Hindi Speech Recognition and Online Speaker Adaptation

Ganesh Sivaraman
Dept. of Electrical & Electronics
BITS Pilani K.K. Birla Goa Campus
Zuarinagar, Goa, India

K Samudravijaya
School of Tech. & Computer Science
Tata Inst. of Fundamental Research
Colaba, Mumbai, India

ABSTRACT

Speaker Adaptation is a technique which is used to improve the recognition accuracy of Automatic Speech Recognition (ASR) systems. Here, we report a study of the impact of online speaker adaptation on the performance of a speaker independent, continuous speech recognition system for Hindi language. The speaker adaptation is performed using the Maximum Likelihood Linear Regression (MLLR) transformation approach. The ASR system was trained using narrowband speech. The efficacy of the speaker adaptation is studied by using an unrelated speech database. The MLLR transform based speaker adaptation technique is found to significantly improve the accuracy of the Hindi ASR system by 3%. It was also observed that the improvement in accuracy is dependent upon the recognition accuracy of the un-adapted system.

General Terms

Automatic Speech Recognition, Hindi Speech Recognition

Keywords

Automatic Speech Recognition, online speaker adaptation, Maximum Likelihood Linear Regression (MLLR), Hindi Speech recognition.

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems provide a user-friendly interface to computers. ASR systems are being increasingly used in a variety of special tasks such as voice dialing in mobile phones, voice operated aids to handicapped, spoken document summarization, information retrieval etc. Yet, the accuracy of ASR systems is not high enough to be used widely by the general public as a replacement to a keyboard.

The foremost challenge facing ASR systems is the mismatch between the training and the testing (actual use) conditions. For example, a system trained using speech data recorded in a quiet environment will have poor recognition accuracy when tested with speech data in presence of noise. Another challenge is caused by the variation of pronunciation of words by different people. Yet another variation in speech signal is caused by different voice characteristics of different people due to anatomical differences. Quality of telephone handsets and limited bandwidth of telephone channel aggravate the problem. In order to handle such varied changes in speech signal, an ASR system has to be trained well taking into consideration all possible pronunciations for every word spoken by a large number of people. Thus, a huge amount of training data is essential for training a good speech recognition system.

1.1 Types of ASR

Based on the usage, ASR systems can be classified as Speaker Dependent (SD) and Speaker Independent (SI) systems. SD systems are trained to recognize the speech of only one particular speaker at a time. On the other hand, SI systems can recognize speech from anyone. Only one acoustic model is trained using training data from many speakers; this single model is used for recognition of speech by anyone whose voice it might not have 'seen'. SD systems have higher recognition accuracies than SI systems. Thus, it is preferable to use SD systems. However, in order to train SD systems, a large amount of speaker specific training data would be required which is not possible to have in case of a multi-user application such as telephonic enquiry system.

Speaker adaptation is a technique that uses a new speaker's voice sample to re-train a SI system to recognize the new speaker's speech better. However, it is not practical to demand the user to provide speech data for adapting the ASR system to his voice, in case of a voice interface, to be used by the general public (for example, information retrieval over telephone/mobile channel). A solution would be to use the spoken query (text spoken by the caller) to adapt acoustic models. However, as the enquiry of the user is short (just a few seconds), model adaptation has to be carried out with tiny amount of adaptation data. Since the transcription (text) of the query is not known, such an adaptation is called unsupervised speaker adaptation. Since the adaptation is carried out, on the spot, just using the caller's query speech, the approach is called online speaker adaptation.

While the advantage of online speaker adaptation has been studied for western languages [2,3,4], we are not aware of such a study for Indian language speech. This paper reports a study of improvement in the accuracy of a speaker independent, Hindi speech recognition system with the addition of an unsupervised online speaker adaptation module. The rest of the paper is organized as follows. A brief theory on speech recognition and speaker adaptation is given in section 2. Section 3 provides the experimental details. Section 4 deals with the experimental results and discussions. The conclusions are presented in section 5.

