[1]Dr. S. Deepajothi

[2]Dr. Vuda Sreenivasa Rao

[3]Dr C Ambhika

[4]Vishwanadham Mandala

[5]R V V N Bheema Rao

[6]Dr Shailendra Kumar

[7]Dr. Venkateswara Rao Gera

[8]Dr D Nagaraju

A Comparative Study of Khasi Speech Recognition Systems with Recurrent Neural Network-Based Language Model

Abstract: - This paper offers a comparative analysis of Khasi speech recognition systems utilizing a recurrent neural network-based language model (RNN-LM). Develop different acoustic models (AMs) to evaluate the optimal performance. This paper observed that using RNN-LM performed best than traditional other models. The wave surfer performs data processing followed by collecting the recorder based continuous speech database. Moreover, a minimization of word error rate (WER) in 2.83.8% range for major speech data and 2.4-3.5% for minor speech data. Additionally, two acoustic features are used, and from the experimental results, the Mel frequency cepstral coefficient (MFCC) yielded improved performance than the perceptual linear prediction (PLP).

Keywords: Hidden Markov model, Language model, Perceptual linear prediction, Gaussian mixture model, Acoustic model

[1] [1]*Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur- 603 203, Chengalpattu Dist, Tamilnadu, India*

deepajos@srmist.edu.in

[2]*Associate professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, INDIA.522302*

vsreenivasarao@kluniversity.in

[3]*Associate Professor, Department of AIML, R.M.D Engineering College, RSM Nagar, Kavarapetai*

ambhi durai@gmail.com

[4]*Data Engineering Leader,Indiana University,Bloomington, Indiana*

Vmandala@iu.edu Personal email: Vishwanadh.mandala@gmail.com

[5]*Associate Professor, Dept. of Information Technology,Aditya College of Engineering & Technology, Surampalem, India*

rvvnbrao@gmail.com

[6]*Associate Professor, Department of ECE, Integral University Lucknow, Uttar Pradesh, India. Pin code: 226026*

skumar@iul.ac.in

[7]*Professor, Department of CSE, Kallam Haranadhareddy Institute of Technology,*

gvraocse777@gmail.com

[8]*Professor, Department of CSE, Sri Venkatesa Perumal College of Engineering and Technology, Puttur, Andhra Pradesh, India*

raj2dasari@gmail.com

1. INTRODUCTION

Speech recognition to less resource language groups attained little attention in recent years [1]. Many well-resource languages, such as Mandarin, English, Hindi, and Japanese have had automatic speaker recognition (ASR) systems developed. A large volume of speech and text information is frequently needed to the development of the ASR systems. The major world's languages are under-resourced, which means they have less access to needed resources for the development of ASR. Research on the development of speech technological are useful approaches to addressing the above challenging problem and introducing technologies and resources in several language groups as possible [2].

The well-known throughout the world for its rich linguistic and cultural diversity is north east India. Presently, scientists and language teachers recognise this as one ‘s dedication to make a contribution to audio and describing the language of just about every society in attempt to uphold and revitalise one ‘s language [3]. Khasi, a member of the Austro-Asian language family, is widespread spoken about in Meghalaya's Khasi and Jaintia hills municipalities, as well as along the state's and nation's territory. These same dialects are the sub division of language [4]. The major contribution of the research is described as;

- ❖ This paper describes a comparative study on Khasi recognition of speech systems using a RNN-LM.
- ❖ RNN-LM performed superior to the traditional N-gram model, PLP and LPREFC.

2. LITERATURE SURVEY:

Incorporating RNN-based LM from the previously published research studies has shown significant improvement towards recognition performance. RNN-LM towards Hindi speech recognition system showed significant improvement over the traditional N-gram LM [5]. J. Ashraf *et al.* introduced HMM that achieved less WER of 10.6% with speech corpus from 10 speakers and the vocabulary size, 52 words [6]. Upadhyaya *et al.* developed Deep Neural Networks for speech recognition system in Hindi language. Here, 1000 balanced phonetically utterances recorded by 100 speakers from 54 Hindi phone numbers are included in the dataset. The minimal WER obtained is 11.63 % [7].

