Implementation and Performance Evaluation of Speaker Adaptive Continuous Hindi ASR using Tri- phone based Acoustic Modelling

Ankit Kuamr , Mohit Dua ,Computer Engineering Department National Institute of Technology Kurukshetra, India

Rahul Shivhare Computer Science Department Dr. K.N.M.I.E.T. Modinagar, India

Abstract— Speech interface to computer is the next big step that computer science needs to take for the general users. Speaking in our native language is a natural and effortless task which carried out with great speed and ease. Speech recognition will play an important role in taking technology to common man. The need is not only for speech interface but speech interface in local language like Hindi. In this paper, we evaluate the performance of Hindi Automatic Speech Recognition (ASR) by using two most popular feature extraction techniques namely Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) at front- end of ASR. The Hidden Markov Model (HMM) was used at back- end of an ASR for Hindi language. The proposed system has been implemented for the continuous Hindi speech using tri-phone based acoustic modelling. For speaker adaptation Maximum Likelihood Linear Regression (MLLR) technique has been implemented in this paper. All experiments were done by using HTK 3.4.1 (Hidden Markov Toolkit) on Linux environment (Ubuntu 12.05).

Keywords— Hindi Speech Recognition; Speaker Adaptation; automatic speech recognition; HMM; MFCC; MLLR

I. INTRODUCTION

An ASR is the process of parameterization of speech signal at front- end and likelihood evaluation of these features at back- end [1]. In general words speech recognition is the process of converting speech sound in to text form. During the last few decades lots of research has been done in the field of ASR for the language like English, Japanese etc. These languages have got mature ASR engine by now, while speech recognition for Indian languages is still in its infancy stage.

In a multilingual society like India where 1652 native languages are use and 17 other official languages are recognized by the constitution of India [2]. The use of speech recognition in Hindi language in a man- machine communication is great boon to the society because Hindi is a first (majority) or second language for every Indian. The major population in our country is not aware with English in one or other way i.e. reading or writing. The majority of population lives in rural areas and most of them is unfamiliar with computer and English language. In such situation keyboard become a barrier between man and machine. Speech recognition is an attempt towards reducing the gap between common man and machine. An ASR plays a significant role in such conditions and the need of our Indian society.

Based on speaking style, ASR system can be classified in to four categories: - (1) Isolated speech recognition, (2) Connected speech recognition, (3) Continuous speech recognition and (4) Spontaneous speech recognition. First two ASR systems are not feasible to handle the real time speech of human being. In this paper, we deal with continuous Hindi speech recognition which is only achieved by phone level acoustic modelling. We implemented ASR system for Hindi language by using mono-phone based acoustic modelling as well as tri-phone based acoustic modelling. In tri-phone based acoustic modelling preceding and succeeding phones are grouped with the middle phone to improve the performance.

For the implementation of speaker independent ASR system huge number of speakers is needed and as the speakers increase the computational load of the ASR system is also increase. This type of system is not a good choice in real time environment. To cope with this problem speaker adaptation technique is employ in this paper. Speaker adaptation [3] is the process of changing acoustic model parameters in such a way that ASR system recognizes the different users by using few utterances from that person. In this paper, MLLR technique is used for the speaker adaptation.

All the investigations are based on the experiments conducted in our paper. All experiments were conducted in general field condition on Hindi language. Rest of the paper is organized as follows: section II presents the motivation behind this work, architecture and working are discussed in section III, and section IV presents the overview of speaker adaptation techniques. In section V, an experimental comparison of continuous Hindi ASR system with various conditions is presented. Finally, the paper concludes with a brief discussion of the experimental results.

II. MOTIVATION

India is a multilingual society where 17 official and 1652 other languages are used [2]. The majority of Indians are not aware to English in one or other way (reading or writing) as English is not the first language. It would be great boon to society by giving benefits of information technology to common man with the help of speech as an interface between man and machine. Except that research in Indian languages (like Hindi, Telugu etc.) are still in its starting stage and there is no ASR system for Hindi language with high accuracy.

