Bengali Speech Recognition: An Overview

Mashuk Arefin Pranjol, Jahidul Hasan, Farhin Rahman, Saiadul Arfain, Bushra Yesmeen Anika
Tanjib Ahmed, Ehsanur Rahman Rhythm, Rajvir Ahmed Shuvo, Md. Abdullah Al Masum Anas
Md Humaion Kabir Mehedi, Shadab Iqbal
and Annajiat Alim Rasel
School of Data and Sciences, BRAC University
66, Dhaka 1212, Bangladesh
Email: {mashuk.arefin.pranjol, jahidul.hasan, farhin.rahman, saiadul.arfain, bushra.yesmeen.anika,
tanjib.ahmed, ehsanur.rahman.rhythm, rajvir.ahmed.shuvo, md.abdullah.al.masum.anas,
humaion.kabir.mehedi, shadab.iqbal}@g.bracu.ac.bd,
annajiat@gmail.com

***Abstract*—This study outlines the notable efforts of creating of automatic speech recognition (ASR) system in Bengali. It describes data from the Bengali language's existing voice corpus and the major reports that have contributed to the recent research scenario. It provides an overview of dataset or corpus that has been created for bengali ASR, challenge faced to create bengali ASR as well as techniques used to build Bengali ASR system. ASR techniques for the Bengali language have made significant progress in recent years. Our article contains studies from 2016 through 2020. We examined the results of these investigations, as well as the strategies used to accomplish this goal, for Automated voice recognition. We examined these publications to obtain a feel of the present state of Bengali ASR. We observed a dearth of sufficient datasets among these researchers, which is important for any automated system. Due to the language's abundance of consonant clusters, the ML system has difficulty interpreting Bengali words. As a result of these modifications, the system now confronts a new set of difficulties in terms of effectiveness and efficiency. Additionally, numerous words have nearly identical pronunciations. These are only some of the issues that the papers we examined face. This research makes use of a variety of techniques, including linear prediction coding, Mel Frequency Cepstral Coefficient, Hidden Markov Model, Neural Network, and Fuzzy logic. Bengali ASR will require further investigation shortly. While recent research is encouraging, ASR of other languages, such as English, is far from perfect and efficient.**

***Index Terms*—Bengali ASR; Challenges; Techniques; Bengali speech corpora;**

I. Introduction

Automatic Speech Recognition, or ASR for short, is a technique that enables individuals to use their voices to communicate with a computer interface in a way that resembles regular human speech in its most advanced forms. ASR systems are commonly utilized in automated equipment to identify speech commands. It is used to create chatbots for smartphones and other devices. Speech recognition systems are utilized in call centers to provide automated responses to customers. Speech recognizers may also be used to detect crimes planned through phone calls, as well as hate speech delivery.

Throughout 228 million people around the world speak Bengali as their first or second language [1]. Many people in West Bengal and other parts of India speak Bengali, as well as people who came from these places as well as the people who lived there before them. A good ASR system for Bengali will benefit a large number of people due to the language's large number of speaker groups. Bengali ASR research saw a resurgence in the 1990s. Since approximately the year 2000, Bengali speech has been recognized. In 2002, A. Karim et al. [2] published a system for recognising spoken letters in Bengali. Artificial neural networks were used by K. Roy et al. to develop a Bengali voice recognition system [3]. There was a phoneme identification system developed in 2003 by M.R. Hasan [4] and a continuous voice recognition system developed in 2003 by K.J. Rahman [5], both of whom used artificial neural networks (ANNs). Google has unveiled a working speech recognition and voice search service for Bengali and other languages. The purpose of this study is to outline recent research on Bengali ASR in order to make it easier for academics to keep up with the latest developments. x(n) is the input for the system. Finally, feature extractions are performed to lower the input vector's dimensionality yet preserve the distinguishing properties for recognition. Analytical model computes acoustic signal's (x1...xN) likelihood of being noticed based on sequence of words (w1...wN). The probability of a suggested word sequence, P r, is provided by the language model (w1...wN). This is a list of words with their phonetic transcriptions, as well as a list of terms that are appropriate for a certain situation. All three elements of the decoder are combined and the recognised text is sent as output y(n).

The contents of this paper are structured as follows. Section II discusses related works, in section II we reviewed all the available dataset or corpus, section III presents issues to consider when building ASR, section IV explains techniques to create an effective Bengali ASR and section

V analysis part discusses significant advancements in the previous decade.

