Hindi Number Recognition using GMM

Himanshu Rai Goyal
School of Computing
Graphic Era University,
Dehradun -248002, Uttarakhand, India.

Shashidhar G.Koolagudi
School of Computing
Graphic Era University,
Dehradun -248002, Uttarakhand, India

Abstract

This paper aims at designing and implementation of Hindi number recognition system using the microphone and mobile recorded speech. Spectral features known to represent phonetic information are used as the features to characterize different Hindi digits. Gaussian mixture models (GMM) are used to develop the digit recognition system. This paper focuses on the ten basic Hindi digits where '0' is pronounced as 'shunya' to '9' is pronounced as 'no'. Data has been collected separately from male, female and child speakers using microphone and mobile phone device. The experimental results show that the overall accuracy of digit recognition is 98.9\% in the case of microphone recorded speech and 96.4\% in the case of mobile phone recorded speech.

Keywords-- Gaussian mixture models (GMM), Mel frequency cepstral coefficients (MFCC), Hindi digit microphone database (HDMD), Hindi digit telephonic database (HDTD).

1. INTRODUCTION

Speech recognition is the process of transforming human speech into equivalent text. Automatic speech recognition (ASR) is defined as the process of interpreting human speech as a sequence of words by computer [6]. ASR basically deals with building a system for mapping acoustic signals to a string of words [1]. In general, all ASR systems aim to automatically extract the string of spoken words from the given input speech signals as illustrated in Figure 1

During the past decade, research in automatic speech recognition has produced increasingly better results for a wide range of tasks. At the same time, the evolution of telephone and the widespread use of telephones are urging telephony service providers to apply speech technology [2]. Application such as mobile banking system where person authentication can be granted based on the digits he pronounces over a phone, voice dialing and so on. This system may also be helpful in interactive voice response system (IVRS) where physically challenged and illiterate people may access the information regarding agricultural needs, weather etc over phone. This kind of application demands a high level of accuracy in order to be of any use. However, speaker independent recognition of telephone speech is more difficult than clean speech recognition because besides speaker variability problems, we also have to deal with a potentially very large variability in channels and microphones, with many different kinds of channel and environmental noise [10][11].

Figure1. A Generic Speech Recognition System

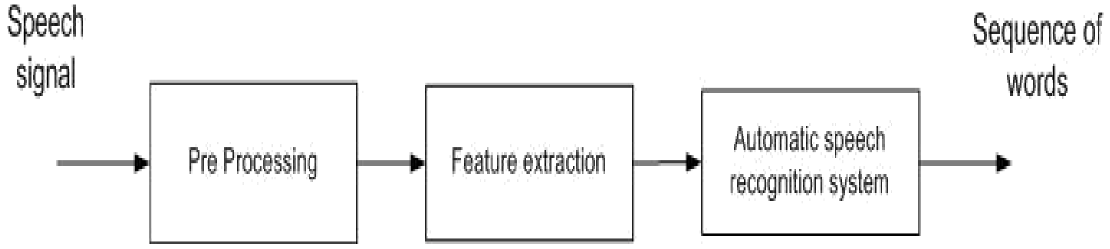

TABLE I
Hindi digits and their pronunciation

Hindi Digit	Hindi Pronunciation
0	Shunya
1	Eyk
2	Dho
3	Teen
4	chaaR
5	Paanch
6	Chey
7	Saath
8	Aat
9	No

In the past decade, much works have been done in the field of speech recognition for Hindi language. Tarun Pruthi *et al*. describe a speaker-dependent, isolated word recognizer for Hindi. Features are extracted using LPC and recognition is carried out using HMM. System was designed for two male speakers only. However the system is giving good performance, but the design is speaker specific and uses very small vocabulary [7]. An Isolated word speech recognition tool for Hindi language is designed by Gupta using continuous HMM. The system uses word acoustic model for recognition. Again the word vocabulary contains Hindi digits. Recognizer gives good results when tested for sound used for training the model. For other sounds too, the results are satisfactory. System is highly efficient but vocabulary size is too small [3]. To overcome from this problem kuldeep and aggarwal was tried by using a vocabulary size of 30 words. They got good results but it is again not gender and region specific [5].