There are number of areas where speech recognition proves their superiority over other. For example, speech can be used for information retrieval at railway stations, airports, bus stations and government offices etc. by serving customer with answer to their spoken queries. Speech recognition would be beneficiary to those people who are physically challenged (e.g. voice controlled wheel-chair, typing on computer etc.). According to [1], speech would be used as shortcut to open a particular file / folder instead of traversing many level of hierarchy e.g. "OPEN SPEECH PROJECT" as well as a portable application as the size of computer reduces from desktop to laptop and laptop to tablet, the use of keyboard becomes the challenging task. In all above scenario speech will be the competitive alternative.

III. WORKING OF ASR

ASR is the process of taking speech utterances as a input captured by a microphone, a telephone or any other transducer and convert it into most probable text sequence which was represented by the acoustic data [5]. State- of- art ASR systems consists four modules: signal processing module (pre-processing, feature extraction), acoustic models (HMM), language model, decoder. The basic architecture of ASR is inspired by the human auditory system.

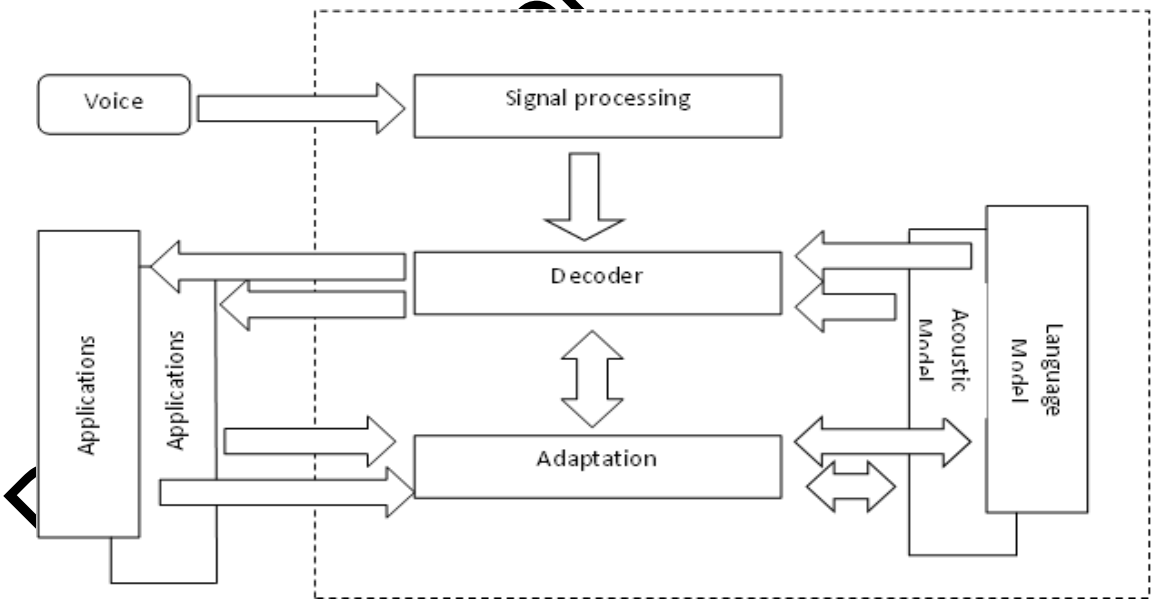

Figure1: Architecture of ASR [7]

Speech utterances of word sequence W is decided by speaker mind and delivered through his / her text generator [6]. Speaker vocal apparatus contain the signal processing component and produce the speech waveform which passes

through the many noise channels. At last decoder decode the acoustic signal X into most probable word sequence W* as close as possible to the original word sequence W [6].

