# Marathi Isolated Digit Recognition System using HTK

Devyani S. Kulkarni[1], Ratnadeep R. Deshmukh[1], Vandana L. Jadhav Patil[2], Pukhraj P. Shrishrimal[1] ,Swapnil D. Waghmare[1], Aaron M. Oirere[1]
[1]Dept of CS and IT, Dr. B.A.M.U, Aurangabad,
[2]Deogiri College, Aurangabad, MS-India

## ABSTRACT

This paper proposes a system of isolated digit recognition for Marathi language using HTK approach. The database used contains 800 utterances of 40 individuals. Among them 20 are female and 20 are male. For training the acoustic features of the database powerful MFCC i.e. Mel frequency cepstral coefficients technique is used. We have used word level model to recognize the Marathi isolated digits. The result analysis of the system shows 99.75% recognition with 48.75% accuracy.

## General Terms

Hidden Markov Model Toolkit, Mel Frequency Cepstral Coefficient

## Keywords

HTK, HMM, Acoustic model, digit recognition, Grammar, MFCC

## 1. INTRODUCTION

Speech recognition is such a field of computer science that deals with designing a computer system that recognizes the words spoken by humans [1]. There are so many approaches for classification and recognition which is based on variety of techniques like template based approach, statistical approach, learning approach, knowledge based approach, artificial intelligence approach etc. among these one of the most successful technique is use of algorithms based on Hidden-Markov Model. We have used the same for building Marathi isolated digit recognition system [2].

### Marathi Language

Marathi which is an Indo-Aryan language is spoken by the Marathi people who live in western and central part of India. Marathi language is similar to National language Hindi. Both the languages are been derived from Sanskrit and uses the Devanagari script for writing. Marathi is having fourth largest number of native speakers in India as Marathi is spoken in the Complete Maharashtra state which covers a vast geographical area which consists of 34 different districts. Standard Marathi is the official language of state of Maharashtra [3].

This paper is organized as follows. Section 2 gives the literature review of speech recognition systems developed in Indian languages using HTK; Section 3 puts light on what is HTK (Hidden Markov Model) Toolkit; Section 4 describes the details of the speech database used for the research work; Section 5 explains the Acoustical Analysis, Section 6 describes the Training Phase; Section 7 explains how the recognition was carried out; Section 8 explains the performance analysis carried out for the developed system, Section 9 gives the conclusion and future work of the developed system.

## 2. RELATED WORK

Jitendra Singh Pokhariya et al. (2014) in their paper presented a work done for building a speech recognition system for Sanskrit language. The system was trained for recognizing 50 Sanskrit words. The database used was developed by taking speech samples from 10 speakers. The system showed the overall accuracy with 5 states of HMM topology as 95.2% and for 10 states 97.2%.

Patil A. S. (2014) in his paper presented the implementation of HMM based speaker independent isolated word speech recognition system for Ahirani language. The particular system was trained with 20 Ahirani words. The database contains data recorded from 10 speakers. The system was tested by using data collected from another 10 speakers. The recording of database was done in room environment. The system has given 94 % accuracy.

Shweta Tripathy et al (2013) in their paper have presented their work about developing the speech recognition system for Hindi language. For feature extraction various techniques like MFCC (Mel Frequency Cepstral Coefficient), LPC (Linear Predictive Coding) were used and HMM (Hidden Markov Model) was used as the classifier.

Annu Choudhary et al (2013) have discussed the work of implementing the Speech Recognition system (ASR) for isolated words and connected words of Hindi language and working of HTK i.e. Hidden Markov Model Tool which is based on Hidden Markov Model (HMM) The system was trained for 100 distinct Hindi words initially. After testing the recognition results the system had shown the overall accuracy 95% and 90% for isolated words and connected words respectively.

Babita Saxena et al (2015) in their paper presented a baseline digits speech recognizer for Hindi language. For the database they had collected the speech samples in different environment, so the recordings contain various noises like vehicle horns, door opening etc. After that all these audio recorded data taken from 8 speakers was used to train the acoustic model. The vocabulary size of the recognizer was 10 words. Authors had used HTK toolkit for building acoustic model and evaluating the recognition rate of the recognizer. The efficiency of the recognizer developed on recorded data, is shown at the end of the paper and possible directions for future research work are suggested.