II. Literature Review

The study, gives a quick overview of important initiatives to construct an ASR system for the Bengali language. They summarized the research done in Bengali ASR in last decade. Badhon et al. assessed 15 research publications that worked in 2020 [6]. Sultana and Palit [7] conducted an assessment of various prevalent voice recognition algorithms for the Bengali language in 2014. The datasets and specific methodology used in the research were represented in the study.

III. Review of bengali dataset/corpus

Corpus (plural corpora), also known as text corpus, is an organised collection of texts that can be used for research purposes. Corpus linguistics uses them for statistical analysis and hypothesis testing as well as checking for occurrences and validating linguistic rules in specific regions. And a dataset is a collection of data.

For the research of Bengali language ASR there already have been developed some datasets or corpus. Some of the most used datasets/corpus are mentioned below.

A. Bengali.AI | Datasets by Google

1) Bengali Text to Speech Dataset

This data collection provides multi-speaker high-quality Bengali transcribed audio data. The data package is made up of wave files and a TSV file. There are two zip files, one for each local, each of which contains the file line index.tsv and the wave files. A fileID by and a transcription are included in the line index. This data set has been carefully quality reviewed was gathered by Google [8].

2) Bengali Automatic Speech Recognition Dataset

There are 196k utterance in Bengali ASR dataset. The data package is made up of wave files and a TSV file. This tsv dataset contains a FileID, an anonymised UserID, and audio transcription. This data set has been carefully quality reviewed and collected by google [8].

B. Shruti Bengali Bengali ASR Speech Corpus

SHRUTI, an acoustic-phonetic corpus, aims to provide speech data for the development and evaluation of ASR systems. IITKGP's Communication Empowerment Lab collaborated with Media Lab Asia to generate the text corpus. The IITKGP recorded, transcribed, and verified the speaker's remarks for accuracy.

SHRUTI has a total of 7383 distinct sentences. There were 34 speakers from West Bengal who spoke the colloquial Bengali dialect. Male speakers make up 75%, whereas female speakers make up 25% of the total. The speakers in this corpus range in age from twenty to forty years old. The speaker's dialect region is the geographical area in West Bengal, India, in which a speaker's dialect is located. The text in the SHRUTI prompts (included in the file "shruti train transcription") is composed of phrases created at IITKGP. The text is taken from the Anandabazar patrika, a story book. To cover phonetic variants of the Bengali language, four large news domain articles were compiled. Sports, politics, general news, and geography are the domains. The corpus was deliberately developed to be useable for typical ASR systems. In addition, the most frequently uttered phrases were gathered and recorded for this purpose. In total, 34 speakers recorded phrases over the course of two sessions. Compact phonetic phrases were created to cover the majority of the frequently used words in Bengali. Every speaker read a different amount of sentences. Sentences were transcribed into ITRANS format [5], [9]. A total of 7383 sentences, 49 phonemes, 22012 words, and 21.64 hours of recording time comprise the corpus.

C. Bengali Raw Speech Corpus by Central Institute of Indian Languages

LDC-IL The Bengali Speech data collection is made up of several datasets such as word lists, sentences, running texts, and date formats. West Bengal and Tripura have Bengali as their official languages. Greater usage of Bengali has led to the language's expansion in terms of vocabulary as well as the variety of styles and registers. LDC-IL Bengali Speech data is gathered from the Standard Colloquial (Central Bengal) and Barendri areas (North Bengal) [10]. The dataset was of 81.2 GB (48 kHz | 16-bit wav) whose total duration is 128:46:59 hours. 73,470 audio segments were contributed by 476 speakers.

D. SUST Bengali Emotional Speech Corpus

SUBESCO is a Bengali language audio-only emotive speech corpus. The corpus has a total duration of more than 7 hours and contains over 7000 utterances, making it the biggest emotional speech corpus known for this language. The gender-balanced set included twenty native speakers, each recording ten words imitating seven different moods. Fifty university students took part in the corpus examination. All but one of the audio clips in this corpus were evaluated four times by male and female reviewers, except for the Disgust emotion. The overall recognition percentage was stated to be better than 70% for human perception testing. Inter-rater reliability and consistency were found to be high based on intra-class correlation coefficient scores and Kappa statistics [11].

IV. Challenges and scopes for Bengali speech to text research

It is a difficult task to develop an efficient and successful speech recognition system in the Bengali language. The primary barrier to beginning work on Bengali voice recognition is the lack of a publicly available, high-quality dataset. Implementing a speech recognition system for Bengali is more difficult than other languages. The Bengali language has several dialects and accents. Along with the common challenges in every language, like recording

devices, noise, speech patterns, and non-standard pronunciations, Bengali has a number of unique challenges that make it more difficult to develop speech recognition engines for Bengali. Bengali is a tonal language, which means that the pitch of the voice changes depending on the tone of the word. This makes it more difficult for the speech recognition engine to distinguish between words with the same pronunciation but different tones. Consonant clusters in Bengali make it difficult for the engine to recognise words. As a result, we demonstrate the challenges associated with a Bengali speech recognition system in the following subsections.

