International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I7P227

Volume 70 Issue 7, 269-274, July 2022
© 2022 Seventh Sense Research Group®

Original Article

Investigating Khasi Speech Recognition Systems using a Recurrent Neural Network-Based Language Model

Fairriky Rynjah[1], Bronson Syiem[2], L. Joyprakash Singh[3]

[1, 2, 3]*Department of Electronics and Communication Engineering, North Eastern Hill University, Shillong, Meghalaya, India.*

[1]fairriky.rynjah@gmail.com

Received: 15 May 2022 Revised: 01 July 2022 Accepted: 10 July 2022 Published: 25 July 2022

Abstract *- The language model (LM) plays a vital role in automatic speech recognition systems (ASRs), and it remains a challenging task, particularly with low/under-resourced languages. Khasi language being an under-resourced language, very little study has been done on the Khasi speech recognition system. To date, no Khasi speech recognition system has been developed using a recurrent neural network-based language model (RNN-LM). This paper presents an investigation of Khasi speech recognition systems using an RNN-LM. In the study, different acoustic models (AMs) are built. The study shows that RNN-LM performs better compared to the traditional N-gram model. Further, using RNN-LM, a reduction of word error rate (WER) in the range of 2.8-3.8% for more speech data and 2.4-3.5% for lesser speech data are observed. In addition, two acoustic features are studied, and from the experimental results, it is found that the Mel frequency cepstral coefficient (MFCC) yields better performance than perceptual linear prediction (PLP). The investigation is performed in the two most widely spoken dialects of the Khasi language.*

Keywords *- Acoustic model, Deep neural network, Language model, Under-resourced language, Word error rate.*

1. Introduction

Speech recognition for under-resourced languages has gotten prefatory attention in the past few years [1]. The ASR systems have been developed in many rich-resource languages, including English, Mandarin, Japanese, and Hindi. The development of ASR systems often necessitates a large volume of speech and text data. The world's languages can be categorized as under-resourced, meaning they lack or have minimal access to the resources required for developing technologies like ASR. Researchers working on speech technology development for under-resourced languages are exploring various possibilities for dealing with this challenge and establishing resources and technologies in as many languages as feasible [2]. Khasi language is an under-resourced language, and very little study has been done on the Khasi speech recognition system. So far, no Khasi speech recognition system has been developed using an RNN-LM. This paper proposes an investigation of Khasi speech recognition systems using RNN-LM.

The frequently used technique for building acoustic models for speech recognition systems is the hidden Markov models (HMMs) [3, 4]. With the advancement in technology, the implementation of ASR using neural networks has outperformed the traditional models. Deep neural networks (DNNs) are known for their learning and generalizing abilities for the input features, whereas HMM is known for its sequential modelling. However, DNN alone will be ineffective because it only accepts fixed-size input. HMM aids in taking dynamic information, and DNN aids in learning complex characters [5]. DNN is a multi-hidden-layer feed-forward artificial neural network. Each hidden layer uses a nonlinear function to translate the feature input from the layer below to the current layer, as shown in Equation 1 [6].

$$y = \frac{1}{1 + e^{-(b + xw)}} \qquad (1)$$

Where y, b, x and w denote the output unit, the bias, the input feature, and the weights between connections, respectively. The architecture of the DNN model is shown in Figure 1. The deep neural network is discriminatively trained in each training step using back-propagation derivatives of a minimization problem that analyses the mismatch between the projected and discovered outputs. Supervised and unsupervised learning are the two methods to pre-train a DNN [7, 8].

