Implementation of Efficient Speech Recognition System on Mobile Device for Hindi and English Language

Gulbakshee Dharmale[1], V. M. Thakare[3]
Research Scholar[1], Professor[3]
Computer Science Department
SGB Amravati University
Amravati, INDIA

Dipti D. Patil[2]
Associate Professor
Information Technology Department
MKSSS's Cummins College of Engineering for Women
Pune, INDIA

***Abstract*—Speech recognition or speech to text conversion has rapidly gained a lot of interest by large organizations in order to ease the process of human to machine communication. Optimization of the speech recognition process is of utmost importance, due to the fact that real-time users want to perform actions based on the input speech given by them, and these actions sometime define the lifestyle of the users and thus the process of speech to text conversion should be carried out accurately. Here`s the plan to improve the accuracy of this process with the help of natural language processing and speech analysis. Some existing speech recognition software's of Google, Amazon, and Microsoft tend to have an accuracy of more than 90% in real time speech detection. This system combines the speech recognition approach used by these softwares and joined with language processing to improve the overall accuracy of the process with the help of phonetic analysis. Proposed Phonetic Model supports multi-lingual speech recognition and observed that the accuracy of this system is 90% for Hindi and English speech to text recognition. The Hindi WordNet database provided by IIT Mumbai used in this research work for Hindi speech to text conversion.**

Keywords*—Automatic Speech Recognition (ASR); Mel Frequency Cepstral Coefficient (MFCC); Vector Quantization (VQ); Gaussian Mixture Model (GMM); Hidden Markov Model (HMM); Receiver Operating Characteristics (ROC)

I. INTRODUCTION

Automatic speech recognition (ASR) has taken a big leap in recent years. Companies like Google, Amazon, Microsoft, Apple and many others have developed complicated speech recognition algorithms to improve the accuracy of speech recognition and to reduce error rates and delays to a sufficiently low level, such that these systems can be used in real time scenarios like while driving a car, or places where typing not possible and the user needs to communicate verbally with the device. Google's Assistant, Samsung's Bixby, Apple's Siri, Amazon's Alexa and Microsoft's Cortana are examples of such high accuracy systems.

System designers have picked up a concept of ASR further, by adding context-sensitive support into ASR, by which the system can not only recognize the voice, but also act on the commands provided by the user based on existing and previous user interactions which the device. For example, if a speaker asks the device "Who is the PM of India", then the device will respond with "Mr. Narendra Modi", and if speaker continues asking, "What is his age?", then the device will respond "67 years", thus the ASR system understands that the "he" in the context is "Mr. Narendra Modi". This is just one example of modern-day ASR systems, and these systems can do a lot more than just answering simple queries. They can be used to set up reminders, do automatic restaurant bookings, check flight status and much more.

The accuracy of recognition of such systems is high and can classify the input voice data into text data with more than 95% accuracy in real-time environments. That means, there`s limited scope for further improvement in terms of raw speech to text conversion from a research point of view. But the accuracy of these systems can be further improved with the help of phonetic analysis. Phonetics is a field of audio processing to text, transliteration, where similar sounding output text is produced for a given input voice data. Suppose that sentence is, "What is the plane status?", this sentence for a traveler will mean, "What's the flight status?", while for a person who wants to know the status of the plain surface (like archeologists), will mean "What's the plain status?". Such examples are where phonetic comes into play. Researchers have studied phonetics and their applications into ASR, and have tried to improve the accuracy of such systems, in this paper, trying to perform the same task, but with a more advanced Hidden Markov model (HMM) based method [1]. Automatic Speech Recognition carried out in two phases, training and testing phase. In the training phase of automatic speech recognition, parameters of the categorization model are projected using a large number of training classes. The features of a test speech are mapped with the skilled speech model of each class in testing phase.

The next section describes the recent techniques for ASR, and how they have improved over the years, followed by our proposed phonetic model for improving the speech to text contextual quality, and concluded from the results and some interesting observations of the system.

II. Related Work

Speech recognition is an emerging area of research with different methodologies to get a high level of precision. MFCC based arrangement of 39 measurements is a standout amongst the most utilized methodologies for extricating the component from the organized information signals [2]. Henceforth, need to investigate the impact of the MFCC based arrangement of 52 measurements since it isn't utilized for the procedure of extraction yet.