P. Smit *et al.* develops DNN for speech recognition of continuous to Serbian by Kaldi that gained best WER value of 1.86%. This method shows a growth of DNN with 48.5% WER and 62.39% WER for GMM-HMM [8]. V. Manohar *et al.* devised a speech recognition system with various databases, and its result gained high reduction in WER [10].

T. Mikolov *et al.* developed speech recognition system for different languages (Arabic, English, Swedish, Finnish) in which RNN-LM gained best when compared to the traditional N-gram LM [11]. Similarly, B. Popovic *et al.* performed a comparison between RNN-LM and feed forward NN for ASR, and its findings showed the optimal performance [12].

3. IMPORTANCE OF KHASI LANGUAGE

Khasi is an O-Asiatic language of the Mon-Khmer branch and spoken in Meghalaya [12]. It is widely used in Meghalaya. According to the Meghalaya Statistical Manual 2008 and Population of Meghalaya by Language 2001, about 48.6% of the total population speaks Khasi. Languages differ depending on the geographic region and the local population [13]. Bareh classified 11 Khasi languages based on this. Khasi (Sohra dialect) is considered the standard Khasi language.

The Nongkrem dialect is spoken everywhere around the city of Shillong. This language is also spoken in most parts of Shillong city [14]. There are many differences between the Khasi Nongkrem dialect and the standard Khasi dialect. The standard dialect, known as Sohra dialect, was widely spoken in and around the East Khasi hills of Meghalaya when a British government moved the capital from Sohra to Shillong in when they ruled India. The standard language is like an intermediate language that speaks to other Khasi dialects [15].

4. MODEL OF SPEECH RECOGNITION SYSTEM

A speech recognition device intended for human interaction that has understanding of speech analysis and variables that can be used to achieve the best possible reaction [16]. Figure 1 describes the training an ASR system includes two major modules such as language model (LM) and acoustic model (AM).

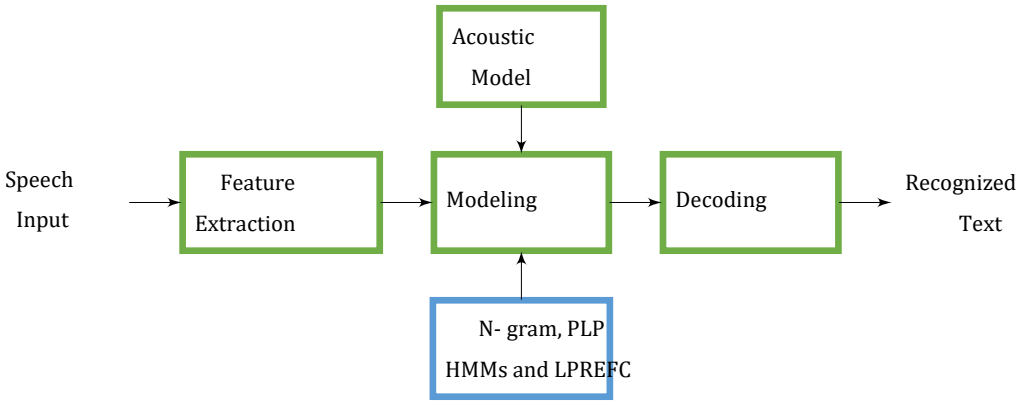

Figure 1: Pictorial representation of an Automatic Speech Recognition System

4.1 Hidden Markov Models (HMMs):

HMM is a commonly used technique for creating acoustic models for the speech recognition systems [17]. With technological advancements, neural networks have outperformed classic ASR models. DNN are well-known for their ability to learn and generalize input features, whereas HMM is known for sequential modeling. However, DNN will be unsuccessful on its own because it only accepts fixed-size inputs. HMM helps with dynamic data collection, while DNN helps with sophisticated learning [18].

4.2 Perceptual Linear Prediction (PLP):

PLP represents a seamless brief spectral range equalised and condensed in a manner analogous to the human auditory cortex, getting closer to Michael characteristics [19]. It PLP gives useful settlement at higher frequency, implying an aural filtration financial institution tactic, while still producing orthogonal output data similar to mel frequency cepstrum evaluation [20]. Because it uses sequential forecasts for spectroscopic sculpting, it is called linear perception prognostication. The linear prediction analysis and spectral analysis combination is PLP.