The main objective of this paper is to design Hindi digit recognition system from spoken digit using GMM models. Spectral features are extracted from the speech signals of the digits. Using these features Gaussian mixture models are trained. During validation the same features from the speech signal of unknown digit are extracted and given as input to the trained models. Based on the probability of feature vector belonging to the model, classification of input feature vector into one of the digit categories is done.

2. DATABASE

The database of isolated Hindi digit has been collected from two perspectives, one using microphone (HDMD) and other using mobile (HDTD) handset. Thirty five speakers containing, fifteen male, fifteen female and five children from different age groups and origin have given their voice during recording of the database. Isolated Hindi digits from 0 to 9 are recorded from all the speakers in multiple sessions. Every speaker pronounces each digit fifteen times. Therefore there are 35 *speakers* X 10 *numbers* X 15 *sessions* = 5250 utterances with is recorded with the short pause between the utterances. The speakers are chosen from different background so that different pronunciation patterns of Hindi digits are captured. The waveforms recorded continuously from different speakers are manually segmented and stored separately based on digit and speakers. Database is recorded with the sampling frequency of 16 KHz and 16 bits per sample. Speakers are chosen from different states of India from age group of 7 - 60 year. A distance of 2 - 3 inches was maintained between microphone and mobile from speaker at the time of recording. Mobile phone made by Nokia (given specification) and microphone made by Intex (given specification) used for recording the database. Hindi digits and their patterns pronunciation are given in Table I

TABLE II
Average digit recognition performance(Speaker: Child, Database: Mobile(in %))

	shunya	eyk	dho	Teen	chaaR	paanch	chey	saath	aat	No
shunya	47	0	0	0	33	0	0	20	0	0
eyk	0	67	0	0	0	0	33	0	0	0
dho	0	0	100	0	0	0	0	0	0	0
teen	0	0	0	87	13	0	0	0	0	0
chaaR	0	0	0	0	80	0	0	20	0	0
paanch	0	0	0	0	74	13	0	0	13	0
chey	7	0	0	0	7	33	53	0	0	0
saath	0	0	0	0	40	7	0	27	27	0
Aat	0	0	0	0	20	0	0	47	33	0
no	0	0	0	0	0	47	0	20	0	33

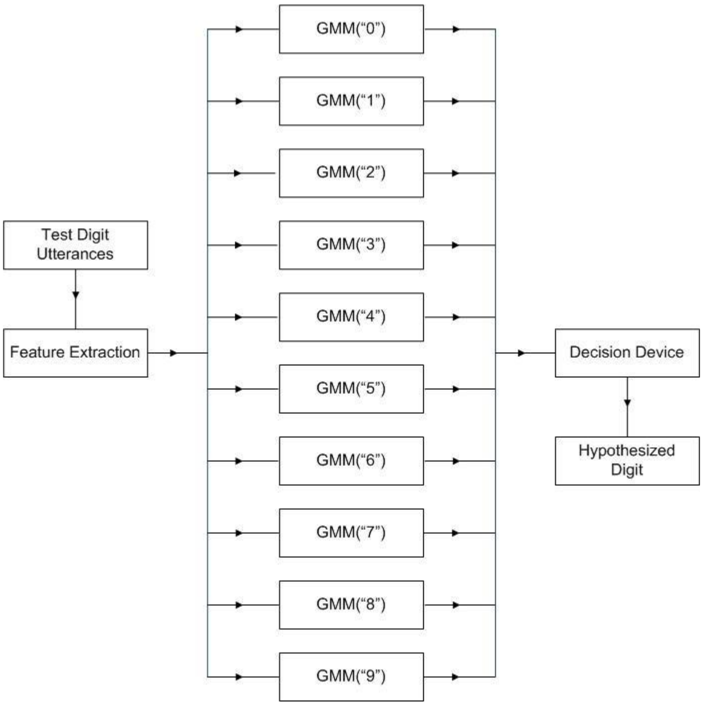

Figure2. GMM model block diagram of digit recognition

TABLE III
Average digit recognition performance (Speaker: Female, Database: Mobile (in %))

