Word/ Text Display (Kannada) | Recognition Rate in % |
|---|---|---|
| 0 | ಸೊನ್ನೆ | 100% |
| 1 | ಒಂದು | 100% |
| 2 | ಎರಡು | 90% |
| 3 | ಮೂರು | 100% |
| 4 | ನಾಲ್ಕು | 100% |
| 5 | ಐದು | 100% |
| 6 | ಆರು | 90% |
| 7 | ಏಳು | 90% |
| 8 | ಎಂಟು | 100% |
| 9 | ಒಂಭತ್ತು | 100% |

## 4. EVALUATION OF PERFORMANCE AND RESULTS

The Fig.15 shows the Overall result of Speech recognition system. This shows that, the overall data used for training is 2500 samples, out of which 2495 samples or digits are recognized correctly. SENT says that sentence level recognition rate is 99.80% and the WORD says that Word recognition rate is also 99.80%. N is the total number data set used for training is 2500, H is the total number of data correctly recognized is 2495, S is the number of substitution error is 5, I is the insertion error and D is the deletion error. By this concluded that the system performance is 99.80% of accuracy. The system is tested with a test data set and live data samples, and system was successful

HResults -A -D -T 1 -e ??? sil -I ref.mlf hmmlist.txt rec.mlf

This command is executed in MATLAB; HResults are the supporting tool in HTK which is used to compute the performance of the system with respect to reference network of the system. The accuracy and computation time says the performance of the speech recognition system. The accuracy of the system is measured in terms of word error rate (WER) [16]. The computation time says that the time taken by the system to recognize the uttered word.

$$WER = \frac{S+D+I}{N} \quad \ldots(3)$$

1. PA = (N-D-S-I)/N *100 or H-I/N *100%
   = (2500-0-5-0)/2500 = 99.80%.
2. WER = 100% - PA =100% - 99.80 = 0.20%.

The final touch to the system is done by MATLAB i.e. The HTK is interface with the MATLAB.

