Continuous Hindi Speech Recognition using Monophone based Acoustic Modeling

Ankit Kumar
Department of Computer Engg
National Institute of Tech
Kurukshetra, India

Mohit Dua
Department of Computer Engg
National Institute of Tech
Kurukshetra, India

Tripti Choudhary
Department of Elect & Comm
Vishveshwarya Inst. of Tech
Greater Noida, India

ABSTRACT

Speech is a natural way of communication and it provides an intuitive user interface to machines. Although the performance of automatic speech recognition (ASR) system is far from perfect. The overall performance of any speech recognition system is highly depends on the acoustic modeling. Hence generation of an accurate and robust acoustic model holds the key to satisfactory recognition performance. In this paper, we compare the performance of continuous Hindi speech recognition system with different vocabulary sizes and feature extraction techniques. Mel frequency cepstral coefficient (MFCC) and perceptual linear prediction (PLP) both are used as a feature extraction techniques in our proposed system. Monophone based acoustic modeling is done by Hidden Markov Model (HMM) at the back-end of an ASR system. HTK 3.4.1 toolkit is used for the implementation of this system. The system is trained for 70 different Hindi words. The experimental result shows that our system is able to achieve 95.08% accuracy, when we use MFCC as a feature extraction technique.

General Terms

Pattern Recognition, Speech Recognition, Signal Processing, Communication systems

Keywords

Hindi Speech recognition; Automatic speech recognition; HMM; MFCC

1. INTRODUCTION

Automatic speech recognition is the process of taking speech utterances and converting it into text sequences as close as possible [1]. An ASR provides the natural interface for easy use of the machines. The communication in native language like Hindi with machine provides the great boon to society in the country like India. The Major population in our country is not aware with English in one or other way i.e. reading or writing. So, Automatic Hindi speech recognition is the need of our society. There are number of applications of automatic Hindi speech recognition which can be useful in public areas such as information retrieval system at railway stations, airports, bus stations, and government offices etc. by serving the customer with answer to their spoken quires.

Based on speaking style, ASR is divided in to four categories: (1) Isolated word recognition, (2) Connected word recognition, (3) continuous speech recognition, and (4) spontaneous speech recognition. Connected word speech recognition systems are also able to process the continuous speech but with little pause in between each word. This drawback make infeasible for practical use of this system. Unlike this system, continuous speech recognition systems are able to process the continuous speech of human being.

Continuous Hindi speech recognition is the state of art and gives better results in compare to the other techniques. Continuous speech recognition can be achieved by two ways: (1) speech recognition by Monophones, and (2) speech recognition by Triphones. In this paper, we developed the continuous Hindi speech recognition system based on Monophone acoustic modeling. HMM is used for acoustic modeling at the back-end of an ASR system. For feature extraction MFCC and PLP both are used in front- end of an ASR system.

The rest of paper is organized as follow: basic components of ASR (Section 2), literature review of related work (Section 3), Monophone based acoustic modeling (Section 4), and experimental results are discussed in this section (Section 5), and at last concludes the proposed work with future direction.

2. BASIC COMPONENTS OF ASR

State-of-the-art ASR systems consist of five basic modules: the signal processing components (i.e., pre-processing and feature extraction), the set of acoustic models (i.e. HMM), the language model (i.e. N-gram estimation), the pronunciation dictionary with a word lexicon and search engine for final decoding as shown in Fig. 1 [2].

2.1 Pre-processing

The speech signal is an analog signal; to process digitally we have to convert it into digital form. Analog to digital convertor is used for this purpose. After this, signal has to be preprocessed. Signal preprocessing involves some crucial steps like Background noise elimination, Pre-Emphasis Filtering, Framing and Windowing.

Fans, footsteps, opening and closing of doors, etc., are the some sources of background noise. Recording in noise free environment and close speaking microphone helps in minimizing the background noise. The signal is filtered using a simple high pass FIR filter in pre-emphasis step with pre-emphasis parameter value 0.95. In framing, the pre-emphasized speech signal is blocked into frames of N samples. Window means the portion of speech waveform to be processed. Each frame is multiplied by a window to reduce the edge effect of every frame segment.

