Performance Analysis of Hybrid Automatic Continuous Speech Recognition Framework for Kannada Dialect

Praveen Kumar P S
Research Scholar
Department of Electronics and Communication Engineering
Siddaganga Institute of Technology
Tumakuru, Karnataka, India.
Email: pravin227@gmail.com

H S Jayanna
Professor and Head
Department of Information Science and Engineering
Siddaganga Institute of Technology
Tumakuru, Karnataka, India.
Email: jayannahs@gmail.com

***Abstract*—This paper demonstrates the execution investigation of the automatic continuous speech recognition system for Kannada language using hybrid modelling techniques. The well-known modelling techniques, in particular, deep neural system (DNN), hidden Markov display (HMM), subspace Gaussian mixture model (SGMM) and maximum mutual information (MMI) have been combined to form the hybrid modelling for speech recognition. The persistent Kannada speech information is gathered from 1600 speakers (960 males and 640 females) of the age bunch in the scope of 8 years-80 years. The speech information is acquired from different geographical regions of the Karnataka state under certifiable condition. It comprises of 20,000 words that spread 30 locale. The point of this paper is to examine the execution of hybrid modelling techniques with regards to Kannada speech recognition. Kaldi toolbox is utilized for the implementation of this system, in which Mel frequency cepstral coefficient (MFCC) is used as a feature extraction procedure. The word error rate (WRR) is the error metric used to determine the efficiency of the automatic speech recognition (ASR) system. The experimental results demonstrate that the recognition rate got through the combination of DNN and HMM is better over other hybrid ASR modelling strategies.**

***Index Terms*—Automatic speech recognition, Continuous Kannada speech, Kaldi toolkit, Hybrid modelling techniques, Mel frequency cepstral coefficient (MFCC), DNN, HMM, SGMM, MMI, Word error rate (WER).**

I. Introduction

Over the most recent 50 years, we have seen consistent improvement in speech recognition (SR) [1]. This advancement can be classified into two components: (i) the utilization of hidden Markov model (HMM) in modelling the temporal varieties in the speech [2] and (ii) the expanding computational 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. It is the known fact that the research on ASR is essentially centred around English and other European dialects [3]. It may be said very well that no considerable advancement has been made 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. Our research centres around overcoming these constraints to assemble a sensibly decent, substantial vocabulary, continuous speech recognition (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 hybrid modelling techniques are the combination of state-of-the-art modelling techniques trained on a few thousand hours of speech. They have significantly reduced the WER. They accomplish word-level correctnesses of almost 90% for vocabulary sizes of around 200,000 for the English speech database [3]. Such ASR frameworks are currently generally utilized for business and entertainment purposes. A productive and robust ASR framework for the Kannada language is the need of our country. By utilizing this any layman gets the advantages of data innovation. Speech recognition is the challenging job undertaking from the most recent couple of decades and still, there is no exact framework which goes about as an interface among man and machine. There are as yet numerous potential outcomes in this field.

Because of the absence of standard speech databases, ASR in Kannada language has not seen any progress. This inspired us to take up the work on building the speaker-independent large vocabulary ASR framework.

The remaining of the paper is sorted out as pursues. Section II portrays the related works in the field of ASR. Section III talks about the phoneme attributes of Kannada language and in Section IV, we portray the databases utilized in building the Kannada ASR framework. Building blocks of an ASR framework and ventures in building a Kannada ASR are talked about in section V. The details of conducting experiments and the evaluation results of continuous Kannada speech recognition (CKSR) system is provided in Section VI. As the ending part of the paper,

we would like to conclude and briefly discuss our future research bearings in Section VII.

II. Related work

The CSR has been a functioning field of research for quite a while. Numerous experiments have been conducted in the recognition of dialects like Punjabi, Hindi, Tamil, Telugu and Kannada and so on [4] [5] [6] [7]. The research related to SR in Hindi utilizing kaldi is accounted in [8]. Numerous toolkits are accessible to the researchers for research in the field of SR like the Sphinx, HTK, Julius and Kaldi.. As of late, Kaldi is a standout amongst the most well known and the most recent toolkit for the specialist who works in the field of ASR which is written 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 [9].

The work in [10] 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 [11] yield scores with HMM emission probabilities. The neural network is prepared under mean square error and the HMM is trained under maximum likelihood estimation. A best in the class HMM framework is joined with a time delay neural network (TDNN) incorporated in a Viterbi framework. A various leveled TDNN structure is portrayed that parts training into subtasks relating to subsets of phonemes. This structure makes training of TDNNs on expansive vocabulary tasks reasonable on workstations. It was demonstrated that the joined framework, regardless of the low precision of the various leveled TDNN, accomplishes a WRR reduction of 15% according to cutting-edge HMM framework.

