Triphone Model Based Novel Kannada Continuous Speech Recognition System using Kaldi Tool

Anand H.Unnibhavi, D.S.Jangamshetti, Shridhar K.

Abstract*: Accent is one of the issue for speech recognition systems. Automatic Speech Recognition systems must yield high performance for different dialects. In this work, Neutral Kannada Automatic Speech Recognition is implemented using Kaldi software for monophone modelling and triphone modeling. The acoustic models are constructed using the techniques such as monophone, triphone1, triphone2, triphone3. In triphone modeling, grouping of interphones is performed. Feature extraction is performed by Mel Frequency Cepstral Coefficients. The system performance is analysed by measuring Word Error Rate using different acoustic models. To know the robustness and performance of the Neutral Kannada Automatic Speech Recognition system for different dialects in Kannada, the system is tested for North Kannada accent. Better sentence accuracy is obtained for Neutral Kannada Automatic Speech Recognition system and is about 90%. The performance is degraded, when tested for North Kannada accent and the accuracy obtained is around 77%. The performance is degraded due to the increasing mismatch between the training and testing data set, as the Neutral Kannada Automatic Speech Recognition system is trained only for neutral Kannada acoustic model and doesn't include north Kannada acoustic model. Interactive Kannada voice response system is implemented to identify continuous Kannada speech sentences.

***Keywords*: ASR, NKASR, Linear Discriminant Ananlysis, Maximum Likelyhood Linear Transform, Speaker Adaptive Training.**

I. INTRODUCTION

Dialects of a given language are the differences in speaking styles of that language. Acoustic space spanned by phonemes for native speakers will shift when speakers are non-native. Voice onset time, voice stop release time, durations of the sound units and pitch contours also play an important role while identifying the dialect [1]. The dialect specific information is present in speech at different levels. At the segmental level, the dialect specific information can be observed in the form of unique sequence of the shapes of the vocal tract for producing the sound units. The shape of the vocal tract is characterized by the spectral envelope. In this work, spectral envelope is represented by Mel frequency cepstral coefficients (MFCC). Kannada is a Dravidian language which has around 20 dialects [2]. These dialects are categorized in to 4 groups as 1) Coastal dialects, this group comprises of Mangalore Kannada, Halakki, Barkur, Havyaka, Kundagannada, Sirsi Kannada and Ankola Kannada. 2) Northern Dialects this group covers Dharwad Kannada, Gulbarga Kannada and Vijayapura Kannada. 3) South-eastern, this group has Gowda Kannada, Tiptur Kannada, Rabakavi and Nanjangudu Kannada. 4) South Karnataka dialects this group includes Bangalore Kannada, Soliga Kannada, Kurumba Kannada and Banakal Kannada. Dialect variations in the language presents challenges for continuous performance of speech systems. In a given language, diversity among many dialects gives important information for speech researchers [3]. For some phonic words, there are different accents, change in pronunciation and intonations. Hence these variations and different dialects of Kannada language results in poor performance in Kannada Automatic Speech Recognition (ASR) system. In this research work, Neutral Kannada Automatic Speech Recognition (NKASR) system is implemented and results in good performance when tested for neutral Kannada accent. It is also shown that dialectal variation [4] of the same language results in degradation of the accent specific NKASR when tested for North Kannada accent. The paper is organised as follows : Section 2 describes the Literature survey, Section 3 deals with details of feature extraction, Section 4 gives the implementation details of the proposed method, Section 5 presents the results of ASR experiments and Section 6 discusses conclusion and future scope.

II. LITERATURE SURVEY

This section deals with the survey on the recent literature on speech recognition.

Mohamed G. Elfeky *et al.* [4] have shown that ASR performance typically decreases when evaluated on a dialectal variation of the same language that was not used for training its models. Similarly, models simultaneously trained on a group of dialects tend to under perform when compared to dialect-specific models. Empirical analysis and extensive experiments support the finding, and prove that country-based automatic selection of the speech recognizer outperforms user's selection. Authors have shown that automatically selecting the recognizer based on the user's geographical location helps improve the user experience.

Revised Manuscript Received on June 30, 2020.