The working of typical ASR system is described in figure 1. In signal processing module, an speech utterances is converted into sequence of feature vectors *X= {x1, x2 ,…. xT}.* Except that signal pre-processing is also done here, including background noise removal, windowing, framing, pre-emphasis etc. Normally, 0.95 is the value of pre-emphasis parameter generally used [8]. According to [9], speech is the sequence of uttered phonemes, approximately 12 phonemes per second. For easy computation speech is broken down into small sets (windowing and framing) because speech signal is quasi-stationary at that time. Generally, Hamming window (25-30 ms) is applied at every short time interval (10 ms) to generate 50-70% overlapping with adjacent frames. Clearly, these extracted feature vectors plays a vital role for high accuracy ASR system. For the extraction of these features MFCC, PLP are the techniques which are currently in use. However, [1] each of them has several drawbacks in various conditions such as noisy environment, typical field conditions etc.

The decoder produces the maximum probability word sequence W* after processing the extracted features vector with the help of acoustic and language modelling. The mapping from each sub-word units to acoustic observations is done in acoustic modelling [10]. Being the main component of ASR, acoustic model takes most of the computational load of the system. HMM is the natural choice for the acoustic modelling. The role of language model is to produce the probability of each sequence of word by accepting the various competitive hypothesis of word from acoustic model [5]. The decoder also provides the required information needed for adaptation to adaptation module to modify the acoustic parameters so that performance of ASR system will increase [6]. Bayesian probability theory is used for classification as given in equation bellow:

$$W^* = Argmax_{\,w} P(O/W).\ P(W) \qquad (1)$$

Here, O is the observation sequence and W is the word sequence.

IV. ADAPTATION METHODS

Due to the mismatch between training and testing conditions, the performance of an ASR system degrades rapidly. To overcome this problem, speaker adaptation techniques are used [11], in which we transform the parameters in such a way so that they adapt new conditions. Acoustic model adaptation is the process of changing the acoustic model parameters used for speech recognition in such manner that ASR system recognizes the different target user by using a few utterances from the target user [2]. Speech dependent recognition systems have high accuracy for one speaker but this system become worse for different target user. By using speaker adaptation technique we can improve the speech recognition accuracy by using few utterances of target user.

In speaker dependent system is based on ML estimation using the EM algorithm, which cannot precisely estimate the model parameters. As a result, recognition accuracy would be much worse. Speaker adaptation aims to overcome these problems. An adaptation model should improve recognition accuracy with a small amount of data.
Maximum a posteriori (MAP) estimation is used in statistical modeling and particularly, used for speaker adaptation. It estimates model parameters more robustly than ML estimation in a condition when the amount of data is small, and its estimates asymptotically approach ML estimations as the amount of data increases [3]. Maximum Likelihood Linear Regression (MLLR) based on a linear mapping between the acoustic feature spaces of different target speakers. MLLR is one of the most popular adaptation methods because it is easy to use and it performs well in most cases.

Generally, MLLR is robust and well suited to unsupervised incremental adaptation. There are two main variants of MLLR: unconstrained and constrained.

In MLLR, the mean vectors of the Gaussian distributions in the HMMs, $\mu = (\mu_{1\ldots}, \mu_n)$, where n is the dimension of a feature vector, are updated according to the following transformation:

$$\boldsymbol{\mu = A\,\mu + b} \quad \textbf{(2)}$$

Here, A is an n x n matrix, and b is a n-dimensional vector.

V. EXPERIMENTAL SETUP AND RESULTS

In the first data preparation step, the speech utterances were sampled at 16 KHz with hamming window size 25 ms to obtain the 39 acoustic features [12]. The proposed system used the vocabulary of 50 to 150 different Hindi words. The system database was taken from the short story of jaat maharaja surajmal from jaatland.com. The speech sounds were recorded with the help of unidirectional microphone by close talking (2-4 cm). To obtain 39 acoustic feature vectors, we used MFCC and PLP as a feature extraction technique. In data preparation step, for speaker dependent system each word was recorded 15 times, 10 times for training and 5 times for testing the base system. For speaker adaptive system 2 utterances of each word were recorded, 1 for training and 1 for testing purpose. The speaker adaptation experiments were performed on a set of speech data recorded by the 7 different male speakers, in which 6 were from same age group of 25-30 years old and one speaker was near about 60 years old. At the back- end of ASR system acoustic modeling was done with the help of 5 states HMM model. All experiments were performed by using open source tool HTK 3.4.1 with Ubuntu 12.05. The various experiments with their results are as:

5.1 Experiment with different vocabulary sizes

In this experiment, we developed an ASR system with different vocabulary sizes from 50 words to 150 words, and system performance is compared between speaker adaptive, and speaker dependent speech recognition system. Speaker adaptation of Hidden Markov Models (HMMs) using the Maximum Likelihood Linear Regression (MLLR) method was implemented in this system. The graph shows in figure 2 that when we applied speaker adaptation technique with minimal training data we get satisfactory results. In this experiment we use MFCC as feature extraction technique in front-end, and HMM in the back-end of ASR system. The HMM model of 5 states were used for this experiment with 39 acoustic features. Results show that the accuracy of SD ASR system is high in the case of speaker dependent system.

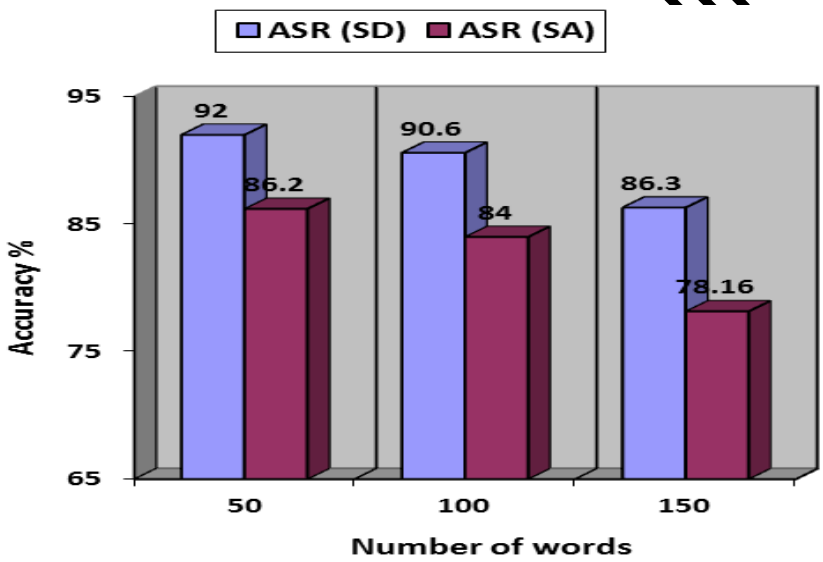

Fig 2: Accuracy of ASR with different vocabulary size

5.2. Experiment with speaker adaptation using 50 words

In this experiment speaker adaptation using Maximum Likelihood Linear Regression (MLLR) technique is applied. Acoustic modeling is done by tri-phones and single mixture HMM model is used at back-end. MFCC and PLP are used at front-end for feature extraction. For the results of this section speech utterance of seven different speakers were recorded. Each word was recorded two times one for training and one for testing. In this experiment we use 50 word vocabulary size and test utterance contain the same sentence as we use before for checking the performance for 50 words system. Out of seven speakers six are from 25-32 year age group and one is 60 year old. The baseline system is made for the author of this dissertation and lies in the age group 25-32 years. The result shows that the dedicated system gives worse result for the speaker 4 (62 years old) not come in the age group of authors.

TABLE I

No. of Speakers	% Accuracy of different techniques	
	MFCC	PLP
1	86.20%	83.68%
2	84.00%	82.18%
3	79.20%	77.00%
4	62.00%	60.00%
5	84.26%	81.78%
6	82.18%	79.12%
7	85.19%	83.00%

Table 1: Performance comparison of speaker adaptive system using 50 words

5.3 *Experiment with speaker adaptation using 100 words*

In this experiment speaker adaptation technique is applied on 100 words vocabulary. As in the last experiment based on 50 words, we use single mixture HMM model and tri-phone based acoustic modeling for the training purpose of this system. Comparative results are shown below for different feature extraction techniques. For system performance testing, same utterance is used which we used for 100 words GMM mixture model i.e. 10 sentences approximately containing 75 different words. In this experiment speaker 4 (i.e. not in age group) again gave not satisfactory results as compared to others.

