Journal of Autonomous Intelligence (2024) Volume 7 Issue 5
doi: 10.32629/jai.v7i5.1019

ORIGINAL RESEARCH ARTICLE

Effective speech recognition for healthcare industry using phonetic system

Gulbakshee Dharmale[1,*], Dipti D. Patil[2], Tanaya Ganguly[3], Nitin Shekapure[4]

[1] *Department of Information Technology, Pimpri Chinchwad College of Engineering, Pune 411044, India*

[2] *Department of Information Technology, MKSSS's Cummins College of Engineering for Women, Pune 411052, India*

[3] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur 522302, India*

[4] *Department of Production Engineering, All India Shri Shivaji Memorial Society (AISSMS) College of Engineering, Pune 411001, India*

*** Corresponding author:** Gulbakshee Dharmale, gul12dharmale@gmail.com

ABSTRACT

The automatic speech recognition helps to achieve today's demands such as flexibility in patient care, efficiency, medical records. ASR allows more effective use and combination of process management devices and systems. Because speech interaction is contactless, they can be seamlessly combined into a current hardware environment. This paper presents the phonetic system that implemented to improve the automatic speech recognition with higher accuracy for increasing performance. The system obtains input speech by a mic then works on the tried speech to recognize the spoken word. After that, it passes the ensuing text to the HMM classifier. The HMM classifier compares occurrence of the accredited word with probability map. The word with the highest probability of occurrence gets selected. It then substitutes accredited word with this utterance; this process is carried out for the entire accredited text. The phonetic system directly obtains and translates speech to text by providing 8% improvement in the accuracy of the system. Smart text independent multi-lingual SMS system is developed using phonetic system, which allows the user to convert their voice into text and send message. STIM SMS system can offer a very spirited substitute to traditional keyboard.

Keywords: healthcare industry; Fourth Industrial Revolution; machine learning; automatic speech recognition; Hidden Markov Model (HMM); mobile computing; neural network

ARTICLE INFO

Received: 22 July 2023
Accepted: 9 October 2023
Available online: 2 April 2024

1. Introduction

The concept of automatic discourse recognition is applied for recognizing words and phrases from any spoken language and with the help of machines. This helps to convert the spoken or speech input into the machine-readable text. The speech recognition-working model is based on two things—one is an acoustic model and the other one is language modeling. Acoustic model is used to provide and establish a connection between linguistics and audio signal. After this, the language models help to match for word sequences as per sounds for differentiating between words based on sound similarities. Mostly hidden Markov get used to recognize the different temporal patterns for speech, resulting in better accuracy.

Artificial intelligence (AI) has made revolutionary advances in recent years. AI developments, in turn, owe their rapid adoption and success to the massive increase in computational capacity provided by cloud computing. It enables the creation of dependable industrial and consumer products with a high market adoption rate. As a result,

we now see and use speech applications across a variety of prominent platforms, like WhatsApp, Waze, and others.

Artificial intelligence (AI) has no clear definition; it can be thought of as a bridge between human intelligence and the desired outcomes. For AI to be sustainable, it requires the assistance of algorithms, codes, mathematical approaches, and computer programs[1]. For example, there is a lot of data accessible, and we only needed data that met particular criteria. While there is a lot of data available, physical verification of data may take a long time, and it may be prone to inaccuracies. There will be no errors, and it will be efficient, effective, and speedier.

There has been a huge improvement in different speech recognition algorithms have made it possible to accomplish high performance and computational assets. Computational assets; assume an imperative role in the speech recognition framework dependent on mobile devices; it can be categorized into two fundamental techniques. The client-server technique is the primary strategy for execution. Google Assistant, Microsoft's Cortana and SIRI are all using the client-server approach. 2nd approach attempts to perform the ASR as an individual system on the smart phones, Pocketsphinx by CMU is the most common example[2].

In the medical diagnosis system, disease detection is a time-consuming yet necessary task. Machine learning is critical for diagnosing and forecasting diseases at different stages. It is a highly ad hoc and rapid way for analyzing disease utilizing laboratory and clinical symptoms, and it aids medical professionals in formulating a more successful diagnostic strategy for such illnesses[3] Speech recognition system helps to diagnosing diseases and analyzing medical records.

In healthcare, protecting patient privacy during speech recognition, which involves sensitive medical data, is vital. Strategies include data encryption, anonymization, secure data transfer, on-device processing, and data minimization. Obtaining user consent, implementing access controls, auditing, and de-identification techniques are vital. Compliance with data protection regulations, careful selection of third-party vendors, and clear data retention policies are essential. Additionally, regular staff training ensures privacy is maintained while achieving accurate speech recognition in healthcare[4,5].

Nowadays, a lot of work on speech recognition is usually performed using conventional figuring gadgets like PCs and laptops. Many Industry 4.0 types, such as manufacturing, transportation, and logistics, as well as consumer segments including retail, smart home, healthcare, and others, will benefit from this. Wearable gadgets, small smart devices without a haptic interface, and situations where the user has both hands occupied, such as driving or managing complicated industrial machines or robots, will all benefit from smart voice[6,7].

Mobile computing permits the transmission of information within different gadgets including PCs, cell phones over a remote connection. The client can access the network and communicate with one another anywhere and whenever required. The mobile technology allowed clients to share documents, e-mail communication, audio-video conferencing, internet access, etc[8]. The mobile application is used to perform a certain task. A personal computer-PC can be used to create a program called a software application. Such application that keeps running on cell phones like iPhone, mobile phones, and tablets is called as a mobile application. These applications help to take advantage of the mobile phones and subsequently, the mobile application world is meeting new and inventive applications going from simple calculators to discourse to-content converters[9].