2. THEORY OF SPEECH RECOGNITION AND SPEAKER ADAPTATION

An overview of the basic concepts of Automatic Speech Recognition and speaker adaptation of acoustic models is provided in this section.

2.1 Automatic Speech Recognition

Speech Recognition is a specific case of pattern recognition. In pattern recognition, a set of reference patterns are stored and the test patterns are compared for matching with the reference patterns for recognition. The speech recognition system implemented here employs Hidden Markov Models (HMM) [1] for representing speech sounds. A HMM is a stochastic model; it does not store a set of reference patterns. A HMM consists of a number of states, each of which is associated with a probability density function. The parameters of a HMM comprises of the parameters of the set of probability density functions, and a transition matrix that contains the probability of transition between states. A lot of training data consisting of well-chosen application specific sentences are recorded from various people. Speech signals are analyzed to extract features useful for recognizing different speech sounds. These features and the associated transcriptions are used to estimate the parameters of HMMs. This process is called ASR system training. In case of Continuous Speech Recognition, the goal is to determine that sequence of words whose likelihood of matching the test speech is the highest. The training procedure involves the use of forward – backward algorithm. The recognition is done using Viterbi decoding.

2.2 Speaker Adaptation

Speaker adaptation is a technique which reduces the difference between training and testing conditions by transforming the acoustic models using a small set of speaker specific speech data. It is also necessary to have the correct word transcriptions for the adaptation data for robust adaptation. There are two types of speaker adaptation. In supervised adaptation, the text of the adaptation speech is known to the system. Supervised adaptation provides good improvement in the recognition rates as it is same as re-training the system. However, this process is not practical in a system designed for a telephonic enquiry that is used by practically everyone, and the user does not have the patience to provide enough adaptation data. Unsupervised adaptation is the approach adopted in this paper. The system automatically adapts itself to the present speaker at the same time as he keeps using the system. The system uses the first sentence spoken by a new speaker as the adaptation data. The (possibly incorrect) output of the SI system is assumed as the correct transcription. Using this transcription and the adaptation data, the system transforms its acoustic models in order to recognize the speech of the current user better. Then, it re-recognizes the unknown utterance with adapted models, hopefully resulting in better recognition accuracy. Since the speaker adaptation is carried out as and when a speaker is using the system, this approach is called online speaker adaptation.

A popular speaker adaptation method that needs small amount of data is Maximum Likelihood Linear Regression (MLLR). The MLLR method assumes that the SI acoustic models can be transformed into speaker adapted models by a simple linear transformation.

$$\mu_{new} = A_n \mu_{old} + b_n.$$

As shown in the above formula, the mean vectors of the SI models (μ_{old}) can be transformed linearly into the mean vectors of the adapted models (μ_{new}). The transformation matrix A_n and the vector b_n are the parameters to be found by maximizing the likelihood of the adaptation data [2,3]. Thus, MLLR method puts the linear regression into the Baum-Welch estimation framework. On transformation, the sound clusters are shifted and rotated to better represent the new speaker's voice as illustrated in Figure 1. There are two types of MLLR transforms: Single class and Multi class. In single class MLLR, all the acoustic models are transformed by a single global transformation matrix. In the multiple class MLLR transformation, the phonemes are grouped into several classes. The transformation matrices are computed for each class of phonemes and each class is transformed with the transformation matrix corresponding to that class. The advantage of multiple classes MLLR is that the different degrees of *speaker specific* changes in different phoneme classes can be modelled. However, the number of parameters to estimate is larger; this demands a large amount of speech data from the new speaker. Since the focus of this work is on online speaker adaptation based on just single input utterance (typically lasting a couple seconds), we employed the single class MLLR transform based adaptation. Fgure1 shows a 2-dimensional representation of the shifting and rotation of the sound clusters for both single and multiple class MLLR transformations. A global MLLR transformation allows speaker adaptation of all the sounds even if there are only a few of them present in the adaptation data [4]. In this paper, we have used such a global MLLR transform because of scarcity of adaptation data.