Kuldeep Kumar et al (2011) have developed a speech recognition system for Hindi language. They have used Hidden Markov Model Toolkit (HTK) for developing the system. The system recognizes the isolated words with the help of acoustic model. The system was trained for 30 Hindi words. Training data had been collected from eight speakers. The experimental results of the developed system showed the overall accuracy 94.63%.

Sharmila et al (2012) in their paper describes the making of a speech recognition system for Hindi language recognition with Hidden Markov Model Toolkit (HTK). HTK recognizes the isolated digits using acoustic digits model. They have trained the system with 10 Hindi digits. The data used for

Training data was collected from twenty four speakers. The system gives good accuracy in the range 93 - 100% [4].

## 3. HTK

HTK is a software toolkit which is used for building and manipulating systems that uses Hidden Markov Models which was developed by the Speech Group of Cambridge University Engineering Department. HTK software includes a software library as well as a number of tools (programs) which performs tasks such as coding data, various styles of HMM training including embedded Baum-Welch re-estimation, Viterbi decoding, results analysis and editing of HMM definitions[5]. Primary use of HTK is for speech recognition research. HTK has also been used for number of other applications which includes research into speech synthesis, character recognition and DNA sequencing etc.

There are 4 important stages of processing in HTK which includes Data Preparation, Training, Testing/ Recognition and Analysis.

## 4. DATABASE

The database selected contains text corpus consists of 10 Marathi Isolated Digits. The table below shows the word set with their English Pronunciation. The database contains speech samples of 40 speakers from Aurangabad district out of which 20 are Male and 20 are Female in Noisy environment. The speakers were asked to speak the 10 Digits with 3 utterances of each word. Total 30 speech samples from each speaker were collected. The database contains total 1200 utterances in all of the 10 Digits from 40 speakers.

Database contains the following isolated Marathi Digits:

**Table 1 Speech Database**

| | | | |
|---|---|---|---|
| **Zero** | Shunya | **Five** | Pach |
| **One** | Ek | **Six** | Saha |
| **Two** | Don | **Seven** | Sat |
| **Three** | Tin | **Eight** | Aatha |
| **Four** | Char | **Nine** | Nau |

For developing the Database PRAAT software and SENNHEISER PC 360 headsets was used for recording the speech. PC 360 headsets are having noise cancellation facility and the signal to Noise Ratio (SNR) is less. The sampling frequency was set to 16 KHz with 16 bit in Mono sound type.

## 5. ACOUSTICAL ANALYSIS:

In Acoustical Analysis, the speech files obtained are represented in more compact and efficient way by extracting features of speech files. The system uses the Mel Frequency Cepstral Coefficients (MFCCs) to extract features from speech files. It is the well known and most widely used feature extraction method in speech domain. In this stage, the acoustic signal is converted into a sequence of acoustic feature vectors. Figure 1 illustrates the different stages that take place in the feature extraction process.

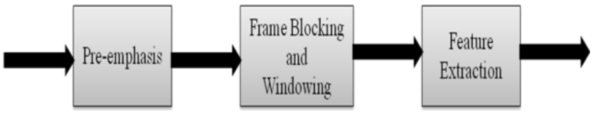


**Figure 1: General feature extraction process**

In the MFCC computation, the HTK tool HCopy and the configuration file play an important role in the parameterization of the speech signal into a sequence of feature vectors. HCopy parameterizes the source speech data according to the configuration file, and copies the target speech data into the output file. This is schematized in Figure 2.

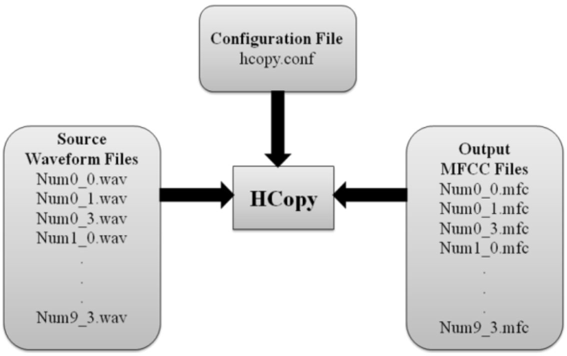


**Figure 2: Parameterization of the speech data by HCopy; and list of the source waveform files and its corresponding MFCC files generated**

The whole speech signal processing and the characteristics of the generative model are controlled by configuration parameters. The MFCC extraction process of Figure 16 will be followed in the description of the most important configuration parameters. Table 2 shows the setting of those configuration parameters related to such MFCC extraction process.

