# Continuous Bengali Speech Recognition Based On Deep Neural Network

Md. Alif Al Amin*, Md. Towhidul Islam†, Shafkat Kibria‡ and Mohammad Shahidur Rahman§
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet-3114, Bangladesh
Email: *alifalamin4@gmail.com, †tuhintowhidul9@gmail.com, ‡shafkat80@gmail.com, §rahmanms.bd@gmail.com

***Abstract*—Nowadays, deep learning is the most reliable approaches in the field of speech recognition to do the Acoustic modeling. Working with a language like "Bengali" that is not very resource-rich in terms of availability of parallel data (i.e. speech with aligned text) is a challenging problem. Also, there are lots of approaches going with deep learning to achieve better performance in Bengali Language without benchmarking a specific corpus. So, the achieved results are biased. In this paper, DNN-HMM and GMM-HMM based models have been used, which have been implemented in Kaldi toolkit, for continuous Bengali speech recognition benchmarking on a standard and publicly published corpus called SHRUTI. Previously, the best word error rate (WER) had been achieved on SHRUTI was 15% using CMU-SPHINX based GMM-HMM and this study has been shown that using Kaldi based feature extraction recipes with DNN-HMM and GMM-HMM acoustic models have achieved performances WER 0.92% and WER 2.02% respectively. Another finding of this study is, the WERs of both models are very close because the size of the corpus is small.**

***Index Terms*—Speech Recognition; DNN-HMM; GMM-HMM; Word Error Rate(WER)**

## I. Introduction

About 260 millions people speak Bengali as their native language [1]. It is a noticeable amount of the world's population. But it is a matter of regret that quantity as well as quality research works on Bengali speech recognition is very few. Nowadays speech is the smartest communication medium between human and a machine. This communication will be more effective if and only if the human is comfortable to communicate with machines through their mother language. Effective researches on Bengali speech recognition is a crying need.

Speech recognition is a technique of translating a speech data into a related document or decoding human voice in machine-readable content. In speech recognition researches, deep learning approach is a very hot topic nowadays. There must be something special that every NLP researchers are trying to use deep learning. Deep learning approaches are being popular as it is outperforming many previously developed models for speech recognition. This is the age where CPU's are replaced by GPU's. So, it is very easy to train huge models. Multi-threading process can be used to multiple GPUs and CPU's. It encourages our researchers to use machine learning approaches.

Many pieces of research on speech recognition are going on but maintaining the decoding accuracy is challenging to the researchers. Feature extraction, speaker normalization, acoustic modeling, language modeling etc. are challenging works in speech recognition. However, in latest days ASR systems are getting very elite performances with the utilization of open research apparatuses like Kaldi, SPHINX, CMU LM, and HTK etc.

So, the building of the ASR system is much easier than before. Among those toolkits, Kaldi is the most popular toolkit and has got the most varieties of implementation of acoustic models [2]. That's why Kaldi has been selected as the working toolkit on this study. GMM-HMM models and DNN-HMM-based models have been used which are implemented in Kaldi [3], [4]. There are two types of DNN implementation in Kaldi. The first one is DBN (Deep Belief Networks) with pre-trained RBM (Restricted Boltzmann Machine) implemented by Karel. The other one is DNN (greedy layer-wise supervised training) implemented by Dan. The second one is the latest implementation of DNN. So the second one has been used. GMM-HMM-based models were the best performing acoustic models but now DNN models are outperforming GMM-HMM-based models [5], [6].

Several researches in 2009 showed that DNN acoustic model outperformed the best published recognition results on TIMIT, which is a benchmark dataset in English Language for testing new algorithms to Speech Recognition [7], [8]. There are several other corpora available on English Language for evaluating latest algorithms performances. DNN based acoustic modeling have outperformed on most of the corpora that have more than 100 hours of data [4]. Whereas, there are no such benchmarking dataset for Bengali Language. Over last few years, there are several new approaches have been attempted on several different published and unpublished datasets and achieved several better performances, which are biased because the results are not benchmarking on a specific dataset [9] "in press" [10]–[12].

By using SHRUTI Bengali speech corpus, 21.64 hours phonetic transcripted data has been processed for building a Bengali transcripted standard corpus [13]. Then The corpus has been prepared for Kaldi for training. Total 143 dimensional feature vectors has been extracted using MFCC feature extraction method. LDA, MLLT and SAT have been used

978-1-5386-9111-3/19/$31.00 

for speaker adaptation. Two hybrid models (GMM-HMM and DNN-HMM) have been applied for continuous Bengali speech recognition and performance have been shown by the mentioned models are very satisfactory.