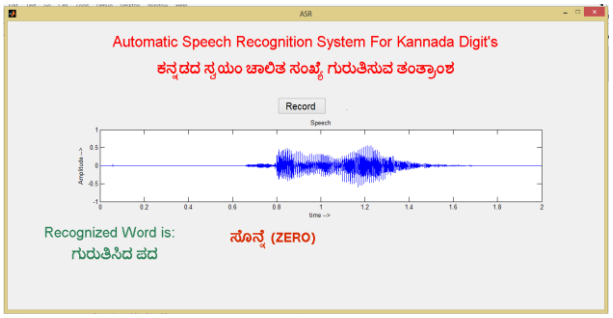


**Fig.5: GUI for the Recognition of Isolated digit 'ಸೊನ್ನೆ'.**

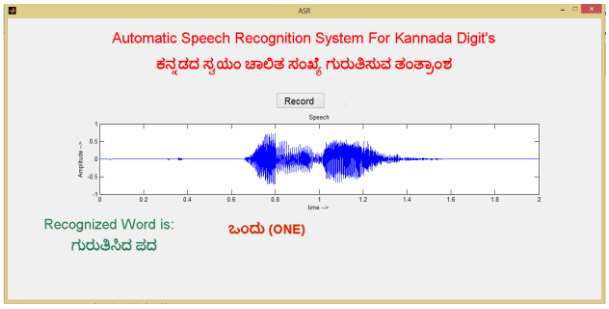


**Fig.6: GUI for the Recognition of Isolated digit 'ಒಂದು'.**

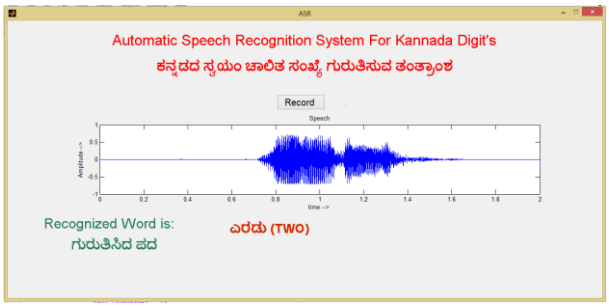


**Fig.7: GUI for the Recognition of Isolated digit 'ಎರಡು'.**

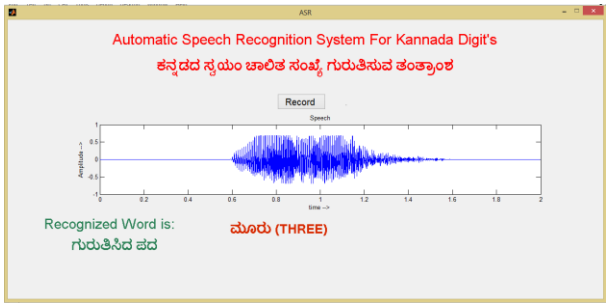


**Fig.8: GUI for the Recognition of Isolated digit 'ಮೂರು'.**

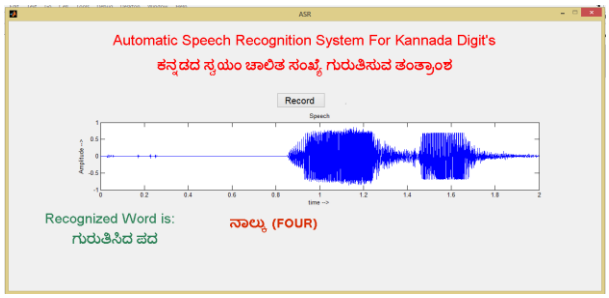


**Fig.9: GUI for the Recognition of Isolated digit 'ನಾಲ್ಕು'.**

The interfacing of HTK to the MATLB creates the easy way of communication between user and HTK. The Table 1 shows the uttered words and their recognition rate. The Fig.5 to Fig.14 shows the graphical user interface of the experimental results in implemented speech recognition system of isolated digits in the Kannada language.