**Table 2: Configuration parameters related to MFCC extraction process**

| Configuration Parameters | Value |
|---|---|
| SOURCEKIND | WAVEFORMAT |
| SOURCEFORMAT | WAV |
| TARGETKIND | MFCC_0_D_A |
| TARGETRATE | 100000.0 |
| WINDOWSIZE | 250000.0 |
| USEHAMMING | T |
| SAVECOMPRESSED | T |
| SAVEWITHCRC | T |
| PREEMCOEF | 0.97 |
| NUMCHANS | 26 |
| NUMCEPS | 12 |

## 6. TRAINING PHASE

For a fast and precise convergence of the training algorithm the HMM parameters must be initialized with the training data corpus. We have done this initialization with the tool HInit. This command line initializes the HMM by time-alignment of the training data with a Viterbi algorithm. The initialisation process by HInit tool is can be showed by following diagram.

The HCompv tool performs a “flat” initialisation of a model. Every state of the HMM is given the same mean and variance vectors: these are computed globally on the whole training corpus. The initialisation process by HInit tool is can be showed by following Figure 3.

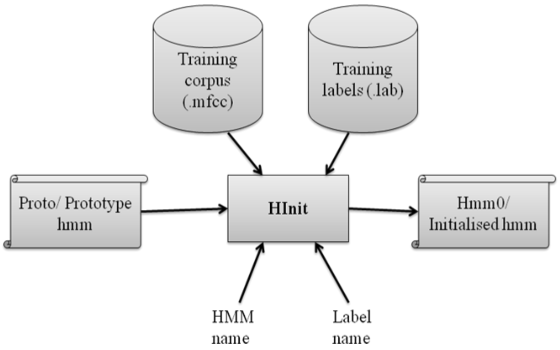


**Figure 3: Initialsation from a prototype**

Training procedure is done with HTK tool HRest. This procedure has to be repeated several times for each of the HMM to train. Each time, the HRest iterations (i.e. iterations within the current re-estimation iteration…) are displayed on screen, indicating the convergence through the change measure. As soon as this measure do not decrease (in absolute value) from one HRest iteration to another, it's time to stop the process. In our recognizer we did it for 5 times.

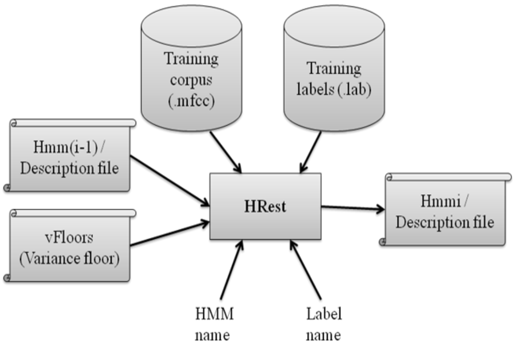


**Figure 4: Recognition process of an input signal**

## 7. RECOGNITION

The diagram follow describe the recognition procedure: Recognition procedure:

An input speech signal input.sig is first transformed into a series of "acoustical vectors" (here MFCCs) with tool HCopy, in the same way as what was done with the training data (Acoustical Analysis step). The result is stored in an input.mfcc file (often called the acoustical observation). - The input observation is then process by a Viterbi algorithm, which matches it against the recognizer's Markov models. This is done by tool HVite: (one file at a time) [6].

## 8. PERFORMANCE ANALYSIS

The percentage correct is calculated with the help of following equation

**Percent correct =** $\dfrac{N-D-S}{S}\times 100\%$ **(1)**

And

**Percentage Accuracy=** $\dfrac{N-D-S-I}{S}\times 100\%$ **(2)**

Where,

S= the number of substitution errors

D= deletion errors

I=insertion errors

N= is the total number of labels in the reference transcriptions.