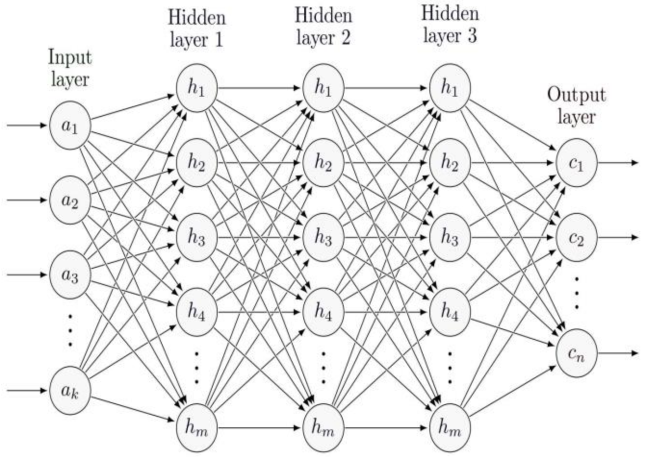

Fig. 1 Architecture of deep neural network

The language model (LM) is an integral part of an ASR system. The LM provides context for distinguishing between phonetically identical words and sentences. Language models depend on acoustic models (AM) to transform analog speech waves into digital and discrete phonemes that make up the foundations of words. The standard N-gram model's purpose is to predict the next word in a textual data set given the context. This model uses a limited-length context by a value of *N-1*, which is challenging. This issue can be overcome using RNN-LM since neurons can represent history with recurrent connections that limit the context length [9, 10].

2. Related Works

Several published works have inspired this study. Despite significant advances in the state-of-the-art ASR systems, a challenging task still exists, specifically with under-resourced languages. Syiem et al. studied Khasi speech recognition systems using different spectral features and HMM states. The study focused on one dialect (standard Khasi), and they did not incorporate RNN-LM [11]. It is observed from the previously published research studies that there has been significantly improved recognition performance while using the RNN-LM. Shraf et al. developed an isolated speech recognition system using HMM and obtained a minimal WER of 10.6% with the speech corpus taken from 10 speakers and a vocabulary size of 52 words [12]. Upadhyaya et al. used deep neural networks to build a Hindi speech recognition system based on the AMUAV Hindi speech database. The lowest WER was reported to be 11.63 % [13]. Popovic et al. developed a DNN-based continuous speech recognition system for the Serbian language using the Kaldi and obtained an optimal WER of 1.86% with three hidden layers. The system also showed an improvement in DNN with a WER of 48.5% over GMM-HMM with a WER of 62.39% [14]. Mikolov et al. performed a speech recognition system with different standard databases, and their outcomes showed more reduction in WER [9]. Smit et al. conducted a speech recognition system for four languages (Finnish, Swedish, Arabic, and English). Their results showed that RNN-LM performed better than traditional N-gram LM [15]. Amberkar et al. investigated Speech-to-text conversion using RNN, and their studies improved recognition performance [16]. Using RNN-LM in the Hindi speech recognition system significantly improved the traditional N-gram LM [10]. Similarly, Gandhe et al. compared feed-forward neural network (NN) and RNN-LM for ASR, and their findings showed better performance with RNN-LM [17].

3. Khasi Language

The Khasi language is an Austro-Asian language belonging to the Mon-Khmer branch, spoken by the native people of the state of Meghalaya [11, 18]. The Khasi language is widely spoken in the state of Meghalaya. As per the Statistical Handbook of Meghalaya 2008, the population of Meghalaya language 2001, around 48.6% of the population speaks Khasi language [19]. Dialects differ to some extent depending on geographical region and local inhabitants. Bareh postulated 11 Khasi dialects based on this [20]. Khasi (Sohra dialect) is considered the standard Khasi dialect [11]. Nongkrem dialect is spoken in almost all the areas surrounding Shillong city and within the city. There exist some variations between the Nongkrem dialect and the Khasi standard dialect. The standard dialect is also known as the Sohra dialect, widely spoken by the people of the East Khasi Hills district and other parts of Meghalaya. The standard dialect is like an intermediary dialect for communication with other dialects of the Khasi language [19].

4. Database Description

A speech database must be developed to build an automatic speech recognition system. The initial step taken was preparing text corpora to be recorded from the native language speakers with two widely spoken dialects which are important for creating a language model. The sentences for preparing the text corpus were collected from local newspapers, books, and the Khasi-English dictionary. More efforts have been made to form sentences with 6 to 18 words. In the case of the Nongkrem dialect, each sentence has been modified as per the pronunciation. There is not much variation in the dialect, but the tone of each word differs from speaker to speaker. The standard procedure was followed for collecting both the text corpus and the speech data. The number of occurrences of each tone was enhanced to record all kinds of co-articulation effects in the spoken language while reducing manual recording efforts [23]. Recordings of audio files were correctly done to omit noisy background and long silence. Approximately 21 hours and 4 hours of speech files for standard Khasi and Nongkrem dialects were recorded conversationally. While recording data, proper positioning of the speakers was maintained to record the speech through the microphone without distortion and with minimal background noise. The selected sentences of each dialect were cross-checked with the data obtained for transcription and then properly assessed to enhance overall transcription accuracy.