Gaussian Model and MFCC based arrangement of 39 measurements are utilized for creating Bangladeshi Dialect Recognition. Iterative Expectation-Maximization (EM) calculation is used to prepare this framework [3]. This framework was tried in different areas of Bangladesh like Barisal, Noakhali, Sylhet, Chapai Nawabganj and Chittagong. There was comparatively performed between Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). The outcomes demonstrated an exactness of 100% with GMM adjustment with MFCC of 39 measurements.

An ASR framework for Bangla letters in order was exhibited in 2014. In this framework, a little-measured database was shaped in which the extricated highlights were spared as reference layouts with the assistance of MFCC39 based framework [4]. After the separated highlights are spared, Dynamic Time Warping (DTW) calculation was utilized for the examination of constant information coefficients and the spared ones. To build the exactness of the yield of DTW, K-Nearest Neighbors (K-NN) was utilized. This gave it a precision of 90%.

In Curvelet based technique Automatic Speech recognition in the noisy environment along with input speech signals disintegrated at various other frequencies, channels, applying available descriptions of Curvelet conversion to reduce estimated obstacles and size of feature vectors. Also, it has more accuracy, fluctuating size of a window because they observed much suitable for immobile signals [5]. The distinct Hidden Markov model can be used for better speech recognition and classification also it considers as time distribution of word signal. HMM Method achieved maximum accurate output in terms control 63.8% recognition rate, scientific phrases 86% and the identification rate is 80.1%. The arithmetic output describes that signal recognition precision improves by using isolated Curvelet transforms than other regular methods.

Nikita Dhanvijay [6] presents an automatic speech recognition system for the Hindi language. This system takes the Hindi audio along with its textual labels as input and converts spoken word or sentence to the text. Data is collected from 20 speakers with two iterations. These recorded data are trained by acoustic model. This model was trained for a vocabulary size of 45 words by 20 speakers. This system uses the HTK toolkit to train the input data and estimate results. This system shows recognition rate 98.09 % for word and about 94.28 % for sentence.

Rajat Haldar [7] implemented Multilingual Speech recognition system. This research work is divided into two steps. In the first stage, Artificial Neural Network is used for speech recognition and Language Recognition of Chhattisgarhi, Bengali, Hindi, and English speech signal. In the second stage, the combination of Particle Swarm Optimization (PSO) technique and Artificial Neural Network is used for Speech recognition and Language Recognition of Chhattisgarhi, Hindi, Bengali, and English speech signal. Then comparison has done based on error and recognition rate. Speech recognition and Language recognition have done by using Radial Basis Function Neural Network (RBFNN). Multilingual Language Recognition with PSO gives excellent end result as compared to without PSO. Likewise, in Speech Recognition with PSO gives very good results as compared to without PSO.

Gaurav Kumar Leekha [8], has developed speech recognition systems to attain high performance in the perspective of Indian languages mostly Hindi. In this work, HTK open source tool kit is used to present the practical problems in constructing HMM. Three basic steps are used to implement HMM with HTK.

The First step is recording signals with recording software like Audacity or HSLab. After signal recording feature extraction is done using MFCC. In this step recorded.

WAV file converted into MFCC by applying Hcopy command. In the third step, HMM training is performed. Dictionary preparation and transcript preparation are the most important step. The accuracy of a system is evaluated by varying vocabulary size. Outcomes indicate that accuracy is more for smaller vocabulary size.

Ms. Jasleen Kaur [9] developed an ASR system that can recognize the English words in English pronunciation used by Punjabi people. In this research work, an Acoustic model and language model has developed for commonly used English words in north-west Indian English pronunciation.

Implementation of this work involves the preparation of data and phonetic dictionary along with the development of acoustic and language model. In data preparation step, speech recordings of 500 frequently spoken English words are collected from 76 Punjabi speakers. The text corpus consists of grammar for 500 English words. After this, Phonetic transcriptions are used for developing a phonetic dictionary. Then, an acoustic model and language model are developed. The CMU Sphinx system supports in training and recognition stage. If this system is trained by 128 GMMs then best performance of it is 85.20 %.

Malay Kumar [10], implemented a new advance in the Hindi speech recognition system by assembling different feature extraction techniques of ASR systems such as PLP, MFCC, LPCC then combines an output of it used voting technique ROVER. Each feature extraction techniques have some merits and faults in different conditions and environments. By assembling these feature extraction techniques in one system, the implemented system will execute better in noisy and clean environments. Experimental results have been shown that the combination system is better than the individual ASR system also observed improved performance against traditional ASR systems. Table 1

describes a comparative study on the accuracy of different speech recognition techniques.