Accuracy of discourse acknowledgement relies on the design of a signal to deal with the background in which it is conversed or recorded. Phonetic system applied to improve the discourse acknowledgement accuracy. Phoneme is the most reduced component of phonetics; it is made through vowels just like consonants. The essential component of a word is phoneme. In the word discontinuity, the given word separated into the individual phonemes. Subsequently, the discourse identifier is just expected to perceive

specific arrangement of various phonemes inside the word as opposed to perceiving all the distinctive individual words[10].

In response to growing concerns about medical data privacy and the limitations of centralized data methods, federated learning (FL) emerges as a promising solution. This innovative work introduces two key concepts: Federated Weakly Supervised Segmentation (FedWSS) and the Federated Drift Mitigation (FedDM) framework. FedWSS focuses on segmentation tasks and leverages the benefits of FL to handle weakly supervised data, where supervision signals may be imprecise. FedDM addresses the challenges unique to FL, particularly in weakly supervised scenarios, such as local drift during client-side optimization and global drift during server-side aggregation. These challenges are effectively managed through Collaborative Annotation Calibration (CAC) and Hierarchical Gradient De-conflicting (HGD). Chen et al.[11] and Zhu et al.[12] presented this work which offers a pioneering solution for training segmentation models in a federated learning context, addressing issues related to data privacy and accuracy, which holds great promise for advancements in medical data privacy and machine learning.

Speech recognition and keyboard are the two input methods available on mobile devices to perform various tasks like sending messages, internet access and calling contacts. Use of the right input method reduces communication time. STIM SMS system is implemented using a phonetic system for Hindi and English language on mobile devices[13].

In this article, the introduction section presents ASR on mobile devices. The related work of speech recognition on mobile devices is elaborated in the section two. The third section gives detail information about the phonetic system and STIM SMS system. The fourth section describes observational results. Conclusion is stated in section five.

2. Related work

Jeeva Priya et al.[14] applied phonetic level discourse acknowledgement in Kannada using HTK to create a programmed discourse acknowledgment framework to perceive Kannada words expressed ceaselessly. It utilizes an open source discourse acknowledgment device, HTK. With an emphasis on the travel industry application, a speaker autonomous, moderate sized jargon articulation corpus on a word reference amount of 250 has been made for this reason. The discourse acknowledgment speed is checked for male and female utterers for disconnected and continues acknowledgment. It has been observed that male utterers perceived better than female utterers[14].

Bolla et al.[15] build gadget exchanging by means of voice directions. Accordingly, the emphasis is on gadget exchanging and the controlling Bluetooth modem to send an SMS if there should arise an occurrence of crisis, all constrained by voice directions. To be explicit, there are three fundamental goals in this venture. With regards to importance, the first is to plan and build a voice empowered gadget-changing framework to help physically tested and old individuals. Apart from this, it also needs to build up an android application for alarming if there should arise an occurrence of crisis and for mailing reason with the assistance of the voice acknowledgment framework. Lastly, the goal is to give a corresponding component among client and predefined number by SMS messages with the help of Bluetooth modem in the event of crises[15].

Karpagavalli and Chandra[16] used the Hidden Markov Tool Kit to develop isolated-phoneme, uttered autonomous and word acknowledgment systems for the Tamil language. With the small vocabulary size both phoneme and word reproduction provides negligible error rate and high acknowledgment accuracy.

Speech recognition systems are classified into two categories. The first level extracts features using Mel Frequency Cepstral Coefficients and Linear Predictive Cepstral Coding. Hidden Markov Models, Support Vector Machines, Neural Networks, Gaussian Mixture Models, are used in the second level of classification. After the speech signal has been effectively fragmented, classification methods are utilized to classify the

relevant fragmented words or phonemes[17–19].

Thalengala and Shama[20] developed a separate ASR for Kannada. There are two types of dictionaries: (a) syllable level dictionaries (b) phone level dictionaries. The Kanada news archive is being utilized to create pronunciation dictionaries. They achieved 60.2 percent and 74.35 percent general word recognition accuracy for monophone and triphone acoustic versions, respectively. They stated that selecting an appropriate acoustic model based on vocabulary size can improve the performance of an ASR system[20].

As of the last decade, researchers have started working on the local lingo in India. Though the research work on discourse acknowledgment for local lingo is in full function mode, it is still, in the emerging stage. Study and research on Indian languages with speech recognition accuracy and technology used is shown in **Table 1**.

Table 1. Related work on speech recognition for local lingo in India.

Name of Author and Publication Year	Local lingo	Feature extraction technique used	Selected classifier	Accuracy
Patil et al., 2016[21]	Hindi	MFCC	VQ-GMM	93%
Aggarwal and Dave, 2012[22]	Hindi	MFCC	Segmental HMM, Genomic HMM, Hybrid HMM	-
Supriya et al., 2017[23]	Marathi	MFCC	GMM-HMM	80%–90%
Narkhede and Nemade, 2018[24]	Marathi	MFCC	LPC, DWT & ANN	78%
Malewadi and Ghule, 2016[25]	Marathi	MFCC & LFZI	SVM	-
Kalamani et al., 2018[26]	Tamil	CSD-NE, SS-NE	FCM with EM-GMM	Improved accuracy from 1.2 to 4.4%
Mannepalli et al., 2015[27]	Telugu	Prosodic-NNC, MFCC-GMM	Gaussian Mixture Model	77%–80%
Bhowmik and Mandal, 2018[28]	Bengali	MFCC	Bidirectional LSTM	Accuracy of classification 98.9%
Tailor and Shah, 2016[29]	Gujarati	MFCC	HMM	95.1% and 95.9% Accuracy in noisy environment and in lab respectively
Londhe and Kshirsagar, 2018[30]	Chhatisgarhi	MFCC	HMM, ANN and SVM	Accuracy for isolated word recognition using SVM and ANN 94.24% and 99.84% respectively.
Koolagudi et al., 2017[31]	Dravidian	SDC and MFCC	ANN	Classification accuracy of Dravidian lingo such as for Malayalam Kannada, Telugu and Tamil is 72%, 73.6%, 68.8% and 65.1% respectively
Bharali and Kalita, 2018[32]	Assamese	MFCC	VQ, I-vector and HMM	90%–100%
Zia and Zahid, 2018[33]	Urdu	MFCC	Bidirectional LSTM, RNN	Word fault rate is 0.68
Guglani and Mishra, 2018[34]	Punjabi	PLP and MFCC	MPE, N-gram model	MFCC outperformance than results of PLP
Mittal and Singh, 2019[35]	Punjabi	MFCC	CD-Untied, CI CD-Tied, D_DelInterp	Highest accuracy is given by CD_Untied with rate 81%
Kadyan et al., 2018[36]	Punjabi	MFCC	DNN-GMM and HMM-GMM	Performance in DNN-GMM improved 4%–5% compared to HMM-GMM

