A Robust Speaker Independent Speech Recognizer for Isolated Hindi Digits

Vedant Dhandhania, Jens Kofod Hansen, Shefali Jayanth Kandi, and Arvind Ramesh

Abstract—**In this paper we present a speaker independent speech recognizer for isolated Hindi digits. Speech samples are collected from 30 individuals representing 5 distinct age groups from 15 to 40 years. For training the Hidden Markov Model (HMM) we use a total of 1000 utterances from 20 individuals. Optimal features such as MFCC, ΔMFCC and MFCC are used to train a HMM model. We aim to find the best combination of these features which yields the highest recognition rate along with the optimal number of hidden states of the HMM. Using MFCC and ΔMFCC as the feature vectors and 8 hidden states, an average recognition rate of 75% is achieved on a dataset of 500 utterances.**

Index Terms—**Hindi, MFCC, HMM, Speaker Independent, Recognition.**

I. Introduction

Hindi is one of the most widely spoken languages in India. More than 400 million people speak Hindi in the Indian subcontinent. It is also spoken in countries like Fiji, Singapore, Mauritius, UAE, and etc. which are outside the subcontinent. Literacy is as low as 65% in most states in India [1]. Hindi speech recognition systems would play a vital role in acquiring information from the masses. However, very little work has been done in developing robust systems that can successfully recognize Hindi words across a wide age group. Recognizing spoken Hindi digits (see Table I) would have wide applications in the field of ticketing, banking, handling search queries, etc. In this paper, we propose a speaker independent isolated digit word recognition system.

The contribution of the paper is two-fold. Firstly, a Hidden Markov model (HMM) is trained with a dataset covering a wide age group resulting in a robust speaker independent system for Hindi digit recognition. Secondly, the optimal number of states for HMM and the optimal number of features to represent the speech signal are found, which results in the highest accuracy for Hindi digits.

This paper is divided into 6 Sections. Section 1 gives a brief introduction on the topic and the motivation to pursue this research. Section 2 describes the contributions of related works pertaining to this field. The systematic data collection is described in Section 3. In, Section 4 we present the methodologies used and then we proceed to evaluating the results in Section 5. The final Section is the conclusion.

II. Related Work

A speaker dependent system is one which is trained on a specific speaker and recognizes the speech of that speaker with high accuracy. On the other hand, speaker independent systems have the capability of recognizing speech from any new speaker with the new speaker training the systems. It is well known that speaker independent systems are more difficult to design than speaker dependent systems.

Several speech recognition systems have been proposed for the isolated digit recognition in the Hindi language. A speaker dependent system using the Discrete Wavelet Transform is proposed in [2]. A success of 84% is achieved using the Daubechiesb8, 5- Level Decomposition (db8, Lev 5).

Saxena et al. proposed a microprocessor based Speech Recognizer using a novel zero crossing frequency feature combined with a dynamic time warping algorithm [3]. An overall success of 95.5% was reported with the implementation in MATLAB. The above systems involved training and testing on similar data leading to high performance. The number of speakers was limited to two in the experiments.

Swaranjali, a speaker dependent system uses a Linear Predictive Coding- Vector Quantization (LPC-VQ) front end for processing speech signal and an HMM model for recognition [4]. A success of 84.49% was achieved. It was suggested to obtain a wide public dataset to increase the performance and make it more robust.

Mishra et al. proposed a connected Hindi digit recognition system using robust features such as Mel Frequency Perceptual Linear Prediction (MF-PLP), Bark Frequency Cepstral Coefficient (BFCC) and Revised Perceptual Linear Prediction (RPLP) [5]. A success of 99% was achieved using the MF-PLP feature extraction and training Hidden Markov Models (HMMs). Pre-defined 36 sets of 7 connected digits uttered by 35 speakers was used in training and the 5 other speakers for testing. The performance for this system might be high as pre defined sets are used with a fix number of known digits in each set.

Apart from English, successful results have been proposed in word digit recognition in Japanese [6], Thai [7] and Italian [8]. Owing to their success we too evaluate the possibility of developing a robust system for Hindi digit word recognition.

Manuscript received May 30, 2012; revised August 25, 2012.

V. Dhandhania, S. J. Kandi, and A. Ramesh are with Manipal Institute of Technology, Manipal, India (e-mail: arvindnr89@gmail.com).

J. K. Hansen is with the University of Southern Denmark, Denmark

TABLE I: HINDI DIGITS ALONG WITH THE PRONUNCIATION.