TABLE I. COMPARATIVE STUDY ON THE ACCURACY OF DIFFERENT ASR TECHNIQUES

Sr. No.	References.	Feature extraction Techniques	Classification Techniques	Accuracy of Recognition (%)
1	[3]	MFCC39	HMM-Based Classifier	98
2	[4]	MFCC	HMM Based Classifier	89.7
3	[5]	MFCC	HMM	80.1
4	[6]	MFCC	HMM with GMM	94.28
5	[7]	Radial Basis Function Neural Network (RBFNN)	Artificial Neural Network	95
6	[8]	MFCC	HMM	77
7	[9]	MFCC	GMM	85.20
8	[10]	MFCC, PLP, and LPCC	HMM	96

III. PROPOSED WORK

The block diagram of the proposed work can be shown as in Fig. 1. First, the input speech is given to a standard speech to text (STT) conversion engine like Google STT, Amazon STT, Microsoft STT or Apple STT. In this experiment, Google's and Amazon's STT are found easy to use and give very good accuracy after text conversion.

Once the converted text is obtained, it is passed to the contextual HMM engine. The contextual HMM engine contains a probabilistic connected, trained Markov map of the user's specific words.

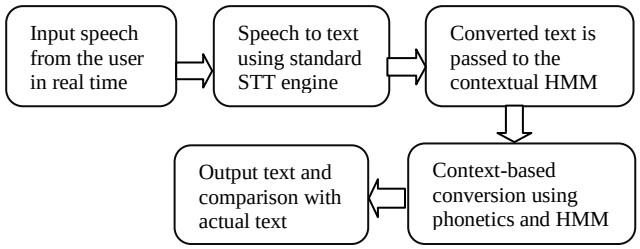

Fig. 1. Architecture of Proposed Phonetic Model.

TABLE II. A SAMPLE HMM TRAINED MARKOV MAP

Base word	Continuing word	Trained probability
Student	Marks	0.9
Student	Mass	0.2
Python	Code	0.8
Python	Snake	0.4
Java	Location, code	0.8
Java	Location, course	0.2
Microsoft	Code	0.5
Microsoft	Course	0.8

An example of such a connected HMM model is shown in Table 2. From this table, it can be seen that for faculty, words like "student marks" have a higher probability of occurrence than "student mass", which indicates that even though the words "marks" and "mass" sound similar phonetically and might be mixed while speech to text conversion, but the developed system would know that the user needs to know the "student marks", and has little interest in "student mass". Similar comparisons are trained in the database, and the system is made to self-learn from correctly converted words so that the training needed is not very large in terms of data collection, and the system remains lightweight to be applicable in real time situations.

Suppose there is a speech to text conversion with 'n' output words, w1, w2, w3 ... wn. Then for each bi-gram, tri-gram, and quad-grams, it could find the matching phonemes and their respective probabilities of the match from the generated table. Assume each word has 'k' random phonemes; hence there are ' n x k ' different combinations of words for the given sentence. Also for bi-gram, tri-gram and quad-grams,

'n x k x 2 ', ' n x k x 3 ' and ' n x k x 4 ' combinations obtain respectively. For each combination of the sentence, the HMM probability is evaluated. The max probability of each combination is determined and then used for the particular bi-gram, tri-gram, and quad-gram. These changed grams are then again put in the sentence and probabilities are re-evaluated. This process continues until to obtain highest and saturated probabilities for each of the grams. These grams then connected in order to get the final processed sentence which is given in the output to the user. This increases the delay of processing, but due to the recent power and speed upgrades in Smartphones, the delay is minimal and is infinitesimal in real time experience.

The other interesting part, the algorithm self-updates the HMM probability map, by checking user's response to a converted text. Example, if the converted text is correct, the user does not manually change it, but if the text is incorrect, then the user would manually correct it, and these corrections or non-corrections are used in order to update the HMM probability map.

The results shown in the next section indicate that the proposed system improves the accuracy of real-time speech to text conversion by 7%. The paper is then concluded by making some interesting observations about the proposed method, and ways to improve the system's performance in the future using advanced techniques like Deepnets.

