# Creation and Instigation of Triphone based Big-Lexicon Speaker-Independent Continuous Speech Recognition Framework for Kannada Language

**Praveen Kumar P S, H S Jayanna**

***Abstract*: This paper proposes a framework that is intended to do the comparably accurate recognition of speech and in precise, continuous speech recognition (CSR) based on triphone modelling for Kannada dialect. For designing the proposed framework, the features from the speech data are obtained from the well-known feature extraction technique Mel-frequency cepstral coefficients (MFCC) and from its transformations, like, linear discriminant analysis (LDA) and maximum likelihood linear transforms (MLLT) are obtained from Kannada speech data files. At that point, the system is trained to evaluate the hidden Markov model (HMM) parameters for continuous speech (CS) data. The persistent Kannada speech information is gathered from 2600 speakers (1560 men and 1040women) of the age bunch in the scope of 14 years-80 years. The speech information is acquired from different geographical regions of the Karnataka (one of the 29 states situated in the southern part of India) state under degraded condition. It comprises of 21,551 words that spread 30 locales. The performance evaluation of both monophone and triphone models concerning word error rate (WER) is done and the obtained results are compared with the standard databases such as TIMIT and aurora4. A significant reduction in WER is obtained for triphone models. The speech recognition (SR) rate is verified for both offline and online recognition mode for all the speakers. The results reveal that the recognition rate (RR) for Kannada speech corpus has got a better improvement over the state-of-the-art existing databases.***

***Keywords: Automatic speech recognition, Continuous speech, Kannada dialect, Kaldi toolkit, monophone, triphone, HMM, WER.***

## I. INTRODUCTION

The objective of ASR is to perceive speech independent of uttering style, dialect, region, grammar and so on. Essentially, there are three ways to deal with speech recognition to be specific, Acoustic-Phonetic methodology, Pattern Recognition approach and Artificial Intelligence (AI) approach [1]. Over the most recent 50 years, we have seen consistent improvement in SR.

This advancement can be classified into two components: (i) the utilization of HMM in modelling the temporal varieties of the speech and (ii) the expanding computing intensity of present-day PCs. In the previous fifteen years alone, we have seen some ease, business intuitive SR applications created by Apple, Google, Microsoft, Amazon, and so forth. One of the notable facts observed is that the research on ASR is essentially centered around English and other Western dialects. It is observed that there is no considerable work done for Indian dialects, particularly South-Indian dialects, for example, Kannada. One of the fundamental disadvantages of building up a Kannada SR framework is the inaccessibility of standard speech and content corpora. The aim of our research centers around removing these constraints to assemble a sensibly decent, substantial lexicon, CSR framework for Kannada. The SR researchers around the globe have recognized the proficiency of state-of-the-art modelling techniques in building ASR frameworks. The speech is perceived as text-dependent i.e., the probability of occurrence of phonemes depends on past and the subsequent phoneme, constantly triphone models are adopted. We know that there are enough triphones that has all the acoustic-phonetic varieties in the dialect, clustering is done to group the models that are having comparable states. This process limits the amount of triphones thus the searching process is streamlined. With this technique of triphone modelling with clustering, it is conceivable to accomplish worthy RR.

A productive and robust ASR framework for the Kannada language is the need of our country. The SR is the challenging job undertaking from many years and till today it's very difficult to find such a system which exactly links human beings and the machine. Yet there are numerous potential outcomes in this field. Plenty of work has been completed in the speech recognition field majority of which concentrating on English and other European dialects. Presently the interest is shifted in the direction of recognizing the provincial dialects. Here we present the work that is centred around the recognition of Kannada dialect, the official language of the state of Karnataka is being exhibited.

Because of the absence of standard database of speech, ASR in the Kannada language has not seen considerable progress.

Revised Manuscript Received on December 12, 2019.

**Praveen Kumar P S,** Research Scholar, PhD, Department of Electronics and Communication Engineering, Siddaganaga Institute of Technology, Tumkur.

**Dr H S Jayanna,** Professor and Head, Department of Information Science and Engineering, Siddaganaga Institute of Technology, Tumkur, Karnataka, India.

This inspired us to take up the task on developing the speaker-independent large vocabulary ASR framework for Kannada dialect.

The remaining of the paper is sorted out as pursues. Section II portrays the related works in the field of ASR. Section III explains the CSR acoustic model. The clustering of triphone and its working mechanism is discussed in section IV and Section V, the implementation of ASR framework for Kannada dialect is given. The Kannada phoneme characteristics are presented in section VI. The details of data collection, data preparation and call-flow for online recognition are provided in Section VII. The method of conduction of the experiments and obtained results are portrayed in section VIII. As the ending segment of the article, we would like to conclude and concisely present our future research aspects in Section IX.