In the healthcare industry, documentation is essential, but it doesn't have to take up valuable time. The use of speech recognition technology increases productivity among medical professionals throughout the day.

This allows medical professionals to visit more patients during the day because they won't have to stay late at work to do paperwork.

Patient care is improved as a result of increased productivity. Doctors and other medical professionals can spend more time visiting patients and concentrating on their care if they have less time to spend on paperwork. This raises the standard of care and enables medical professionals to spend more time with each patient.

Open Set Recognition (OSR) in the context of speech recognition aims to accurately identify known diseases and identify previously unseen diseases as an "unknown" class in medical scenarios. However, current OSR methods often require collecting data from various locations to create large centralized training datasets, which can pose significant privacy and security risks. These risks can be effectively reduced by adopting federated learning (FL), a widely-used cross-site training approach. In this context, we introduce the concept of Federated Open Set Recognition (FedOSR) and propose a new framework called Federated Open Set Synthesis (FedOSS) to address the primary challenge in FedOSR: the absence of samples representing unknown classes for all expected clients during the training phase. The FedOSS framework primarily relies on two components, namely Discrete Unknown Sample Synthesis (DUSS) and Federated Open Space Sampling (FOSS), to create virtual unknown samples that help in learning decision boundaries between known and unknown classes[37].

Automatic speech recognition creates notes that can be quickly added to patients' medical records by recording all spoken phrases. The accuracy of medical records is crucial, especially when several services are provided during a single visit. A misdiagnosis or inadequate treatment options may result from incomplete or erroneous medical data. The accuracy and completeness of medical records are helped by autonomous speech recognition.

3. Phonetic system

Phonetic system helps to get better accuracy of discourse acknowledgment. In this system, ASR classifier is first taught, and then, the state likelihood framework is developed. This framework is made up of chains of repeatedly occurring words associated by the means of the likelihood of mutual occurrence. Preparation corpus comprises of an enormous arrangement of words chosen from online references such as twitter, UCI printed dataset archive however other web-based social networking datasets. The word associations are then put in the record through forward ordering. Then, the record is additionally utilized for examination. Whenever the user wants to undertake another discourse test, it is first changed over into content by utilizing MFCC based HMM grouping strategy; the subsequent content is then given to the contextual HMM map. At that point, the event of the perceived words matches with the likelihood map, which has been created initially. The probabilistic associated and trained Markov map of the user's precise phrases is included in the contextual HMM map. In the sentence, "How old are you?" words like "how old" have a higher likelihood of occurrence than "how would," indicating that even if the words "old" and "would" sound alike phonetically and may be merged during dialogue recognition, the implemented system would identify that the client needs to know "how old," and is less concerned with "how would." Similar assessments are taught in the record; certainly the system is configured to self-learn from successfully transformed words, because the training required is not particularly extensive in terms of data set provision. As a result, the system remains light in weight in order to appropriate in actual state.

Assume there is a discourse to content transformation with 'm' yield words, say, x1, x2, x3, ..., xy. At that point, for each bi-gram, tri-gram, and quad-grams, it might see the coordinating phonemes also there different prospects of the equivalent from the produced table. Accept each word has 't' irregular phonemes; thus there are 't' totally various combinations of words for the given sentence. Conjointly for bi-gram, tri-

gram and quad-grams, 'm × t × 2', 'm × t × 3' and 'm × t × 4' mixtures are acquired. For each blend of the sentence, the HMM probability is assessed. The most elevated likelihood of each mixture is set, afterward utilized for the real bi-gram, tri-gram, and quad-gram. These adjusted grams are set aside from the spot inside the sentence and potential outcomes are re-thought[38]. On the off chance that the likelihood is lower than a given limit, different words coordinating the phonetic translation of the given word are chosen, which have a higher likelihood of event. The perceived word is then supplanted by this new word, furthermore the procedure is preceded for the whole perceived content. **Figure 1** depicts the architecture of phonetic system for automatic speech recognition. To increase the efficiency of ASR techniques, implemented phonetic system algorithm is explained as below:

1) Get input speech from client, label it as 'S'.
2) Speech is divided into fragments, lets those fragments are S1, S2, S3, ..., Sy.
3) Every fragment is a different spoken word.
4) Discover MFCC features of every fragment, label it as F1, F2, F3, ..., Fy.
5) These features then compared with the stored database using HMM, in result the output words W1, W2, W3, Wy.
6) The sentence is then form by combining the words.
7) Phonetic analysis is applied on expression to check the HMM probabilities.
8) Select words with highest probability and join to produce correct sentence.
9) Produce the output text.