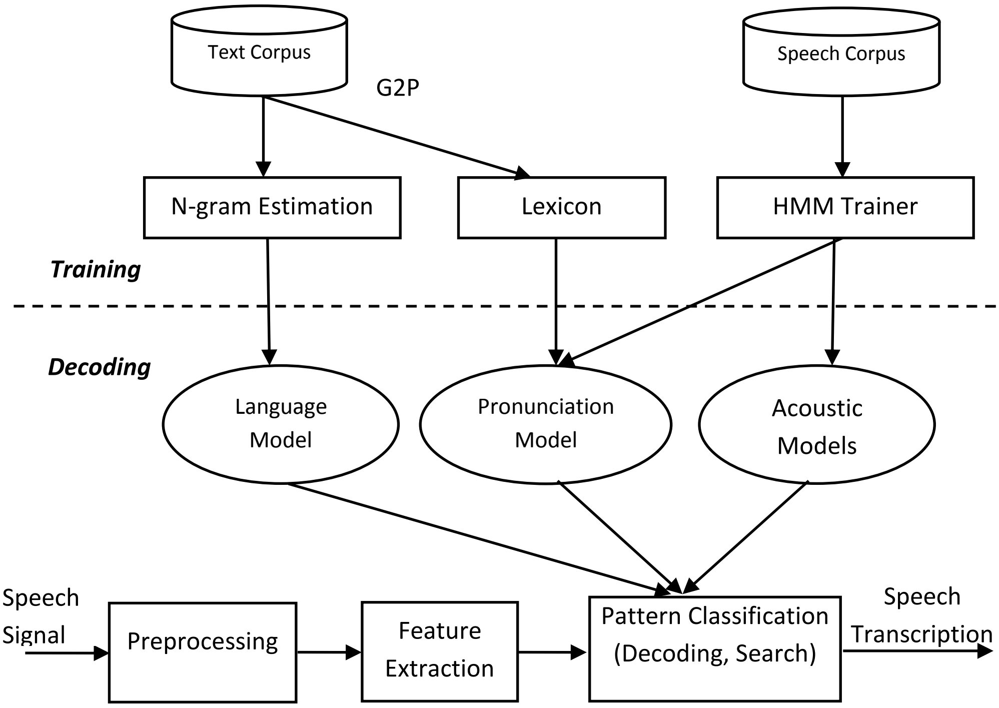


Fig 1: Architecture of Automatic Speech Recognition [2]

2.2 Feature Extraction

Feature extraction in ASR system aims to preserve the information needed to determine the phonetic class, while being invariant to other factors including speaker based differences such as accent, emotion, fundamental frequency or speaking rate, as well as other distortions, for example, background noise, channel distortion, reverberation or room modes. Feature extraction reduces the bandwidth from 16,000 samples per second (speech sampled at 16 kHz) to about 3,900 features per second (39 features per frame × 100 frames per second with 25 ms window size). Clearly, this step is crucial to ASR system, as any loss of useful information cannot be covered in later processing.

Feature extraction is the method of extracting the limited amount of useful information from high dimensional data. In reducing the dimensionality of the speech signal, the typical order is 80:1. Some popular feature extraction techniques are given bellow:

- Linear Predictive Cepstral Coefficient (LPCC)
- Mel Frequency Cepstrum Coefficient (MFCC)
- Perceptual Linear Prediction (PLP)
- PLP Derived from Mel scale Filterbank (MF-PLP)

2.3 Acoustic Modeling

Acoustic models are used to link the observed features of speech signal with the expected phonetics of hypothesis sentence. Being the main component of ASR, acoustic model accounts for most of the computational load and performance of the system. The most typical implementation of this process is probabilistic, making use of hidden Markov models. To generate mapping between the basic speech units (phones, syllables) and the acoustic observations, a rigorous training procedure is followed. Training involves creating a pattern representative for the features of a class using one or more patterns that correspond to speech sounds of the same class. A phonetically rich and balanced database is required to train the acoustic models. To transcript the acoustic features into linguistic units, various representations are used such as whole words, syllables, and context dependent phoneme and context independent phonemes [3].