A. Accent Variations

The Bengali language has a rich variety of accents, which can be classified into two types: (1) regional accents and (2) non-regional accents. Many districts in Bangladesh, like Dhaka, Sylhet, Chittagong, Noakhali, and others, have their own dialects. As a single word is pronounced in different ways in different regions, it becomes a major challenge for Bengali speech-to-text research. For example, বিকাল /Bikāla/ (English:Afternoon) is pronounced as বিয়েল /Biẏēla/ in Chittagong. The same word is pronounced differently in different regions. There are six distinct groups of Bengali dialects, which may be distinguished based on differences in phonology and pronunciation among the various dialects [12]. In an ASR system, dialects add more phoneme patterns and more words to their vocabulary. The use of these dialects should also be taken into consideration in order to create a faultless Bengali ASR system.

B. Sentence Structure

Every language has its own grammar. The language follows the rules of grammar to construct the structure of the sentence. It fundamentally varies by language. The grammar of English is different from the grammar of Bengali. As a result, the structure of the sentence is different as well. For example, in English, the verb comes before the object in a sentence, such as "I love you." (I is the subject, and you is the object) In Bengali, the verb comes after the object. That is, in Bengali, "I love you" is "আমি তোমাকে ভালোবাসি" where the sentence structure is subject + object + verb. In the English language, a preposition precedes a noun. But for Bengali language, if the preposition is needed, the noun or noun-equivalent word is inserted before the preposition. Some prepositions in Bengali are merged with the noun or noun-equivalent word as well (such as, "of Tiger" becoming "বাঘের", where "এর" is merged with the noun, "বাঘ"). Additionally, the Bengali language lacks the use of auxiliary verbs. It is actually clear that because of these differences, the architecture might not perform well in Bengali speech even if it performs well in English. It might be necessary to take sentence structures into consideration in models like the n-gram and others.

C. Different Phonemes and Consonant Clusters

The Bengali language has a large number of alphabets. It features 14 vowels and 29 consonants. 7 of those 14 vowels are Nasal vowels which contains a wide variety of diphthongs and inherent back vowels. In Standard Bengali as spoken in Dhaka and other Eastern dialects, /r/ and /ɽ/ are frequently phonetically unclear, and both can be phonetically realised as [ɾ] or [ɹ]. For example, পড়া /Paṛā/ (English:read) and পরা [Parā] (English:wear) has the same pronunciation but the meaning is different. Again, Consonant clusters are used often in Bengali speech, making it difficult to distinguish word boundaries. In many cases the system can not identify the whole word and detects the boundary in the middle of the word. For example, in the Bengali word সম্মান /Sam'māna/ (English:Respect) where it might detect সম /Sama/ (English:same) and মান /Māna/ (English:value) as two separate words. As a result, it causes inaccuracy in detection.

D. Insufficient Dataset

One of the major challenges in Bengali speech-to-text research is the lack of enough datasets. The availability of resources and datasets in the Bengali language is very low. The size of a dataset is often responsible for poor performance in ML projects. Poor ML performance is frequently caused by insufficient datasets. Lack of sufficient datasets in the Bengali language complicates things since models are data-hungry and their performance is strongly dependent on the amount of training data available. The diacritics and consonant conjuncts in Bengali make word construction difficult [13]. As a result, an inadequate word database will lead to words that have been omitted from the database being incorrectly identified, particularly in ASR systems that employ word pattern matching.

E. Homophones

There are many words in Bengali language which have almost similar pronunciation but the meaning is actually different. For example, the word কোন /Kōna/ (English:any) and কোণ /Kōṇa/ (English:angle) both have similar utterances, but the meaning is not the same. Generally, while speaking and listening, we understand the meaning from the formation of the sentence. This difficulty, however, cannot be overcome at the auditory or linguistic levels for the system. A language model and pronunciation dictionary are required to solve this. The language model is used to predict the probability of the next word given the previous words, and the pronunciation dictionary is used to map the pronunciation of the word to the corresponding word in the language model. We can use recurrent neural network models over previously predicted words to solve this. The problem of homophones can also be addressed with n-grams as well [13].