Anand H. Unnibhavi, Dept. of Electronics and Communication Engineering Basaveshwara Engineering College Bagalkot, India-587103. . Email: anandhu.rampur@gmail.com

D.S.Jangamshetti, [2]Dept. of Electrical and Electronics Engineering Basaveshwara Engineering College, Bagalkot, India-587103 Email: asdj1229@gmail.com

Shridhar K, Dept. of Electronics and Communication Engineering Basaveshwara Engineering College Bagalkot, India-587103. . Email: shridhar.ece@gmail.com

Xuesong Yang *et al.* [5] have shown that the performance of ASR system increases when the mismatch between the training and testing scenarios decrease. The traditional approach pools data from several accents in training phase and builds a single model in multi-task fashion. In this paper, authors have explored an model which combines accent classifier and multi-task acoustic model. This model when experimented on American English Wall Street Journal and British English Cambridge corpora, yields best performance than the multi-task acoustic model baseline.

Maryam Najafian and Martin Russell [6] have implemented an ASR system using Hidden Markov Model (HMM) that compensates the effects of accents. Due to the discriminative nature of DNNs, HMM systems based on Deep Neural Networks (DNNs) outperforms traditional systems, that uses Gaussian Mixture Models (GMMs), Acoustic models are constructed in an ASR system using standard pronunciations in the acoustic training database. In a test dialect, speakers of standard pronunciations usually show a higher ASR performance. Author Louis ten Bosch [7] have attempted to relate an ASR based distance measure between dialects with a phonologically inspired distance. The correlation between both distances equals 0.70.

The survey reveals that performance of ASR degrades when tested for dialectal variation of the same language. In this paper, a system is developed to recognize continuous speech recognition system for neutral Kannada and the same is tested for different dialects of Kannada language.

III. FEATURE EXTRACTION

This section deals with the details of feature extraction. Speech segments of a given speech sample of duration 20-40 msec remains almost stationary, further Fourier transformation can be applied on these stationary speech segments. The spectral features extracted from the frames help in identifying the phones of the speech sample that distinguish two words. In this work, MFCC are used as feature extractor. The Mel-scale is a perceptual scale of frequencies which is dependent on the sensitivity of the human ear. The discrete audio signal is $s_i(n)$, where i is the frame number, n is the number of samples. The periodogram of the audio signal [8] is computed by

$$P_i\left(k\right) = \frac{1}{N}\left(\sum_{n=0}^{N-1} s_i\left(n\right)h\left(n\right)e^{-j2\Pi kn/N}\right)^2 \quad 0 \le k \le N-1 \quad (1)$$

$h(n)$ is a N sample long analysis window, N is the size of the DFT. An upper frequency $f[m]$ and a lower frequency $f[0]$ is chosen in Hz. Fig.1 shows Mel-Scale Filter Bank.

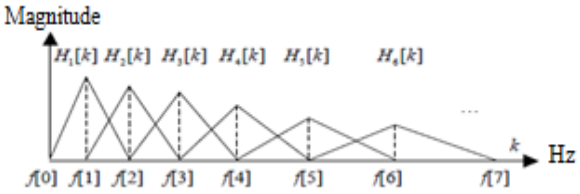


Fig.1: Mel-Scale Filter Bank

In Fig. 1. X-axis represents frequency in kHz and Y-axis is the amplitude. The upper and lower frequency boundaries are M points uniformly placed in the Mel-scale, and are given by [8]

$$f\left[m\right] = \left(\frac{N}{F_s}\right)B^{-1}\left(B\left(f_0\right) + m\frac{B\left(f_M\right) - B\left(f_0\right)}{M+1}\right) \quad (2)$$

Fs is sampling frequency in Hz, $B(f)$ is the transform from Hertz scale to Mel-scale and $B^{-1}(b)$ [8] is the inverse of Mel Scale to Hertz scale given by.

$$B\left(f\right) = 1125\ln\left(1 + f/700\right) \quad (3)$$

$$B^{-1}\left(b\right) = 700\left(e^{\left(b/1125\right)} - 1\right) \quad (4)$$

The human ear captures differences and changes in lower frequencies better than changes in higher frequencies. Therefore in Mel-scale, equally placed points in Hertz-scale will be denser on the lower frequencies than the higher. To find out how much energy is present in each triangular in mel-bin, a triangular filter $H_m[k]$ is applied for each point M mel-bins. The final two steps in finding MFCC are computing the log-energy for speech samples $S[m]$ [9], and finding Discrete Cosine transform (DCT) for each of the M-filters of $c[n]$ [9].