The literature in [12] portrays the usage of an SR framework in Assamese dialect. The database used for this exploration work comprises a dictionary of 10 Assamese words. The models for speech recognition is prepared to utilize HMM, VQ and I-vector strategy. The two new combination strategies are presented in this exploration by joining the above mentioned three techniques. In the first strategy, the recognition results of HMM, VQ and I-vector technique combined together to improve the recognition rate. The authors named this technique as fusion-1 techniques. They proceed further in the quest to improve the recognition rate and they combined the outputs of HMM, VQ, I-vector and the fusion-1 technique and they called it as fusion- 2 technique. The demonstrated results show that the fusion-2 technique outperforms all other existing techniques including fusion-1 technique. The work in [13] 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, this paper has two goals: The first is to gather a medium vocabulary isolated word speech database, which will act as the database for the Amazigh speech researchers. The second goal is to build up an 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 8.20%. From the extensive literature survey, we found that the work on continuous Kannada speech recognition (CKSR) is not remarkable. As a result of which, we would like to verify the performance of the state-of-the-art techniques for continuous Kannada speech. Since we didn't know the performance of state-of-the-art techniques for CKSR, we conducted some experiments by creating our own database of 2400 speakers collected across the Karnataka state (in India) 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. Kannada Phoneme Characteristics

Kannada is a Dravidian dialect talked prevalently by Kannada individuals in India, mostly in the territory of Karnataka. Kannada is the oldest of the four noteworthy Dravidian dialects with an abstract convention. Kannada letter set is famously known as *Aksharamale* or *Varnamale* and the current *Varnamale* list comprises of 50 characters. Keeping in mind the end goal to make the acknowledgment framework good to the prior *Varnamale* set, 50 characters have been considered in the present work. The 50 essential characters are grouped into three classifications. They are *Swaras (vowels)*, *Vyanjanas (consonants)* and *Vogavahakas (part vowel, part consonants)*. There are fourteen vowels and are called *swaras*. Table I demonstrates the graphemes of vowels and the relating ITRANS (Indian dialect Transliterations). The ITRANS of the corresponding alphabets is in the brackets.

IV. The Corpus Collection and Data Preparation

There can be various reasons that can definitely change the execution of the SR framework. The reasons resemble session fluctuation, intra-speaker and between speaker inconstancy. The Bharat Sanchar Nigam Limited (BSNL) provided the telephone facility for interactive voice response system (IVRS) call flow. The continuous speech data is collected from 1600 speakers (960 males and 640 females) of the age group in the range of 8 years-80 years. The speech data is obtained from various parts of the Karnataka state under real-world environment. It

TABLE I
CLASSIFICATION OF KANNADA ALPHABETS AND THEIR RESPECTIVE 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	

consists of 20,000 words that cover 30 districts. The ratio 60:40 (60% of male speakers and 40% of female speakers) is maintained throughout the process of data collection. The speech data are collected across the entire state of Karnataka covering every district in the state since there is diversity in speaking the Kannada language from region to region of Karnataka state. The continuous Kannada speech data obtained from the speakers are subjected to transcription from word level to the phoneme level. The tool used for the transcription is the Indic transliteration (IT3 to UTF-8) tool. The tags for non-lexical sounds which are also known as silence phones used during the transcription of the speech data. The few among the continuous speech sentences in Kannada dialects are shown in Table II. These sentences are well-known as Kannada gaadegalu/naannudigalu. The collected speech data is subjected to manual transcription at word level and authenticated by supervisors.

V. THE PORTRAYAL OF KANNADA 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 formulation of phoneme set for the Kannada language.

A. Extraction of speech features through MFCC

The way toward obtaining valuable information from speech data files and expelling the redundant information, for example, speaker-based and condition-based data are called feature extraction. This stage requires much consideration since the execution of the ASR framework vigorously relies upon this stage and any loss of helpful data can't be recovered later. Before extracting the features from the speech signal, the speech signal should undergo prepossessing step. Speech pre-preparing incorporates a few stages as (1) noise detection and elimination, (2) pre-emphasis, (3) framing and (4) windowing The presence of noise in the speech data retards the efficiency of the CKSR system. Before extracting the features from the speech signal the noise should be removed. After noise removal the resultant signal should undergo pre-emphasis step to boost the high frequency component.