2.4 Pronunciation Model

During recognition, the sequence of symbols generated by the acoustic component is compared with the set of words present in the lexicon to produce optimal sequence of words that compose the system's final output. Thus a lexicon (or dictionary) is used to provide the mapping between words an phones (or sub-word units). It contains information about which words are known to the system and also how these words are pronounced, i.e., what their phonetic representations look like. Usually a standard pronunciation also called canonical pronunciation (or base form) is used which can be found in ordinary dictionaries.

2.5 Language Model

ASR systems use n-gram language models to guide the search for correct word sequence by predicating the likelihood of the n^{th} word, using the $n-1$ preceding words. The probability of occurrence of a word sequence W is calculated as:

$$\begin{aligned} P(W) &= P(w_1, w_2, \ldots\ldots, w_{m-1}, w_m) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)\ldots \\ &\quad \ldots P(w_m|w_{m-n+1}w_{m-n+2}\cdots w_{m-2}w_{m-1}) \end{aligned} \quad ..(1)$$

Common feasible n-gram models are tri-grams $(n = 3)$, where $P(w_3|w_1 w_2)$ is modeled for words w_1, w_2 andw_3, and bi-grams$(n = 2)$, where $P(w_2|w_1)$ is modeled.

2.6 Decoder

A decoder is a part of the speech recognizer which performs actual recognition with the help of trained acoustic and language models. The decision on the final transcription (recognized words) must be taken by combining and optimizing the information of acoustic and language models. For continuous speech recognition, Viterbi decoder can be implemented using the token passing algorithm [4].

2.7 RELATED WORK

Pruthi et al. [5] have developed a speaker-dependent, real-time, isolated word recognizer for Hindi. Developed system uses a standard implementation. Linear predictive cepstral coefficients are used for feature extraction and recognition is carried out using HMM. System was designed for two male speakers. The recognition vocabulary consists of Hindi digits (0, pronounced as "shoonya" to 9, pronounced as "nau").

K. Kumar et al. [6] developed a small vocabulary, isolated Hindi speech recognition with high performance 94.63%. This system is speaker independent. For training, 5 male and 3 female speakers are used. Vocabulary size of system is 30 words. MFCC is used as a feature extraction technique and at back end HMM is used. HTK toolkit is used to develop this system.

Mishra et al. [7] in 2011 proposed a Hindi ASR system for connected digit recognition. To build this speaker independent system, 40 different speakers are used in which 23 are female and 17 are male speakers. All speakers are age group of 18-26 years. After speech recording some noises is added artificially. In this paper different feature extraction technique is used such as BFCC, RPLP, MFCC, PLP, MF-PLP in front end and HMM is used at back end. They receive high accuracy 99%, when MF-PLP is used as a feature extraction technique.

Shweta et al. [8] in 2013 proposed a speaker independent, continuous Hindi speech recognition system for with different vocabulary sizes. In this paper, Gaussian mixture HMM model with various states was used for training and recognition. In this paper, MFCC and perceptual linear prediction (PLP) with heteroscedastic discriminant analysis (HLDA) was used as a feature extraction technique. HTK and Sphinx was used to implement this system. Overall accuracy 93% was achieved with MFCC at front end and 8 states GMM (Gaussian mixture model) at beck end, when the vocabulary size was 600 words.

In 2011, R. k. Aggarwal et al. [9] proposed an ASR system, in which different state GMM was used to train the ASR system. This speaker independent system was built for 100 – 400 vocabulary size. The best performance of ASR was observed when 4 state GMM was used. This system gives 88% accuracy with 400 vocabulary size. To reduce the MFCC features, HLDA feature reduction algorithm was used in front end of the ASR system.