$$S\left[m\right] = \ln\left(\sum_{k=0}^{N-1} NP_i\left(k\right)H_m\left[k\right]\right) \quad 0 < m < M \quad (5)$$

$$c\left[n\right] = \sum_{m=0}^{M-1} S\left[m\right]\cos\left(\Pi n\left(m+1/2\right)/M\right) \quad 0 < n < M \quad (6)$$

From every frame only first 13 coefficients will help to improve the decoding. It is because human ear is more sensitive for lower frequencies and less sensitive to higher frequencies. The first 13 coefficients correspond to lower frequencies and remaining coefficients correspond to higher frequencies, which never help in improving the decoding result. Therefore only first 13 coefficients are considered for the recognition task [10]. Cepstral Mean Variance Normalization (CMVN) is the feature transformer applied after the extraction of MFCC's. CMVN minimizes the effect of difference in variable environments like ambient noise, recording equipment and transmission channels. CMVN $\left(\hat{x}_t\left(i\right)\right)$ [8] is calculated by

$$\hat{x}_t\left(i\right) = \frac{x_t\left(i\right) - \mu_t\left(i\right)}{\sigma_t\left(i\right)} \quad (7)$$

Where $x_t(i)$ is the i^{th} component of the original feature vector at time t and the mean $\mu_t(i)$ and standard deviation $\sigma_t\left(i\right)$ are calculated [8] over some sliding finite window of length N is given by.

$$\mu_t\left(i\right) = \frac{1}{N}\sum_{n=t-N/2}^{t+N/2-1} x_n\left(i\right) \quad (8)$$

$$\sigma_t^2(i) = \frac{1}{N}\sum_{n=t-N/2}^{t+N/2-1}\left(x_n(i) - \mu_t(i)\right)^2 \quad (9)$$

The first and second order deltas $\Delta + \Delta\Delta$ [8] of the MFCCs can be calculated to add dynamic information to the MFCCs. For an acoustic feature vector ***x,*** the first order deltas are defined as

$$\Delta X_t = \frac{\sum_{i=1}^{n} w_i\left(X_{t+i} - X_{t-i}\right)}{2\sum_{i=1}^{n_i} w_i^2} \quad (10)$$

where n is the window width and w_i the regression coefficients. The second order delta parameters are derived in the same fashion as,

$$\Delta^2 X_t = \frac{\sum_{i=1}^{n} w_i\left(X_{t+i} - X_{t-i}\right)}{2\sum_{i=1}^{n_i} w_i^2} \quad (11)$$

The combined feature vector [8] becomes as

$$X_t = \begin{bmatrix} X_t & \Delta X_t & \Delta^2 X_t \end{bmatrix} \quad (12)$$

IV. IMPLEMENTATION

Implementation details of the proposed method are discussed in this section. It deals with Corpus collection in which speaker selection and recording is performed. It also deals with acoustic modeling, language modeling and decoding of speech data.

A. Corpus Collection, Speaker Selection and Recording

In this work, four different female speakers were selected for recording sentences. Each speaker was asked to utter 300 neutral Kannada sentences, with a total of 1200 (4 speakers x 300 sentences = 1200 sentences) sentences are recorded using wave surfer software. Also twelve more sentences of north Karnataka accent of the same first twelve sentences of neutral Kannada accent are recorded. The uttered sentences are sampled at frequency of 16 kHz with 16 bit depth. Short sentences with a minimum of 5 and maximum of 10 words are selected for creating pronunciation dictionary and word level dictionary. The sentences are carefully selected so that they are meaningful and do not contain any offensive or sensitive words. Phonetic rich sentences are needed for robust estimation of the statistical model parameters of context sensitive phonemes. A set of sentences is considered to be phonetically rich if it contains all permissible triphones of the language in sufficient quantity. If there are M phonemes in a language, there can be M^n triphones.

B. Acoustic Modelling

The Acoustic Model estimates the likelihood function $P(X / w;\theta)$ where θ is initial state distribution , X is acoustic features and w is probable word sequence. These are HMM parameters found through training the HMM. HMM models the uncertainty between acoustic features and corresponding transcriptions. There are 53 unique phones generated for neutral Kannada data base. In this work a total of 1100 audio file are used for training and 100 audio files are used for testing. The training utterance transcription is converted from words to phones using a phonetic dictionary. An example of Monophone training is shown in Fig. 2.