Speech is basically a non-stationary signal. To make it stationary the speech signal is split into shorter frames of N samples, with next frames isolated by M samples (M<N) with this the contiguous frames are separated by N-M samples. Each frame size is of 20-30 ms duration with the overlapping of 10 ms. If the size of the frame shorter than this size is taken, the number of samples in the frames won't be sufficient to get the dependable data. With substantial size frames, it can cause considerable change in the data inside the frame. Windowing is performed to limiting the distortions at the starting and towards the ending of the frame. Then the frame and windowing function are multiplied together. The Hamming window is used to perform windowing. The process of extracting MFCC features from the speech samples starts with taking fast Fourier transform. The speech samples are converted

TABLE II
List of Kannada gaadegalu/naannudigalu recorded from the people across the state of Karnataka

English Version of Gaadagalu	Kannada Version of Gaadegalu
veida sul:l:aadaru gaade sul:l:aagadu	ವೇದ ಸುಳ್ಳಾದರು ಗಾದೆ ಸುಳ್ಳಾಗದು
ad:ikege hooda maana aane kot:t:aruu baaradu	ಅಡಿಕೆಗೆ ಹೋದ ಮಾನ ಆನೆ ಕೊಟ್ಟರೂ ಬಾರದು
kai kesaraadare baayi mosaru	ಕೈ ಕೆಸರಾದರೆ ಬಾಯಿ ಮೊಸರು
maatu bel:l:i mauna ban:gaara	ಮಾತು ಬೆಳ್ಳಿ ಮೌನ ಬಂಗಾರ
ad:d:a good:eya meile diipa it:t:a haage	ಅಡ್ಡ ಗೋಡೆಯ ಮೇಲೆ ದೀಪ ಇಟ್ಟ ಹಾಗೆ
akki meile aase nen:t:ara meile priiti	ಅಕ್ಕಿ ಮೇಲೆ ಆಸೆ ನೆಂಟರ ಮೇಲೆ ಪ್ರೀತಿ
end nd e ban:daaga kand nd u muchchikon:d:d:an:te	ಎಣ್ಣೆ ಬಂದಾಗ ಕಣ್ಣು ಮುಚ್ಚಿಕೊಂಡ್ಡಂತೆ
attegon:du kaala sosegon:du kaala	ಅತ್ತೆಗೊಂದು ಕಾಲ ಸೊಸೆಗೊಂದು ಕಾಲ
bekkige chellaat:a ilige praand a san:kat:a	ಬೆಕ್ಕಿಗೆ ಚೆಲ್ಲಾಟ ಇಲಿಗೆ ಪ್ರಾಣ ಸಂಕಟ
chin:te illadavanige san:teiluu nidde	ಚಿಂತೆ ಇಲ್ಲದವನಿಗೆ ಸಂತೇಲೂ ನಿದ್ದೆ
tun:bida kod:a tul:ukuvudilla	ತುಂಬಿದ ಕೊಡ ತುಳುಕುವುದಿಲ್ಲ

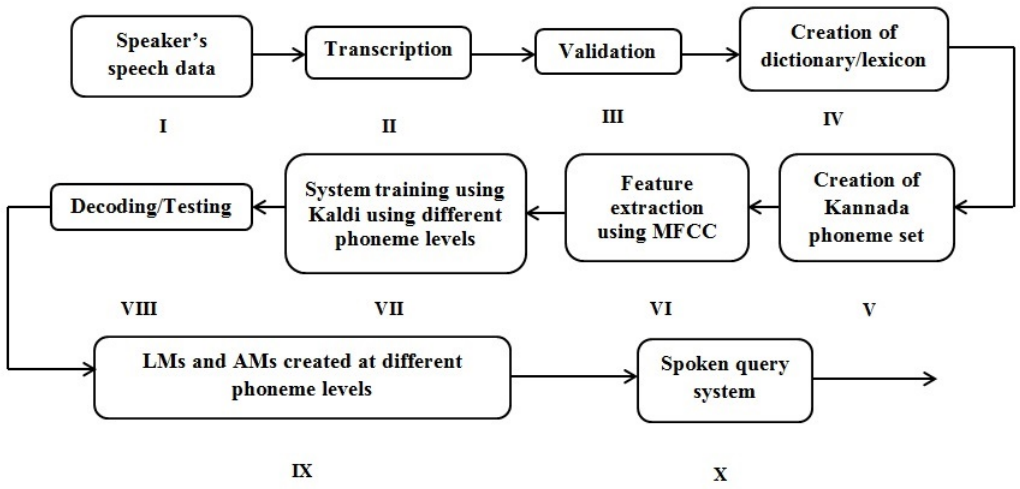

Fig. 1. The Block diagram of continuous speech recognition system for Kannada language

