HINDI DIGITS RECOGNITION SYSTEM ON SPEECH DATA COLLECTED IN DIFFERENT NATURAL NOISE ENVIRONMENTS

Babita Saxena[1] and Charu Wahi[2]

Department of Computer Science, Birla Institute of Technology, Noida
babita.gs@gmail.com
charu@bitmesra.com

ABSTRACT

This paper presents a baseline digits speech recognizer for Hindi language. The recording environment is different for all speakers, since the data is collected in their respective homes. The different environment refers to vehicle horn noises in some road facing rooms, internal background noises in some rooms like opening doors, silence in some rooms etc. All these recordings are used for training acoustic model. The Acoustic Model is trained on 8 speakers' audio data. The vocabulary size of the recognizer is 10 words. HTK toolkit is used 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.

KEYWORDS

HMM, Acoustic Model, Digit Speech Recognition, Grammar

1. INTRODUCTION

In the last few years, Hidden-Markov-Model based (HMM) algorithms have been the most successful techniques used for speech recognition systems. Using the same, the experiments are conducted for building a Digit Speech Recognition(DSR) for Hindi. Thus, for building a DSR, acoustic characteristics like pitch, formant frequencies etc have to be computed. These characteristics are captured and a model is built based on these. These models are further used for recognition purposes.

In this paper we present our work on building acoustic model for Hindi Digits. Hindi belongs to the Indo Aryan family of languages and is written in the devanagari script. There are 11 vowels and 35 consonants in standard Hindi. In addition, 5 Nukta consonants are also adopted from Farsi/Arabic sounds.

David C. Wyld et al. (Eds) : CSITY, SIGPRO, DTMN - 2015
pp. 23–30, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50303

This paper is organized as follows. Section 2 gives the related work. Section 3,4,5 describes database, feature extraction process the Acoustic model preparing procedure respectively. Section 6 describes grammar structure used for decoding. Section 7 tells how the utterances are recognized. Section 8 discusses the Observation and Results of speech recognition on this system. Conclusion and Future work is stated in Section 9 and Section 10 respectively.

2. RELATED WORK

Now-a-days research work is being carried out for Hindi Digits. Some speech recognition systems have been proposed for the isolated digit recognition in the Hindi language.

Sharmila et al. proposed hybrid features for speaker independent Hindi speech recognition system. In this paper Mel-frequency cepstral coefficients (MFCC), Perceptual linear prediction (PLP) coefficients along with two newly modified hybrid features are used for isolated Hindi digits recognition. Two modified hybrid features Bark frequency cepstral coefficients (BFCC) and Revised perceptual linear prediction (RPLP) coefficients were obtained from combination of MFCC and PLP. Experiments were performed for both clean as well as on noisy data. In this experiment six different noises: Car noise, F16 noise, Factory noise, Speech noise, LYNX noise and Operation room noise have been added to clean Hindi digits database at different SNR levels to get noisy database. The recognition performance with BFCC features was better than that with MFCC features. RPLP features have shown best recognition performance as compared to all other features for both noisy and clean databases.[2]

Dhandhania et al. proposed a speaker independent speech recognizer for isolated Hindi digits. They aimed to find the best combination of features which yields the highest recognition rate along with the optimal number of hidden states of the HMM. Using MFCC and delta-MFCC as the feature vectors and 8 hidden states, an average recognition rate of 75% is achieved on a dataset of 500 utterances.[3]

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). 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 predefined sets are used with a fix number of known digits in each set.[4]

Saxena et al. proposed a microprocessor based Speech Recognizer using a novel zero crossing frequency feature combined with a dynamic time warping algorithm. 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.[5]

Apart from English, successful results have been proposed in word digit recognition in other languages like Japanese, Thai, and Italian. Owing to their success we too evaluate the possibility of developing a robust system for Hindi digit recognition in natural noise environment.