Gaurav et al. [10] proposed a speaker independent continuous speech recognition system for Hindi. The main objective of this paper is to build ASR system for teaching geometry in primary schools. This system gives 88.81% accuracy, when Julius is used as a recoznizer at the back end of ASR system. For training speech utterance of 30 different individual is used, in which 18 male and 12 female speakers are used. Each individual utter same word 5 times, MFCC and HMM is used as a feature extraction and classifier for this system.

2.8 MONOPHONE BASED ACOUSTIC MODELING

The process of acoustic modeling in ASR can be described as comparison of test pattern and stored reference pattern in the form of feature vector and producing the speech unit in the text form [11]. The word based acoustic mode has few limitations such as rapidly reduction of performance with increased size of database, recognition for known word only, and large computational overhead [12]. To overcome these limitations, phone based acoustic model is used. Phone is the small part of speech sounds which contain the useful information. Linguistic classify the speech sounds or phoneme in to number of categories as shown in fig:

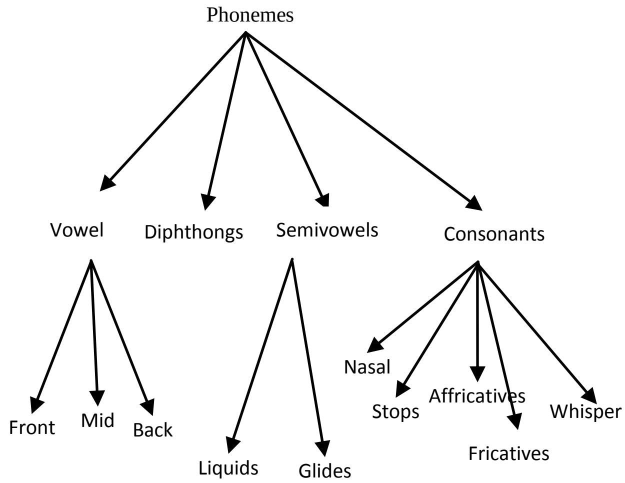
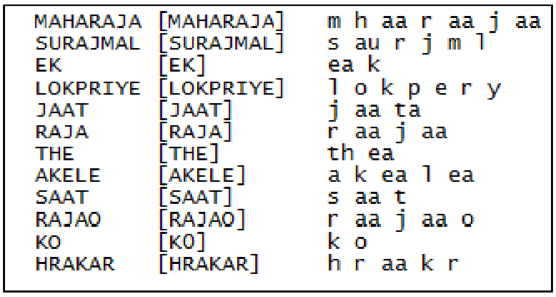


Fig 2: Speech sound classification

Hindi language is broadly divided into vowel and consonants. Hindi language contains 13 vowels and 36 consonants. Thus, Hindi language contains 49 characters in total. The Monophone based acoustic modeling starts with training the recorded speech data. For training of our system we use 40 monophone HMM model like (अ a, आ aa, औ ao etc). For acoustic modeling, the Hindi words are divided in to sequences of phones. Some entries of our proposed system lexicon are given bellow:



```
MAHARAJA  [MAHARAJA]   m h aa r aa j aa
SURAJMAL  [SURAJMAL]   s au r j m l
EK        [EK]         ea k
LOKPRIYE  [LOKPRIYE]   l o k p e r y
JAAT      [JAAT]       j aa ta
RAJA      [RAJA]       r aa j aa
THE       [THE]        th ea
AKELE     [AKELE]      a k ea l ea
SAAT      [SAAT]       s aa t
RAJAO     [RAJAO]      r aa j aa o
KO        [KO]         k o
HRAKAR    [HRAKAR]     h r aa k r
```

Fig 3: Some lexicon entries of proposed ASR system

However, the monophone based acoustic model is not context dependent as compare to triphone based acoustic model. Triphones are context dependent monophones with its left and right hand side. Phones are depends on preceding and succeeding phones and this ability of capture the variation with respect to context improves the overall performance [11].