to frequency domain by taking fast Fourier transform. The speech samples are sent through the Mel-filter Bank, placed linearly up to 1KHz and then placed logarithmic after 1 KHz The yield of Mel-filter bank is subjected to the computation of log of energy to make the framework robust to slight variations in the input. As the last step of obtaining MFCC features the resultant Mel spectrum is converted back to time-domain by taking inverse cosine transform. This outcomes in Mel-Frequency Cepstral Coefficients. Totally 39 MFCC features are extracted in Kaldi out of which 13 are MFCC features, 13 delta features and 13 delta-delta features.

B. The Training and Testing

The 80% and 20% of validated speech information are utilized for training and testing respectively. The LMs and AMs were built independently for every district and for overall continuous speech data. Every speaker uttered 10 continuous sentences in the Kannada language for each session.

C. The Acoustic Model (AM)

The acoustic model measurably refers to the connection between the phone labels and the features of the speech data. The acoustic modelling is utilized for identifying the speech phoneme. Over the most recent three decades, the statistical method based on HMM turned into the most dominant technique for acoustic modelling [14]. Being the primary segment of ASR, the majority of the computational burden and execution load are upon this model. There are likewise some different methodologies for acoustic modeling for example, (1) HMM, (2) DNN, (3) SGMM and so on.

1) SGMM: The ASR systems based on the GMM-HMM structure usually involves completely training the individual GMM in every HMM state. A new modeling technique is introduced to the SR domain is called SGMM [15]. Present day ASR frameworks utilize left-to-right HMMs to collect and model the temporal variations in each and every phone. All the states in HMM models a probability density function specified by a Gaussian

mixture model (GMM) [16]. As of late, GMMs have been supplanted by DNNs [17] in probability density function of the state. In this manner, it gives rise to DNN-HMM acoustic models [18] [19].

2) DNN: The GMM-HMM-based acoustic modeling approach is inefficient to model the speech data that lie on or near the data space. The major drawbacks of GMM-HMM-based acoustic modeling approach are discussed in [20]. The artificial neural networks (ANN) are capable of modeling the speech data that lie on or near the data space. It is found to be infeasible to train an ANN using the maximum number of hidden layers with back-propagation algorithm for a large amount of speech data. An ANN with single hidden layer failed to give good improvements over the GMM-HMM-based acoustic modeling technique. The DNN consists of the maximum number of input hidden layers and an output layer to model the speech data to build ASR systems. The posterior probabilities of the tied states are modeled by training the DNN. This yielded the better performance in recognition compared to conventional GMM-HMM acoustic modeling approach.

D. The Language Model (LM)

The language model is actually a statistical model, which encourages the ASR framework to recognize words or expressions that sound comparable. The language model takes the yield of the acoustic model for further processing. The language model works on all the vocabulary words and with the assistance of the elocution model it builds up the most plausible word sequence. Normally, word N-gram models are utilized as the LM, which models the conditional density function of a word given the past $N-1$ words. Along these lines, the joint likelihood of any hypothesis, word arrangement can be approximated as the product of these conditional probabilities. The Language model takes the output of acoustic model for processing. Language model operates on all the vocabulary words and with the help of pronunciation model develop the most probable word sequence.

VI. Experiments and Results

The design of the machine in which the trial was performed is Ubuntu 18.04 LTS (64-bit working framework), Intel Core i7 processor with 3.70 GHz clock speed. The exploratory results were accounted for from the verbally expressed corpus made up of 10 Kannada sentences are spoken by 1600 speakers. The MFCC features and their derivatives were utilized for the formation of models. Kaldi utilizes the FST-based system and the IRSTLM toolbox was utilized to construct the LM. The guidelines used to create the LMs are as follows:

- Dictionary: The dictionary file used in Kaldi is named as lexicon.txt. It is built according to IT3-UTF:8 and ILSL12 format.
- Silence phones: The text “sil”and “SIL”were termed as silence phones.
- Optional silence phone: The text “sil”is termed as an optional silence phone.
- Non-silence phones: 168 non-silence phones were used to build the lexicon and the LMs.

The guidelines used to create the AMs are as follows:

- 3000 leaves were used for SGMM.
- 3000 Gaussians were used for SGMM.
- 3 jobs were used for training and decoding.
- 2 hidden layers were used for DNN.