English Digits	Hindi Pronunciation	English Pronunciation	English Digits	Hindi Pronunciation	English Pronunciation
0	“Shoonya”	“Zero”	5	“Paanch”	“Five”
1	“Ek”	“One”	6	“Chaeh”	“Six”
2	“Do”	“Two”	7	“Saat”	“Seven”
3	“Teen”	“Three”	8	“Aath”	“Eight”
4	“Chaar”	“Four”	9	“Nau”	“Nine”

III. DATA COLLECTION

To the best knowledge of the authors no public dataset is available for recorded isolated Hindi words. Our data is divided into the training data and testing data. The speakers that are used to create the training data are different from the ones used for creating the testing data. This ensures that training and testing are not done on similar data. We briefly describe how we created our datasets in this section.

A. Training Data

The data collection was done by selecting subjects across various age groups and sex. 5 distinct age groups were selected. These were as follows: 15-20 years. 20-25 years, 25-30 years, 30-35 years, 35-40 years. Two native males and two native female speakers were chosen from each age group. Each digit was individually uttered 5 times by the twenty participants. Hence, there is a total of 1000 utterances used for training the HMM. Each speech recording was sampled at 16 kHz and size of each sample was kept at 16 bits.

B. Testing Data

For the construction of the testing database, one male and one female speaker was chosen from each of the age groups described in Section 3.1. Each of the ten digit words is again uttered 5 times each, by the participants. Hence a total of 500 utterances are used as the testing data.

IV. METHODOLOGY

Fig. 1 shows the block diagram of the proposed system. In this section we discuss each of the blocks individually.

A. Pre-processing

Each speech excerpts is sampled at 16 kHz, mono channel PCM WAV, 16 bits quantization. The speech signal is pre-emphasized and is then framed by a Hamming window.

B. Feature Extraction

A short time feature vector is needed to represent the speech signal in most speech recognition algorithms. Standard feature vectors such as the Mel Frequency Cepstral Coefficients (MFCC) are the most popular ones. The MFCC describes the frequency content of a speech signal and is widely used in most of the speech recognition systems. We use this feature to represent our speech signal and the temporal derivates mainly the ΔMFCC and the MFCC. The filter bank is constructed using 13 linearly spaced filters where frames are of 25ms in length and a hop size of 10ms in length. Well known formula for the Δ MFCC and the MFCC is shown in 1.

$$\mathrm{d_t} = \sum_{i=1}^{N} \frac{i(c_{t+i} - c_{t-i})}{2\sum_{i=1}^{n} i^2} \qquad (1)$$

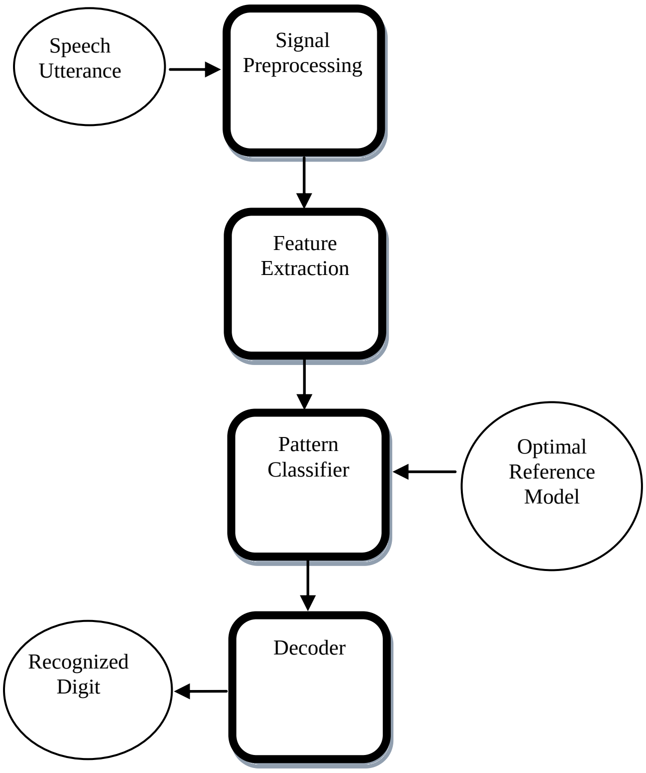

Fig. 1 Flowchart of the system.

C. Hidden Markov Model

The Hidden Markov Model (HMM) is trained as a standard left-to-right model discrete HMM, where we vary the number of states per digit to find the optimal numbers. A K-means clustering method is used for building a vector quantization code book, and again we try to vary the amount of centroids for optimal performance. The K-means centroids are build from all the MFCC vectors in the training set. Several HMMs are trained, one for each digit. For this, we use the Baum-Welch algorithm. Afterwards, the test data is evaluated on each HMM using HMM decoding. The utterance is assigned to a digit by the HMM giving the largest log-likelihood.

V. RESULTS

We evaluate the performance of our systems and analyze the most optimal configuration. Our evaluation is initially based on finding the optimal number of states for our HMM. Our second evaluation is based on evaluating the optimal number of features which provides the best results.