End

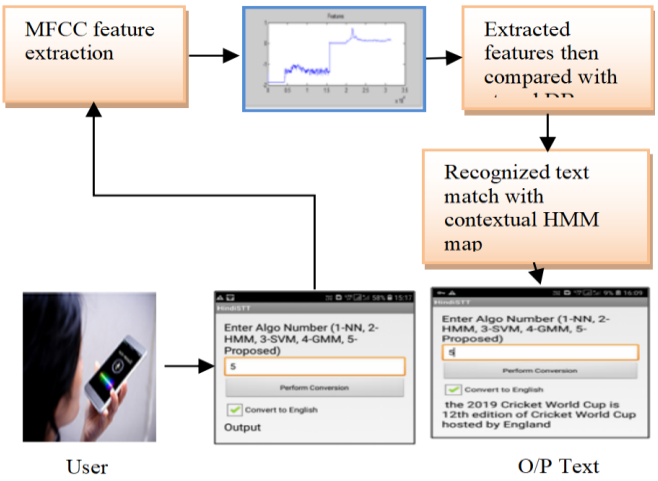

Figure 1. Automatic speech recognition process using phonetic system.

3.1. Hindi WorldNet database

Hindi WordNet is used as a database in phonetic system for Hindi language. This system manages the lexical data in provision of word implications and can communicate as a glossary based on the psycholinguistic standards. The words are clustered together based on their resemblance of meaning in the Hindi WordNet. Two words can be interchanged if they are synonymous with each other. This can be done to eliminate uncertainty in cases where homonyms or a single word has multiple meaning[38]. In this system, each entry includes the following elements where Synset is basic element of this system;

a) Synset: It is a group of similar words, i.e., synonyms which represent one lexical concept.
b) Position in ontology: Ontology is a hierarchical association of ideas, more specifically, a classification of actions and entities. Each Synset is mapped into some position in the ontology.
c) Gloss: It explains the idea about words. It is divided into two components. The first component is text definition, which describes the idea indicated by the Synset. The second component is example sentence, which explains the usage of the words in the sentence.

3.2. Smart text independent multilingual SMS system

Since the last few decades, the percentage of mobile phone users has been in a rise and everyone wants to access new applications, which are modified and better versions of present one. The SMS user might want a faster SMS application, which can convert their voice into text then send the SMS. This STIM SMS application is divided into small sub-units. A user has two options; first to choose language as either English or Hindi. After clicking on the 'select contact' tab, one can select contacts manually or second, use speech as an input. If client chooses a manual alternative, then a service is called to get all contacts in the contact list that exist in the mobile phone. In the event that the user choses speech is choose as the input method, at that point the speech to text dialog box opens to asks for 'speak now' and a mic picture is displayed. After completion of talking, the programming interface takes a couple of moments to process the information. After converting speech to text, converted text matches with HMM-map table to finds the correct sequence of words having highest probability. The process of applying phonetic system for increasing accuracy of discourse acknowledgment is explained in above algorithm. Finally, correct output text is shown in the message box. Once both the fields (select contact and message) are filled, by clicking on 'send message' button, message administration is called, thereafter SMS can be sent to the recipient.

To produce efficient output text, phonetic model is applied to recognize words by which the recognized word matches with the contextual HMM map. It searches for the highest probability of occurring phoneme in a sentence certainly forms the output text with the highest probability and proper matching of words. Implemented phonetic system is explained in details in the next section.

In the HMM mapping phase, the Hidden Markov Model is used to sample the inputted speech by applying forward calculation which is connected to advance variable and fractional perception is stored until time t with model λ. The forward factor is characterized in Equation (1)[39].

$$\alpha_t(i) = P(O_1, O_2, \dots, O_t, m_t = S_i \quad (1)$$

The Viterbi calculation is utilized to calculate the maximum probability of the given sequence, which is expressed through inward conditions of the sequence for that. It is expressed in the given formula;

$$\delta t(i) = \max_{m1,m2,\dots,mt-1} P[m_1, m_2, \dots, m_t, i, O_1, O_2, \dots, O_t|\lambda] \quad (2)$$

where O is observed sequence, mt is actual state at time t, t is the time at which the signal is sampled.

Baum-Welch calculation is used to select the factors are to calculate highest probability, which is defined in following equation.

$$\xi t(i,j) = P(mt = Si,\ mt + 1 = Sj|O,\lambda) \quad (3)$$

Once output text is ready, then it is shown on the preferred place.

4. Experimental results

Mainly, two methods are used for the evaluation of accuracy. The first method figures out the error rate of speech recognition for entire text, which was entered with speeches from a mobile device. The standard metric used for the evolution of speech recognition systems first calculates the error rate of final transcriptions of both voice and keyboard inputs, followed by measuring the word fault rate (WER) or word error rate.

Using the android speech engine, we constructed a back propagation NN, HMM, SVM, and GMM-based calculation to compare phonetic system with accuracy of existing ASR classifiers. The speech engine splits the input speech signal into segments, each segment being a user-spoken phrase, allowing the input sounds to be analyzed in real-time. The sound samples are then extracted from each of the words by an MFCC extraction unit. These characteristics are then compared to the typical IIT Bombay Hindi Word network of words. The IIT Hindi Word system has a complete list of Hindi words that were utilized as a database for this method. The feature matching procedure is carried out using the proper classification

technique, which is implemented in its standard form[40]. From experimental observation it shows that existing speech recognition system outperforms with phonetic system.

Phonetic system is used to implement smart text speaker independent SMS system for both Hindi and English languages. Experimental results of Phonetic system and STIM SMS system is explained in the next section.