## II. LITERATURE SURVEY

The work in [2] goes for adding to the Amazigh language ASR. The researchers have considered and understood an ASR framework, utilizing a domain completely dependent on the Amazigh-Tarifit language. In this structure, they have first developed the Amazigh-Tarifit speech database, which was utilized to survey and cause the aftereffects of this work to experience a test. Indeed, the work presented has 2 goals: The first goal is to gather a medium lexicon isolated word speech database, that act as the database for the Amazigh speech researchers. The second goal is building up the Amazigh ASR system utilizing this speech database which consists of 187 unmistakable confined words. The speech database was recorded by 55 people (30 males and 25 females) from Amazigh-Tarifit local speakers. The framework was assessed on a speaker-free methodology. The tests were completed putting together basically with respect to two parameters: the GMM, and tied states (senones). The WER accomplished was 8.19%.

In [3] the authors introduce their work on building a large lexicon CSR framework for Tamil utilizing DNN. The state-of-the-art methodologies, to be specific, MLLT and speaker-versatile training is utilized to develop the DNN based speech recognition framework. They have utilized 7 long periods of Tamil speech recorded from 40 speakers which has the lexicon size of 14,231 words, among that five hours long stretches of information was utilized for training. The results show that the SR frameworks accomplish a phone error rate (PER) of 25.01% and WER of 3.49%, individually. The proposed modelling methodology has shown the substantial improvement when it compared to mono-phone acoustic model. The authors in [4] has come-up with the ideology of speaker-independent, CSR framework for Hindi dialect which comprises of 600 words vocabulary. In this work, the researchers utilized HMM model for training and recognition. The features of the speech data are obtained from the techniques such as MFCC and perceptual linear prediction (PLP) along with heteroscedastic discriminant analysis (HLDA). The speech database has the voice samples of 57 distinct speakers wherein 36 are the males and 21 are females. The HTK and Sphinx were utilized to execute this framework. The accuracy of 92.98% was accomplished with MFCC at the front end and with 8 states GMM (Gaussian mixture model).

The work in [5] demonstrates a hybrid continuous-SR framework that prompts enhanced outcomes on the speaker dependent asset administration task. This hybrid framework, called the consolidated framework, depends on a mix of standardized neural network [6] yields best results. A best in the class HMM framework is joined with a time delay neural network (TDNN) incorporated in a Viterbi framework. A various level of TDNN structure is portrayed which separates the training task into sub-tasks relating to subsets of phonemes. It was demonstrated that the joined framework, regardless of the low precision of the various levelled TDNN, accomplishes a WRR reduction of 15% according to cutting-edge HMM framework. The author in [7] investigated the exhibition of SR system with the customary Cepstral features when utilizing the linear feature transforms. This combination of features is used to model the DNN-HMM system. The test results uncover that the combination of cepstral features, utilizing neural nets, can incredibly boost the efficiency of ASR framework regarding utilizing the cepstral includes alone. The author then applied this approach on various assignments, i.e., the noise-free speech database (DARPA-WSJ), the Aurora-4 corpus and a real-time open-vocabulary. The experimental results show that the proposed combination of cepstral features has obtained WER improvement of 18% compared to the individual cepstral features The CSR has been a functioning field of research for quite a while.

Numerous experiments have been conducted in the recognizance of dialects such as Punjabi, Hindi, Tamil, Telugu, Kannada and so on [8] [9] [10]. The research related to SR in Hindi utilizing kaldi is accounted in [11] [12] [13]. Numerous toolkits are accessible to the researchers in the area of SR namely Sphinx, HTK, Julius and Kaldi. As of late, Kaldi is a standout amongst the most well-known and the most recent toolkit for ASR and it is scripted in C++ programming language. The upsides of SR application created utilizing Kaldi produce excellent systems and are quick enough for the applications of real-time recognition [14]. Through the extensive literature survey, we come to the conclusion that the work on continuous Kannada speech recognition (CKSR) is not remarkable. As a result of which, we would like to verify the behaviour of the state-of-the-art techniques for continuous Kannada speech. Since we didn't know their performance for CKSR, we conducted some experiments by creating our own database of 2600 speakers collected across the Karnataka state in the real-world environment. The transcription and the validation are done for all the speaker wave files. We built our own phoneme level lexicons.

## III. CSR ACOUSTIC MODEL

The acoustic model (AM) in an SR system creates the essential units of speech in the composed structure regarding a specific input signal [15]. The signal which acts as input is grafted up into overlapping periods of 10 ms with a 5 ms. At that point from each frame, 39 MFCC [1] co-efficient are extricated. The obtained features are contrasted with deference with the prepared AM.