## II. Related Works

Speech Recognition researches on many languages are very much ahead than Bengali. There are very few works on Bengali speech recognition. A team worked on continuous Bengali speech recognition. They used the Application Programming Interface (SAPI) of Microsoft Corporation [14]. But the corpus size is very small consisting of 270 words. For one to one English to Bangla relationship, their rate of recognition is 58.22% and for one to many relationships, their recognition accuracy rate is 74.81%.

A team of researchers did a nice work on Isolated and Continuous Bangla Speech Recognition [15]. They used a Hidden Markov model (HMM) classifier and they tried to recognize both Isolated and Continuous words. They used unique 100 Bangla words. For isolated word and speaker dependent and speaker independent rate of recognition respectively are 90% and 70%.

Another research is implemented Back Propagation Neural Network for recognizing just Bengali digits [16]. Speaker dependent accuracy was 96.3% and independent accuracy was 92%. The accuracy of their work was good enough but the corpus was very little for real-world implementation. Automatic recognition of real numbers was implemented by a brilliant team using CMU-SPHINX. Their accuracy was 85% for personal computer and 75% for android mobile [17].

Recently, some excellent works in several languages have been done. A work on continuous Hindi speech recognition is done by a research team. Their dataset consists of 1000 unique sentences [18]. Their WER (Word Error Rate) was better than many previous works on Hindi.

There is a good research work done on continuous Serbian speech recognition. They have 90 hours of speech data and 21000 of utterances [19]. They got very satisfactory results. Their WER of GMM-HMM is 2.19% and DNN(for 3 hidden layers) is 1.86%.

## III. Feature Extraction

Mel-Frequency Cepstral Co-efficient(MFCC) has been selected for feature extraction technique. MFCC can mimic human voices and perceptions. It is the most popular feature extraction techniques. PLP is also a very popular technique for feature extraction. PLP performs better for a noisy dataset [20]. As the dataset was clean enough, MFCC is preferable for the experiment.

After using conventional MFCC, More improved feature extraction technique has been used. We have got 13-dimensional vectors across 11 frames to get the 143-dimensional feature vectors. A conventional MFCC derivation has been shown in Figure 1.

After that, LDA (Linear Discriminant Analysis) has been applied for de-correlation and dimensionality reduction. For more precise features MLLT has been used over it. For normalization of inter-speaker variability used fMLLR. Now an improved feature extraction technique is created by using MFCC on top of LDA + MLLT + fMLLR [21]. In Figure 2, Improved feature extraction technique has been shown.

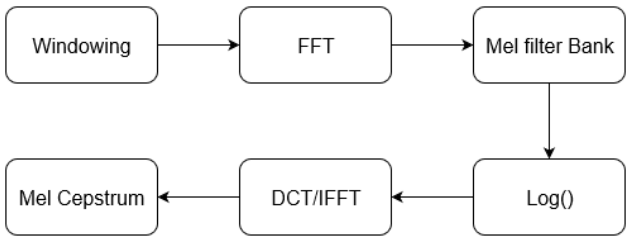


Fig. 1. Mel Frequency Cepstral Coefficients Derivation

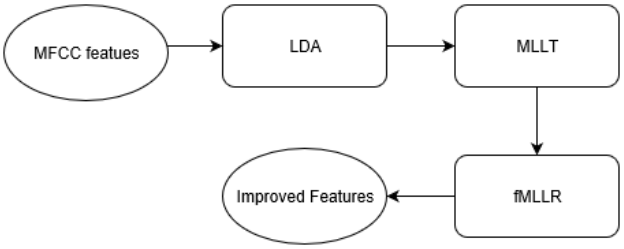


Fig. 2. Improved Output from MFCC

## IV. GMM-HMM Modeling

Hidden Markov model is the most effective and simplest classifier having so many applications as well [22]. In the context of speech recognition, speech data has been taken and extracted the features out of it. Then a system consisting of an acoustic model (HMM), phonetic model, language model and search space can generate a predicted output from the given training data.