**WER = 100% - Percentage of word Accuracy ------------ (3)**

Equation (1) is used to calculate the sentence correction rate and the word correction rate. Equation (2) is used to calculate the word accuracy. Finally, equation (3) is used to calculate the word error rate [7].

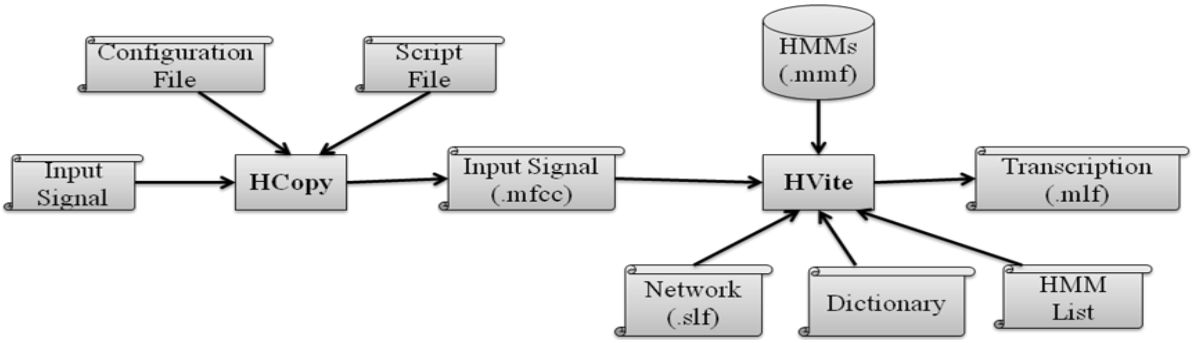
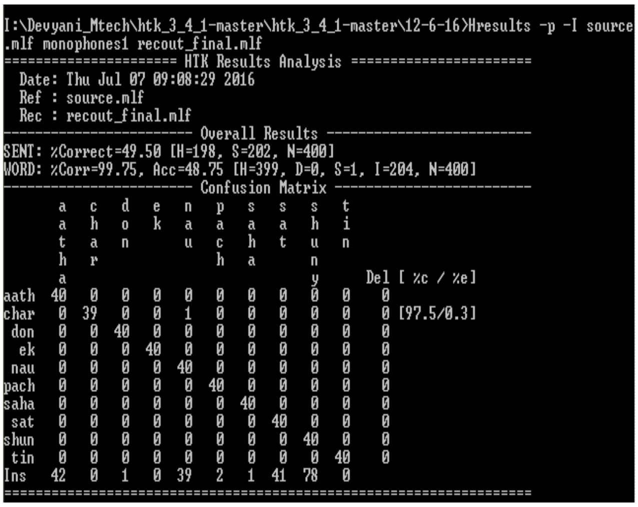


**Figure 5: Parameterization of the speech data by HCopy; and list of the source waveform files and its corresponding MFCC files generated**

The system was trained with 800 utterances spoken by 20 males and 20 females. We have tested the system in 3 different scenarios 1st by taking overall samples, 2nd by taking only female samples and 3rd by taking only male samples.