### A. Monophone AMs

The procedure of making of monophone AMs begins with the readiness of the training and testing of speech files. The information in these files contains utterances of numerous speakers and the comparing transcripts encoded utilizing the phone set for the dialect. These transcripts are subjected to the training module that uses Baum-Welch estimation [1]. The procedure begins with a default model HMM for each phone which is iteratively tuned by the input speech data files and the interpretations/transcriptions. Formation of the monophone HMMs anyway requires determining the number of states preceding training. In the experiments conducted, we proposed 3-state HMM topology for non-silence phones and a 5-state typology for silence phones. Anyway, the monophone based models can't catch the variation in the phone as for the specific circumstance. It was observed that the phones tend to fluctuate contingent upon the previous and the next successive phones and this angle should be caught inside the AMs to boost the execution.

### B. Triphone AMs

By monitoring the efficiency issue engaged with utilizing monophone based AMs alongside the inspiration of implementing context-based information in the phone models, drove us to attempt triphone based AMs. The triphone based modelling deals with obtaining the context information at the expense of training considerably more essential units. In the experiments that we conducted, the phones which are located before and after the present phone are used for modelling the phoneme model. The individual phones inside words are modelled by considering their previous and succeeding phoneme that come in the word limits are modelled as diphones. The significant advances associated with triphone modelling are equivalent to that of monophone modelling aside from that it requires undeniably more training information for fruitful modelling. Generally, to model the monophone, let us assume P words are associated with the effective training. Then to train triphone model it would require $(P)^3$ words. This stayed an enormous issue since such rich training data isn't accessible in Kannada. Regardless of whether adequate training information is given, there is no assurance that each conceivable triphone would happen with such low recurrence. Every one of these variables makes triphone modelling a troublesome suggestion for Kannada.

## IV. GROUPING OF TRIPHONES

So as to address the above discussed issues for triphone modelling, they are grouped together based on the phonetic likenesses [16]. This lessens the successful number of units. It means that as opposed to modelling each triphone, the phoneme clusters were modelled. The clustering procedure utilized for the design is choice tree-based clustering. The phonetic decision trees wherein each hub is related to a specific inquiry in regards to the phonetically comparable setting is grouped and made tied state HMM models of them.

## V. IMPLEMENTATION OF ASR FRAMEWORK

The procedure of continuous speech recognition for Kannada language includes numerous modules as demonstrated in Figure 1. As shown in the block diagram of ASR system, the first step is to collect the speakers speech data through the interactive IVRS call flow system provided by BSNL. The next step is the transcription and validation of the collected speakers speech data in accordance with the various types of noise formats associated with each speech file which is nothing but the authentication of transcribed speech data. Then the Lexicon/Vocabulary is created followed by the formation of the set of phonemes for Kannada dialect.

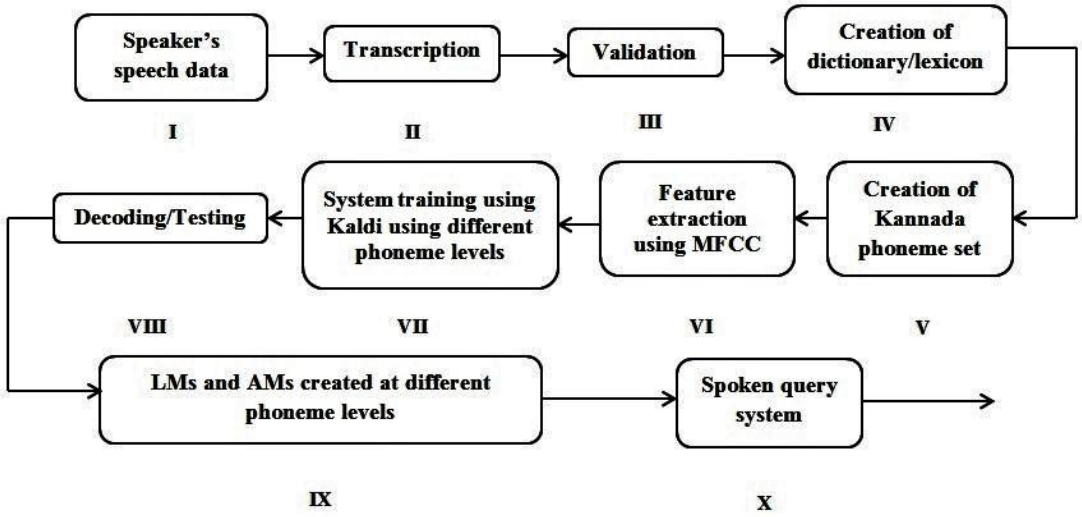


**Fig. 1. The Schematic of CSR framework for Kannada dialect**

### A. *Extraction of features in Kannada speech*

The speech features present in continuous Kannada speech are extracted using the one of the well-known feature extraction