Details of development data for both dialects are shown in Tables 1 and 2, respectively. Two files from each dialect were taken as an example to illustrate the organization of each wave file with the corresponding transcription file. For instance, in the wave file 0136-005-06767 in the standard Khasi dialect, the first four digits (0136) represent the speaker ID, and the second three digits (005) represent the dialect ID, and the last five digits (06767) represent the sentence ID. Similarly, the Nongkrem dialect follows the same format**.** Table 3 depicts the structure of filenames used in the experiment.

Table 1. Database description details for standard Khasi dialect

Recording tool	Zoom H4N Handy Portable Digital Recorder
Sampling frequency	16 kHz
Speaker distance from microphone	30 cm
Channel	Mono
Wave file duration	4-6 seconds
Speakers age	18-55 years
Language	Khasi
Dialect	Sohra (Standard)
Total No. of speakers	241
No. of Male speakers	131
No. of Female speakers	110
Wave file per speaker	50
Duration of entire speech data	Approximately 21 Hrs
Total No. of sentences	12050
Total No. of words vocabulary	119069

Table 2. Development data details for Nongkrem dialect.

Recording tool	Zoom H4N Handy Portable Digital Recorder
Sampling frequency	16 kHz
Speaker distance from microphone	30 cm
Channel	Mono
Wave file duration	4-6 seconds
Speakers age	18-55 years
Language	Khasi
Dialect	Nongkrem
Total No. of speakers	42
No. of Male speakers	17
No. of Female speakers	25
Wave file per speaker	50
Duration of entire speech data	Approximately 4 Hrs
Total No. of sentences	2098
Total No. of words vocabulary	9175

Table 3. Structures of file names used in the experiment.

Dialect	Wave files	Transcription files
Standard Khasi	0139-05-06767.wav	0139-05-06767.trans
	0222-05-11064.wav	0222-05-11064.trans
Nongkrem	0242-06-12236.wav	0242-06-12236.trans
	0248-06-12542.wav	0248-06-12542.trans

5. Experimental Setup

The experiment was performed using the Kaldi ASR toolkit in the Ubuntu 18.04 long-term support (LTS) platform. Using Kaldi as a toolkit for real-time speech recognition is advantageous since it generates lattices of high quality and is fast enough to handle real-time recognition [21]. Recipes in Kaldi provide a detailed description of the steps necessary for creating a recognizer for speech databases and making speech recognition software available to scientists and programmers [22]. Training an ASR system involves two main modules, an AM and LM, as shown in Figure 2. The AM module uses acoustic feature vectors extracted from the wave files as input. For this study, MFCC and PLP [23] have been used for two different acoustic features. Feature vectors were extracted using a Hamming window of 25 ms size shifted by 10 ms.

On the other hand, LM uses transcription labeled files that correspond to the wave files. Before building the LM, firstly, the pronunciation dictionary/lexicon has to be created, and this was performed by breaking down the word sequences as shown in Table 4. In the present work, five different acoustic models for each dialect, namely monophonic-based GMM-HMM, Tied state-based GMM-HMM (Tri1), linear discriminative analysis (Tri2), adaptive speaker training (Tri3), and hybrid HMM-DNN have been developed. 39- Dimensional feature vectors were used as input for training the monophonic and Tri1 systems. However, Tri2, Tri3, and HMM-DNN use only 13-Dimensional features as input. While training the HMM-DNN system, the number of hidden layers (HL) varied from 1 to 7, and the optimal value was observed with HL=3. Additionally, two different LMs were built (viz. N-gram and RNN-LM). For N-gram LM, the value of N has been set from 2 to 4 (i.e., bi-gram, tri-gram, and 4-gram), and the optimal result was obtained with bi-gram LM. Furthermore, the RNN-LM has been trained with different HL ranging from 50-300 to observe the minimal value of perplexity, and from the study, it was found with HL=200 for both the dialects.

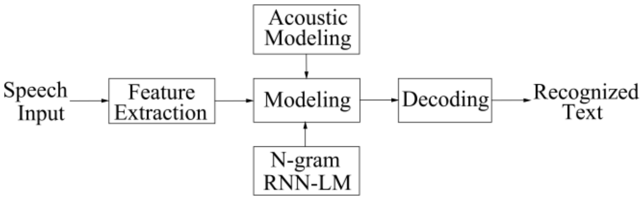

Fig. 2 Block diagram of an automatic speech recognition system

Table 4. Illustration of dictionary/lexicon used in the experiment.