3. DATABASE

The speech data is collected from 10 speakers in varying noise environments. The recording is done in every speakers' home with respective noise in their rooms. The respective noise refers to vehicle horn noises in some rooms, internal background noises in some rooms like opening doors, silence in some rooms, etc. All these recordings are used for training acoustic model. The audio data consists of Hindi digits recordings of 10 speakers. In all 10 digits in Hindi are recorded by the speakers from age group from 20 to 40 yrs. The utterances are recorded in 48 KHz, stereo 16 bit format. The recordings of 8 speakers are used for training acoustic model. The recording is done using Sony Xperia L headset on laptop model - Dell Inspiron 1440. It is then channel separated and down sampled to 16 KHz and then single channel is used to train acoustic models. The database of 10 speakers is divided into a training database of 8 persons and a testing database of 2 persons.

4. FEATURE EXTRACTION

The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc.

The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and Mel Frequency Cepstral Coefficients (MFCC) accurately represent this envelope. MFCCs are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since.[6]

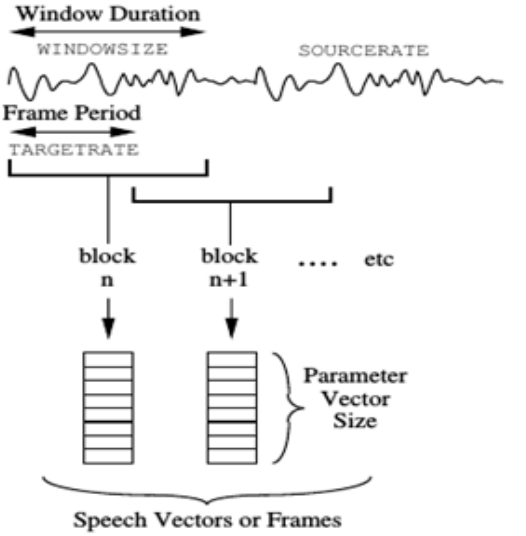

Fig.1: Feature Extraction Process

The primary feature vector size of 13 is considered along with their delta and delta-delta features. Thus in all 39 feature vectors are used.

5. ACOUSTIC MODEL

The acoustic model is built using HTK 3.4 toolkit. The acoustic model is HMM based since HMM based statistics for building acoustic model is very popular and have been shown to give better results than any other techniques. In a statistical framework for speech recognition, the problem is to find the most likely word sequence

$$\hat{W} = \arg\max_{w} p(W / A) \quad (1)$$

With a Bayesian approach to solving the above problem, we can write

$$\hat{W} = \arg\max_{w} p(A / W)\, p(W) \quad (2)$$

Equation 2 above gives two main components of a speech recognition system, the acoustic model and the language model. One type of language model is the grammar, which is a formal specification of the permissible structures for the language. The deterministic grammar gives the probability of one if the structure is permissible or of zero otherwise. Furthermore, the probabilistic relationship among a sequence of words can be directly derived & modeled from the corpora with the stochastic language model.[7] Language model is used for dictation type purposes systems & grammar is used for command & control systems or small vocabulary systems.[8] Since digit recognizer is a small vocabulary system, grammar is used for decoding.

The Acoustic Model preparation is described ahead with details of training steps. The acoustic model is made by training the recorded audio data out of which mfcc feature vectors are extracted and their delta and delta-delta features are considered. The hmm models are over 61 context independent phonemes. Each phone HMM definition file represents a single stream single-mixture diagonal covariance left-right HMM with five states.

Prototype models for 61 phonemes are built using flat start approach. These models were further refined by applying nine iterations of the standard Baum-Welch embedded training procedure. These models are then converted to triphone models and two iterations of Baum-Welch training procedure are applied, then there states are tied using decision tree based approach and two iterations of Baum-Welch training procedure are applied. Now the number of mixtures is incremented to 14 and seventeen iterations of the standard Baum-Welch training procedure were applied.