technique known as MFCC. It is the minimized portrayal of the speech data. The speech signal is subjected to pre-emphasis step before going for the extraction of features. Basically, the speech signal is a nonstationary signal, to make the non-stationary speech signal into stationary signal, the given speech signal is divided into several frames. Each frame is of duration 20 ms-40 ms duration with each frame overlapping of around 20-30%. Then the isolated frames are subjected to windowing and FFT. Later the FFT coefficients are passed through triangular filters that are placed according to the Mel-scale. The weighted aggregate in the filters of the filter bank represents the spectral magnitude in that channel. Later to decorrelate the obtained features of speech data, the discrete cosine transform (DCT) is applied. The features of speech data obtained after applying DCT are known as MFCC. The additional advantage of applying DCT is that it has the impact of smoothing the range.

### B. *The Language Model (LM)*

To construct the LM, just interpretations of training information with all conceivable acoustic variations is needed. It characterizes the grammar of a dialect by probabilistic worth and limits the search procedure also It predicts the probability of explicit words happening in a steady progression in a certain dialect. The n-gram model that is unigram, bigram or trigram LM can be used as a Statistical model. The greater part of the LMs is stochastic. The number of events of a word in an example source content is given by the unigram model. The measurements of an event of a word given the past word are given by the bigram model. The trigram model relies upon past two words and it is commonly utilized for a large lexicon SR framework. The recognition precision, for the most part, relies upon the measure of the database used to prepare the LM.

### C. *Training and Testing*

Training the speech files with great phonetic parity and inclusion is required. The translations of the training information are utilized to create a LM and alongside acoustic information, the AM is built. The search algorithm utilized is Viterbi deciphering calculation that takes a system depicting the suitable word successions known as LM, a lexicon characterizing how every word is articulated and the set of HMMs called AM as input and yields as a content document. The testing stage is to look through the expressed word arrangement within the words that are present in the lexicon.

## VI. KANNADA PHONEME CHARACTERISTICS

If you Kannada is a south Indian language generally spoken in the territory of Karnataka. It is one of the Dravidian languages like Telugu, Tamil, and Malayalam. The local speakers of the state of Karnataka are called Kannadigas, number around 38 million, making it the 27th most communicated in language on the planet. Kannada as a language has experienced adjustments since BCs. In view of the changes it very well may be ordered into 4 sorts: Purva Halegannada (earliest starting point up to the tenth century), Halegannada (tenth century to twelfth century), Nadugannada (twelfth century to the fifteenth century), Hosagannada (fifteenth century onwards). Hosagannada or Kannada language utilizes forty-nine phonemic letters, ordered into three gatherings they are: Swaragalu/vowels: There are thirteen vowels, which are the autonomously existing letters and 2 Swaras (Vowels) contingent upon the time used to articulate. They are Hrasva Swara and Deerga Swara Vyanjanagalu/Consonants: The total number of consonents in the Kannada duialect are thirty-four. They are partitioned into two kinds Vargeeya and Avargeeya and Yogavaahakagalu: there are two yogavaahakagalu are Anuswara and Visarga. Essential Language Rule in Kannada
is the point at which a needy consonant consolidates with a free vowel; an Akshara is framed as indicated below: Vyanjana +Vowel (matra)=Letter (Akshara). From this it is understood rthat the consolidation of consonents and vowels gives rise to Kaagunitha (Kannada letters). Table I shows the classification of Kannada phonemes and the depicting ITRANS (Indian dialect Transliterations). The ITRANS of the respective phonemes indicated inside the braces.

**Table-I: Analysis of Kannada characters and their corresponding ITRANS**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Vowels** | ಅ (a) | ಆ (aa) | ಇ (i) | ಈ (I) | ಉ (u) | ಊ (U) | ಋ (Ru) | ೠ (RU) | ಎ (e) | ಏ (E) | |
| | ಐ (ai) | | ಒ (o) | | | ಓ (O) | | ಔ (ou) | | | |
| **Yogavahakas** | ಅಂ (aM) | | | | | ಅಃ (aH) | | | | | |
| **Structured consonants** | ಕ (ka) | ಖ (kha) | ಗ (ga) | ಘ (gha) | ಙ (nga) | | | | | | |
| | ಚ (cha) | ಛ (Cha) | ಜ (ja) | ಝ (jha) | ಞ (nja) | | | | | | |
| | ಟ (Ta) | ಠ (Tha) | ಡ (Da) | ಢ (Dha) | ಣ (Na) | | | | | | |
| | ತ (ta) | ಥ (tha) | ದ (da) | ಧ (dha) | ನ (na) | | | | | | |
| | ಪ (pa) | ಫ (pha) | ಬ (ba) | ಭ (bha) | ಮ (ma) | | | | | | |
| **Unstructured consonants** | ಯ (ya) | ರ (ra) | ಲ (la) | ವ (va) | ಶ (sha) | ಷ (Sha) | ಸ (sa) | ಹ (ha) | ಳ (La) | | |