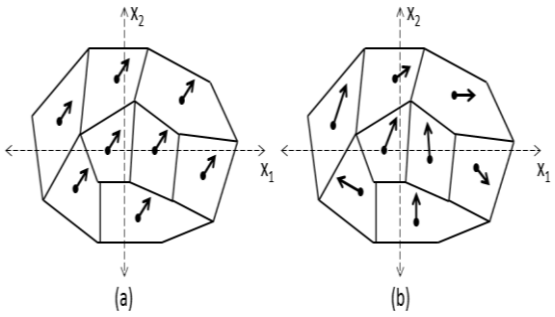

Figure 2. Illustration of the magnitude and direction of shift of centers of 8 classes of phonemes due to MLLR transformation (a) Single (global) transformation – All classes are shifted by same amount (b) Multiple transformation – 8 different MLLR transformation matrices transform each class independently.

3. EXPERIMENTAL DETAILS

The experiment was carried out using a Hindi speech recognition system developed using the Sphinx speech recognition toolkit [5,6]. A tutorial on implementing a Hindi ASR system can be found at [7]. The 39-demensional feature vector was comprised of 13 Mel Scale Cepstral Coefficients, their first and second derivates. The basic acoustic units were context dependent phonemes (triphones) modelled by left-to-right, 5-state, semi-continuous HMMs. The output probability distributions of states were represented by Gaussian mixture densities; 256 global Gaussian density functions were used to generate Gaussian mixtures for every state of every phoneme. The language model used in the system was backoff trigram grammar.

3.1 Speech databases

In this subsection, we will briefly describe two speech databases that were used in this work.

The experiment on speaker adaptation was carried out on a Speaker Independent (SI) Hindi speech recognition system. The SI system was trained using the database containing 924 randomly selected grammatically correct sentences in Hindi. These sentences were spoken by 92 different speakers, of which 56 were males and 46 females. The total number of words in the pronunciation dictionary was 2731; some lexical entries had multiple pronunciations associated with them. The speech data was recorded over telephone channel, and hence band limited to 4 kHz.

In order to test the efficacy of online speaker adaptation, we used another Hindi speech database, called the "rlwRes" database. This multi-speaker, continuous speech database consists of spoken queries related to railway reservation availability. The database consists of 1581 sentences spoken by 92 speakers, both male and female. The speech data had been recorded using a desktop microphone system in a quiet environment. This wideband speech was down sampled to 8 kHz, and then used as adaptation cum test data. The transcriptions of all "rlwRes" sentences were used to derive the backoff trigram language model used for decoding. The perplexity of the language model was low due to the fact that the utterances were queries related to a specific task.

3.2 Speaker adaptation and evaluation methodologies

The online speaker adaptation experiment was carried out by repeating the following procedure for every utterance of the test set ("rlwRes" database). An utterance from the test set was recognized using the Speaker Independent speech recognition system. Then, the SI system was adapted to the test speaker's voice as follows. Using the transcription provided by the SI system and the test feature vector sequence, the MLLR transformation algorithm generated the transformation matrices A_n and b_n. These matrices were used to transform the acoustic models on the fly so as to model the test utterance better in the maximum likelihood sense, and thus adapt to the new speaker's voice. The same test utterance was re-recognized using the speaker adapted models. This process was repeated, one by one, for all the sentences of the "rlwRes" database. Finally, the recognition accuracies, before (SI system) and after adaptation (speaker adapted system), were computed. The accuracy was calculated for percentage correct word matches and complete sentence matches.