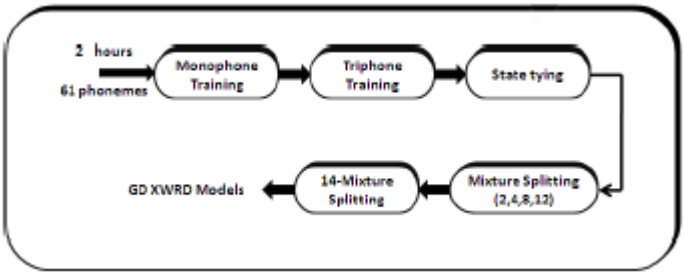

Fig.2: Acoustic model training procedure

The speech data is recorded in varying noise environment and pronunciation mistakes were taken care of, a few errors were found after recording. These errors were compensated by changing the transcription accordingly as per the pronunciation, by manually listening to the speech.

6. GRAMMAR

A speech recognition grammar contains one or more rules. Each rule defines a set of language constraints that a speech recognition engine uses to restrict the possible word or sentence choices during the speech recognition process. Speech recognition engines use grammar rules to control the elements of sentence construction using a predetermined list of recognized word or phrase choices.[9]

HTK provides a grammar definition language for specifying simple task grammars such as this. It consists of a set of variable definitions followed by a regular expression describing the words to recognize. For the digit recognizing application, a suitable grammar might be

$digit = एक | दो | तीन | चार | पाँच |
छः | सात | आठ | नौ | शून्य ;
({SENT-START} {$digit} {SENT-END})

where the vertical bars denote alternatives, the square brackets denote optional items and the angle braces denote one or more repetitions. The complete grammar can be depicted as a network as shown in Fig. 3.

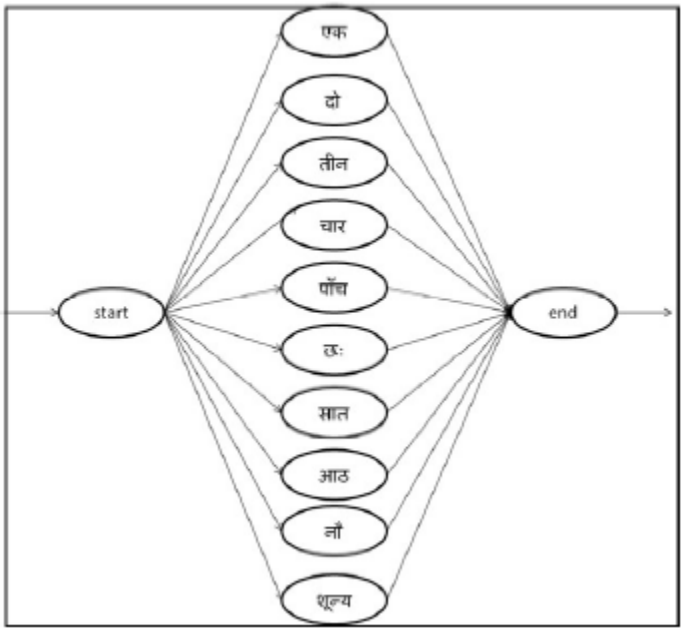

Fig.3: Grammar for Digit Recognition

7. DECODING

HTK's viterbi decoder Hvite is used for decoding the utterances. HVite is a general-purpose Viterbi word recogniser. It will match a speech file against a network of HMMs and output a transcription for each. When performing N-best recognition a word level lattice containing multiple hypotheses can also be produced.

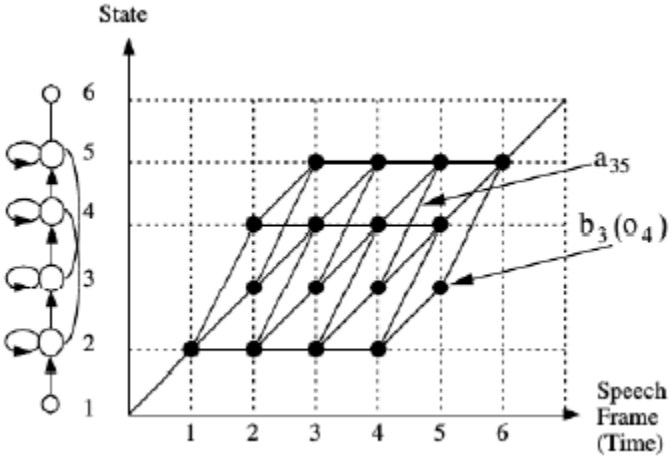

Fig.4: Viterby decoding process

8. OBSERVATION AND RESULT

The performance of DSR is measured in terms of recognition rate. The system is tested on