TABLE II

No. of Speakers	% Accuracy of different techniques	
	MFCC	PLP
1	84.00%	82.67%
2	80.00%	81.00%
3	78.00%	72.00%
4	60.00%	56.00%
5	83.00%	80.00%
6	79.00%	77.00%
7	82.68%	81.12%

Table 2: *Performance comparison of speaker adaptive system using 100 words*

5.4 *Experiment with speaker adaptation using 150 words*

By following the same procedure, the speaker adaptation results are study by using 150 vocabulary sizes. All scenarios are same for testing the system performance i.e. HMM model and tri-phone based acoustic modeling is used for the results. Speaker 7 gives the better results in this case. Approximately 77 % accuracy is achieved when we use MFCC and 72% accuracy is achieved by PLP feature extraction techniques.

TABLE III

No. of Speakers	% Accuracy of different techniques	
	MFCC	PLP
1	78.16%	76.37%
2	75.50%	72.00%
3	77.12%	73.40%
4	50.00%	45.23%
5	81.12%	76.37%
6	75.50%	71.00%
7	80.00%	78.16%

Table 3: Performance comparison of speaker adaptive system using 150 words

CONCLUSIONS

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. Tri-phone based acoustic modeling is new in Hindi ASR system. Automatic recognition of speech is a challenging task due to the various sources of variability like speaker, environmental and linguistic variability. Speaker adaptation is a good choice to avoid the computational load of an ASR system as well as speech coder. By using MLLR adaptation technique is this paper we get the satisfactory result by using MFCC in compared to PLP technique of feature extraction. When we use MFCC instead of PLP we get the 3-4% high accuracy rate. This work motivates to enhance the vocabulary size as well as the number of speakers by using different adaptation techniques in near future.

REFERENCES

[1] R. K. Aggarwal and M. Dave, "Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system," Telecommunication Systems, vol. 52, no. 3, pp 1457-1466, Springer, 2013.
[2] Pukhraj P. Shrishrimal et al., "Indian Language Speech Database: A Review", International Journal of Computer Applications, Volume 47, No. 5, June 2012.
[3] Koichi Shinoda, "Speaker adaptation techniques for automatic speech recognition", Proceeding of APSIPA ASC'2011, Oct, 2011.
[4] Alexander et al., "Survey of current speech technology", Communication of ACM, Vol. 37, No. 3, March, 1994.
[5] R. K. Aggarwal and M. Dave, "Using Gaussian mixture for Hindi speech recognition system," International Journal of Speech processing, image Processing and Pattern Recognition, vol. 4, no. 4, December 2011.
[6] Xuedong Huang et al., Spoken language processing- A guide to theory, algorithm and system development, Prentice Hall Inc., 2001.
[7] J. Gao et al., "A unified approach to statistical language modeling for chinese", International conference on Acoustic, Speech and Signal Processing, pp. 1703-1706, Istanbul, 2000.
[8] C. Becchetti and L. P. Ricotti, Speech Recognition Theory and C++ Implementation, 3rd ed., vol. 2, John Wiley & Sons, pp 121-141.
[9] D. O'shaughnessy, "Acoustic analysis for automatic speech recognition," Proceeding of the IEEE, vol. 101, no. 5, May 2013.
[10] R. K. Aggarwal and Mayank Dave, "Discriminative Techniques for Hindi Speech Recognition System", ICISIL, pp.261-266, Springer, 2011.
[11] Jean- Paul Haton, "Automatic Speech Recognition: A Review" Enterprise Information Systems V, 6-11, 2004.
[12] S. Young et al., The HTK Book. Available: http://htk.eng.cam.ac.uk.