A. Effect of Features Combination

As mentioned earlier, we use the MFCC, ΔMFCC and the ΔΔMFCC as our acoustic features. We performed different experiments with 7 possible combinations of these 3 features. The number of hidden states is kept constant at 6. The accuracy based on these features is shown in Table II. By adding more features the performance increases considerably. It is evident that using the MFCC along with the ΔMFCC yields the highest recognition rate of 72 %.

TABLE II: PERFORMANCE OF DIFFERENT FEATURES.

Features	Dimension	Accuracy (in %)
MFCC	13	69
ΔMFCC	13	48
ΔΔMFCC	13	39
MFCC + ΔMFCC	26	72
MFCC + ΔΔMFCC	26	67
ΔMFCC + ΔΔMFCC	26	54
MFCC+ΔMFCC+ ΔΔMFCC	39	70

B. Effect of Different Number of States

It is evident from Section 5.1 that the MFCC with the MFCC feature vector yields the highest performance. Using this feature dimension we evaluate the best number of hidden states for the highest accuracy. We vary the states from 5 to 10. The results for the recognition are shown in Table III. It is evident that using 8 as the number of hidden states yields the highest recognition accuracy.

TABLE III: PERFORMANCE OF DIFFERENT NUMBER OF HIDDEN STATES

Number of Hidden states	Accuracy (in %)
5	69
6	72
7	73
8	75
9	71
10	68

C. Recognition of Individual Words

The recognition accuracy of the ten digit words using the optimal parameters using MFCC and ΔMFCC as the features and 8 hidden states in the HMM is as shown in Fig. 2. The digits zero, three, eight and nine have very high recognition while the digits four and five have poor recognition rates.

VI. CONCLUSION

In this paper, we created a robust database for a speaker independent isolated Hindi digit word recognition system. An HMM was trained and tested and optimal number of acoustic features were found to best represent the speech signals. An accuracy of 75% was achieved using MFCC and MFCC as the features and 8 as the number of hidden states. This system can be extended to a connected digit word recognition system using algorithms proposed in [9]. Also, acquiring more training data from various individuals will increase the accuracy further.

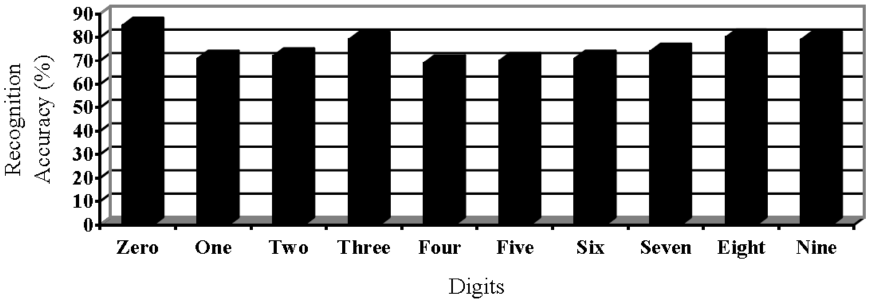

Fig. 2. Recognition accuracy for each digit

REFERENCES

[1] Provisional Population Totals : India : Census 2011 . [Online]. Available: http://www.censusindia.gov.in/2011-prov-results/indiaatglance.html

[2] S. Ranjan, "A Discrete Wavelet Transform Based Approach to Hindi Speech Recognition," in *Proceedings of the International Conference on Signal Acquisition and Processing (ICSAP)*, pp-345-348, August 2010.

[3] A. Saxena and A. Singh, "A Microprocessor based Speech Recognizer for Isolated Hindi Digits," in *IEEE ACE*, 2002.

[4] T. Pruthi, S. Saksena, and P. K. Das, "Swaranjali: Isolated Word Recognition for Hindi Language using VQ and HMM," in *Proceedings of the International Conference on Multimedia Processing and Systems (ICMPS).*

[5] A. N. Mishra, M. Chandra, A. Biswas, and S. N. Sharan, " Robust Features for Connected Hindi Digits Recognition," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 4, no. 2, June, 2011.

[6] H. Kawai and N. Higuchi, "Recognition of connected digit speech Japanese collected over the telephone network," in *Proceeding of the 5th International Conference on Spoken Language Processing,* Sydney Australia, pp 341-344. 1998.

[7] A. Deemagarn and A. Kawtrakul, " Thai connected Digit Speech Recognition using Hidden Markov Models," in *Proceedings of the 9th Conference on Speech and Computer*, St. Petersburg, Russia, September 2004.

[8] P. Cosi, J. Hosom, and F. Ravera. "High performance Italian continuous digit recognition," in *Proceedings of International Conference on Spoken Language Processing*. Beijing, China, pp. 242-245.2000.

[9] L. R. Rabiner, J. G. Wilson, and F. K. Soong, "High performance connected digit recognition using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 37, no. 8, 1989.