4.1. Precision of phonetic system

The accuracy of the phonetic system is contrasted with the standard outcomes of GMM method. In existence implementation, the GMM exceeds other speech recognition classifiers by providing very high classification accuracy. The accuracy of GMM is considered the best because it is evaluated under real-time outdoor and indoor noisy conditions. To measure accuracy of this system, it has been tried on One Plus 7T-a high end android Smartphone, on Samsung Galaxy J6 with moderate specification, and on a Samsung Galaxy J2 having lesser specification. It, then, assessed the distinctive ongoing sentences on all gadgets, at that point assessed the deferral and exactness of the framework. In addition, note that for every sentence, it took 5 to 30 combinations of words. To standardize the outcomes across both the examinations, all the cell phones had a similar system association rate throughout assessment. The rate of effectively grouped words with GMM and phonetic framework are assessed at the mean of the accurately ordered qualities in every combination of words in the expressions. These blends change from having middle size words to enormous size words in each and every expression. In phonetic framework, a standard methodology for distinguishing proof of discourse utilizing language preparing is required. Delay, precision of phonetic system is then contrasted with output of GMM method, which is shown in **Table 2**. **Figures 2** and **3** explain the precision of phonetic system and GMM for Hindi and English language correspondingly and **Figure 4** shows the comparison of delay in speech recognition using GMM and phonetic system. This system assists multi-lingual discourse acknowledgment. It is observed that the accuracy of speech recognition is above 90% for both English and Hindi language by using this system.

Table 2. Accuracy and delay of speech recognition between GMM and phonetic system.

No. of Words	ASR delay by phonetic system (ms)	ASR delay using GMM (ms)	ASR Accuracy of phonetic system (%)	ASR Accuracy of GMM (%)
5	0.015	0.023	100.00	100.00
7	0.017	0.024	100.00	100.00
9	0.022	0.024	100.00	100.00
12	0.023	0.026	100.00	91.67
15	0.028	0.026	93.33	86.67
20	0.027	0.027	95.00	90.00

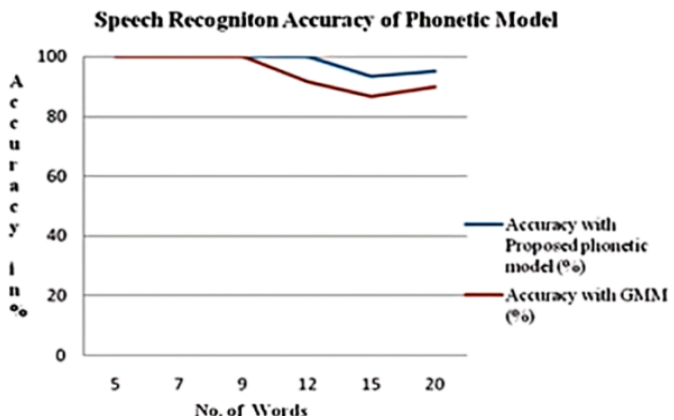

Figure 2. Evaluation of ASR accuracy of GMM and phonetic system for Hindi language.

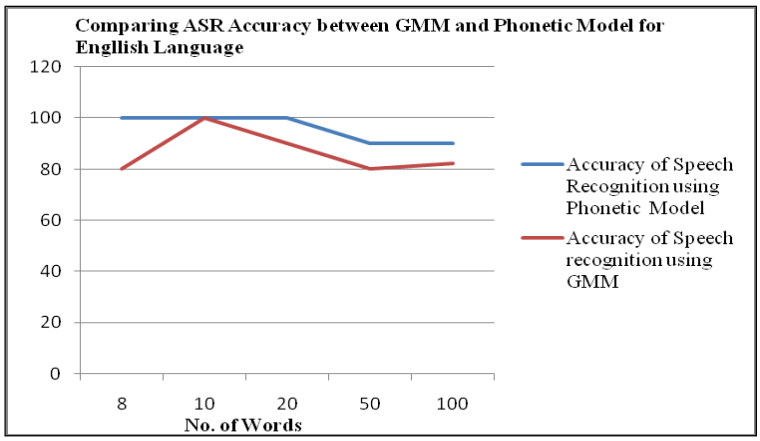

Figure 3. Evaluation of ASR precision of GMM and phonetic system for English language.

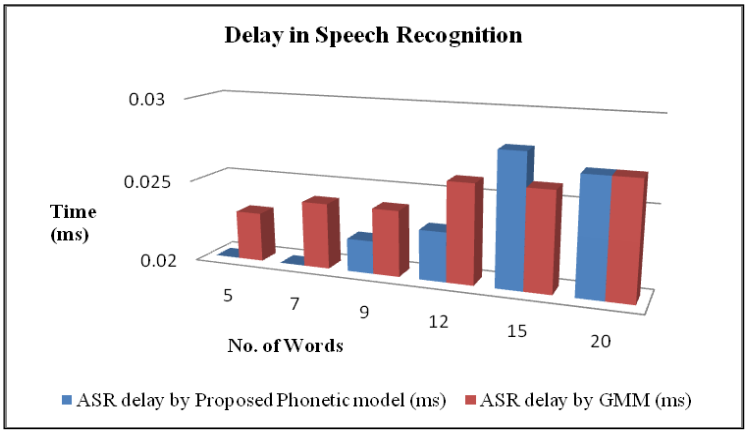

Figure 4. Comparing delay in ASR between phonetic system and GMM.

It took 5 to 12 sentences for each combination of sentences to normalize the data across both comparisons. During the evaluation, network connection speed of all smartphones was the same, the values of successfully categorized by GMM and phonetic system are evaluated as the mean of the correctly classified values across all combinations of words in the sentences. These combinations of words range from modest to long in each of the sentences. The phonetic system improves the efficiency of automatic speech recognition algorithms.

4.2. ROC curve

Recipient operating characteristics (ROC) curve is planned to clarify precision of phonetic framework and GMM. The ROC curve is commonly utilized in signal identification hypothesis to speak to trading among fault positive rate, true positive rate of GMM and phonetic system[41]. The ROC curve for phonetic system and GMM is plotted using four potential results, effectively perceived words, for example true positive (TP), words that are incorrectly perceived are called true negatives (TN), words that are not verbally articulated but recognized are called false positives (FP), hence spoken words that are not perceived are called false negatives (FN). These four variables are used to calculate the true positive rate, false positive rate, exactness, affectability, and explicitness. In the ROC curve, the TP rate is plotted on the Y-axis and the FP rate is plotted on the X-axis. Different ROC spotlights show positive and negative affirmation. The flawless conversation acknowledgment is represented by the upper left point (0, 1). Rate of accurately perceived word by phonetic system and GMM is classified, then the ROC curve is plotted in **Figure 5**. It is clear from the figure that the rate of effectively perceived expressions of the phonetic framework is superior to GMM.