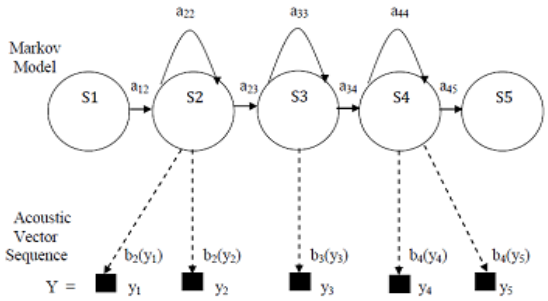


Fig. 2: Monophone training

Triphone training is better then monophone training, and is mathematically proved in the previous work [11] as it captures the context of the single middle phone very efficiently. To reduce the dimensionality, acoustically similar triphones are tied together in a process called state-tying. In Kaldi, the state-tying is performed using decision trees. Each triphone is modelled by a HMM. The model parameters that are estimated in the acoustic training are $\theta = [\{a_{ij}\},\{b_j()\}]$, where $\{a_{ij}\}$ corresponds to transition probabilities and $\{b_j()\}$ to output observation distributions. In Kaldi, the acoustic training uses Viterbi training for updating the Gaussian variables and model parameters θ. GMMs are able to describe multimodality (Speaker, accent and gender). The output observation is given by

$$b_j(X) = \sum_{m=1}^{M} c_{jm} N\left(x; \mu^{jm}, \Sigma^{jm}\right) \quad (13)$$

where c_{jm} is the prior probability for component m of state (S_j) and $N\left(x, \mu^{jm}, \Sigma^{jm}\right)$ is the Gaussian distribution with parameters μ^{jm} and Σ^{jm} corresponding to the mean and covariance of state (S_j). In the training of a GMM the aim is to update the mean and covariance.

C. *Language Modelling*

The language model contains information of the likelihood of words to co-occur. The prior probabality $P(w)$ [8] is determined by the language model. The prior probability $P(w)$ for a word sequence $w = w_1, w_2, w_3 \ldots, w_K$ is given by

$$P(w) = \prod_{k=1}^{K} P\left(w_k / w_{k-1}, \ldots\ldots w_1\right) \quad (14)$$

For large vocabulary speech recognition, the probability of a word is often modelled using the n-gram language model. The n-gram model assumes that the probability of a word to occur is dependent on the n number of words occurring before it. For large vocabulary, n typically ranges between two and four.

$$P(w) = \prod_{k=1}^{K} P\left(w_k / w_{k-1}, w_{k-2} \ldots\ldots w_{k-n+1}\right) \quad (15)$$

The model is trained by counting occurrences of word sequences in the training data and estimating probabilities [8] from the maximum likelihood.

$$P\left(w_k / w_{k-1}, w_{k-2}\right) \approx \frac{C\left(w_{k-2} w_{k-1} w_k\right)}{C\left(w_{k-2} w_{k-1}\right)} \quad (16)$$

Where $C\left(w_{k-2} w_{k-1} w_k\right)$ represents the total number of occurrences of the word sequence $w_{k-2}, w_{k-1} w_k$ and $C\left(w_{k-2} w_{k-1}\right)$ the total number of occurrences of the words $w_{k-2} w_{k-1}$ respectively.

D. *Speech Decoding*

In Speech decoding the most probable word sequence is found by *w** [8]

$$w^* = argmax \left[P\left(w / X\right)\right] \quad (17)$$

P(*w*/*X*) is computed by applying Bayes rule and the above equation is given by

$$w^* = argmax \left[P\left(X / w\right) P\left(w\right)^{w_{lm}}\right] \quad (18)$$

where *P*(*X*/*w*) is the acoustic model and *P*(*w*) is the language model, and final output is regulated by language model weight w_{lm}. To maximize the above equation, Kaldi solves the search task using word lattices. Kaldi uses Weighted Finite State Transducer (WFST) to combine the information from the acoustic model and the language model.

V. TESTING PROCEDURE

In this work, four different female speakers are asked to utter 300 neutral Kannada sentences each, total of 1200 (4 speakers x 300 sentences = 1200 sentences) sentences are recorded using wave surfer software. Also twelve more sentences of north Karnataka accent of the same first twelve sentences of neutral Kannada accent are recorded. The uttered sentences are sampled at frequency of 16 kHz with 16 bit depth. NKASR is implemented for which the model is trained with 1100 audio files of Neutral Kannada accent, and 100 untrained Neutral Kannada accent speech files are used for testing NKASR. The same NKASR is tested for 100 more untrained audio files of which 12 audio files are north Kannada accent and 88 audio files are of neutral Kannada accent.