3. DATABASE PREPARATION

For the design and development of European languages ASR systems, large and standard databases are available which were prepared by various agencies. For example, TIMIT and ATIS are two of the most important databases that are used to build acoustic models of American English in ASRs [13]. But to prepare such kind of standard databases for Indian languages, no much effort has been done so far.

For the development of this system, we first prepare the database of speech utterances. System uses the database of 70 different Hindi words. The speech sounds were recorded with the help of unidirectional microphone and the distance between mouth and microphone was minimum (2-4 cm) during recording. Recording was done in room environment and each word was recorded 10 times in separate file with .wav extension. So, 10*70 = 700 speech files were process for the training of our speaker dependent speech recognition system. Labeling is done with the help of wavesurfer and audacity is used for recording the speech sounds. All recording is done by Panasonic unidirectional mike.

4. EXPERIMENT RESULTS

In this section, we show the results of different experiment of our work. The various experiments with their results are as:

4.1 Experiments with different vocabulary sizes

In this paper, the performance of an ASR system was computed with different vocabulary size (10 words – 70 words). When we use the small size of vocabulary, we achieve the higher performance. This fact is supported by the figure 3. Monophone based HMM is used in this experiment to train the ASR model. MFCC is used as a Feature extraction technique at the front-end of an ASR system. Up to 30 words vocabulary size we get the 100% accuracy. After this performance degrade as the size increase. We achieve 95.08% accuracy with 70 word vocabulary size. For testing we 10 speech files are used and each file contain the 7-12 continues words.

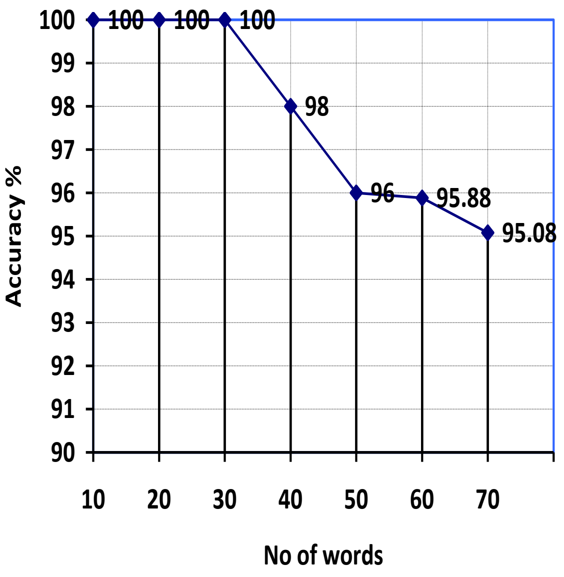


Fig 4: Accuracy of ASR with different vocabulary size

4.2 Experiments with different feature extraction techniques

In this experiment, performance of continuous Hindi speech recognition is compared between MFCC and PLP as feature extraction techniques. 5 states HMM with 39 feature vector are used for the monophone based acoustic modeling. When we use 30 words vocabulary size then we gets 100.00% accuracy with MFCC and PLP as feature extraction technique. When we increased the size of vocabulary then MFCC gives better results in compare to the PLP. Performance comparison is show in table below

Vocabulary Size	Monophone Based Acoustic Modeling	
	MFCC	***PLP***
30 Words	100.00%	100.00%
50 Words	96.00%	84.00%
70 Words	95.08%	85.25%

Fig 5: Performance of ASR with different feature extraction techniques

5. Conclusion

Speech based interfaces or applications are getting popularity in every walk of life as they have started to fulfill the essential needs of civilized or uneducated society, such as inquiries from call centers or telephone booking of railway or cinema tickets. Automatic speech recognition systems have many rapidly growing application areas such as human–computer interactions, document preparation, interactive voice response systems, database access, web enabling via voice and hands-free applications as in car phones or voice-enabled PDAs. Some people seek it out, preferring dictating to typing and others find it embedded in their hi-tech gadgetry, in mobile phones and car navigation systems. ASR deals with the decoding of an acoustic signal of a speech utterance into corresponding text transcription, such as words, phonemes or other language units. This paper plays an important role to realize this kind of communication, and has achieved nearly 95% recognition accuracy with 70 different Hindi words vocabulary. In this paper, we proposed the monophone based continuous Hindi speech recognition. An experimental result shows, we got maximum accuracy when we use MFCC as a feature extraction technique. This work can be enhanced with large vocabulary system.