- 2 seen speakers
- 2 unseen speakers

where seen data means the data is taken from the training corpus itself and unseen data means the data is not from the training corpus. Recognition on the unseen data is carried out using the same acoustic model, lexicon and grammar. The performance is analyzed using HTK toolkit's HResult tool. The percentage number of correct labels recognized is given by:

$$\%\text{Correct} = \frac{H}{N} \times 100\%$$

where H and N are the number of correct labels and total labels respectively. In addition to the recognition rate, the tool also reports the number of insertions, deletions and substitutions for all the test data.

Table 1: Digits recognition rate for 2 speakers in training and test sets

Speaker	Phone level(%)	Word level(%)
Training	92.82	94.09
Test	86.17	85

The acoustic model and grammar is used for performing the recognition. The phone level recognition rate of 92.82% and 86.17% is observed on training (seen) data and test (unseen) data respectively. The word level recognition rate of 94.09% and 85% is observed on training (seen) data and test (unseen) data respectively.

9. CONCLUSION

In this paper, we described our experiments with the Hindi digits speech recognition. A baseline digits recognizer was developed and the results found are quite encouraging.

10. FUTURE WORK AND SCOPE

Our future work will be to further refinement of the word accuracy and supporting. A number of further experiments may be tried to achieve better accuracy. Some of them can be described as under.

- Continuous digits recognition experiments can be performed.
- Training corpus may be increased.
- Lexicon can be increased by adding 2-digit numbers.

REFERENCES

[1] Steve Young, Gunnar Ever, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Vaitchev, Phil Woodland, "The HTK Book", copyright 2001-2002 Cambridge University Engineering Department

[2] Sharmila1, Dr. Achyuta, N. Mishra, Dr. Neeta, Awasthy, "Hybrid Features for Speaker Independent Hindi Speech Recognition", International Journal of Scientific & Engineering Research, Volume 4, Issue 12, December-2013

[3] Vedant Dahndhania, Jens Kofod Hansen, Shefali Jayanth Kandi & Arvind Ramesh, "A Robust Speaker Independent Speech Recognizer for Isolated Hindi Digits", International Journal of Computer & Communication Engineering, Vol 1, No. 4, November 2012

[4] A.N.Chandra, Mahesh Chandra, Astik Biswas, S.N.Sharan, "Robust Features for Connected Hindi Digits Recognition", International Journal of Signal Processing & Pattern Recognition, Vol. 4, No. 2, June 2011

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

[6] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[7] Xuedong Huang & Li Deng, Microsoft Corporation, "Chapter 15, An Overview of Modern speech Recognition, Handbook of Natural Language Processing" C5921_C012 Page 343, 2009-9-9

[8] http://en.wikipedia.org/wiki/Acoustic_model

[9] http://msdn.microsoft.com/en-us/library/hh378438%28v=office.14%29.aspx

AUTHORS

Ms. Babita has received her MCA from Uttar Pradesh Technical University, Lucknow (UP)-India in 2008. Presently she is pursuing M.Tech from Birla Institute of Technology, Noida -India. She has published more than 3 research papers in the area of Speech, Signal and Image Processing at National/International level. Her areas of interest are Speech and Signal Processing.

Ms. Charu Wahi is currently a Ph.D. candidate in the Department of Computer Science and Engineering, Birla Institute of Technology, Ranchi. She received her B.E. degree in 2000 and M.Tech - Computer Science in 2008. She is currently working as Assistant Professor in Department of Computer Science and Engineering, Birla Institute of Technology, Ranchi, Noida Campus. Her research areas include routing, security, quality of service especially in mobile ad-hoc networks