4.3 Deep Neural Network

A DNN consists of several hidden layers, where each hidden layer uses dynamic targets to first transform the representation input from the current layer to the lower layer, as shown in Equation 1 [21].

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

Where, the output unit is y, b denotes the bias, the weights between connections is w and input feature is x. Figure 2 describes the architecture of the DNN model. The deep neural network is adversarial trained using back-propagation variants of a minimization problem that analyzes the mismatch between projected and discovered outputs. The two techniques for pre-training a Deep Convolutional neural System are supervised or unsupervised learning [22].

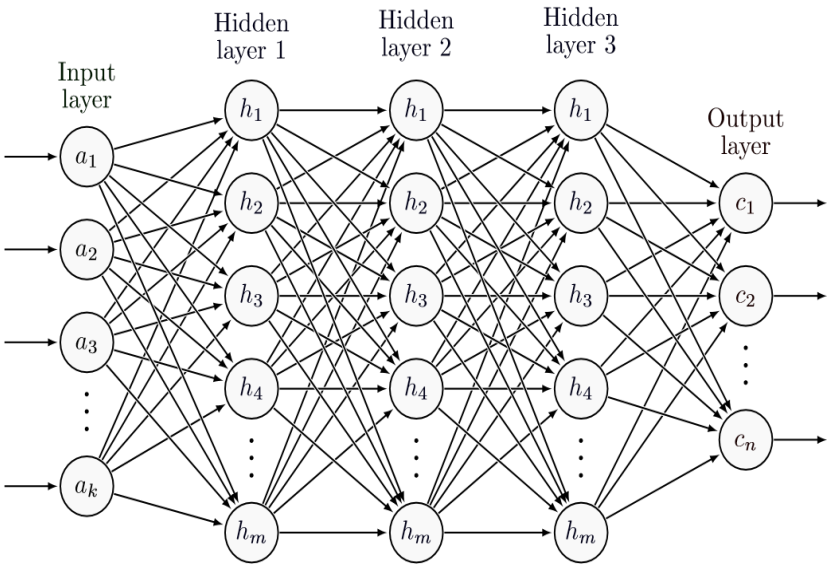

Figure 2: Basic DNN structure

4.4 Linear predictive reflection coefficients (LPREFC):

LPCC characteristics seem to be companies issued properties actually derived from coefficients of LPC. These cepstrals are indeed the explanatory variables of the IFT frequency band log - linear amplitude, that have proven to be more an efficient and consistent typically subject for voice recognition execution [23].

4.5 Language Model

The LM is an essential component of the system for automatic speech recognition. The LM provides information to help distinguish between phonetically identical sentences and words [24]. Language models rely on AM to convert analog speech waves to digital and discrete phonemes that serve as the foundation of words. The goal of a standard N-gram model is to predict the next word in a text dataset based on its context. This example uses a finite length N-1 value context, which is challenging. This problem can be solved by using RNN LM because the neurons can represent the history through connections, so the length of the context is infinite [25]. This paper presents a comparative study of Khasi information recognition systems using LM based on RNN.

5. DESCRIPTION OF DATASET

A database of information must be created to build a speech recognition system. The first step is to prepare a set of documents, recorded by native speakers of two widely spoken languages, which is essential to create a language model. We followed standard procedures to collect text and audio data [26]. It increases the number of events of each note to record the various effects of the spoken language and reduces the manual writing tasks. Audio files have been properly recorded to eliminate background noise and long periods of silence. Conversations were recorded for 21 hours and 4 hours of conversation in Khasi and Nongkrem Standard languages respectively [27]. When recording objects, we place the speakers so that the microphone can capture speech without problems and with less background noise. Selected sentences from each language are checked against the data from the text and reviewed to improve the accuracy of the text. Table 1 provides more information about language development.