## VII. THE CORPUS COLLECTION AND DATA PREPARATION

This section gives the bit by bit methodology for making straight forward ASR utilizing our very own arrangement of information through Kaldi toolkit. For our trial, our very own database is made which comprises of 2600 speakers. Every speaker speaks the 10 short Kannada sentences. Accordingly, 26000 continuous Kannada speech database is readied that is phonetically adjusted. For training of the model, 80% of the data are picked and the remaining are utilized for testing. At last, acoustic meta-information of every speaker is to be made that is utilized for preparing and testing the AMs. The preparation of data is partitioned into acoustic information and dialect information. Meta-information used for acoustic information given in the Table II are compulsory for Kaldi ASR: The Bharat Sanchar Nigam Limited (BSNL) granted the service of an interactive voice response system (IVRS) call stream. The speech information is obtained from different places of the Karnataka state under the certifiable condition. The word level to the phoneme level transcriptions are done for the Kannada speech data files that are obtained from the speakers.

**Table-II: Data preparation files and their functionalities in Kaldi**

| File name | Format | Function |
|---|---|---|
| *spk2gender* | *speaker_ID gender* | This file informs about speakers gender.<br>Speaker ID is a unique name of each speaker (sometime also referred as recording ID) |
| *wav.scp* | *uterrance_ID pathoftherecorded.wav* | It provide the path of recorded audio files sentence along with speakers ID |
| *text* | *uterrance_ID transcription* | This file contains every utterance matched with its text transcription |
| *utt2spk* | *uterrance_ID speaker_ID* | This has the mapping of the utterance of particular speaker |
| *corpus.txt* | *transcription* | It contain all the utterance transcription that are use for building the model |
| *lexicon.txt* | *word phone_1 phone_2* | This contain the phone transcriptions of every word |
| *nonsilence_phones.txt* | *phone* | This contain all the phones that are used for preparing the database |
| *silence_phones.txt* | *phone* | This contain the silence and short pause phone |

### A. *The call-flow structure for SQS*

In the SQS, the speaker is prompted to speak the Kannada sentence. In case Kannada sentence not clear then it will demand the customer to impel next sentence from the database. It is likewise perceived effectively; at that point the framework will demand the speaker to provoke the following sentence from the database. In like manner, the process is continued till the last sentence in the database. If any Kannada sentence isn't seen really, that will be drawn nearer again by the customer for multiple times. Regardless of the way that in case it doesn't see, by then, the system says Too awful...! Endeavor again afterwards...! The schematic depiction of SQS for perceiving persistent Kannada speech has shown up in Figure 2

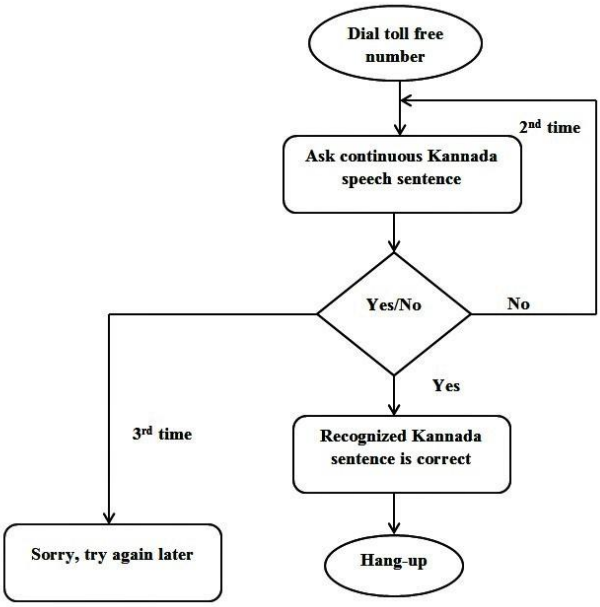


**Fig. 2. The call-flow structure of SQS for online recognition of continuous Kannada speech**

## VIII. EXPERIMENTS AND RESULTS

All our experimentations were implemented on Ubuntu 18.04 LTS (64-bit working system) platform, Intel Core i7 processor with 3.70 GHz clock speed. The exploratory outcomes are represented from the verbally communicated corpus made up of 10 Kannada sentences are spoken by 2600 speakers. The MFCC features and their subsidiaries are used for the formation of models. Kaldi uses the FST based framework and the IRSTLM toolbox was used to build the LM. The Table III exhibits the differing WERs acquired at different phoneme levels. It is found in the table that, the monophone has the base WER of 8.56% and the WER relating to triphone is 6.09%. The distinctive WER of standard TIMIT database is likewise portrayed in the table. From the table, it is discovered that the triphone modelling method has given a better precision differentiated than monophone demonstrating strategy. Clearly the CKSR database has the better RR when contrasted with that of TIMIT database.

