An improved Hindi speech Recognition system by using I-Rover

Rahul Rana,
Department of Computer science, Govt. Engineering College, Ajmer (Raj.), India
rahulrana_63@rediffmail.com

Ravinder Singh,
Assistant professor, Department of Computer science, Govt. Engineering College,
Ajmer (Raj.), India

Devarshi Mishra
Department of Computer science, Govt. Engineering College, Ajmer (Raj.), India

Abstract— Speech technology and systems in human computer interaction have witnessed a stable and remarkable advancement over the last two decades. These technologies enable machines to respond correctly and reliably to the voices of humans, and provide useful and great services with high values. But there are limited work done for Hindi language and that is also not able to provide a good accuracy Hindi system. In this paper, we proposed a Hindi speech recognition system by using iROVER (improved Recognizer Output Voting Error Reduction), a system combination technique. This system is modification of the previously generated Hindi system by using old Rover system combination approach. By this proposed system, word error rate of previously generated Hindi system by traditional ROVER approach, will be reduce and combination of ASR's will produce efficient result.

Index Terms— Automatic speech recognition system, Hindi ASR, iROVER, *MFCC*, *PLP*.

1 INTRODUCTION

SPEech is the most important, common, basic and efficient form of communication for people to interact with each other, It is also a natural and quick way of exchanging the information among humans, if used to interact with computers can overcome many limitations. That is why speech recognition is in research for many years and has attracted many researchers across the world. Speech-to-Text or automatic speech recognition can be described as a system which converts speech into text. There are many applications of automatic speech recognition system some of them are in health care instruments, banking devices, aircraft devices, robotics etc. There is lot of work is done in this field but mostly in European languages. Although some significant work has been done for South Asian language including Hindi but none of them have given satisfactory results. In the field of Indian language speech recognition, various researchers have tried to examine the different aspects of speech. This paper aims to propose an efficient Hindi speech recognition system using iROVER (an improved system combination technique) for ensembeling output of individual ASR system which got by extracted features of ASR systems by different extraction techniques (MFCC, PLP and LPCC) .It is basically the modification of the previous work done by the use of traditional ROVER approach which will improve WER.

This proposed system will provide less world error rate and also will be beneficial for the increasing vocabulary size. Having improved ROVER (Recognizer Output Voting Error Reduction) this system overcome limitations of the previously developed system [6].Besides accuracy in clean environment and large vocabulary it can also performed well in noisy environment. Paper has been prepared in following order Section 2 presents architecture of ASR and its function. Section 3 explains system description and iROVER and proposed model combination.

2 ASR ARCHIETECTURE

ASR system basically works in two steps; in first step preprocessing is done with feature extraction, while second step covers acoustic modeling, language model, pattern recognition or transcription. The block diagram of the ASR is shown below with its entire module:

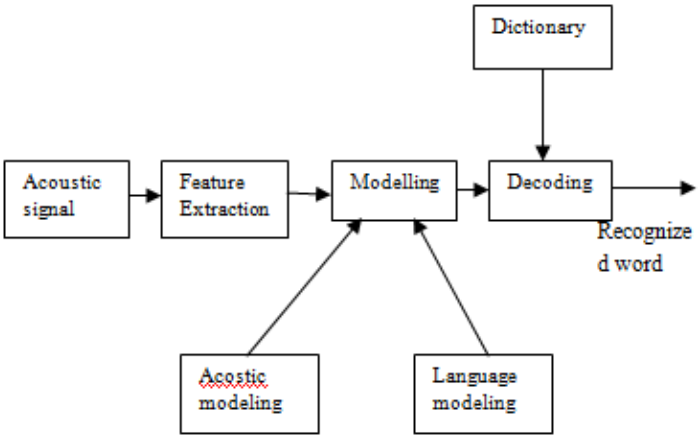

Fig. 1: ASR block diagram.

2.1 Digitized Preprocessing

Transformation of the acoustic signal into the digital signal is very important before processing of the signal because recorded signal is in the form of analog signal and analog signal cannot be directly processed by ASR systems. When this converted signal is passed, it is passed through the first order filters to spectrally flatten the signals. The result of this step is to increase the magnitude of higher frequency as compared to lower frequency. This process, known as pre-emphasis, because of increment in the magnitude of higher frequencies with respect to the magnitude of lower frequencies. The side by side step is to block the speech-signal into the frames with frame size ranging from 10 to 25 milliseconds and an overlap of 50%–70% between consecutive frames.