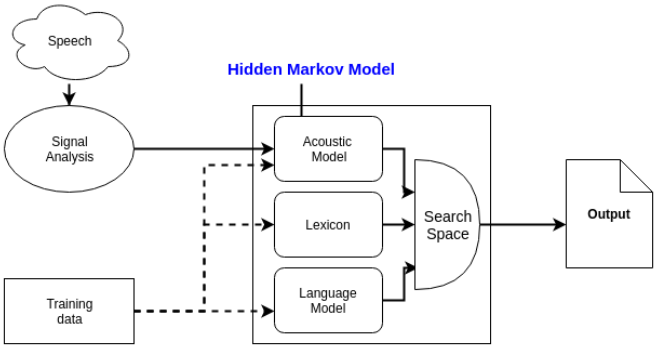


Fig. 3. Conventional Hidden Markov Model

The whole process is shown in Figure 3. After extraction of features, a sequence of fixed size acoustic vector has been

found, $Y_{1:T} = y_1, y_2, y_3.....y_T$ For decoding the audio file to the sequence of words, it needs to determine as a sequence of words (Equation 1).

$$W_{1:L} = w_1, w_2, w_3.....w_L \tag{1}$$

$W$ denotes the most likely word sequence. $W^*$ is given by $P(W|Y)$ (Equation 2).

$$W^* = argmaxP(W|Y) \tag{2}$$

$P(W|Y)$ has been calculated by using Bayes theorem (Equation 3).

$$P(W|Y) = argmaxP(Y|W) * P(W)/P(Y) \tag{3}$$

Now $W^*$ can be written (Equation 4). As the other values of right side have been determined.

$$W^* = argmaxP(Y|W) * P(W) \tag{4}$$

$P(Y|W)$ can be determined from acoustic model and $P(W)$ from language model. Every words has been decomposed into a sequence of phones, $(Q_k)_w = (q_1)_w, (q_2)_w(q_k)_w$ for acoustic model. Now the relation can be derived (Equation 5).

$$P(Y|W) = sum(P(Y|Q) * P(Q|W)) \tag{5}$$

Now it can move through a Markov chain with the best transition probability. For the output distribution, it can calculate multivariate Gaussians. That's why its called GMM-HMM model.

Nowadays, researchers are working with sGMM (subspace Gaussian Mixture Models) [23]. This types of modeling allow better representation of data and results for the small amount of data. sGMM has also been used in this experiment.

## V. DNN-HMM Modeling

The setup of GMM-HMM is used for the basis for training the DNN-HMM model. It has been mentioned that Dan's implementation of DNN has been used which is greedy layer-wise supervised training. First of all, The network has been initiated with the input layer, one hidden layer, and a softmax layer (figure 4).

Then the system has been trained for a small amount of time (5 iterations). After that, the soft-max layer has been removed and another hidden layer has been added with different sets of weight. This process has been done for 3 times to get DNN. After all of it, we got a shiny DNN. About 20 iterations have been performed for training and then more 10 iterations with a constant learning rate. At first, the learning rate was 0.02 and finally, it constant in 0.004. There are many implementations of DNN for acoustic modeling. This is one of the common approaches to DNN implementation. GMM has been replaced from GMM-HMM model with DNN for building DNN-HMM hybrid model.

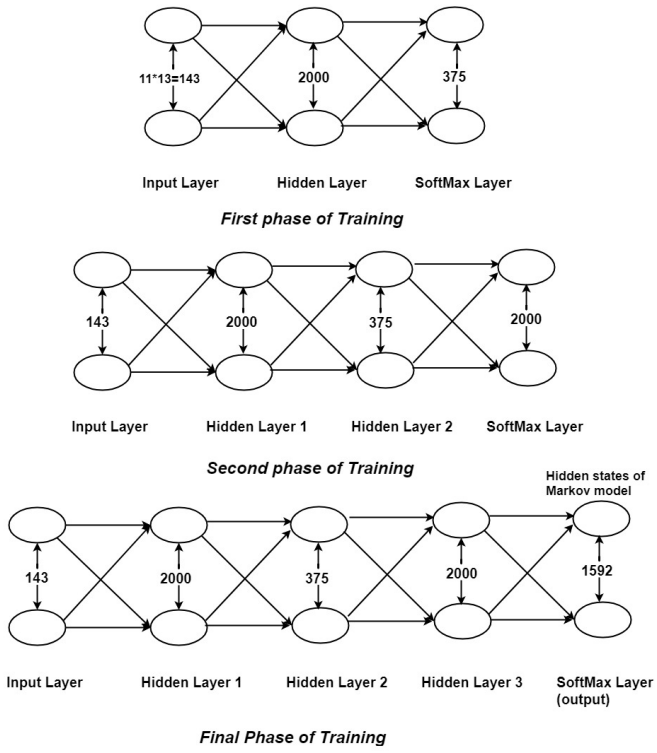


Fig. 4. DNN implementation

## VI. Data Preparation

In this study, SHRUTI Bengali speech corpus has been used [13]. This corpus has been created by some of the students of Indian Institute of Technology, Kharagpur. They have worked with the recognition of vowel, words and some other corpus analysis. The whole corpus size and speaker details are shown in the tabular form.