Table I: Data analysis for the Khasi Standard language

Tool for recording	Zoom H4N Handy Portable Digital Recorder
Dialect	Sohra (Standard)
Male Speakers	131
Sampling frequency	16 KHz

Wave files per speaker	50
Speaker distance from the microphone	30 cm
Wave file duration	4 - 6 secs
Language	Khasi
Channel	Mono
Speakers Age	18- 55 years
Total Speakers	241
Female Speakers	110
Period of Speech data	Approximately 21 Hrs
Whole Sentences	12050
Entire Vocabulary	119069

Table II: Text corpora details

Language	Khasi
Dialect	Sohra (Standard)
Amount of sentences	6000
Amount of words	16310
Number of unique words	1750

6. EXPERIMENTAL ARRANGEMENT

The resulting settings were created with the Kaldi ASR toolbox on the Ubuntu 18.04 Long Term Support (LTS) platform. Kaldi is favored as a toolset for on-the-fly information recognition because it provides higher-quality screens and responds quickly [21]. Kaldi's recipes demonstrate how to build a voice database browser as well as approaches for developing speech recognition software for scientists and the programmers [22]

The AM model uses, namely, acoustic feature vectors as input. For this study, both various acoustic features, namely, PLP and LPREFC [23] have been used. Feature vectors were extracted by Hamming window with a size of 25 ms and a shift of 10 ms. LM, on the other hand, uses a text file that is similar to a waveform file. Before compiling the LM, a pronunciation dictionary must first be created, by analysing single word sequences, as shown in Table III. The transcription file corresponding to the wave file is plotted in Table IV.

Table III: Describing of lexcion employed for investigation

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

Table IV: TRANSCRIPTION FILE CORRESPONDING TO THE WAVE FILE

La kine iohi ba ki samla ki hap la ban shaniah shu tang ka ia sorkar

In this work, we develop five different noise models for each language, namely GMM-HMM on a single call, GMM-HMM (Tri1), and Linear discriminant analysis (Tri2). , Speaker Adaptation Training (Tri3) and hybrid HMMDNN. A 39-dimensional vector is employed as the input for training Monophone and the Tri1 systems. Though, Tri3, Tri2, and the DNN only use 13-dimensional features. When DNN system is trained, the number of hidden layers (HL) is varied in the range 1-7 and the best value is found at HL=3. In addition, two different LMs (i.e., Ngram and RNN-LM) are developed. RNN-LM was trained with different HLs in the range of 50-300 to observe the lowest surprise value and found HL=200 for both languages.

In this investigation, two separate databases (development) were employed. The speech data utilized for training and testing in Standard Khasi is 12050 and 3900, respectively, whereas the training and testing files in Nongkrem are 2098 and 620. As described in Section 6, many models have been developed, and the results are compared in Table V. Figure 3 shows a comparison of WER (in%) for N-gram LM and RNN-LM evaluated using several models for the nongkrem dialect. Using a monophonic sample, the test results show poor performance for both languages. This may be due to the invariance of left-right phonetic contexts, as discussed in [22].

Table V: Assessment of WER (in %) of RNN-LM computed from various models for standard khasi language

Model	***N-gram***		***RNN-LM***	
Monophone	22.88	23.64	20.11	20.70
Tri2 (LDA+MLLT)	16.31	17.22	13.22	13.94
LDA+MLLT+SAT (Tri3)	14.82	15.69	11.14	11.99
DNN	11.16	11.49	7.31	7.68