.

2.2 Feature Extraction

Feature extraction aims to find a set of properties of an utterance that have acoustic correlations to the speech-signal, that is parameters that can somehow be computed or estimated through processing of the signal waveform. Such parameters are known as features. The feature extraction process is likely to discard irrelevant information to the task while keeping the useful one. It involves the process of measuring some important characteristic of the signal such as energy or frequency response (i.e. signal measurement), augmenting these measurements with some perceptually meaningful derived measurements (i.e. signal parameterization), and statically conditioning these numbers to form observation vectors.

2.3 Language modeling

Language model is the single largest component trained on million of words, consisting of millions of parameters and developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary. ASR systems utilize *n*-gram language models to guide the search for correct word sequence by predicting the likelihood of the *n*th word on the basis of the *n*−1 antecedents words. The probability of occurrence of a word sequence *W* is calculated as:

$$P(W) = P(w1, w2. \ldots\ldots\ldots\ldots\ldots\ldots . wm\text{-}1, wm)$$

$$= P(w1).\ P(w2 \mid w1).\ P(w3 \mid w1w2) \ldots\ldots\ldots\ldots\ldots\ldots$$

$$P(wm \mid w1w2w3 \ldots\ldots\ldots\ldots\ldots wm\text{-}1).$$

While constructing of n-gram language models for large vocabulary speech recognizers, two problems are being faced. Large amount of training data generally leads to large models for real time works. Another is the sparseness trouble, which is being faced during the training of domain specific models. Language models are cyclic and non-deterministic.

2.4 Accoustic modeling

Important component for an ASR is acoustic model and it accounts for most of the computational load and performance of the system. It is used to join the observance features of the speech signals with the expected phonetics of the hypothesis sentence. The Acoustic model is developed for detecting the spoken phoneme. Its creation involves the use of audio recordings of speech and their text scripts and then compiling them into a statistical representation of sounds which make up words. There are many models for this purpose, but Hidden Markov Model (HMM) is the most widely used and accepted technique because of its efficient algorithm for training and recognition. It is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters. This algorithm is often used due to its simplicity and feasibility of use. Strong Independency assumption in HMM states that frames are independent, given a state. As a result, it lacks an ability to deal with a feature which straddles over several frames. Features such as delta coefficient, segmental statistics and modulation spectrum have been developed which can deal with phenomena of straddling. Aggarwal and Dave [11] have reviewed the variety of modifications and extensions adopted for the HMM based acoustic models in the form of refinements such as variable continuance models, discriminative techniques, connectionist approach (HMM+ANN) to overcome the limitations of traditional HMM and advancements such as margin based methods, wavelets and dual stream approach.

2.5 Recognition

Once all the sub-processes of preprocessing of analog signal is completed, classifier component recognizes the test samples based on the acoustic specifications of word. The categorization problem can be stated as finding the most probable sequence of words W given the acoustic input *O* (Juraf'sky & M'artin, 2009), which is computed as:

$$P(W/O) = \frac{P(O/W).P(W)}{P(O)}$$

Given an acoustic observance sequence *O*, classifier finds the sequence *W* of words which maximizes the probability $P(O \mid W).P(W)$. The quantity $P(W)$, is the prior probability of the word which is estimated by the language model. $P(O \mid W)$ is the observation likelihood, called as acoustic model. The value *P(W),* usually referred to as the *Language Model (LM)* depends on high-level constraints and linguistic knowledge about the allowed word strings for a specific task. The value *P(O|W)* is known as the *Acoustic Model (AM).* It describes the statistics of sequences of parameterized acoustic observations in the feature region given the corresponding uttered words.

3 SYSTEM IMPLENTATION

In this section, implementation of the speech system based upon the proposed system architecture has been presented. Various tools and techniques have been proposed by research-

ers for the implementation of speech recognition systems. Each approach has some advantages and disadvantages, means some ASR system performs better in some environment while its performance degraded in other environment. For example, the feature extraction technique PLP outperforms MFCC, when training and testing conditions are different. But with similar training and testing conditions, MFCC is better than the PLP. Both the techniques are computationally expensive. LPCC works well in clean environment but its performance gets degraded in noisy environment; it takes low computation power and little time to extract the features. Both the feature extraction techniques MFCC and PLP perform better than LPCC. Besides of above discussed techniques for feature extraction BFCC, RPLP, MF-PLP are being used for robust feature selection. And also, it has been found that F0 contour is the most essential characteristic to differentiate various tones and MFCC & PLP fail to provide it. The proposed combination system is shown in Fig 2. The proposed system encapsulates three individual ASR systems and the system will produce output using voting technique iROVER (improved ROVER).