Word sequence	Phone sequence
jingsngewtynnad	j i ng s ng e w t i n n a d
bhabriew	bh a b r i e w

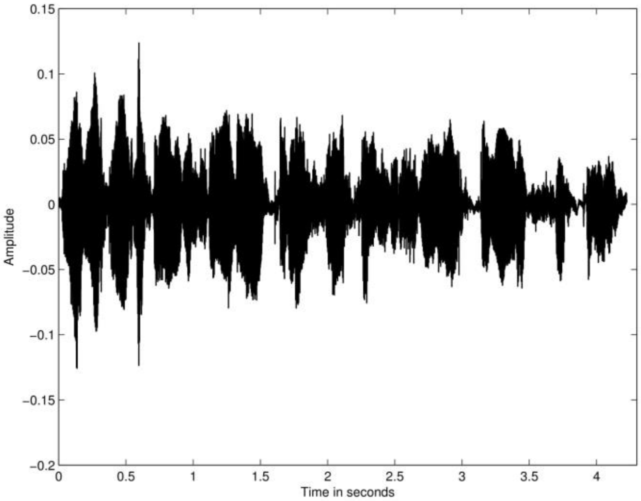

Fig. 3 The short selected speech file waveform corresponds to the sentence "la iohi ba kine ki samla ki la hap ban shu shaniah tang ia ka sorkar."

6. Results and Discussion

In this study, two different databases (development) have been used. As discussed in section. 5, different models have been built, and the comparisons of the results are shown in Tables 5 and 6, respectively. Using the monophonic model, the experimental results gave a poor performance for both dialects. It might be attributed to the current insufficient variation of phones in terms of the left and the right contexts, respectively, as stated in [22]. Further observations were made by incorporating tied state triphone, LDA, and SAT systems. Though much improvement was obtained, the results were not satisfactory.

Furthermore, more improvements were obtained using the hybrid HMM-DNN system. It might be because the hybrid system is trained as a discriminative classifier [21]. Re-scoring the traditional systems, particularly the monophonic system with RNN-LM. Though the results obtained were unsatisfactory, some reduction in WER can be seen. The WER is evaluated using Equation 2 [24]. In addition, using RNN-LM, reduction of WER in the range of 2.8-3.8% for more speech data and 2.4-3.5% for fewer speech data were observed compared to traditional N-gram LM, and the results are shown in Figures 4 and 5. In terms of features used, it can be observed that MFCC provided better performance irrespective of models and dialects. This may be because of the non-linearity of the speech signal, as stated [25].

$$WER(\%)=\frac{S+D+I}{N}*100 \tag{2}$$

Where N, I, S, and D denote the number of words in the test set, number of insertion errors, substitutions, and deletions, respectively.

Table 5. Comparison of WER (%) for N-gram LM and RNN-LM evaluated from different models for standard Khasi dialect.

Model	N-gram		RNN-LM	
	MFCC	PLP	MFCC	PLP
Monophone	23.87	24.65	21.01	21.80
Tri1 (Triphone)	15.61	16.12	12.32	12.91
Tri2 (LDA+MLLT)	14.82	15.53	11.32	12.14
Tri3 (LDA+MLLT+SAT)	13.92	14.59	10.24	10.97
DNN	10.06	10.39	6.21	6.58

Table 6. Comparison of WER (%) for N-gram LM and RNN-LM evaluated from different models for the Nongkrem dialect.

Model	N-gram		RNN-LM	
	MFCC	PLP	MFCC	PLP
Monophone	27.84	28.09	25.33	25.66
Tri1 (Triphone)	18.45	18.62	15.34	15.55
Tri2 (LDA+MLLT)	18.13	18.29	14.84	15.06
Tri3 (LDA+MLLT+SAT)	17.91	18.07	14.51	14.69
DNN	12.88	13.11	9.31	9.69

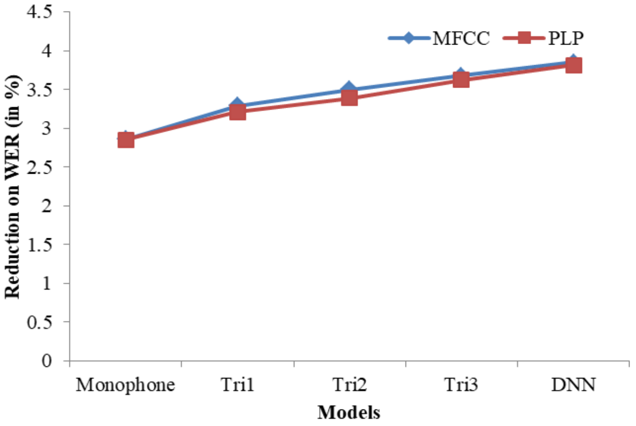

Fig. 4 Reduction on WER (%) obtained by re-scoring with RNN-LM for standard Khasi dialect evaluated from different models.