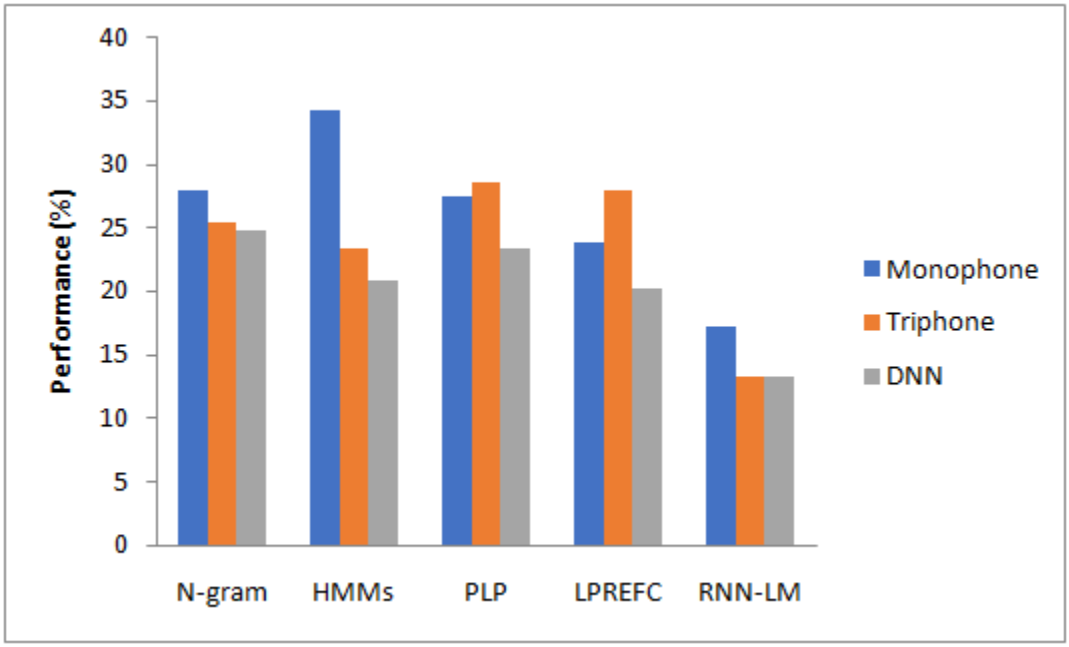

Figure 3: Assessment of WER (in %) RNN-LM for nongkrem dialect

7. RESULTS AND DISCUSSION

Additional observations were made by combining the constrained binary system, SAT and LDA. Although much progress has been made, the results are not satisfactory. In addition, further improvements can be achieved by using a hybrid HMM-DNN system. This is possible because the hybrid system is trained to discriminate. Also benchmarking the traditional system, especially the monophone system and the RNN-LM, although the results

were less than good, it was found that the single word error rate (WER%) was significantly reduced, and the evaluation results are presented Equation 2 [24] . In addition, compared to the traditional N-gram LM, by re-evaluating the RNNLM, it can be seen that the WER is reduced by 2.8 - 3.8% for more information and 2.4-3.5 % for more informative data, as Figure 4 and Figure 5 show. For the conditions used, MFCC can be seen to perform better regardless of model and language. This may be caused by the nonlinearity of the information signal, as in [25].

$$WER(\%) = \frac{(D+S+I)}{N} * 100 \tag{1}$$

Where, the term N denotes the amount of the words test, I denote the insertion errors, S denotes substitutions, and D denotes deletions.

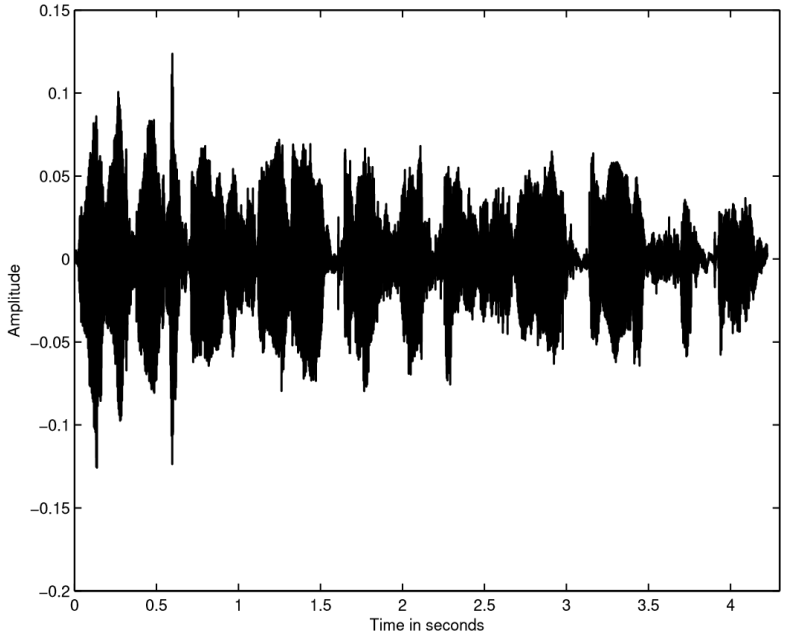

Figure 4: Waveform of short selected speech correspond to sentence “La iohi ba kine ki samla ki la hap ban shu shaniah tang ia ka sorkar”