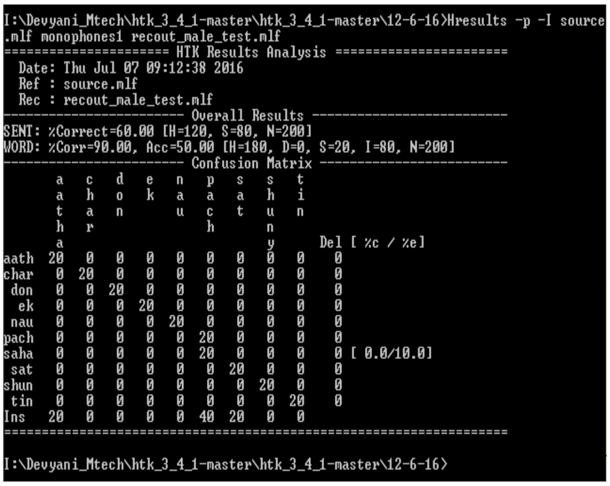
The system was tested with 400 test samples out of which 200 were female samples and 200 were male samples. It is observed that HTK provides overall 99.75 % recognition at word level. The statistics of result analysis provided by HTK is given below.



```
I:\Devyani_Mtech\htk_3_4_1-master\htk_3_4_1-master\12-6-16>Hresults -p -I source
.mlf monophones1 recout_final.mlf
====================== HTK Results Analysis =======================
  Date: Thu Jul 07 09:08:29 2016
  Ref : source.mlf
  Rec : recout_final.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=49.50 [H=198, S=202, N=400]
WORD: %Corr=99.75, Acc=48.75 [H=399, D=0, S=1, I=204, N=400]
------------------------ Confusion Matrix -------------------------
       a   c   d   e   n   p   s   s   s   t
       a   h   o   k   a   a   a   a   h   i
       t   a   n       u   c   h   t   u   n
       h   r               h   a       n
       a                               y     Del [ %c / %e]
aath  40   0   0   0   0   0   0   0   0   0    0
char   0  39   0   0   1   0   0   0   0   0    0 [97.5/0.3]
 don   0   0  40   0   0   0   0   0   0   0    0
  ek   0   0   0  40   0   0   0   0   0   0    0
 nau   0   0   0   0  40   0   0   0   0   0    0
pach   0   0   0   0   0  40   0   0   0   0    0
saha   0   0   0   0   0   0  40   0   0   0    0
 sat   0   0   0   0   0   0   0  40   0   0    0
shun   0   0   0   0   0   0   0   0  40   0    0
 tin   0   0   0   0   0   0   0   0   0  40    0
Ins   42   0   1   0  39   2   1  41  78   0
===================================================================
```

**Figure 6: Final output of overall test samples**

And also tested by taking only male samples. We have taken 200 utterances taken by 20 males. Then we got 90.00 % recognition rate. As shown below in Figure 7. The statistics of result analysis provided by HTK is given below.

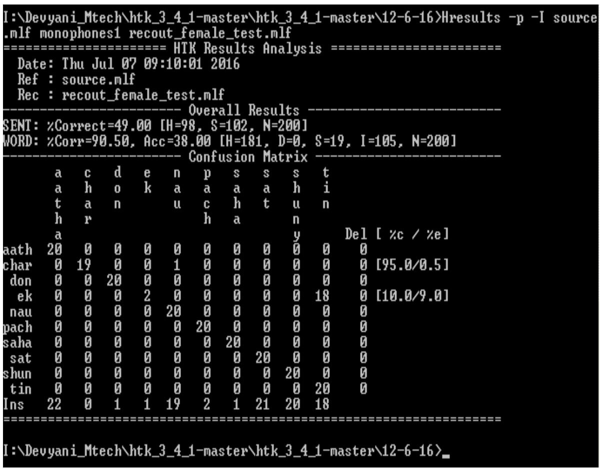


```
I:\Devyani_Mtech\htk_3_4_1-master\htk_3_4_1-master\12-6-16>Hresults -p -I source
.mlf monophones1 recout_male_test.mlf
====================== HTK Results Analysis =======================
  Date: Thu Jul 07 09:12:38 2016
  Ref : source.mlf
  Rec : recout_male_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=60.00 [H=120, S=80, N=200]
WORD: %Corr=90.00, Acc=50.00 [H=180, D=0, S=20, I=80, N=200]
------------------------ Confusion Matrix -------------------------
       a   c   d   e   n   p   s   s   t
       a   h   o   k   a   a   a   h   i
       t   a   n       u   c   t   u   n
       h   r               h       n
       a                           y     Del [ %c / %e]
aath  20   0   0   0   0   0   0   0   0    0
char   0  20   0   0   0   0   0   0   0    0
 don   0   0  20   0   0   0   0   0   0    0
  ek   0   0   0  20   0   0   0   0   0    0
 nau   0   0   0   0  20   0   0   0   0    0
pach   0   0   0   0   0  20   0   0   0    0
saha   0   0   0   0   0  20   0   0   0    0 [ 0.0/10.0]
 sat   0   0   0   0   0   0  20   0   0    0
shun   0   0   0   0   0   0   0  20   0    0
 tin   0   0   0   0   0   0   0   0  20    0
Ins   20   0   0   0   0  40  20   0   0
===================================================================

I:\Devyani_Mtech\htk_3_4_1-master\htk_3_4_1-master\12-6-16>
```