The uppermost point (1,1) represents heigest affectability and explicitness, hence phonetic framework is better in performance than GMM.

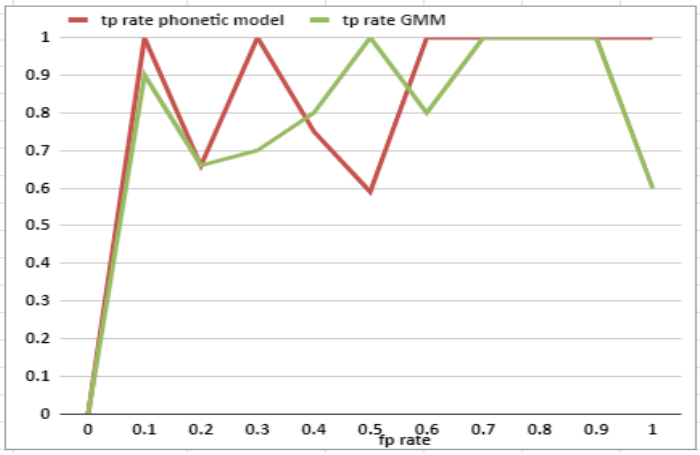

Figure 5. Recipient operating characteristics curve for phonetic system and GMM.

4.3. Result of STIM SMS system

This STIM SMS system is tested on different android mobile devices such as Samsung note, Samsung J6, Samsung J2 to send messages. The result is obtained by comparing with message input by keyboard in contrast giving input using speech recognition. Accuracy beyond the calculated number can be determined for positive authentication[40]. The following formula is used to calculate accuracy;

$$A = \frac{Correctlyrecognizedword}{Totalnumberofword} \times 100\% \quad (4)$$

Accuracy and speed are used to measure performance of speech recognition. Word fault rate is a general measure of machine learning or ASR system. It can be calculated as:

$$WER = \frac{(R + P + Q)}{N} \quad (5)$$

where R is the no. of replacement, P is the no. of removal, Q is the no. of inclusion, and N is the no. of words in the reference.

Accuracy of STIM SMS system for Hindi and English language is calculated for 2 to 100 words sentences on different android mobile is depicted in **Figure 6**.

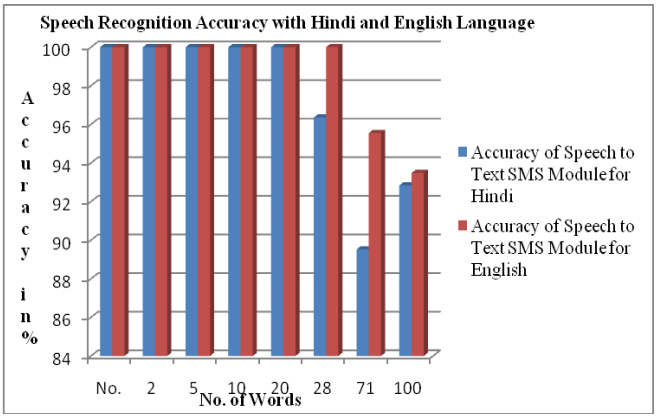

Figure 6. Accuracy of STIM SMS system for Hindi and English language.

It is observed that accuracy of STIM SMS system for Hindi and English is 97.34% and 99.25% respectively as given in **Figure 6**. Time taken to enter and edit SMS using speech recognition decreases as compared to typing SMS using keypad with 99.25%. Comparison of time taken to enter input by STIM SMS system and keyboard is shown in **Figure 7**.

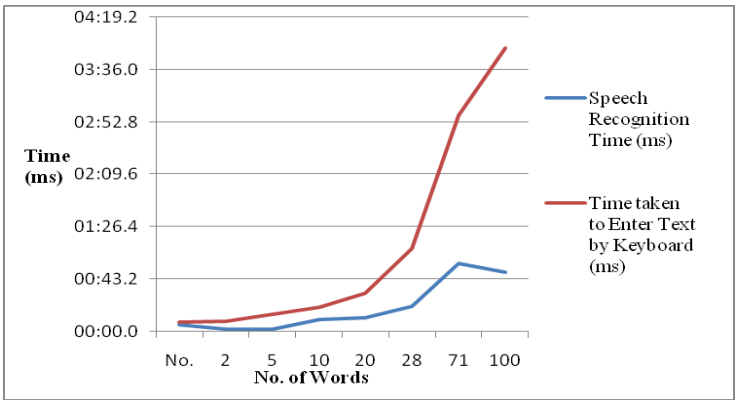

Figure 7. Comparison of time taken to enter input by STIM SMS system and keyboard.

The vowel formants for men, ladies and youngsters differ significantly from one another. The varieties do freely relate to converse vocal tract length, at any rate for certain vowels. The vowels, for which the distinctions in formants didn't relate to vocal tract length, are high front vowels, for instance, which have a Helmholtz reverberation in the pharygial hole, which has resonances at the frequencies which just relies upon one over the square base of the volume of the cavity, and the area parameters which is defined as equation given below.

$$Frequency = \sqrt{\frac{Area}{Length + Volume}} \tag{6}$$

Various speakers with different age group and gender used STIM SMS system, however the result is shown in the given **Table 3** and **Figure 8**. From results, it is stated that the STIM SMS system works faster however accurate in any environment also is independent of the speaker, i.e., it is robust and speaker independent.

Table 3. ASR Accuracy of STIM SMS system with different speakers.