| Unique Words | Utterances | Speaker | Male | Female |
|---|---|---|---|---|
| 22012 | 13025 | 34 | 26 | 8 |

Whole corpus size is 21.64 hours of speech data. About 75% of whole dataset has been separated as training data and 25% data as testing data. Among the 34 speakers, 20 male and 5 female speakers(total 25) have been taken as o training data speakers and for testing data speakers rest of the speakers has been selected (6 male speakers and 3 female speakers, total 9).

The default SHRUTI corpus was phonetic transcribed and was also in English alphabets. A process needed to be done that the system can give the output in the Bengali Language. Firstly, English phonetic transcribed sentences are converted into Bengali alphabets. Then the lexicons are needed to be mapped with their corresponding words. In figure 5, pre-processing of the whole data has been shown.

A practical example has been shown in figure 6 of pre-processing of transcription of speech data.

Preparation of data for Kaldi is a sequential work as follows.

- Non-Silence Unique Phones

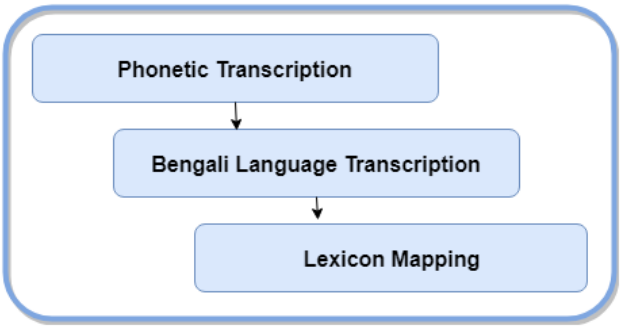


Fig. 5. Bengali Transcription

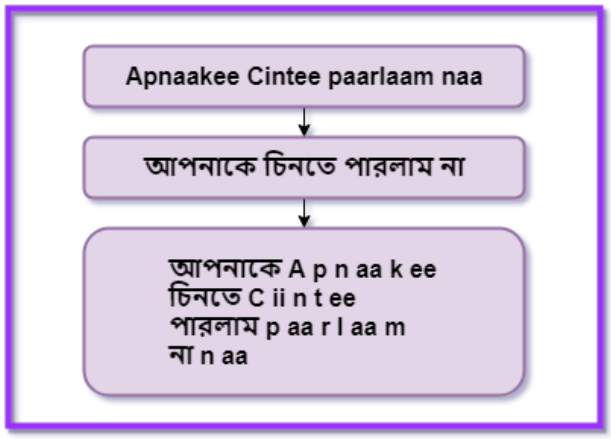


Fig. 6. Practical Bengali Transcription

- Building Lexicon
- Corpus
- Speaker to Utterance Mapping
- Utterance to Speaker Mapping
- Speaker to Gender Mapping
- Defining Path for Utterances

### A. Unique Phones

There are 49 unique non-silence phones by which all the words can be generated.

### B. Lexicon

All the words have been constructed using the non-silence phones and silence phone which has been described in the lexicon file.

### C. Corpus

This file contains words that will be the output when the system needs to decode any utterances.

### D. Speaker to utterance and utterance to speaker mapping

Every utterance used in the corpus needs to be mapped with its corresponding speaker. The same process has been applied in speaker to utterance also.

### E. Speaker to Gender Mapping

A file consists of the gender info of every speaker. 'm' letter has been used for denoting male speaker and for female 'f' has been used.

Now the system and corpus configuration has been shown below.

- Sampling Rate 16000 Hz
- Feature Extraction MFCC
- Vocabulary size 22012
- Most of the utterances are different content based like sports, movies, political contents etc. Ngram-3 language model has been used and was created by the labels from the train speeches using SRILM toolkit.
- Every utterance has two or three Bengali sentences.