	shunya	eyk	dho	teen	chaaR	paanch	chey	saath	aat	no
shunya	100	0	0	0	0	0	0	0	0	0
eyk	0	89	0	4	0	0	7	0	0	0
dho	0	0	100	0	0	0	0	0	0	0
teen	24	0	0	76	0	0	0	0	0	0
chaaR	0	0	0	0	98	2	0	0	0	0
paanch	7	0	0	0	13	64	0	16	0	0
chey	0	36	0	2	0	0	62	0	0	0
saath	0	0	0	0	38	0	0	49	13	0
Aat	0	0	0	0	24	20	0	9	47	0
no	0	0	38	0	0	13	0	0	0	49

Digit utterances are recorded basically in two 1) Plain speech: speech utterance directly recorded from microphone and 2) Mobile speech: speech utterance recorded from the conversations through mobile phone handsets.

3. FEATURE EXTRACTION

In this work for digit recognition mel frequency cepstral coefficients (MFCCs) are used. From the literature, it is known that spectral features properly capture phonetic information [11]. Around 10-13 feature are sufficient to represent four formants required to characterize major Hindi phones. Out of various spectral features mel frequency cepstral coefficients (MFCCs) are widely used as they represent nonlinear hearing mechanism of human beings. The Mel scale is nonlinear logarithmic scale resembling the way that the human ears process sound compensation of mel scale frequencies is performed using the Equation below.

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

Steps performed to obtain MFCCs from speech signal are as follows:

1) Fourier transform of a given vowel signal is obtained to get spectrum.

2)Powers of the above spectrum within the triangular overlapping windows are computed according to mel scale.

3) Logs of the power at each of the mel frequencies are calculated.

4) DCT (Discrete cosine transform) of the list of mel log powers is calculated.

5) The amplitudes of resulting spectrum give the MFCCs [4] [9] [8]

4. Development of digit Recognition model

In this work, Gaussian mixture models are used as classifiers digit recognition with the help of spectral feature. Ten GMMs are developed to capture characteristics of 10 digits from 'shunya' to 'no'. Each mixture contains 32 components. The number of Gausses present in the mixture model is known as the number of components. They give number of clusters in which data points are to be classified within each class. The components within each GMM capture finer level details among the feature vectors of each digit. Depending on the number of data points, number of components may be varied in each GMM. Presence of few components in GMM and trained using large number of data points may lead to more generalized clusters, failing to capture specific details related to each class. On the other hand over fitting of the data points may happen, if too many components represent few data points. The complexity of the models increases, if they contain higher number of components [13]. Digit recognition using pattern classifier is basically a two stage process. In the first stage vowel recognition models are developed by training the models using feature vectors extracted from the speech utterances of known vowels. This stage is known as supervised learning. In the next stage, testing (evaluation) of the trained models is performed by using the speech utterances of unknown vowels. Scheme of categorizing the utterances of unknown digits shown in figure 2. The feature vectors obtained from unknown speech utterances are given to all trained GMM models to compute the probability of particular feature vector belonging to the specific model. Based on the sum of these probabilities for all feature vectors the category of unknown digit is decided.

5. Experimental Results and Discussion

The test results are shown in the form of confusion matrix in various tables. Table II, Table III and Table IV shows the results of the digit identification performance of child spoken utterances. In this case, the GMM models are trained with child utterances and tested with the utterances of different children’s utterances those are not used in training. The numbers shown in the diagonal of the given confusion matrix show correct classification and the miss classification percentages are shown in other positions. For instance as shown in Table II 47\% of 'shunya' utterances are recognized correctly where as 33\% and 20\% of the same utterances are recognized as 'chaaR' and 'saath' respectively.