VI. RESULT ANALYSIS

In the earlier work, Continuous Kannada (neutral Kannada) Speech Recognition System was developed [12] using HTK tool kit. MFCC was used as feature extractor and HMM was used as classifier. Each phoneme is represented by tristate HMM with each state being represented by Gaussian model. Single GMM do not capture the variations of phone. Gaussian mixture splitting was done to achieve better result. Experimental results show that, good recognition accuracy of 95.17% was achieved for context dependent (Tri-phone) modeling as compared to context independent (Monophone) modeling which is about 85.14% [12]. The same Neutral Kannada ASR is implemented using Kaldi tool kit and the results are satisfactory. The HTK results are better then Kaldi tool result because HTK system is gender dependent [13]. Implementation of NKASR and result analysis is carried out as in the two cases: I. Neutral Kannada ASR System Implementation and II. Neutral Kannada ASR System tested for north Kannada (Bagalkot) accent.

Case I. Neutral Kannada ASR System Implementation

In this work, to increase recognition accuracy different modelling techniques are employed such as i) Triphone Modelling is adopted when compared to monophone modelling, as speech is context dependent [11]. ii) Alignment between HMM states and audio frames is to be done, so that training algorithms such as Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transform (MLLT) and Speaker Adaptive Training (SAT) are used to improve and refine the parameters of the models [14,15]. iii) The Word Insertion Penalty (WIP) is used which adds a fixed value [16] to the accumulated log likelihood each time a new word is entered during the Viterbi search algorithm or adds a value to each token when it transits from the end of one word to the start of the next. The parameters such as WIP and grammar scale factor will have a significant effect on recognition performance.

The word error rate (WER) is the most common measurement tool used to evaluate the performance of ASR systems [12, 16]. The WER represents the minimum edit distance between the highest scoring recognition result with the words in the reference. The reference contains the correct transcription of the spoken utterance.

In this phase, NKASR is implemented. The model is trained with 1100 audio files of Neutral Kannada accent and 100 untrained Neutral Kannada accent speech files are used for testing NKASR.

Figure 3 shows the graph of monophone modelling result in which word accuracy (WA) and sentence accuracy (SA) Vs %WER (Word Error Rate) are drawn. The trained model is able to recognize better word level accuracy of 92.07%, sentence level accuracy of 77% for WIP of 1.0 with minimum %WER of 7.93. Figure 4 shows Triphone 1 [Delta + Delta-Delta] model obtained after monophone alignment. Triphone modelling is able to recognize word with accuracy of 91.22% and sentence level accuracy of 83% for WIP of 1.0 and minimum %WER of 8.78. Figure 5 shows Hybrid triphone 2 model [LDA+MLLT+SAT] in which triphone 1 is aligned that yields word recognition accuracy of 94.62% and sentence level accuracy with 90% for WIP of 1.0, and minimum %WER of 5.38. Triphone 2 alignment is performed to get more homogenous or standardized data. This data is used by triphone 3 [LDA + MLLT + SAT] acoustic model that increases Word Accuracy (WA) upto 94.65% and Sentence Accuracy (SA) up to 90% for WIP of 1.0, and minimum %WER of 5.35 as shown in Fig. 6.