**Table-III: The comparison of WER of CKSR database and TIMIT database**

| Phonemes | WER_1 | | WER_2 | | WER_3 | | WER_4 | | WER_5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CKSR | TIMIT | CKSR | TIMIT | CKSR | TIMIT | CKSR | TIMIT | CKSR | TIMIT |
| mono | 8.56 | 13.09 | 8.89 | 12.63 | 9.02 | 14.01 | 8.91 | 13.98 | 8.71 | 12.58 |
| tri1_600_2400 | 7.34 | 11.59 | 7.57 | 11.85 | 7.26 | 12.03 | 7.87 | 11.81 | 7.51 | 12.21 |
| tri1_600_4800 | 6.73 | 11.23 | 6.85 | 11.81 | 6.97 | 12.89 | 6.64 | 11.84 | 6.82 | 12.03 |
| tri1_600_9600 | 6.51 | 12.00 | 6.45 | 12.79 | 6.49 | 11.91 | 6.32 | 11.28 | 6.24 | 11.83 |
| tri2_600_2400 | 8.49 | 13.35 | 8.86 | 12.99 | 8.57 | 13.61 | 8.45 | 13.87 | 8.89 | 14.12 |
| tri2_600_4800 | 7.89 | 12.81 | 7.71 | 12.85 | 7.92 | 13.77 | 7.64 | 13.81 | 7.61 | 12.49 |
| tri2_600_9600 | 7.53 | 14.85 | 7.42 | 13.81 | 7.32 | 14.26 | 7.18 | 14.91 | 7.49 | 14.44 |
| tri3_600_2400 | 7.24 | 13.00 | 7.38 | 12.96 | 7.50 | 12.21 | 7.12 | 12.81 | 7.07 | 12.03 |
| tri3_600_4800 | 6.39 | 10.99 | 6.32 | 11.21 | 6.27 | 11.54 | 6.29 | 11.83 | 6.31 | 10.98 |
| tri3_600_9600 | 6.12 | 10.56 | 6.09 | 10.64 | 6.18 | 10.84 | 6.33 | 10.58 | 6.22 | 10.81 |

**Table-IV: The comparison of WER of CKSR database and Aurora4 database**

| Phonemes | WER_1 | | WER_2 | | WER_3 | | WER_4 | | WER_5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CKSR | TIMIT | CKSR | TIMIT | CKSR | TIMIT | CKSR | TIMIT | CKSR | TIMIT |
| mono | 8.56 | 14.56 | 8.89 | 15.68 | 9.02 | 17.91 | 8.91 | 16.98 | 8.71 | 14.81 |
| tri1_600_2400 | 7.34 | 15.25 | 7.57 | 15.61 | 7.26 | 14.67 | 7.87 | 14.35 | 7.51 | 14.86 |
| tri1_600_4800 | 6.73 | 14.69 | 6.85 | 15.64 | 6.97 | 16.85 | 6.64 | 17.25 | 6.82 | 18.26 |
| tri1_600_9600 | 6.51 | 15.61 | 6.45 | 16.91 | 6.49 | 12.91 | 6.32 | 14.27 | 6.24 | 15.92 |
| tri2_600_2400 | 8.49 | 16.92 | 8.86 | 18.61 | 8.57 | 16.19 | 8.45 | 16.19 | 8.89 | 15.49 |
| tri2_600_4800 | 7.89 | 14.29 | 7.71 | 14.61 | 7.92 | 16.27 | 7.64 | 15.81 | 7.61 | 14.49 |
| tri2_600_9600 | 7.53 | 15.64 | 7.42 | 15.81 | 7.32 | 16.26 | 7.18 | 15.61 | 7.49 | 15.91 |
| tri3_600_2400 | 7.24 | 14.81 | 7.38 | 15.61 | 7.50 | 15.42 | 7.12 | 14.94 | 7.07 | 16.00 |
| tri3_600_4800 | 6.39 | 13.81 | 6.32 | 13.91 | 6.27 | 13.68 | 6.29 | 13.58 | 6.31 | 13.98 |
| tri3_600_9600 | 6.12 | 14.56 | 6.09 | 13.64 | 6.18 | 13.33 | 6.33 | 14.01 | 6.22 | 13.23 |

The Table IV shows the comparison of WER results with respect to the aurora4 database. Here in this case also CKSR data base extends its dominance over aurora4 database. The comparison of CKSR database with TIMIT database and Aurora4 database w.r.t RR is shown in the Figure 3 and Figure 4 respectively. These plots show that the triphone modelling technique performs better when compared to monophone modelling technique. The efficiency of the SQS framework so designed is verified for online SR precision of CKSR framework.