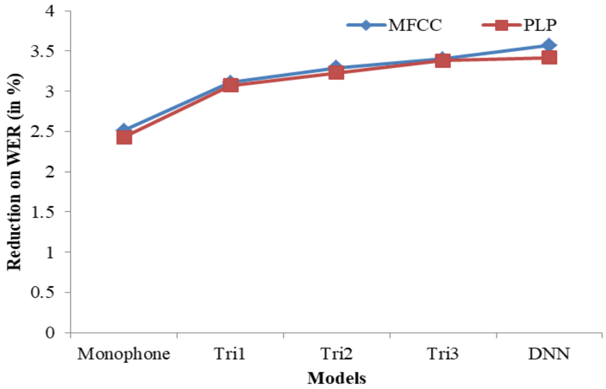

Fig. 5 Reduction on WER (%) obtained by re-scoring with RNN-LM for Nongkrem dialect evaluated from different models.

7. Conclusion

In this experiment, an investigation of Khasi speech recognition systems using RNN-LM was performed. As AM is concerned, HMM-DNN outperformed the baseline GMM-HMM. Further, it was observed that RNN-LM performed better than traditional N-gram LM. Also, it was found that the reduction of WER correlates with the speech data used. Similarly, it was observed that in terms of features used, MFCC provided better performance compared to PLP, irrespective of models and dialects. Future work may involve increasing speech data and incorporating more dialects and machine learning tools.

Funding statement

The Ministry of Tribal Affairs supports the Government of India's research under the National Fellowship for Scheduled Tribes with award letter No. [201819-NFST-MEG-00253].

Acknowledgment

The authors would like to acknowledge research colleagues, friends, and relatives for their outstanding support in giving their precious time in reaching out to native language speakers to collect speech data.

References

[1] L. Besacier, E. Barnard, A. Karpov and T. Schultz, “Automatic Speech Recognition for Under-Resourced Languages: A Survey,” *Speech Communication*, Vol 56, No.1, Pp.85–100, 2014.

[2] F. De Wet, N. Kleynhans, D. Compemello and R. Sahraeian, “Speech Recognition for Under-Resourced Languages: Data Sharing in Hiddenmarkov Model Systems,” *South African Journal of Science*, Vol.113, No.(1/2), Pp.1-9, 2017.

[3] .E. Baum and J.A. Eagon, “An Inequality with Applications To Statistical Estimation for Probabilistic Functions of Markov Processes and To A Model for Ecology,” *Bulletin of American Mathematical Society*, Vol. 73 , Pp. 360–363, 1967.

[4] M. Gales and S. Young, “the Application of Hidden Markov Models in Speech Recognition,” *Foundations and Trends in Signal Processing*, Vol.1, No.3, Pp.195–304, 2007.

[5] V. Manohar, D. Povey and S. Khudanpur, “Semi-Supervised Maximum Mutual Information Training of Deep Neural Network Acoustic Models,” *Interspeech*, Germany, Pp. 2630–2634, 2015.

[6] L. Longfei, Z. Yong, J. Dongmei and Z. Yanning, “Hybrid Deep Neural Network - Hidden Markov Model (Hmm-Dnn) Based Speech Emotion Recognition,” *Humaine Association Conference on Affective Computing and Intelligent Interaction, Switzerland*, Ieee Computer Society, Pp. 312-317, 2013.

[7] F. Seide, G. Li, X. Chen and D. Yu, “Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription,” *Automatic Speech Recognition and Understanding* (Asru), Ieee Workshop, Pp. 24–29, 2011.

[8] G.E. Hinton, S. Osindero and Y.W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation*, Vol.18, No.7, Pp.1527–1554, 2006.