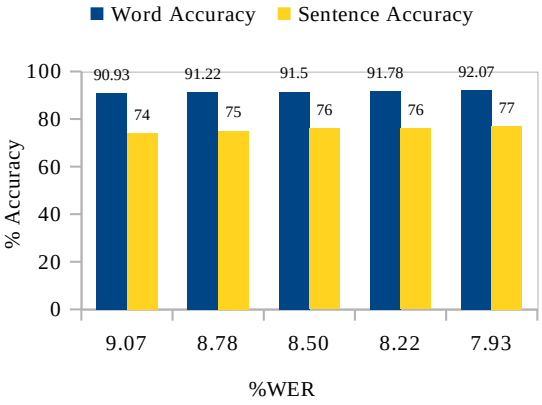


Fig. 3 : WA and SA Vs %WER in Monophone Modelling

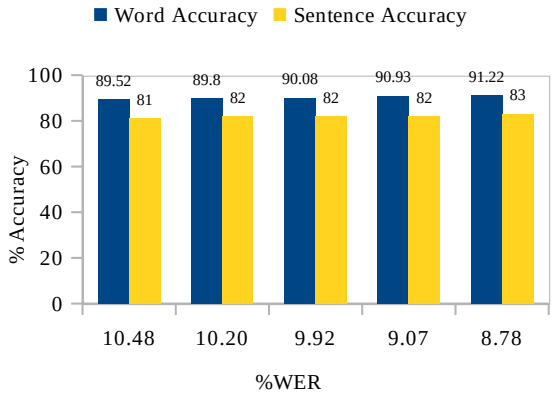


Fig. 4 : WA and SA Vs %WER in Triphone 1 Modelling

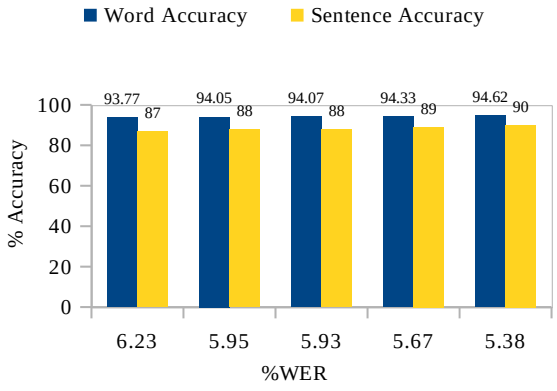


Fig. 5 : WA and SA Vs %WER in Triphone 2 Modelling

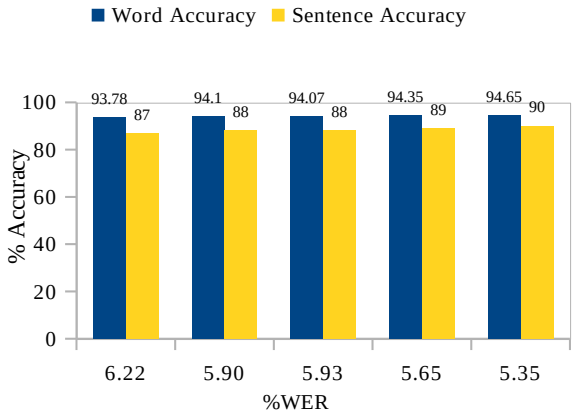


Fig. 6 : WA and SA Vs %WER in Triphone 3 Modelling

Case II. Neutral Kannada ASR System tested for north Kannada (Bagalkot) accent

NKASR system is implemented in case I is accent-dependent, as it is trained only for neutral Kannada accent. To evaluate its robustness, the NKASR system is tested for north Karnataka accent in case II. Following are the results of NKASR system for north Karnataka accent.

The test files consists of 100 audio files out of which 12 audio files are north Kannada accent and 88 audio files are of neutral Kannada accent.

Figure 7 shows the results of monophone modelling recognition. Word level accuracy of about 79.27% and sentence level accuracy of 66% are obtained for WIP of 1.0, and minimum of %WER of 20.73. Figure 8 shows the result of triphone 1 model [Delta + Delta-Delta} obtained after monophone alignment. Triphone word level accuracy of 79.27% and sentence level accuracy of 74% are obtained for WIP of 1.0, minimum of %WER of 21.53. Figure 9 shows the result of hybrid triphone 2 model [LDA+MLLT+SAT]. Word level accuracy of 76.47% and sentence level accuracy of 77% with WIP of 1.0, and minimum of %WER of 23.53 are obtained. The Figure 10 shows the result of north Kannada accent ASR using triphone 3 model [LDA+MLLT+SAT]. It is based on triphone 2 alignment with standardized data resulting in word level accuracy of 75.63% and sentence level accuracy of 77% for WIP of 1.0, and minimum of %WER 24.37 are obtained.

From the above result it is evident that NKASR performance is low for north Kannada (Bagalkot) accent utterances. This is because NKASR system is trained only for neutral Kannada accent, as the model is aware of only neutral Kannada and unaware of north Kannada accent. The overall performance of NKASR system is presented in Table 1.