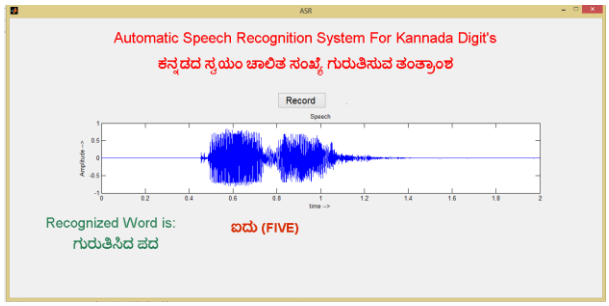


**Fig.10: GUI for the Recognition of Isolated digit 'ಐದು'.**

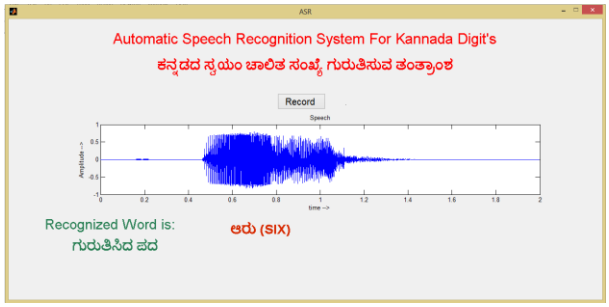


**Fig.11: GUI for the Recognition of Isolated digit 'ಆರು'.**

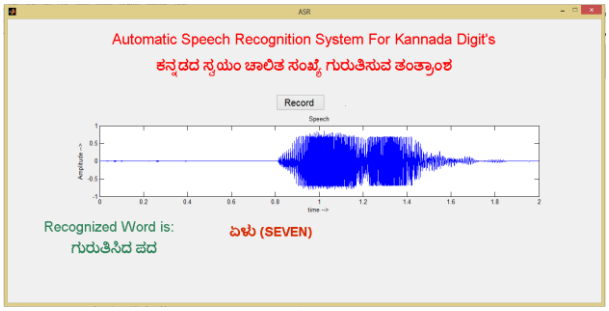


**Fig.12: GUI for the Recognition of Isolated digit 'ಏಳು'.**

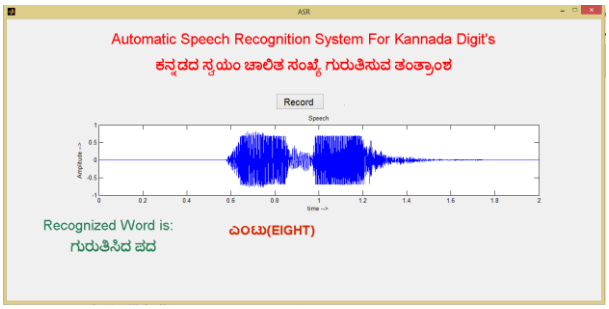


**Fig.13: GUI for the Recognition of Isolated digit 'ಎಂಟು'.**

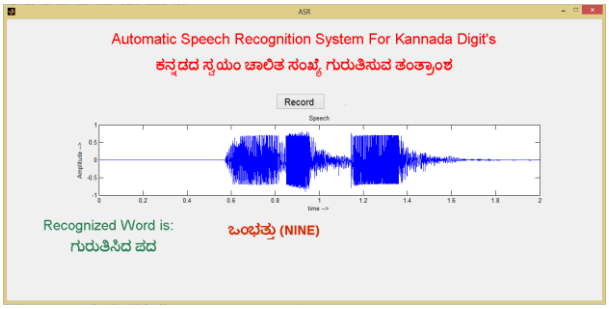


**Fig.14: GUI for the Recognition of Isolated digit 'ಒಂಭತ್ತು'.**

## 5. CONCLUSION AND FUTURE SCOPE

Speech is possible of being an effortless way of communication with computer. MFCC feature extraction method is similar like human auditory system so which gives its best result. The system that can recognize isolated digits in Kannada language was achieved successfully. To achieve most accurate system, with a huge number of vocabulary size need to collect a large amount of data samples from a different speakers.

The future enhancement of the system might be for concatenated words/digits or it can also for continuous speech recognition system. The greatest problem is Co-articulation