**Figure 7: Final output of male test samples**

We have also tested by taking only female samples. We have taken 200 utterances taken by 20 females. Then we got 90.50 % recognition rate. As shown below in Figure 8. The statistics of result analysis provided by HTK is given below



```
I:\Devyani_Mtech\htk_3_4_1-master\htk_3_4_1-master\12-6-16>Hresults -p -I source
.mlf monophones1 recout_female_test.mlf
====================== HTK Results Analysis =======================
  Date: Thu Jul 07 09:10:01 2016
  Ref : source.mlf
  Rec : recout_female_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=49.00 [H=98, S=102, N=200]
WORD: %Corr=90.50, Acc=38.00 [H=181, D=0, S=19, I=105, N=200]
------------------------ Confusion Matrix -------------------------
       a   c   d   e   n   p   s   s   s   t
       a   h   o   k   a   a   a   a   h   i
       t   a   n       u   c   h   t   u   n
       h   r               h   a       n
       a                               y     Del [ %c / %e]
aath  20   0   0   0   0   0   0   0   0   0    0
char   0  19   0   0   1   0   0   0   0   0    0 [95.0/0.5]
 don   0   0  20   0   0   0   0   0   0   0    0
  ek   0   0   0   2   0   0   0   0   0  18    0 [10.0/9.0]
 nau   0   0   0   0  20   0   0   0   0   0    0
pach   0   0   0   0   0  20   0   0   0   0    0
saha   0   0   0   0   0   0  20   0   0   0    0
 sat   0   0   0   0   0   0   0  20   0   0    0
shun   0   0   0   0   0   0   0   0  20   0    0
 tin   0   0   0   0   0   0   0   0   0  20    0
Ins   22   0   1   1  19   2   1  21  20  18
===================================================================

I:\Devyani_Mtech\htk_3_4_1-master\htk_3_4_1-master\12-6-16>_
```

**Figure 8: Final output of female test samples**

## 9. CONCLUSION AND FUTURE WORK

In a country like India, there is huge possibility to use speech recognition as a communication medium with machine. By using speech as an interface between human and machine, each person is able to operate machine easily. This work is a step towards the development of such type of systems. We have done the implementation using word level model for Marathi Isolated Digit Recognition. The proposed Marathi isolated digit recognition system can be further extended for speaker independent recognition. The vocabulary size can be extended. Also we can go for connected and continuous speech system. The result analysis of the system shows 99.75% recognition with 48.75% accuracy.

## 10. ACKNOWLEDGMENT

This work is supported by University Grants Commission under the scheme Major Research Project entitled as "Development of Database and Automatic Recognition System for Continuous Marathi Spoken Language for agriculture purpose in Marathwada Region". The authors would also like to thank the Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing the infrastructure to carry out the research.

## 11. REFERENCES

[1] V. B. Waghmare, R. R. Deshmukh, P. P. Shrishrimal and G. B. Janvale “Development of Isolated Marathi Words Emotional Speech Database” International Journal of Computer Applications 94(4):19-22, May 2014.

[2] Devyani S. Kulkarni, Ratnadeep R. Deshmukh, and Pukhraj P. Shrishrimal. "A Review of Speech Signal Enhancement Techniques." International Journal of Computer Applications 139.14 (2016).

[3] Pukhraj P. Shrishrimal, Ratnadeep R. Deshmukh and Vishal B. Waghmare “Indian Language Speech Database: A Review” .International Journal of Computer Applications 47(5):17-21, June 2012.

[4] Devyani S. Kulkarni, "HTK Based Speech Recognition Systems for Indian Regional languages: A Review", International Research Journal of Engineering and Technology (IRJET), Volume 03 Issue 06, June-2016.

[5] P. C. Woodland et al “Large Vocabulary Continuous Speech Recognition Using HTK”, Acoustics, Speech and Signal Processing, 1994 ICASSP-1994.

[6] D. Jurafsky and J. H. Martin, “HMMs and Speech Recognition,” in Speech and Language Processing, S. Russell and P. Norving, Dorling Kindersley Pvt. Ltd., India, 2000, pp. 261-309.

[7] The HTK Book. By Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev and Phil Woodland.