3.1 System description

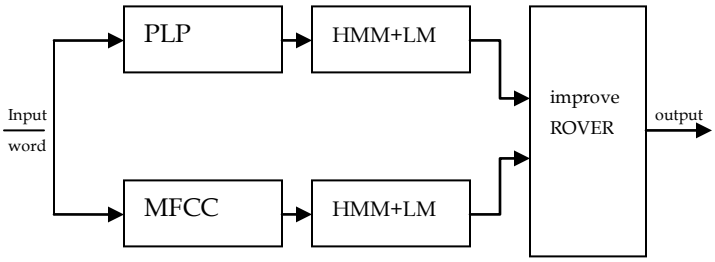

Fig.2: Ensemble system

Above figure of proposed system is shown which will provide efficient, less word error rate approach for two ASR systems.

3.2 Related work

This approach of combining many ASR's for Hindi language to produce less error rate was proposed by[6], although such type of multiple ASR systems was first proposed by Waibel. Where several time delay neural networks were developed for different subsets of confusable consonants and the outputs of these sub-networks were combined to determine the consonant class. Many other also used such approach at different-different places like combining a BU system based on stochastic segment models (SSM) and a BBN system based on Hidden Markov Models. It was a generalization of integrating more than one speech recognition technologies, working on different strategic conditions.

3.3 iRover

It presents an improved system combination technique [4]. It obtains significant improvements over ROVER (Fiscus, 1997), and is consistently better across varying numbers of constituent systems. A classifier is trained on features from the lattices of system, and choose the hypothesis of final word by learning cues to choose the system that is most likely to be correct at each word location. This approach gets the best result published to date on the TC-STAR 2006 English speech recognition evaluation set. In this a system combination method is used that outperforms all previously known techniques and is also robust to the number of constituents systems. The corresponding improvements over ROVER are particularly large for combination when only using two systems.

In this used approach that always outperforms other possible system combination methods. It train a classifier to learn which system should be selected for each end word, using features those shows the characteristics of the component systems. ROVER alignments on the best-1 hypothesis are used for decoding, but many features are taken from the system lattices. The classifier learns a selection strategy (i.e. a decision function) from a development set and then is able to make better selections on the evaluation data then the current 1-best or lattice-based system combination approaches. In this it uses the ROVER alignment as the basis for system combination approach. At first glance the search space used by ROVER is very limited because only the first-best hypothesis from each component system is used. But due to the less oracle error rate, normally less than half of the best system's error rate. For the production of the alignments we use a standard, dynamic programming-based matching algorithm that minimizes the global cost between two hypotheses. The local cost function is based on the time overlap of two words and is identical to the one used by the ROVER tool.

3.4 Classifier

After producing a set of features to characterize the systems, need to have a classifier which is able to train itself with these features that will decide which system will propose the final hypothesis at each slot in the multiple alignment. The target classes include one for each system and a null class (which is selected when none of the system outputs are chosen, i.e. a system insertion). The training data begins with the multiple alignments of the hypothesis systems, which is then aligned to the denotation words. The acquisition target for each slot is the set of systems which match the reference word or the null class if no systems match the reference word. Only slots where there is disagreement between the systems' 1-best hypotheses are included in training and testing.

The preferred classifier for such work is Boostexter (Schapire

and Singer, 2000) using real Adaboost. MH with logistic loss (which outperformed exponential loss in preliminary experiments). Boostexter trains a series of weak classifiers, while also regularly changing the weights of each training sample such that examples that are harder to classify receive more weight. The weak classifiers are then combined with the weights learned in training to predict the most likely class in testing. The main dimensions for model tuning are feature selection and number of iterations, that are choosed on the evolution set.

4. EXPERIMENT

4.1 System Implementation

Development of purposed system need to make a start with collecting the data; Then audacity is used for the purpose of recording. The specification of speech file is 16 KHz sampling rate with 14-16 bits/sec and mono channel, Next phase is preprocessing and extraction of features. In the use of HTK toolkit, HCOPY command is used for preprocessing and feature extraction purpose with a separate configuration file defining parameters for each feature extraction technique. HINIT command is used for acoustic modeling to initialize HMMs and in a separate file prototype is define for each phone, initialization is an important step because successive iteration depends on this step, here need to have stage own topology and number of states can be defined in a prototype file. HREST is used to re–estimate HMMs. For changing the standard grammar to HTK Standard Lattice Format (SLF), HPARSE command is used, because in SLF each word instance and each word-to-word transition is listed explicitly. HTK provides a command called HVITE to decode direct audio input.