The AMs produced at various phoneme levels are as follows:

1) DNN Hybrid Training and Decoding (DNN+HMM)
2) System Combination (DNN+SGMM) with iterations:1, 2, 3, 4.
3) SGMM + MMI Training and Decoding with iterations:1, 2, 3, 4.

The Table III demonstrates the diverse WERs obtained at various phoneme levels. It was seen in the table that, the hybrid combination of (DNN+HMM) has the WER of 4.10% and the WER corresponding to (DNN+SGMM) is 4.21%. At last, the combination of MMI + SGMM technique gives the WER of 4.60%. From the table, it is found that the hybrid combination of (DNN+HMM) has given a superior exactness contrasted and different models. The minimum WER models could be utilized as a part of SQS.

VII. Conclusion

In this paper, we proposed the hybrid modeling strategies for CKSR. We have explained the procedure associated with building an end-to-end, hybrid ASR system for the Kannada language with state-of-the-art tools. With 8 hours of information (with a vocabulary of 30,846 words), the combination of DNN and HMM modeling technique is able to accomplish a WER of 4.10% which is the least WER when contrasted with other hybrid modeling techniques. We intend to gather around 100 hours a speech data for preparing and scale the vocabulary size to about 100,000 words and utilize cross-lingual training to additionally bring down the WER. Some applications demand the greatest conceivable precision but, the accuracy can’t be ensured in the presence of noise. Speakers need to talk particularly all together for the system to function admirably. On the off chance that the speaker has non-standard speech, will in general run words together, or mutter, the training procedure might be long. The accuracy of the ASR system can be further increased if we implement efficient noise elimination algorithms more effectively.

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] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

TABLE III
THE PORTRAYAL OF WER AT VARIOUS PHONEME LEVELS FOR CONTINUOUS KANNADA SPEECH DATABASE

Phonemes	**WER_1**	**WER_2**	**WER_3**	**WER_4**	**WER_5**	**WER_6**	**WER_7**	**WER_8**
SGMM	6.24	**4.87**	6.45	5.34	5.64	5.87	4.93	4.99
SGMM+MMI_it1	4.66	4.89	4.76	5.20	4.72	5.30	4.62	**4.60**
SGMM+MMI_it2	4.65	4.89	5.21	4.89	4.86	5.64	4.83	**4.61**
SGMM+MMI_it3	4.91	**4.70**	4.98	4.79	5.25	4.76	4.71	5.08
SGMM+MMI_it4	4.95	4.77	4.95	4.83	4.85	5.05	5.01	**4.75**
DNN+HMM	4.27	**4.14**	4.29	4.23	4.27	4.54	4.62	4.24
DNN+SGMM_it1	4.40	4.38	4.36	**4.24**	4.77	4.65	4.39	4.25
DNN+SGMM_it2	4.67	4.55	4.87	4.99	4.76	**4.24**	4.90	4.88
DNN+SGMM_it3	4.70	5.22	4.31	4.34	4.82	4.79	**4.28**	4.33
DNN+SGMM_it4	4.49	5.04	**4.26**	4.61	4.85	4.53	4.66	4.89

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[4] S. C. Sajjan and C. Vijaya, "Continuous speech recognition of kannada language using triphone modeling," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 451–455, IEEE, 2016.

[5] F. Patlar and A. Akbulut, "Triphone based continuous speech recognition system for turkish language using hidden markov model," in *12th IASTED International Conference in signal and image processing*, pp. 13–17, 2010.

[6] 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.

[7] M. Kalamani, M. Krishnamoorthi, and R. Valarmathi, "Continuous tamil speech recognition technique under non stationary noisy environments," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 47–58, 2019.

[8] P. Upadhyaya, O. Farooq, 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.

[9] 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.

[10] 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.

[11] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[12] S. S. Bharali and S. K. Kalita, "Speech recognition with reference to assamese language using novel fusion technique," *International Journal of Speech Technology*, pp. 1–13, 2018.

[13] 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.

[14] J. Cai, G. Bouselmi, Y. Laprie, and J.-P. Haton, "Efficient likelihood evaluation and dynamic gaussian selection for hmm-based speech recognition," *Computer Speech & Language*, vol. 23, no. 2, pp. 147–164, 2009.

[15] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, *et al.*, "The subspace gaussian mixture model—a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[16] M. Gales, S. Young, *et al.*, "The application of hidden markov models in speech recognition," *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[17] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[18] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[19] G. Chen, H. Xu, M. Wu, D. Povey, and S. Khudanpur, "Pronunciation and silence probability modeling for asr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] 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.