Age and gender of different speaker	ASR accuracy of English language (%)	ASR accuracy of Hindi language (%)
6 years girl	100	96
40 years men	100	95
30 years women	99.7	96
66 years men	98	94
65 years women	100	95

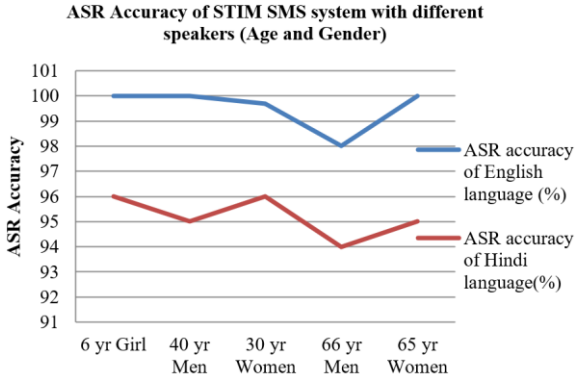

Figure 8. Speech recognition accuracy for different speakers.

4.4. Real-life application

Automatic speech recognition is the most natural mode of human communication. There are some fields

of applications explained as below. Implemented phonetic system is speaker independent. It can be utilized by a much broader number of individuals hence does not require training before usage in any context. It's used in customer support, where an Integrated Voice Response (IVR) technology allows users to navigate menus using voice. Virtual assistants, in-car navigation, likewise other scenarios in which a computer needs to understand a standard set of commands such as "turn up the music," "what is the weather tomorrow," or "where is the nearest pharmacy" typically employ speaker-independent technology. Even compound jobs can be assigned easily and quickly to robots which understand the meaning of input speech.

5. Conclusion

Implemented phonetic system gives precision for both numerical and alphanumerical information tested for SMS applications. It is observed that there is an 8% improvement in evaluating accuracy of the speech recognition. Accuracy of speech recognition increases by further mapping converted text to HMM-Map produces more accurate results. It is observed that speech recognition accuracy for Hindi and English language is 97.34% and 99.25% respectively. Phonetic system is useful in various segments of Industry 4.0 due to its accuracy.

In future, phonetic system can be implemented for multiple national and international languages. Speech recognition time can be minimizing by reducing delay in speech recognition.

Author contributions

Conceptualization, GD and DDP; methodology, GD; software, GD; validation, GD and DDP; formal analysis, GD; investigation, GD; resources, GD; data creation, GD and DDP; writing—original draft preparation, GD; writing—review and editing, NS and TG; visualization, GD; supervision, DDP; project administration, GD; funding acquisition, GD, DDP, TG and NS. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. C Chatterjee I. Artificial Intelligence and Patentability: Review and Discussions. International Journal of Modern Research 2021; 1: 15-21.
2. Lee C. Speech Recognition and Production by Machines. International Encyclopedia of the Social & Behavioral Sciences (Second Edition). Elsevier, Oxford; 2015.pp. 695-702.
3. Singh PD, Kaur R, Dhiman G, et al. BOSS: A new QoS aware block chain assisted framework for secure and smart healthcare as a service. Expert Systems. 2021, 40(4). doi: 10.1111/exsy.12838
4. Lokhande MP, Patil DD, Patil LV, et al. Machine-to-Machine Communication for Device Identification and Classification in Secure Telerobotics Surgery. Chakraborty C, ed. Security and Communication Networks. 2021, 2021: 1-16. doi: 10.1155/2021/5287514
5. Shekapure S, Shekapure N, Dharmale GJ, etal.Clinical Data Analysis of Patient and Recommendation. NeuroQuantology. 2022, 20(6): 6148-615. doi: 10.14704/nq.2022.20.6. NQ22619
6. Shinde AV, Patil DD. A Multi-Classifier-Based Recommender System for Early Autism Spectrum Disorder Detection using Machine Learning. Healthcare Analytics. 2023, 4: 100211. doi: 10.1016/j.health.2023.100211
7. Bobde SP, Mantri ST, Patil DD, et al. Cognitive Depression Detection Methodology Using EEG Signal Analysis. Advances in Intelligent Systems and Computing. Published online 2018: 557-566. doi: 10.1007/978-981-10-7245-1_55
8. Singh P, Kaur R. An integrated fog and Artificial Intelligence smart health framework to predict and prevent COVID-19. Global Transitions. 2020, 2: 283-292. doi: 10.1016/j.glt.2020.11.002
9. Reddy BR, Mahender E. Speech to text conversion using android platform. International Journal of Engineering Research and Applications (IJERA) 2013; 3(1): 253-258.
10. Zhu M, Liao J, Liu J, et al. FedOSS: Federated Open Set Recognition via Inter-client Discrepancy and Collaboration. IEEE Transactions on Medical Imaging. Published online 2023: 1-1. doi:

10.1109/tmi.2023.3294014
11. Chen Z, Yang C, Zhu M, et al. Personalized Retrogress-Resilient Federated Learning Toward Imbalanced Medical Data. IEEE Transactions on Medical Imaging. 2022, 41(12): 3663-3674. doi: 10.1109/tmi.2022.3192483
12. Zhu M, Chen Z, Yuan Y. FedDM: Federated Weakly Supervised Segmentation via Annotation Calibration and Gradient De-Conflicting. IEEE Transactions on Medical Imaging. 2023, 42(6): 1632-1643. doi: 10.1109/tmi.2023.3235757
13. Dharmale G, Thakare V, Patil D D. Intelligent hands free speech based SMS system on Android. 2016 International Conference on Advances in Human Machine Interaction (HMI). Published online March 2016. doi: 10.1109/hmi.2016.7449177
14. Jeeva Priya K, Sree SS, Navya TVS, et al. Implementation of Phonetic Level Speech Recognition in Kannada Using HTK. 2018 International Conference on Communication and Signal Processing (ICCSP). Published online April 2018. doi: 10.1109/iccsp.2018.8524192
15. Bolla DR, Shivashankar, Pavan TS, et al. Voice enabled gadget assistance system for physically challenged and old age people. 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). Published online May 2017. doi: 10.1109/rteict.2017.8256966
16. Karpagavalli S, Chandra E. Phoneme and word based model for tamil speech recognition using GMM-HMM. 2015 International Conference on Advanced Computing and Communication Systems. Published online January 2015. doi: 10.1109/icaccs.2015.7324119
17. Dharmale G, Patil DD, Vilas M. Thakare Implementation of Efficient Speech Recognition System on Mobile Device for Hindi and English Language. International Journal of Advanced Computer Science and Applications 2019; 10(2): 83-87.doi: 10.14569/ijacsa.2019.0100212
18. Patil DD, Wadhai V M. Real-Time Meta Learning Approach for Mobile Healthcare. Advances in Intelligent Systems and Computing. Published online November 20, 2018: 11-23. doi: 10.1007/978-981-13-2414-7_2
19. Dharmale G, Shirsath P, Shinde A, et al. REMICARE—Medicine Intake Tracker and Healthcare Assistant. Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Published online 2023: 273-283. doi: 10.1007/978-981-19-6088-8_25
20. Thalengala A, Shama K. Study of sub-word acoustical models for Kannada isolated word recognition system. International Journal of Speech Technology. 2016, 19(4): 817-826. doi: 10.1007/s10772-016-9374-0
21. Patil UG, Shirbahadurkar SD, Paithane AN. Automatic Speech Recognition of isolated words in Hindi language using MFCC. 2016 International Conference on Computing, Analytics and Security Trends (CAST). Published online December 2016. doi: 10.1109/cast.2016.7915008
22. Aggarwal RK, Dave M. Integration of multiple acoustic and language models for improved Hindi speech recognition system. International Journal of Speech Technology. 2012, 15(2): 165-180. doi: 10.1007/s10772-012-9131-y
23. Supriya S, Handore SM. Speech recognition using HTK toolkit for Marathi language. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). Published online September 2017. doi: 10.1109/icpcsi.2017.8391979
24. Narkhede A, Nemade MU. Efficient Method for Isolated Marathi Digits Recognition using DWT and Soft Computing Techniques. 2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU). Published online February 2018. doi: 10.1109/iot-siu.2018.8519893
25. Malewadi D, Ghule G. Development of Speech recognition technique for Marathi numerals using MFCC & LFZI algorithm. 2016 International Conference on Computing Communication Control and automation (ICCUBEA). Published online August 2016. doi: 10.1109/iccubea.2016.7860099
26. Kalamani M, Krishnamoorthi M, Valarmathi RS. Continuous Tamil Speech Recognition technique under non stationary noisy environments. International Journal of Speech Technology. 2018, 22(1): 47-58. doi: 10.1007/s10772-018-09580-8
27. Mannepalli K, Sastry PN, Suman M. MFCC-GMM based accent recognition system for Telugu speech signals. International Journal of Speech Technology. 2015, 19(1): 87-93. doi: 10.1007/s10772-015-9328-y
28. Bhowmik T, Mandal SKD. Manner of articulation based Bengali phoneme classification. International Journal of Speech Technology. 2018, 21(2): 233-250. doi: 10.1007/s10772-018-9498-5
29. Tailor JH, Shah DB. Speech Recognition System Architecture for Gujarati Language. International Journal Computer Applications. 2016, 138(12): 28-31. doi: 10.5120/ijca2016909049
30. Londhe ND, Kshirsagar GB. Chhattisgarhi speech corpus for research and development in automatic speech recognition. International Journal of Speech Technology. 2018, 21(2): 193-210. doi: 10.1007/s10772-018-9496-7
31. Koolagudi SG, Bharadwaj A, Srinivasa Murthy YV, et al. Dravidian language classification from speech signal using spectral and prosodic features. International Journal of Speech Technology. 2017, 20(4): 1005-1016. doi: 10.1007/s10772-017-9466-5
32. Bharali SS, Kalita SK. Speech recognition with reference to Assamese language using novel fusion technique. International Journal of Speech Technology. 2018, 21(2): 251-263. doi: 10.1007/s10772-018-9501-1
33. Zia T, Zahid U. Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. International Journal of Speech Technology. 2018, 22(1): 21-30. doi: 10.1007/s10772-018-09573-7
34. Guglani J, Mishra AN. Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. International

Journal of Speech Technology. 2018, 21(2): 211-216. doi: 10.1007/s10772-018-9497-6
35. Mittal P, Singh N. Development and analysis of Punjabi ASR system for mobile phones under different acoustic models. International Journal of Speech Technology. 2019, 22(1): 219-230. doi: 10.1007/s10772-019-09593-x
36. Kadyan V, Mantri A, Aggarwal RK, et al. A comparative study of deep neural network based Punjabi-ASR system. International Journal of Speech Technology. 2018, 22(1): 111-119. doi: 10.1007/s10772-018-09577-3
37. Zhu M, Liao J, Liu J, et al. FedOSS: Federated Open Set Recognition via Inter-client Discrepancy and Collaboration. IEEE Transactions on Medical Imaging. Published online 2023: 1-1. doi: 10.1109/tmi.2023.3294014
38. Center For Indian Languages Technology, IIT Bombay. Available online: www.cfilt.iitb.ac.in>hindi_version (accessed on 22 August 2023)
39. Dharmale G, Patil DD, Thakare VM, et.al., Performance evaluation of different ASR Classifiers on Mobile Device. International Journal of Next-Generation Computing-Special Issue, 2021; 12(2).
40. Dharmale G J, Patil DD. Evaluation of Phonetic System for Speech Recognition on Smartphone. International Journal of Innovative Technology and Exploring Engineering. 2019, 8(10): 3354-3359. doi: 10.35940/ijitee.j1215.0881019
41. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006, 27(8): 861-874. doi: 10.1016/j.patrec.2005.10.010