IV. RESULT AND ANALYSIS

The proposed phonetic system has tested on a high-end One Plus 5T android Smartphone, on a moderate specification, Samsung Galaxy A9 Pro, and on a lower specification, Samsung Galaxy Grand Smartphone, and evaluated different real-time sentences on all 3 devices, then evaluated the mean delay and mean accuracy of the system. This accuracy is then compared with the standard results of GMM technique, and tabulated in Table 3 as follows:

TABLE III. COMPARISON OF THE ACCURACY AND DELAY IN SPEECH RECOGNITION WITH PROPOSED PHONETIC MODEL AND GMM

No. of Words	ASR delay by GMM (ms)	ASR delay by Proposed phonetic model (ms)	Accuracy with GMM	Accuracy with Proposed phonetic model (%)
5	0.023	0.15	100.00	100.00
7	0.024	0.17	100.00	100.00
9	0.024	0.22	100.00	100.00
12	0.026	0.35	91.67	100.00
15	0.026	0.38	86.67	93.33
20	0.027	0.48	90.00	95.00

For each combination of sentences, it took 5 to 12 combinations, to normalize the results across both the comparisons. All the Smartphone's had the same network connection speed during evaluation, and the values of correctly classified by GMM and Proposed techniques are evaluated at a mean of the correctly classified values between all the combinations of words in the sentences. These combinations vary from having moderate length words to large length words in each of the sentences. The efficiency of ASR techniques improved by using the results of GMM and applying the proposed phonetic model on it.

As the developed system was trained for a particular user, thus the accuracy is better from the trained user's perspective, but might not be good for a non-trained user's perspective. In this system sstandard procedure for identification of speech using language processing is required. This algorithm for speech recognition does not have multi-lingual support.

Receiver operating characteristics (ROC) curve is plotted to explain the accuracy of GMM and proposed phonetic model [11]. ROC curve generally used in signal detection speculation to represent swapping between true positive rate and false positive rate of classification techniques.

ROC curve for GMM and Proposed phonetic model is plotted among tp rate and fp rate of respective techniques. There are four possible outcomes true positive means correctly recognized words (TP), Incorrectly recognized words are classified as true negative (TN), a word which is not spoken, but recognized it classified as false positive (FP) and spoken words but not recognized counted as false negative (FN). These four instances are used to calculate tp rate, fp rate, accuracy, sensitivity, and specificity. fp rate is plotted on the x-axis and tp rate is plotted on the y-axis in ROC curve. Different points on ROC space shows positive and negative recognition. The upper left point (0, 1) represents perfect speech recognition. ROC curve in Fig. 2 indicates the rate of correctly recognized the words of the proposed phonetic system is improved than GMM.

Sensitivity and specificity of the proposed system are described in the given Fig. 3 which is calculated by given formulae.

Sensitivity =TP/P, P=TP+FP

Specificity= 1-fp rate, fp rate=FP/P

The right most point (1,1) depicts that sensitivity and specificity are high, hence the performance of the proposed phonetic system is higher than GMM.

The results of the system are revealed in given Fig. 4 and Fig. 5 as below.

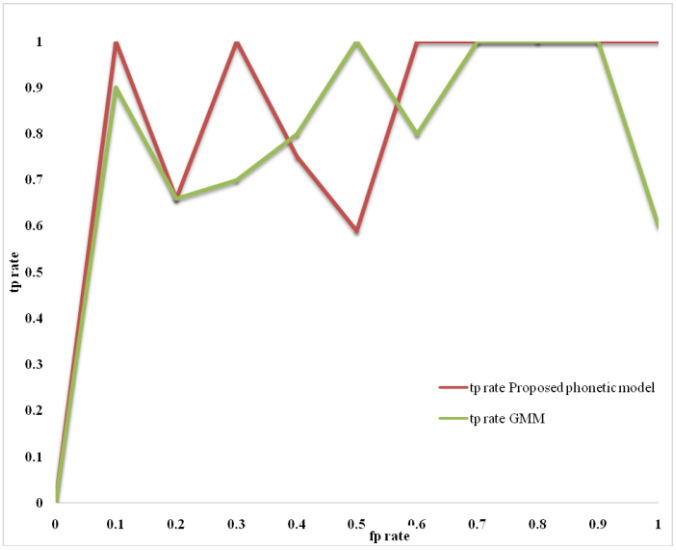

Fig. 2. ROC Curve for Proposed Phonetic Model and GMM.

Fig. 3. Comparison of Accuracy of Proposed Phonetic Model and GMM.

Fig. 4. Speech Recognition using GMM Technique.