Similarly Table III and Table IV show the digit recognition performance of female and male digit utterances respectively. The average digit recognition independently in the cases of child, female and male are about 54\%, 73\% and 69\% respectively.

In this work random data is selected from HDTD, different sets of utterances are used for training and testing. Each utterance of a digit is preprocessed before using it for feature extraction. In this process silence region has been removed from the signal. For evaluating the performance of the speech recognition systems, the feature vectors derived from the test utterances are given as input to all GMM models. GMMs use expectation maximization (EM) algorithm for finding maximum likelihood estimates of parameters in probabilistic models. The output of the each model is given to decision logic. Decision logic determines the digit, based on the highest score among the evidence provided by the recognition models [12].

In all confusion matrices we observe that dho('2') is recognized with 100\%. Maximum miss

TABLE IV
Average digit recognition performance(Speaker: Male, Database: Mobile(in %))

	shunya	eyk	dho	teen	chaaR	paanch	chey	saath	aat	no
shunya	87	2	0	11	0	0	0	0	0	0
eyk	11	65	0	24	0	0	0	0	0	0
dho	0	0	100	0	0	0	0	0	0	0
teen	0	0	0	100	0	0	0	0	0	0
chaaR	2	0	0	0	45	20	0	29	4	0
paanch	0	0	0	0	0	93	0	7	0	0
chey	16	11	0	0	24	0	49	0	0	0
saath	0	0	0	0	4	31	0	52	13	0
Aat	0	0	0	0	2	22	0	7	69	0
no	0	0	67	0	0	0	0	0	0	33

TABLE V

Overall digit recognition performance in the case of Male, Female and Child speaker. LEGEND: C-Child, F-Female, HDMD-Hindi Digit Microphone Database, HDTH-Hindi Digit Telephonic Database (in %)

	HDMD		HDTD		Training with HDTD and Testing with HDMD	Training with HDMD and Testing with HDTD
	Dependent	Independent	Dependent	Independent		
Child(C)	97.60	54.50	92.00	54.00	40.00	49.34
Female(F)	98.80	63.51	99.11	73.33	42.89	50.89
Male(M)	99.46	86.18	99.11	69.11	46.00	43.78
C+F+M	98.68	78.88	96.48	68.28	42.19	57.81

TABLE VI

Overall Digit Recognition Performance (Training: Child and Testing: Individually by Male or Female (in %))

	Training with HDMD(C)		Training with HDTD(C)	
	Dependent	Independent	Dependent	Independent
Male(M)	20.45	23.64	33.11	24.89
Female(F)	58.45	71.86	50.00	58.89

TABLE VII

Overall Digit Recognition Performance (Training: Female and Testing: Individually by Male or Child (in %))

	Training with HDMD(F)		Training with HDTD(F)	
	Dependent	Independent	Dependent	Independent
Male(M)	34.44	36.00	43.78	35.78
Child(C)	57.33	38.00	60.00	70.00

TABLE VIII

Overall Digit Recognition Performance (Training: Male and Testing: Individually by Child or Female (in %))

	Training with HDMD(M)		Training with HDTD(M)	
	Dependent	Independent	Dependent	Independent
Child(C)	20.00	13.00	35.33	16.67
Female(F)	30.22	31.30	50.22	48.44

TABLE IX

Overall Digit Recognition Performance (Training: Child + Male + Female and Testing: Individually by Child, Male or Female (in %))

	Training with HDMD(C+M+F)		Training with HDTD(C+M+F)	
	Dependent	Independent	Dependent	Independent
Child(C)	96.67	57.00	88.67	68.67
Female(F)	98.00	73.15	98.45	78.23
Male(M)	99.78	88.90	95.78	70.28

classifications are with paanch('5'), saath('7') chaaR('4') and aat('8') because all the pronunciation pattern are influenced by the vowel 'aa'. Similar studies have also been done with mobile recorded speech. The consolidation of result is shown in Table V.