**Table-V: Performance analysis of online SR accuracy check by speakers in reality condition**

| | |
|---|---|
| Total number of Kannada sentences | 10 |
| Total number of speakers | 1200 |
| First Attempt | 730 |
| Second Attempt | 223 |
| Third attempt | 143 |
| Total number of recognition | 1096 |
| Recognition rate in percentage | 91.33 |

For the same purpose, the 1200 speakers across the state are approached to examine the framework in noisy conditions. Table V depicts the exhibition assessment of the created SQS. From the table, it is observed that there is a lot of progress in online SR precision contrasted with the prior SQS.

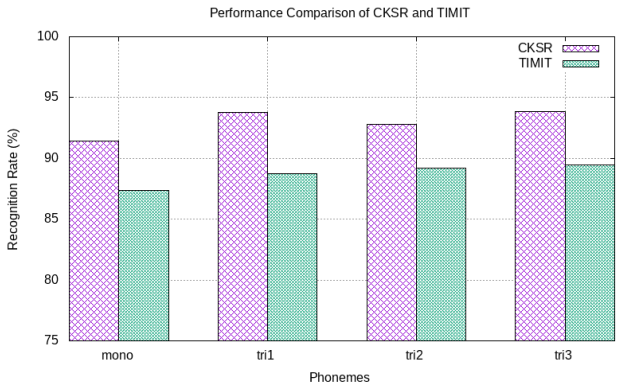


**Fig. 3. The performance comparison of CKSR and TIMIT database**

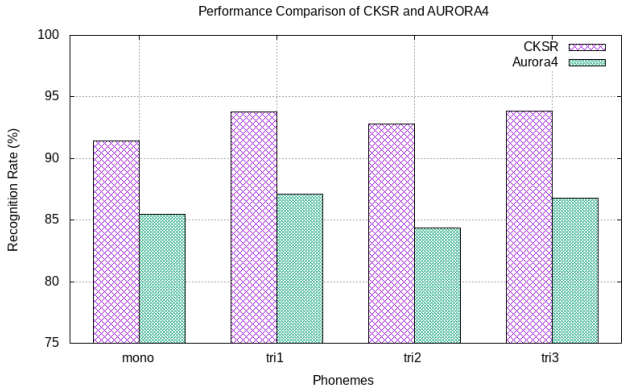


**Fig. 4. The performance comparison of CKSR and Aurora4 database**

Along these lines, it tends to be surmised that there is not much difference in the RRs in online and offline modes. The Figure 5 shows the performance comparison w.r.t online and offline RR of CKSR system. From the plot it can be observed that as the number of speech files keeps increasing, the online recognition efficiency keeps dropping in smaller rate when compared to offline recognition. This is the one point that need to be addressed otherwise the offline and online recognition go hand in hand.

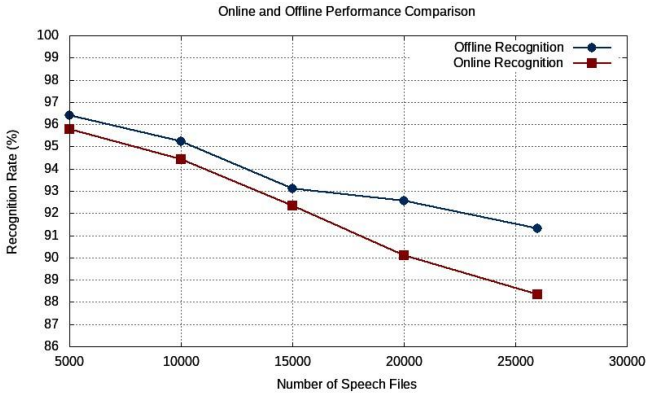


**Fig. 5. The performance comparison of online and offline speech recognition**

## IX. CONCLUSION

In this paper, the CKSR system is built using the well-known techniques such as monophone and triphone modelling. Each word in a sentence is viewed as a succession of phonemes.

By knowing that the quantity of phonemes is limited for any dialect, it is conceivable to structure recognition framework with worthy execution. To lessen the inquiry procedure and to beat the concealed triphones, they are clustered utilizing choice tree-based grouping. Various mixture Gaussian distribution is considered to expand recognition exactness. We expect to assemble around 180 hours a speech information for getting ready and scale the size of the lexicon around 100,000 words and utilize cross-lingual training to furthermore cut down the WER. A few applications request the best possible exactness, be that as it may, the precision can't be guaranteed within the sight of noise. Speakers need to talk especially all together for the framework to work commendably. If the speaker has non-standard speech, fumble words together, or murmur, the training methodology may take a longer duration. The precision of the ASR framework can be additionally expanded on the off chance that we execute productive noise removal methodologies more viably. The expected task is to additionally develop the system execution by extending the quantity of phonemes appropriately and furthermore to the efficiency of CKSR system needs to be improved especially in online recognition mode when the number of speech files keeps increasing.