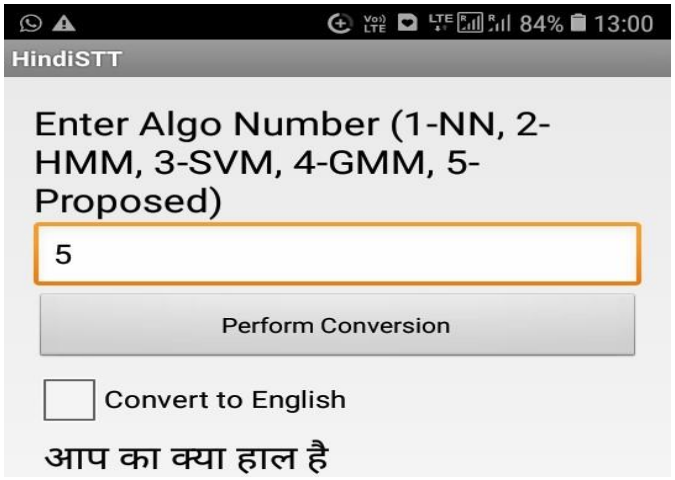

Fig. 5. Speech Recognition using Proposed Phonetic Model.

V. Conclusion and Future Scope

From obtained results, it is observed that the efficiency of existing standard STT systems can be improved by 7%, using a personalized learning HMM map model and thus can be further tweaked on a per-user basis. Proposed phonetic model achieved 90% accuracy. In the future, researchers can work on reducing in delay needed for optimization of STT engine using an HMM map model with the help of advanced artificial intelligence techniques like Deepnets, and Q Learning in order to reduce the search space and then use the optimized results in real time. These results can be tested on a wide variety of linguistics because this research was done in English and Hindi languages, researchers can extend this work on non-Indian languages like French, Chinese and others in order to check its performance and to check the practical applicability of the developed system.

References

[1] M. Manjutha, J. Gracy, Dr. P Subashini, Dr. M Krishnaveni, "Automated speech recognition system–a literature review", International Journal of Engineering Trends and Applications (IJETA), Vol. 4, No. 2, pp. 42-49, Mar-Apr 2017.

[2] S. Chapaneri," Spoken digits recognition using weighted mfcc and improved feature for dynamic time wrapping", International Journal of Computer Applications, Vol. 4, No. 3, PP. 6-12, 2012.

[3] Pronaya Prosun Das, Shaikh Muhammad Allyear, Ruhul Amin, and Zahida Rahman, "Bangladeshi dialect recognition using mel frequency cepstral coefficient, delta, delta-delta and gaussian mixture model" 8th International Conference on Advanced Computational Intelligence, Chiang Mai, Thailand, pp. 359-364, 2016.

[4] Asm Sayem, "Speech analysis for alphabets in Bangla language: automatic speech recognition" Int. Journal of Engineering Research, Vol. 3, No. 2, pp. 88-93, 2014.

[5] Nidamanuru Srinivasa Rao, Chinta Anuradha, Dr. S. V. Naga Sreenivasu,"Curvelet based speech recognition system in noisy environment:A statistical approach", IJCSIT, Vol. 10, No. 3, pp. 57-69, June 2018.

[6] Nikita dhanvijay, prof. P. R. Badadapure, "Hindi speech recognition system using mfcc and htk toolkit", International journal of engineering sciences & research technology, Vol. 3, pp. 690-695, Dec. 2016

[7] Rajat Haldar, Dr. Pankaj Kumar Mishra, "Multilingual speech recognition using radial basis function (rbf) neural network", International Research Journal of Engineering and Technology (IRJET), Vol. 3, No. 5, pp. 2856-2862, May-2016

[8] Gaurav Kumar, Leekha and Prof. R. K. Aggarwal, "Implementation issues for speech recognition techniques in the context of indian languages: a review".

[9] Ms. Jasleen Kaur, Prof. Puneet Mittal, " On developing an automatic speech recognition system for commonly used english words in indian english", International Journal on Recent and Innovation Trends in Computing and Communication, Vol.: 5, No. 7, pp. 87 – 92, 2017.

[10] Malay Kumar, R. K. Aggarwal, Gaurav Leekha and Yogesh Kumar, "Ensemble feature extraction modules for improved hindi speech recognition system", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, pp. 175-181, May 2012.

[11] Tom Fawcett, "An Introduction to ROC analysis", pattern recognition letters 27, pp.861-874, 2006.