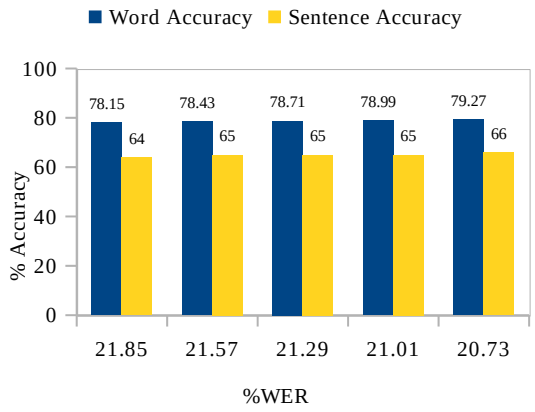
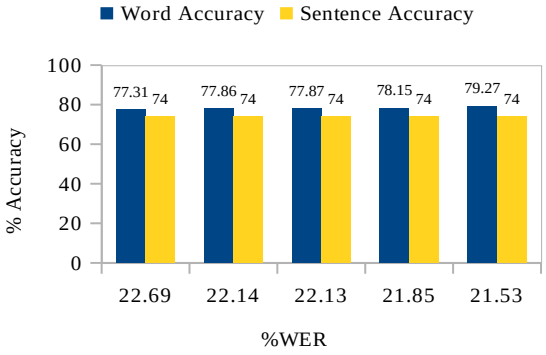


Fig. 7 : WA and SA Vs %WER in Monophone Modelling



Fig, 8 : WA and SA Vs %WER in Triphone 1 Modelling

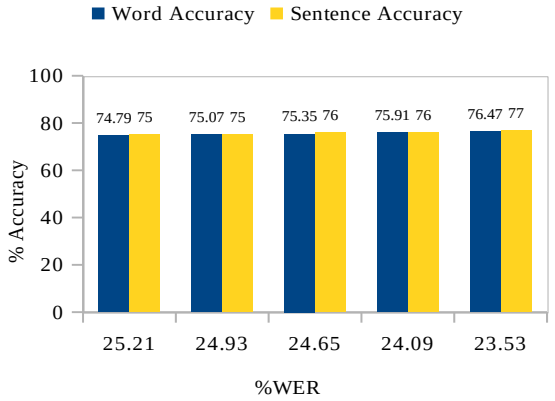


Fig. 9 : WA and SA Vs %WER in Triphone 2 Modelling

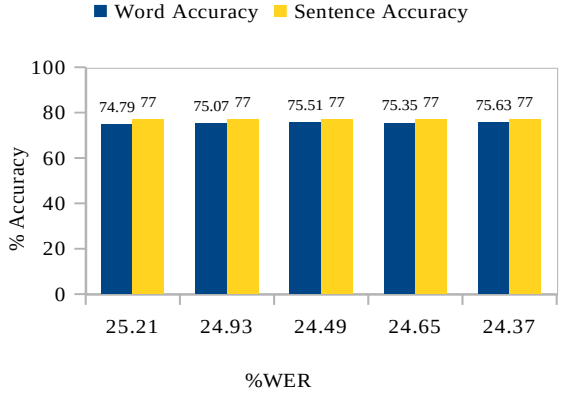


Fig. 10 : WA and SA Vs %WER in Triphone 3 Modelling

Table 1. %WER, %WA, %SA for Different Acoustic Model and Inputs of Different Accent.

Acoustic models	Neutral accent input for NKASR system			North Kannada accent input for NKASR system		
	%WER	%WA	%SA	%WER	%WA	%SA
Monophone	7.93	92.07	77.00	20.73	79.27	66.00
Triphone 1	8.78	91..22	83.00	21.53	79.27	74.00
Triphone 2	5.38	94.62	90.00	23.53	76.47	77.00
Triphone 3	5.35	94.65	90.00	24.37	75.63	77.00

VII. CONCLUSION

In this work, the NKASR system is implemented using Kaldi software for different acoustic modelling such as Monophone modelling and Triphone modelling. The sentence accuracy obtained is 77% for monophone modelling. Triphone 1 model [Delta + Delta-Delta] based on monophone alignment, resulted in best sentence accuracy of 83%. Triphone 2 model [LDA+MLLT+SAT] based on triphone 1 alignment with sentence level accuracy of 90% and Triphone 3 model using triphone 2 alignment giving best sentence level accuracy of 90% as discussed in case I. NKASR is tested for north Kannada accent as discussed in case II. Sentence accuracy for the same acoustic models (monophone, Triphone 1, Triphone 2, and Triphone 3) obtained is 66%, 74%, 77% and 77% respectively. Hence the maximum recognition accuracy obtained for neutral Kannada is 90% and for north kannada accent maximum recognition accuracy is 77%.

NKASR system performance gets degraded for north Kannada accent, because NKASR model is trained only for neutral Kannada accent and the model is aware for this neutral Kannada accent. Thus the future scope of this work is to increase the performance of Kannada ASR system for different dialects of Karnataka.