6. ACKNOWLEDGMENTS

I would like to thanks Mr. R.K. Aggarwal, Mr. Yogesh Kumar, and Ms. Tripti Choudhary for their useful discussions and suggestions during the preparation of this paper.

7. REFERENCES

[1] Aggarwal, R. K. and Dave, M. 2011. Using Gaussian mixture for Hindi speech recognition system, International Journal of Speech processing, image Processing and Pattern Recognition, vol. 4, no. 4.

[2] Aggarwal, R. K. and Dave, M. 2010. An Empirical Approach for Optimization of Acoustic models in Hindi Speech Recognition Systems, 8th International conference on Natural language processing, ICON-2010.

[3] Lee, C. H., Gauvain, J. L., Pieraccini, R. and Rabiner, L. R. 1993. Large vocabulary speech recognition using subword units, Speech Communication, vol. 13, pp. 263–279.

[4] Young, S., Evermann, G. et al. 2009. The HTK book. Cambridge: Microsoft Corporation and Cambridge University Engineering Department.

[5] Pruthi, T., Saksena, S. and Das, P. K. 2000. Swaranjali: Isolated Word Recognition for Hindi Language using VQ and HMM," Paper Presented at International Conference on Multimedia Processing and Systems (ICMPS), IIT Madras, India.

[6] Kumar, K. and Aggarwal, R. K. 2011. Hindi Speech Recognition System using HTK, International Journal of Computing and Business Research, vol. 2, issue 2.

[7] Mishra, A. N. et al., 2012. Robust Features for Connected Hindi digits Recognition, Int. Journal of Signal Processing, Image Processing and pattern Recognition, Vol. 4, No. 2.

[8] Sinha, S, Agrawal, S. S. and Jain, A. 2013. Continuous density Hidden Morkov Model for context dependent Hindi speech recognition, Int. Conference on Advances in Computing, Communication and Informatics (ICACCI), pp. 1953-1958, IEEE.

[9] Aggarwal, R. K. and Dave, M. 2011. Using Gaussian mixture for Hindi Speech Recognition System, International Journal of Signal Processing, Image Processing and pattern Recognition, SERSC Korea, vol. 4, no. 4.

[10] Kumar, Gaurav et al. 2012. Development of Application Specific Continuous Speech Recognition System in Hindi, Journal of Signal and Information Processing, 3,394-401.

[11] Banerjee, Pratyush et al. 2008. Application of Triphone Clustring in Acoustic Modeling for Continuous Speech Recognition in Bengali, 19th international conference on Pattern Recognition, pp. 1-4, IEEE.

[12] Ghai, W. and Singh, N. 2013. Phone based acoustic modeling for automatic speech recognition for punjabi language, Journal of Speech Sciences, vol. 3, no. 1, pp 69-83.

[13] Aubert, X. L. 2002. An overview of decoding techniques for large vocabulary continuous speech recognition, Computer Speech and Language, vol. 16, no. 1, pp. 89–114.

[14] Becchetti, C. and Ricotti, L. P. Speech Recognition Theory and C++ Implementation, 3rd ed., vol. 2, John Wiley & Sons, pp 121-141.

[15] Furui, Sadaoki 2005. 50 Years of progress in Speech and Speaker Recognition Research, ECTI Transaction on Computer and Information Technology, vol. 1, no. 2.

[16] Rudnicky, A. I., Hauptmann, A. G. and Lee, K. 1994. Survey of Current Speech Technology, Communication of the ACM, vol. 37, no. 3.

[17] O'shaughnessy, D. 2013. Acoustic analysis for automatic speech recognition, proceeding of the IEEE, vol. 101, no. 5.