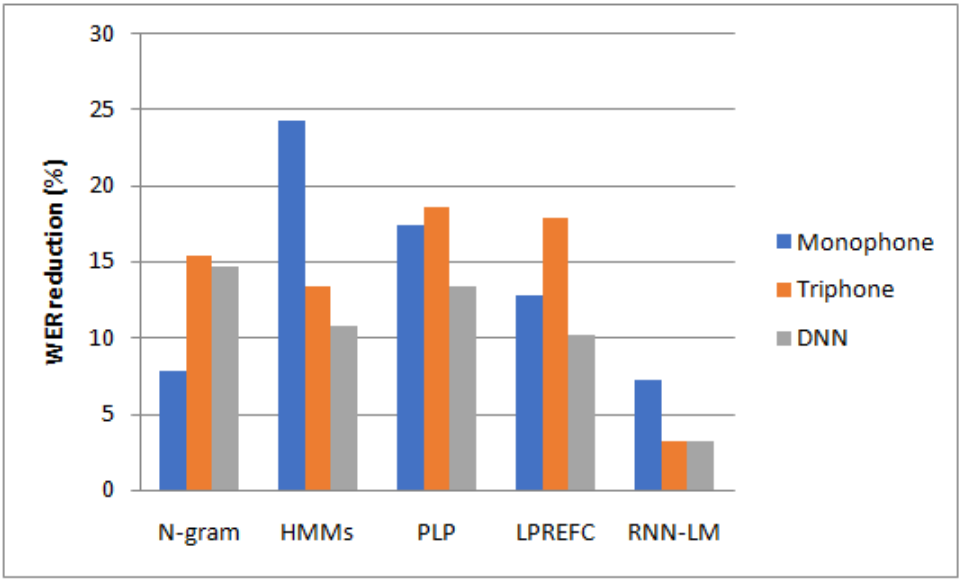

Figure 5: Reduction on WER (in %) performance analysis

8. CONCLUSION

We have investigated speech recognition systems for Khasi dialects using RNN-LM. DNN outperformed the baseline GMM-HMM architecture for several hidden layers, with respect to results obtained. The RNN-LM model outperformed the traditional N-gram model, HMMs, PLP, and Linear predictive reflection coefficients in the study (LPREFC). The wave surfer processes data before collecting the recorder-based continuous speech database. From our research we found that using RNN-LM performs better than traditional N-gram LM. Furthermore, we found that the reduction of WER was related to the information resources used. The standard Khasi dialect contributed to a better reduction in WER than the Nongkrem dialect, which has lesser data. Future work will include adding conversational data and incorporating other languages and machine learning tools.

Conformity to Ethical Standards

Conflict of interest

The authors declare no conflicts of interest.

Animal and Human Rights

This article contains no investigations on animal or human subjects conducted by any of the writers.

Consent with informed knowledge

Informed consent was not appropriate because this was a retrospective review, and no patient information was excluded.

Funding: Not relevant

Statement regarding conflicts of interest: Not relevant

Participation agreement: Not relevant

Consent to publication: Not relevant

Availability of material and data:

Data sharing does not apply to this article because no new data were generated or processed in this investigation.

Availability of code: Not relevant

Conflicting Interests: Not relevant

References:

[1] L. Besacier, E. Barnard, A. Karpov, T. Schultz, ”Automatic speech recognition for under-resourced languages: A survey,” Speech Communication , vol. 56, pp. 85–100, January 2014.

[2] F. de Wet, N. Kleynhans, D. Compemello and R. Sahraeian, ”Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems,” South African Journal of Science, vol. 113, no. 1/2, pp. 1-9, August 2016.