REFERENCES

1. K Sreenivasa Rao and Shashidhar G Koolagudi, “Identification of Hindi Dialects and Emotions using Spectral and Prosodic features of Speech”, systemics, cybernetics and informatics vol. (9), issn : 1690-4524, 2011.
2. https://en.wikipedia.o rg/wiki/Kannada_dialects.
3. Mahnoosh Mehrabani, John H. L and Hansen, “Automatic analysis of dialect/language sets”, Int. Journal of Speech Technologies (2015) 18:277–286.
4. Mohamed G. Elfekya, Pedro Morenoa and Victor Sotob,“Multi-Dialectical Languages Effect on Speech Recognition”, International Conference on Natural Language and Speech Processing, ICNLSP 2015.
5. Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thoma, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, “Joint modeling of accents and acoustics for multi-accent speech recognition”, arXiv:1802.02656v1 [cs.CL] 7th Feb 2018.
6. Maryam Najafian and Martin Russell, “Modelling Accents for Automatic Speech Recognition”, 23rd European Signal Processing Conference (EUSIPCO)©2015 IEEE.

7. Louisten Bosch, “asr, dialects, acoustic/phonological distances”, 6th International Conference on Spoken Language Processing (ICSLP) 2000 Beijing China October 16-20, 2000.
8. Emelie Kullmann, “Speech to Text for Swedish using KALDI”, Master’s Thesis in Optimization and Systems Theory, Royal Institute of Technology School of Engineering Sciences, 2016.
9. Jack Xin and Yingyong Qi,“Mathematical Modeling and Signal Processing in Speech and Hearing Sciences” springer publications; 2014 edition.
10. Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Foreword By-Reddy.“Spoken language processing: A guide to theory, algorithm, and system development”, 2001.
11. http://citeseerx.ist.psu.edu.
12. Anand H. Unnibhavi, D.S.Jangamshetti, Shridhar K, “Continuous Speech Recognition System for Kannada Language with Triphone Modelling using HTK”,International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Vol 8 (3), September 2019.
13. Daniel Povey and Arnab Ghoshal,“The Kaldi Speech Recognition Toolkit” , 2011.
14. https://arxiv.org/pdf/1811.05553.pdf
15. https://www.eleanorchodroff.com/tutorial/kaldi/training-overview.html
16. Hiroaki Nanjo and Tatsuya Kawahara, “A New asr Evaluation Measure and Minimum Bayes-risk decoding for open-domain speech understanding”, Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.

AUTHORS PROFILE

Anand H. Unnibhavi completed his B.E in Electronics and communication Engineering from Vidya Vardhaka college of Engineering Mysore, affiliated to VTU Belagavi Karnataka and obtained his M.Tech in the area of Digital Electronics and Communcation system from Malnad College of Engineering Hassan , affiliated to VTU Belagavi Karnataka. Currently he is pursuing Ph.D in the area of Speech Processing. His Areas of interest are Speech processing, Wireless network. Presently working as Assistant professor in the department of Electronics and Communication Engineering, Basaveshwara Engineering College Bagalkot,Karnataka, India.

Dr. Dakshayani S. Jangamshetti was born in Ilkal, Karnataka, India on 12th February 1964. She obtained her B.E (Electrical) degree from Karnataka University Dharwad in 1985 and M.Tech.(Instrumentation) Degree from IIT Kharagpur in 1989 & Ph.D (Speech Processing) from IIT, Mumbai in 2003. Her areas of interest include speech signal processing, Image processing, Microcontroller and signal systems. She won the “Outstanding IEEE Branch Counselor” award for the year 2014. Presently, she is a Professor of Electrical and Elecronics Engineering department at Basaveshwar Engineering College (Autonomous), Bagalkot.

Dr. Shridhar S. Kuntoji received his B.E(Electronics and Communication Engineering) degree from Gulbarga University Gulbarga, M.Tech (Digital Electronics and advanced communication) from NITK Surathkal, and Ph.D in the area of speech signal processing from Shivaji University Kolhapur in the year 1988, 2000 and 2014 respectively. He joined as Lecture in Electronics and Communication Engineering, Basaveshwara Engineering College Bagalkot, India in the year 1993, where he is currently working as Professor since 2014. He is currently involved in the research area of speech enhancement focusing on people suffering from hearing loss.