The formula for calculation of percentage correct and accuracy are given as follows[9].

$$\textit{Percentage Correct} = (N - D - S)/N \times 100\%$$

$$\textit{Accuracy} = (N - D - S - I)/N \times 100\%$$

where

- N = Number of words or sentences correctly recognized.
- D = Number of unrecognized/missed words (Deletion errors)
- S = Number of times a word was misrecognized as another word (Substitution errors)
- I = Number of extra words inserted between correctly recognized words. (Insertion errors)

4. RESULTS AND DISCUSSION

The performances of speech recognition system before and after online speaker adaptation are presented in Table 1. The percentage of words and sentences correctly recognized before (i.e., without) and after online speaker adaptation are listed in the table. The first and second columns of the table show the recognition accuracy at the word level when word insertions are ignored and considered respectively. These performance figures before adaptation are 84.5%. and 83.8% respectively. We observe that the MLLR transform based speaker adaptation gives an improvement of about 3% in the percentage of correctly recognized words as well as sentences. It may be noted that the word recognition error has reduced from 15.5% to 12.6%, a reduction by a factor of 0.19. The corresponding relative error reduction for sentences is 0.06.

Table 1. Speech recognition accuracies with and without online speaker adaptation.

Performance Measure	Without adaptation	After speaker adaptation	Relative error reduction
Words correct	84.5%	87.4%	0.19
Words accuracy	83.8%	86.8%	0.19
Sentences correct	50.6%	53.4%	0.06

The "rlwRes" database is unrelated to the database used for training the acoustic models. There was little overlap between speakers of the two databases. Yet, the online adaptation of acoustic models to the test speaker just using the test utterance (about 2-3 seconds duration) reduces the word error rate by a factor of 0.19. This is the benefit of online speaker adaptation. However, one should note that MLLR based speaker adaptation of acoustic models is a highly compute-intensive process. So, the approach described in this paper is practical only when speech recognition is performed on a powerful computer, in a batch (offline) mode. In other words, this approach is not suitable for ASR in embedded systems where the computing power is limited and output is expected in real-time.

Although this method of unsupervised adaptation yielded a 3% increase in word accuracy, the efficiency of this adaptation method crucially depends upon the accuracy of the base SI system. If the accuracy of the SI system is not good, the initial transcript provided by the SI system will be erroneous. Consequently, the system will adapt to the new speaker improperly, and there might not be any improvement (or worse degradation) in the recognition accuracy. In fact, online speaker adaptation led to mild deterioration in ASR accuracy when the accuracy of a complex, general purpose SI system was less than 70%. This low accuracy was due to high perplexity of the language model.

5. CONCLUSION

The improvements in word and sentence recognition accuracies after online adaptation (using just one test utterance) shows that MLLR transform based speaker adaptation of Hindi speech models indeed decreases the recognition error by a factor of 0.19. This demonstrates that MLLR transform based adaptation transforms the acoustic models in such a way that the difference between test and train conditions is reduced, resulting in better performance. It is evident that it is possible to successfully adapt the system using just one sentence spoken by the speaker, provided sufficient computing resources are made available.

6. ACKNOWLEDGMENTS

We are very thankful to Tata Institute of Fundamental Research (TIFR) for providing us the necessary facilities and a perfect environment to work.

7. REFERENCES

[1] Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Proceedings of the IEEE 77, 257—286.

[2] Legetter, C.J. 1995. Improved acoustic modeling for HMMs using linear transformations. Ph.D. Thesis. University of Cambridge, (1995)

[3] Leggetter, C.J., Woodland, P.C. 1995. Maximum likelihood linear regression for speaker adaptation of HMMs. Computer Speech & Language 9, 171–185.

[4] Doh, S.-J. 2000. Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression, PhD. Thesis, Carnegie Mellon University.

[5] CMU sphinx – Speech Recognition Toolkit, http://www.cmusphinx.sourceforge.net.

[6] Chan, A., et al. 2003. The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resources.

[7] Samudravijaya, K. 2000. Hindi Speech Recognition. J. Acoustic Society of India 29(1), 385–393.

[8] Sivaraman, G. et.al. 2011. Higher Accuracy of Hindi Accuracy of Hindi Speech Recognition Due to Online Speaker Adaptation, In Proceedings of ICTSM 2011, CCIS 145, 233 – 238.

[9] Steve Young et al. 2002, HTK Book, <http://htk.eng.cam.ac.uk/docs/docs.shtml>