4.2 Pre-making of data

We need to develop our own corpus, due to the unavailability of speech and text corpus because of less initiative for Indian languages by the researchers and for the purpose of recording we used unidirectional microphone. For word detection a sample is taken every 10 milli-seconds. The sound recorder takes the input from the microphone, saves these audio files in the .WAV format and finally forwards them to the succeeding module. It supports conversion of various factors like the sampling rate, the number of channels and the size of the sample as well. For the more than 250 words, data is recorded using above told microphone its distance from speaker is around 7 cm, and recording had been done in clean environment. For training the system voices of eight persons (4 males, 4 females) is used. Every word was recorded five times.

Performance is measured by the help of;

$$\text{Percentage of Correct Words} = \frac{N-D-S}{N} * 100$$

Where N is the total number of words in the test set, D is the number of deletions, S number of substitutions. The *Accuracy evaluation* is computed as:

$$\text{Percentage of Accuracy} = \frac{N-D-S-I}{N} * 100$$

Where I is the number of insertions. The performance of ASR systems in terms of word error rate is evaluated as:

$$WER = \frac{S+D+I}{N} * 100$$

4.3 Result

The proposed system has been tested in clean environment conditions with recorded corpus samples speakers and speakers which does not have their samples. On the basis of performance analysis, the percentage of correct word recognition of the combined system is found more than 97% which is better than previous ASR systems. So, this combination technique gives better result than previous system.

5. CONCLUSION

We have proposed a modified Hindi system over a previously developed system by traditional ROVER technique. We used iROVER, an improved technique for this system. This combination of ASR system used PLP and MFCC method for feature extraction. The proposed result is that this improved system will produce better recognition performance by giving more than 97% accuracy when trained for more than 250 words.

REFERENCES

[1] Evermann and P. Woodland. 2000. Posterior probability decoding, confidence estimation and system combination.In *NIST Speech Transcription Workshop*.

[2] Hoffmeister, T. Klein, R. Schl¨uter, and H. Ney. 2006. Frame based system combination and a comparison with weighted ROVER and CNC. In *Proc. ICSLP*.

[3] Stolcke, H. Bratt, J. Butzberger, H. Franco, V. Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, J. Zheng, and F. Weng. 2000. The SRI March 2000 Hub-5 co nversational speech transcription system. In *NIST Speech Tran Scription Workshop*.

[4] D. Hillardy, B. Hoffmeisterz, M. Ostendorfy, R. Schl¨ uterz, H. Neyz, iROV ER: Improving System Combination with Classification.

[5] Rong Zhang and Alexander I. Rudnicky, Investigations of Issues for Using Multiple Acoustic Models to Improve Continuous Speech Recognition.

[6] Malay Kumar, R K Aggarwal, Gaurav Leekha and Yogesh Kumar, Ensemble Feature Extraction Modules for Improved Hindi Speech Recognition System.

[7] J. Fiscus, "A post-processing system to yield reduced word error rates: Recog nizer output voting error reduction (ROVER)", In Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU"97), Santa Barbara, 1997, pp. 347–352.

[8] M. Kumar, A. Verma, and N. Rajput, "A Large Vocabulary Speech Recogni tion System for Hindi," Journal of IBM Research, Vol. 48, 2004, pp. 703-715.

[9] Hidden Markov Model Toolkit (HTK-3.4.1): http://htk.eng.cam.ac.uk.

[10] Chalapathy Neti*,Nitendra Rajput, Ashish Verma, A Large Vocabulary Con tinuous Speech Recognition System for Hindi.

[11] Aggarwal, R K and Dave, M (2010) Fitness Evaluation of Gaussian Mixtures in Hindi Speech Recognition System, First International Conference on Inte grated Intelligent Computing, SJB Institute of Technology, Bangalore.

[12] F. Reena Sharma and S. Geetanjali Wasson, Speech Recognition and Synthe sis Tool: Assistive Technology for Physically Disabled Persons ISSN 2047-3338.

[13] Kuldeep Kumar and R. K. Aggarwal, Hindi Speech Recognition System Using HTK, International Journal of Computing and Business Research ISSN (Online) : 2229-6166.