## VII. Experiment and Results Analysis

In this study, an open resource(SHRUTI) has been selected for experiments. Improved feature extraction has been given as the input of different types of acoustic models like DNN-HMM, GMM-HMM and sGMM. Word error rate(WER) metric has been used for measuring the system performance.

$$Word\ Error\ Rate(WER) = \frac{S+I+D}{N} \qquad (6)$$

Here (Equation 6) S is the number of substitutions
D is the number of deletions
I is the number of insertions
N is the Number of words in the reference speech
Then a comparative result analysis of different models has been shown in Figure 8. Previously, there is a speech recognition research work has been done on this corpus. Another comparison has been shown between this experiment and the previous in Figure 9.

The working procedure of a Speech recognition system has been discussed in the Figure 7.

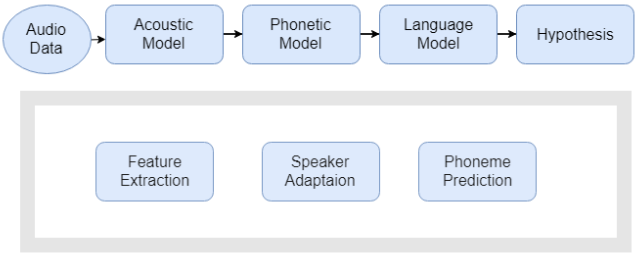


Fig. 7. Speech Recognition System

These tasks mentioned below has been performed sequentially to get the expected performance from acoustic models.

- Feature Extraction (MFCC)
- Mono-phone Model Training
- Audio alignment
- Tri-phone Model training
- Re-align audio and tri-phone
- LDA-MLLT

- LDA-MLLT-SAT(Speaker Adaptation Technique)

GMM-HMM and DNN-HMM models have been used which are developed in Kaldi toolkit in Linux platform. The sentences used in dataset are phonetically compact and designed to cover most of the frequent speaking word in the Bengali language. A tri-gram language model has been used for getting better performance from the output of acoustic models.

A web application has been developed using Kaldi, GStreamer, and some python tools. 30 combinations of different models has been tried like mono, tri1, tri2b, tri3b, tri3b_mmi,tri3b_fmmi etc. Some of the accuracy among those models has been mentioned in this study.

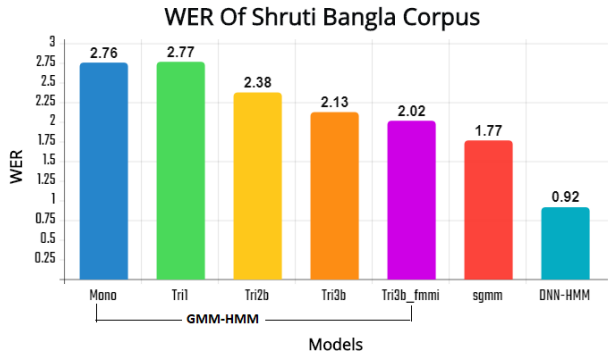


Fig. 8. WER for different types of Models

Now the performances of different types of models have been shown in Figure 8. Several models are compared with each other and DNN-HMM-based model comes on the top. Among GMM-HMM models hybrid tri3b_fmmi model has shown the best performance. Subspace GMM+LDA+MLLT+SAT based model has shown very high performance. It has been mentioned that sGMM is very effective for small size corpus [23]. But in the long run for a large corpus it will not be performing as good as it is showing in the Figure 8.

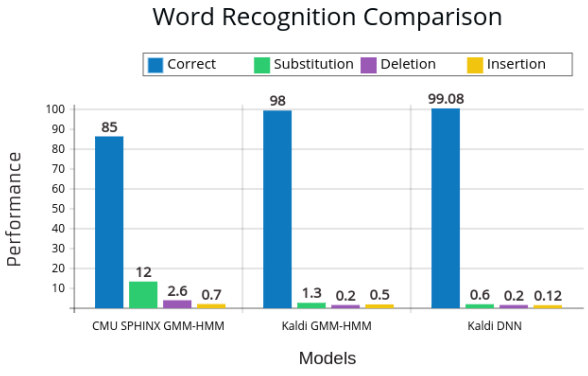


Fig. 9. A comparison of different word error rates (WERs) on SHRUTI

Figure 9 shows the comparison between the CMU-SPHINX (Tri-phone based GMM-HMM tri3b_mllr) based models' WER, which has been achieved by IIT, Kharagpur research group [13], and the Kaldi based models' WERs that have been achieved in this research work. Different types of GMM-HMM besed models has been experimented in this work. But the best performance of GMM-HMM which is tri-phone based tri3b-fmmi model, has been selected for the comparison.