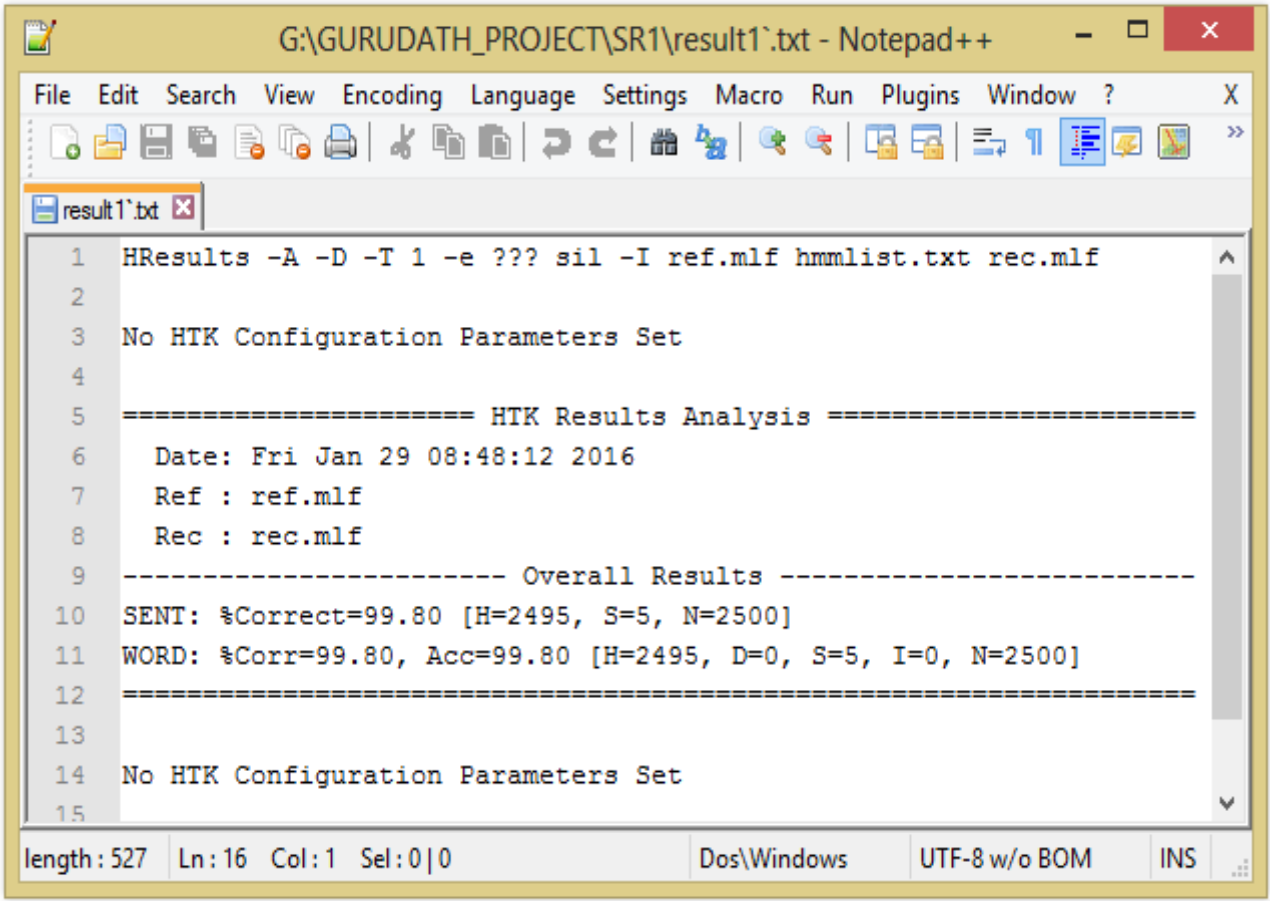


**Fig.15 Result Analysis of ASR**

problem, which reduces the accuracy of the system due to this the misrecognition happen. Need to work on this problem to overcome. The word level model is complex, if the system implementation for large vocabulary the word level will gives poor performance so in order to get good result need to define phoneme level models, which supports and reduce the problem. The current trend is to work with background noise.

The future scope gives the new direction to continue the work on those challenges.

## 6. REFERENCES

[1] Davis, K. H., R. Biddulph, and Stephen Balashek. "Automatic recognition of spoken digits." The Journal of the Acoustical Society of America 24.6 : 637-642,1952.

[2] D. B. Fry, Theoritical Aspects of Mechanical speech Recognition , and P. Denes, The design and Operation of the Mechanical Speech Recognizer at Universtiy College London, J. British Inst. Radio Engr. , 19:4,211-299,1959.

[3] Zue, V., Glass, J., Phillips, M., Seneff, S., The MIT SUMMIT speech recognition system: a progress report. In: Proc. Speech and Natural Language Workshop, Philadelphia, PA, February, pp. 179–189,1989.

[4] Lee, C. H., et al. "Acoustic modeling for large vocabulary speech recognition." Computer Speech & Language 4.2 127-165,1990.

[5] Keh-Yih Su et.al, Speech Recognition using weighted HMM and subspace IEEE Transactions on Audio, Speech and Language.

[6] R. Chengalvarayan and L. Deng, "Use of generalized dynamic feature parameters for speech recognition," IEEE Trans. Speech Audio Process., vol. 5, no. 3, pp. 232–242, 1997.

[7] M. Dua, R. K. Aggarwal, V. Kadyan, and S. Dua, "Punjabi Automatic Speech Recognition Using HTK," vol. 9, no. 4, pp. 359–364, 2012.

[8] P. Saini, P. Kaur, and M. Dua, "Hindi Automatic Speech Recognition Using HTK," vol. 4, no. June, pp. 2223–2229, 2013.

[9] P. V. Bhaskar and S. R. M. Rao, "Telugu Speech Recognition System development using MFCC based Hidden Markov Model technique with Sphinx-4," vol. 2, no. 2, pp. 141–147, 2014.

[10] Y. K. Gedam, S. S. Magare, A. C. Dabhade, and R. R. Deshmukh, "Development of Automatic Speech Recognition of Marathi Numerals - A Review," vol. 3, no. 9, pp. 198–203, 2014.

[11] S. K. Mukundan, "' Shreshta Bhasha ' Malayalam Speech Recognition using HTK," vol. 1, no. 1, pp. 1–5, 2014.

[12] C. Science and S. Engineering, "Isolated English Language Digit Recognition Using Hidden Markov Model Toolkit," vol. 4, no. 6, pp. 781–784, 2014.

[13] M. Anusuya and S. Katti, "Speech recognition by machine: A review," Int. J. Comput. Sci. Inf. Secur., vol. 6, no. 3, pp. 181–205, 2009.

[14] G. Hemakumar and P. Punitha, "Speech Recognition Technology : A Survey on Indian Languages," vol. 2, no. 4, pp. 1–38, 2013.

[15] K. Kumar and R. K. Aggarwal, "International Journal of Computing and Business Research ISSN ( Online ) : 2229-6166," Int. J. Comput. Bus. Res., vol. 2, no. 2, 2011.

[16] Gaikwad, Santosh K., Bharti W. Gawali, and Pravin Yannawar. "A review on speech recognition technique." International Journal of Computer Applications 10.3 : 16-24,2010.

[17] http://wwwlands2.let.kun.nl/members/software/HTKBook_2.0/node5.html.

[18] A. A. M. Abushariah and T. S. Gunawan, "English Digits Speech Recognition System Based on Hidden Markov Models," no. May, pp. 11–13, 2010.

[19] H. Muralikrishna, T. Ananthkrishna, Dr. Kumara Shama "HMM Based Isolated Kannada Digit Recognition System using MFCC," pp. 730–733, 2013.