[9] T. Mikolov, M. Karafiat, L. Burget, J.H. Cernocky and S. Khudanpur, “Recurrent Neural Network Based Language Model,” *Interspeech*, Japan, Pp. 1045-1048, 2010.

[10] M. Dua, R.K. Aggarwal and M. Biswas, “Discriminatively Trained Continuous Hindi Speech Recognition System Using Interpolated Recurrent Neural Network Language Modeling,” *Neural Computing and Applications*, Vol. 31, Pp. 6747-6755, 2018.

[11] B. Syiem, S.K. Dutta, J Binong and L.J. Singh, “Comparison of Khasi Speech Representations with Different Spectral Features and Hidden Markov States,” *Journal of Electronic Science and Technology*, Vol.19, No.2, Pp.1-7, 2020.

[12] J. Ashraf, N. Iqbal, N.S. Khattak and A.M. Zaidi, “Speaker Independent Urdu Speech Recognition,” *International Conference on Informatics and Systems* (Infos), Egypt, Pp. 1-5, 2010.

[13] P. Upadhyaya, S.K. Mittal, O. Farooq, Y.V. Varshney and M.R. Abidi, “Continuous Hindi Speech Recognition Using Kaldi Asr Based on Deep Neural Network,” *Advances in Intelligent Systems and Computing*, Springer, Singapore, Vol.748, Pp.303–311, 2019.

[14] B. M. Popovic, S. Ostrogonac, E. Pakoci, N. Jakovljevic and V. Delic, “Deep Neural Network Based Continuous Speech Recognition for Serbian Using the Kaldi Toolki, Speech and Computer,” *Lecture Notes in Computer Science*, Springer, Greece, Vol. 9319, Pp.186-192, 2015.

[15] P. Smit, S. Virpioja and M. Kurimo, “Advance in Subword-Based Hmm-Dnn Speech Recognition Across Languages”, *Computer Speech and Language*,” Vol.66 , Pp.1-17, 2020.

[16] A. Amberkar, G. Deshmukh, P. Awasarmol and P. Dave, “Speech Recognition Using Neural Network,” *Ieee International Conference on Current Trends Towards Converging Technologies*, Coimbatore, Pp.1-4, 2018.

[17] A. Gandhe, F. Metze and I. Lane, “Neural Network Language Models for Low Resource Languages,” *Interspeech*, Singapore. Pp. 2615-2619, 2014.

[18] B. Syiem and L.J. Singh, “Deep Neural Network-Based Phoneme Classification of Standard Khasi Dialect,” *International Journal of Applied Pattern Recognition*, Vol.6, No.1, Pp.43-51, 2019.

[19] E. Syiem, Ka Ktien Nongkrem Ha Ki Pdeng Rngi Lum Ka Ri Lum Khasi, Lynnong 7, Book Chapter From People’s Linguistic Survey of India, Meghalaya, Vol.19, Pp.135-136, 2014.

[20] S. Bareh, Khasi Proverbs: “Analysing the Ethnography of Speaking Folklore”, Ph.D. Dissertation, Dept. Cultural and Creative Studies,” North Eastern Hill University, Shillong, 2007.

[21] B. Syiem and L.J. Singh, “Exploring End-To-End Framework Towards Khasi Recognition System,” *International Journal of Speech Technology*, Vol.24, No.8 , Pp.419-424, 2021.

[22] J. Guglani and A.N. Mishra, “Continuous Punjabi Speech Recognition Model Based on Kaldi Asr Toolkit,” *International Journal of Speech Technology*, Vol.21 , Pp. 211-216, 2018.

[23] F. Rynjah, B. Syiem, and L.J. Singh, “Khasi Speech Recognition System Using Hidden Markov Model with Different Spectral Features: A Comparison,” *International Conference on Industry Innovations in Science, Engineering and Technology*, 2019.

[24] P. Upadhyaya, S.K. Mittal, Y.V. Varshney, O. Farooq and M.R. Abidi, “Speaker Adaptive Model for Hindi Speech Using Kaldi Speech Recognition Toolkit,” *International Conference on Multimedia, Signal Processing and Communication Technologies* (Impact), Pp. 232-236, 2017.

[25] F. Rynjah, B. Syiem, and L.J. Singh, “Speech Recognition System of Spoken Isolated Digit in Standard Khasi Dialect,” *Proceedings of International Conference on Frontiers in Computing and Systems,” Lecture Notes in Networks and Systems*, Springer, Vol. 404 , Pp. 541–549, 2021.