[3] L.E. Baum, 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,” in Foundations and Trends in Signal Processing, vol. 1, pp. 195–304, 2008.

[5] 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, April 2018.

[6] J. Ashraf, N. Iqbal, N.S. Khattak and A.M. Zaidi, “Speaker Independent Urdu Speech Recognition,” International Conference on Informatics and Systems (INFOS), Cairo, Egypt, 2010.

[7] 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, 2019.

[8] P. Smit, S. Virpioja and M. Kurimo, ”Advance in subword-based HMMDNN speech recognition across languages,”Computer Speech and Language, vol. 66, pp. 1-17, Sepember 2020.

[9] V. Manohar, D. Povey and S. Khudanpur, ”Semi-supervised maximum mutual information training of deep neural network acoustic models,” in Proceedings Interspeech, Dresden, Germany, pp. 2630–2634, September 2015.

[10] T. Mikolov, M. Karafiat, L. Burget, J.H. Cernocky and S. Khudanpur, ”Recurrent neural network based language model,” INTERSPEECH, pp.1045-1048, 2010

[11] B. 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, vol 9319,pp. 186-192, 2015.

[12] Nongbri, T., 2000. Khasi women and matriliny: Transformations in gender relations. Gender, Technology and Development, 4(3), pp.359-395.

[13] Warjri, S., Pakray, P., Lyngdoh, S. and Kumar Maji, A., 2019. Identification of POS tag for Khasi language based on hidden markov model POS tagger. Computación y Sistemas, 23(3), pp.795-802.

[14] Makdoh, K., Lynser, M.B. and Pala, K.H.M., 2014. Marketing of indigenous fruits: a source of income among Khasi Women of Meghalaya, North East India. Journal of Agricultural Sciences, 5(1-2), pp.1-9.

[15] Diffloth, G., 2008. Shafer’s parallels between Khasi and Sino-Tibetan. North East Indian Linguistics, 2, pp.93-104.

[16] Collobert, R., Puhrsch, C. and Synnaeve, G., 2016. Wav2letter: an end-to-end convnet-based speech recognition system. arXiv preprint arXiv:1609.03193.

[17] Picone, J., 1990. Continuous speech recognition using hidden Markov models. IEEE Assp magazine, 7(3), pp.26-41.

[18] Woodland, Philip C., and Daniel Povey. "Large scale discriminative training of hidden Markov models for speech recognition." Computer Speech & Language 16, no. 1 (2002): 25-47.

[19] Revathi, A. and Venkataramani, Y., 2009. Text independent speaker recognition and speaker independent speech recognition using iterative clustering approach. AIRCC's International Journal of Computer Science and Information Technology, 1(2), pp.30-42.

[20] Pollak, P. and Behunek, M., 2011, July. Accuracy of MP3 speech recognition under real-word conditions: Experimental study. In Proceedings of the International Conference on Signal Processing and Multimedia Applications (pp. 1-6). IEEE.

[21] Yu, J., Xie, X., Liu, S., Hu, S., Lam, M.W., Wu, X., Wong, K.H., Liu, X. and Meng, H., 2018, September. Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus. In Interspeech (pp. 2938-2942).

[22] Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S. and Glass, J., 2014, December. A complete KALDI recipe for building Arabic speech recognition systems. In 2014 IEEE spoken language technology workshop (SLT) (pp. 525-529). IEEE.

[23] Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M. and Ney, H., 2007. Speech recognition techniques for a sign language recognition system. hand, 60, p.80.

[25] Chelba, C., Bikel, D., Shugrina, M., Nguyen, P. and Kumar, S., 2012. Large scale language modeling in automatic speech recognition. arXiv preprint arXiv:1210.8440.

[26] Rynjah, F., Syiem, B. and Singh, L.J., 2020. Khasi speech recognition using hidden Markov model with different spectral features: A comparison. Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (I3SET2K19).

[27] Basu, J., Khan, S., Roy, R., Basu, T.K. and Majumder, S., 2021. Multilingual Speech Corpus in Low-Resource Eastern and Northeastern Indian Languages for Speaker and Language Identification. Circuits, Systems, and Signal Processing, 40(10), pp.4986-5013.