## X. ACKNOWLEDGMENT

The authors would like to acknowledge the Visvesvaraya Technological University, Belagavi, Karnataka and the management of Siddaganga Institute of Technology, Tumakuru for extending the help to do the research work.

## REFERENCES

1. L. R. Rabiner, B.-H. Juang, and J. C. Rutledge, Fundamentals of speech recognition, vol. 14. PTR Prentice Hall Englewood Cliffs, 1993.
2. S. El Ouahabi, M. Atounti, and M. Bellouki, "Toward an automatic speech recognition system for amazigh-tarifit language,"International Journal of Speech Technology, pp. 1–12, 2019.
3. A. Madhavraj and A. Ramakrishna, "Design and development of a large vocabulary, continuous speech recognition system for tamil," in 2017 14th IEEE India Council International Conference (INDICON), pp. 1–5, IEEE, 2017.
4. S. Sinha, S. S. Agrawal, and A. Jain, "Continuous density hidden markov model for context dependent hindi speech recognition," in 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1953–1958, IEEE, 2013.
5. C. Dugast, L. Devillers, and X. Aubert, "Combining tdnn and hmm in a hybrid system," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 1 PART II, p. 217, 1994.
6. A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," IEEE Access, vol. 7, pp. 53040–53065, 2019.
7. D. Dimitriadis and E. Bocchieri, "Use of micro-modulation features in large vocabulary continuous speech recognition tasks," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 23, no. 8, pp. 1348–1357, 2015.
8. P. S. Praveen Kumar, G. T. Yadava, and H. S. Jayanna, "Continuous Kannada speech recognition system under degraded condition," Circuits, Systems, and Signal Processing, pp. 1–29, 2019.
9. J. Guglani and A. Mishra, "Continuous punjabi speech recognition model based on kaldi asr toolkit," International Journal of Speech Technology, vol. 21, no. 2, pp. 211–216, 2018.
10. M. Kalamani, M. Krishnamorti, and R. Valarmati, "Continuous tamil speech recognition technique under non stationary noisy environments," International Journal of Speech Technology, vol. 22, no. 1, pp. 47–58, 2019.
11. P. Upadyaya, O. Faroq, M. R. Abidi, and Y. V. Varshney, "Continuous hindi speech recognition model based on kaldi asr toolkit," in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 786–789, IEEE, 2017.
12. R. S. Sharma, S. H. Paladugu, K. J. Priya, and D. Gupta, "Speech recognition in kannada using htk and julius: A comparative study," in 2019 International Conference on Communication and Signal Processing (ICCSP), pp. 0068–0072, IEEE, 2019.
13. M. A. Al Amin, M. T. Islam, S. Kibria, and M. S. Rahman, "Continuous bengali speech recognition based on deep neural network," in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–6, IEEE, 2019.
14. D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, et al., "Subspace gaussian mixture models for speech recognition," in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4330–4333, IEEE, 2010.
15. C. S. Manasa, K. J. Priya, and D. Gupta, "Comparison of acoustical models of gmm-hmm based for speech recognition in Hindi using pocket sphinx," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 534–539, IEEE, 2019.
16. S. J. Young, "The general use of tying in phoneme-based hmm speech recognisers," in [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 569–572, IEEE, 1992.

## AUTHORS PROFILE

**Praveen Kumar P S** received his Master of Technology in signal processing from Siddaganaga Institute of Technology, Tumkur, Karnataka, India. He is presently working as a research scholar and pursuing his PhD in the Department of Electronics and Communication Engineering at Siddaganaga Institute of Technology, Tumkur. He is also a Member of Indian Society for Technical Education (MISTE).

**Dr H S Jayanna** received his PhD from the Indian Institute of Technology, Guwahati. He is currently Professor and Head in the Department of information science and engineering at Siddaganaga Institute of Technology, Tumkur, Karnataka, India. He has published more than 100 papers in international/ national journals and conferences. His areas of interest include Speech processing, Image processing, Multimodal Biometrics, Computer networks, Stenography and watermarking, Network security, Computer Architecture, Computer Organization, Multimedia communication. He has over 20 years of teaching and research experience. He is a fellow of the Institution of Engineers of India (FIEI) and Member of Indian Society for Technical Education (MISTE).