It may be noted that when speaker is common for both training and testing (Dependent) the recognition results are around 99\% however if different speakers are used for training and testing (Independent) the digit recognition performance drops to 70\%. Obviously the digit recognition performance is high in the case of speaker dependent data irrespective of either HDMD or HDTD. In the case of children, digit recognition is poor as the pronunciation pattern highly varies with respect to individual child. Another observation is also true that the vocal tract of children is not properly developed. If we use both(HDMD and HDTD) databases for training and testing then results drastically drop to 50\%. This may be because of several reasons such as speaker independence, gender independence, coding schemes use in the cases of mobile speech and so on. The experiments of Hindi digit recognition are extended to study the effect of speaker, gender and age on the performance. GMMs are trained individually with male, female and child speakers' utterances are used. For example the models trained with male utterances are tested with female and child utterances. The aim of these experiments is to nullify the effect of gender and age on the digit recognition performance. Table VI VII and VIII show the results obtained. The title of the table is self explanatory. The results indicate that the prosodic parameter like pitch has influenced the recognition performance. In all 3 tables mostly child **and** female speakers' utterances have been identified similar as they are high pitch utterances. It is surprising to know that, the digits recorded through mobile devices are recognized with considerably better rate. This trend is clearly seen in the case of models trained with male and female utterances. This shows that there may not be significant effect of coding on digit recognition. Table IX shows the digit recognition results of the GMM models trained with combined utterance of male, female and child speakers. Testing of these models is done individually with male, female and child utterances. The results show a general tread of very high recognition in the case of class dependencies.

6. Summary and conclusion

In this paper digit recognition is done using GMM and 15 MFCCs as features. Two types of data recorded using microphone and mobile telephone are compared. The average recognition in the case of microphone recorded speech is higher than mobile recorded speech this may be due to coding loss. A complete study of telephone recorded data is required to be carried out.

7. References

[1] Jurafsky D. and Martin J.H. *Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Sad- dle River, NJ, USA, 2000.

[2] D. Johnston et al. Current and experimental applications of speech technology for telecom services in europe. In *Speech Communication 23*, pages 5–6, 1997.

[3] R Gupta. *Speech Recognition for Hindi*. M. Tech. Pro ject Re- port, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, 2006.

[4] M. M. Sondhi J. Benesty and Y. Huang. *Springer handbook on speech processing*. Springer Publisher, 2008.

[5] Kuldeep Kumar and R. K. Aggarwal. Hindi speech recogni- tion system using htk. *International Journal of Computing and Business Research*, 2, May 2011.

[6] orsberg M. Why is speech recognition difficult? Gothenburg, Sweden, 2003. Department of Computing Science, Chalmers University of Technology.

[7] S Pruthi T, Saksena and P K Das. Isolated word recognition for hindi language using vq and hmm. In *International Conferenceon Multimedia Processing and Systems (ICMPS)*. IIT Madras,2000.

[8] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersy, 1993.

[9] K. S. Rao and B. Yegnanarayana. Duration modification using glottal closure instants and vowel onset points. *Speech Communication*, 51, JUNE 2009.

[10] Frederico Rodrigues and Isabel Trancoso. Digit recognition using the speechdat corpus.

[11] Anurag Barthwal Mano j Kumar Singh Ramesh Rawat Shashidhar G. Koolagudi, Sujata Negi Thakur and K. Sreenivasa Rao. Vowel recognition from telephonic speech using mfccs and gaussian mixture models. In *Springer*, 2012.

[12] Bhavna Chawla Anurag Barthwal Shashidhar G. Koolagudi, Swati Devliyal and K. Sreenivasa Rao. Recognition of emotions from speech using excitation source features. In *ELSVIER*.ELSVIER, 2012.

[13] Nitin Kumar Shashidhar G. Koolagudi and K. Sreenivasa Rao.Speech emotion recognition using segmental level prosodic analysis. In *IEEE International confrence on device communication*. BIT MESRA, India, IEEE, FEB 2011.

[14] Zheng Hua Tan. *Automatic Speech Recognition on mobile devices and over communication networks*. Springer, 2008.