It is noticeable that Kaldi's tri-phone GMM-HMM based model and DNN-HMM based model are much better performing than CMU SPHINX tri-phone based GMM-HMM model.

## VIII. Conclusion

Main goal of this study was to benchmark the performance of recent approach to speech recognition on a specific standard dataset of Bengali Language. DNN-HMM and GMM-HMM-based models have been used with Kaldis several feature extraction recipes and training approaches, like - MFCC with Mono-phone or Tri-phone or LDA+MLLT or LDA+MLLT+SAT, on a standard and publicly published Bengali Language corpus (SHRUTI) for continuous speech recognition. The performances have been achieved from the both approaches are satisfactory and very close indeed; these are 0.92% WER for DNN-HMM and 2.02% for GMM-HMM-based acoustic modeling on the SHRUTI corpus of 21.64 hours of speech data. It is desirable that more than 100 hours of speech corpus [4], the DNN-HMM approach will outperform the performance of GMM-HMM, otherwise these results will be marginally closer. So, the achieved performances have approved the fact and showed the requirement of large corpus on Bengali language for benchmarking the latest approach in speech recognition.

## Acknowledgement

Authors wish to acknowledge financial support from the Higher Education Quality Enhancement Project (AIF Window 4, CP 3888) For The Development of Multi-Platform Speech and Language Processing Software for Bangla. Authors would like to acknowledge the researchers of IIT, Kharagpur for the open resource of speech data.

## References

[1] G. F. Simons and C. D. Fennig, "Bengali," 2018. [Online]. Available: https://www.ethnologue.com/language/ben

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[3] 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.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[5] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 215–219.

[6] A.-r. Mohamed, G. E. Dahl, G. Hinton *et al.*, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[7] A.-r. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Nips workshop on deep learning for speech recognition and related applications*, vol. 1, no. 9. Vancouver, Canada, 2009, p. 39.

[8] T. N. Sainath, B. Ramabhadran, and M. Picheny, "An exploration of large vocabulary tools for small vocabulary phonetic recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 359–364.

[9] S. H. Sumit, M. Tareq Al Muntasir, R. N. Nandi, and T. Sourov, "Noise robust end-to-end speech recognition for bangla language," in *International Conference on Bangla Speech and Language Processing (ICBSLP)*, vol. 21, 2018, p. 22.

[10] J. R. Saurav, S. Amin, S. Kibria, and M. S. Rahman, "Bangla speech recognition for voice search," in *International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018.

[11] T. Ahmed, M. F. Wahid, and M. A. Habib, "Implementation of bangla speech recognition in voice input speech output (viso) calculator," in *International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018.

[12] S. A. Sumon, J. Chowdhury, S. Debnath, N. Mohammed, and S. Momen, "Bangla short speech commands recognition using convolutional neural networks," in *International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018.

[13] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous auutomatic speech recognition system," in *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*. IEEE, 2011, pp. 51–55.

[14] S. Sultana, M. Akhand, P. K. Das, and M. H. Rahman, "Bangla speech-to-text conversion using sapi," in *Computer and Communication Engineering (ICCCE), 2012 International Conference on*. IEEE, 2012, pp. 385–390.

[15] M. Hasnat, J. Molwa, and M. Khan, "Isolated and continuous bangla speech recognition: Implementation," *Performance and application perspective*, 2007.

[16] M. Hossain, M. Rahman, U. K. Prodhan, M. Khan *et al.*, "Implementation of back-propagation neural network for isolated bangla speech recognition," *arXiv preprint arXiv:1308.3785*, 2013.

[17] M. M. H. Nahid, M. A. Islam, and M. S. Islam, "A noble approach for recognizing bangla real number automatically using cmu sphinx4," in *Informatics, Electronics and Vision (ICIEV), 2016 5th International Conference on*. IEEE, 2016, pp. 844–849.

[18] 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," in *Machine Intelligence and Signal Analysis*. Springer, 2019, pp. 303–311.

[19] B. Popović, S. Ostrogonac, E. Pakoci, N. Jakovljević, and V. Delić, "Deep neural network based continuous speech recognition for serbian using the kaldi toolkit," in *International Conference on Speech and Computer*. Springer, 2015, pp. 186–192.

[20] N. Dave, "Feature extraction methods lpc, plp and mfcc in speech recognition," *International journal for advance research in engineering and technology*, vol. 1, no. 6, pp. 1–4, 2013.

[21] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Interspeech*, 2013, pp. 109–113.

[22] 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.

[23] 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 modela structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.