F. Silent Letters

Every languages has some words where some letters are silent while we pronounce it. In English word Castle

/ˈkasəl/, we don't pronounce the 'T' in the word. Silent letters can be found in the Bengali language as well. Such as, in the word স্বাধীন /Sādhīna/ (English:Independent), there is no utterance of 'ব' (/Ba/). Because of this, speech-to-text systems might fail to recognize the correct spelling of the word as they cannot identify those silent letters. We can use a pre-defined lexicon rule to avoid this.

The challenges outlined above will limit the Bengali ASR's future research potential. Thus, these problems are future research scopes to handle the problems in Bengali ASR. Though all the challenges need to be addressed, improving some of them can help the performance of existing systems. For example, in order to deal with the difficulty of identifying silent letters, we can try a straight word-to-text transition because most silent letters do not have grammatical connections, as opposed to other problems. Otherwise, these silent letters can cause dissimilerity in speech-to-text system output. Speech-to-word matching and phoneme-to-word matching can also be used to resolve irregular letter sequences. It is possible to use a short-term memory in conjunction with a voice character generator to determine the grammatical and preceding dependencies of characters. If more research were conducted, we can easily improve the performance of the Bengali ASR by tackling the challenges.

V. Technique used in Bengali speech recognition

A. Linear Prediction Coding (LPC)

Like the other methods, LPC (Linear prediction analysis) or short-term predictor is one of the them which are the most powerful methods which can be used as a speech to text recognition technique. Ali et al. [26] established a methodology for Bengali word recognition. The primary objectives of this study were to present a way that is suitable for finding out key features from Bengali speech that is not a representative of real-world speech (ii) evaluate the success rate of the proposed method to find out its accomplishment; and (iii) construct an isolated Bengali word recognizer using the proposed feature extraction technique. Four recognition models were used in the study: I MFCC was used for feature extraction and LPC was used for matching; (ii) LPC was used for feature extraction and DTW was used for matching; (iii) MFCC and Gaussian Mixture Model (GMM) were used for feature extraction and posterior probability function was used for matching; and (iv) MFCC was used for feature extraction and LPC was used for matching. The models were then compared graphically to demonstrate detection rate and time. The third model (MFCC plus GMM) had the highest accuracy (84%) out of 100 Bengali words in optimal settings. As beneficial as LPC has been demonstrated to be, it does have drawbacks that it cannot overcome. At first, the audio signal should fit into the source-filter model in some way. A signal that has a variety of sounds may not work as well. The astute reader may have noticed that there is no change in power within a windowed segment. The signal within a windowed section is assumed to be stationary. By the width of the window, a transient or attack will get smeared. Thus, LPC is unsuitable for noises with a high number of short-term noises.

B. Mel Frequency Cepstral Coefficient (MFCC)

Mel is the psychoacoustic unit of measurement for the perceived pitch of a tone; it has no physical meaning. Khan et al. [27] conducted a comparative analysis of three signal research methodologies which are the power spectral, linear predictive analysis, which we discussed earlier, and mel scale cepstral analysis that we will be discussing now. In this method, a software tool was developed for determining the Bengali phonemes. We found from this study that the 3rd signal research methodology of Khan et al. works very well for Bengali phoneme detection. The signal processing front-end, which transforms the speech waveform into some kind of parametric representation, is the most crucial and common component of all recognition systems, according to this study. In an effort to build a Bengali recognition system that employed three key methodologies: speaker-independent, sub-word unit-based, and isolated speech, an attempt was made. It was 96 percent accurate when dealing with a single speaker and 84.24% accurate when dealing with numerous speakers that this system obtained. Despite the fact that it was able to recognize solitary Bengali speech by grouping words, this identification system lacked any linguistic knowledge.

C. Hidden Markov Model

Hidden Markov Model (HMM) is the most common way to model operations that happen in a specific order. A lot of people say that it works well in speech recognition systems [28]. In this model, there are a lot of different things that happen to the infinite states of an observation series are used to create an observation series. How things work: at each step in the process, the state of things change with a probability calculation for the change. They made their own Hidden Markov Model for Hasnat et al. (HMM)-based scheme for classifying patterns. This is how it works: They also said that. Stochastic model for Bengali was added to the plan speech–to-text. Adaptive noise reduction and noise reduction were done. detection of the end point at the signal pre-processing stage. At the time, first and second order coefficients are found during this phase along with Mel Frequency Cepstral Coefficients (MFCC), there was also MFCC used as a tool to extract features from a picture they put in place the system. The Cambridge Hidden Markov Modeling Toolkit is used to do this (HTK). During this study, the average accuracy rate was 85%. This is a good number.

D. Neural Network and Fuzzy Logic

Before, Artificial Neural Network (ANN) was a very renowned way to recognize words. There isn't much use

Table I
Recent works done on Bengali speech recognition

Year	Author	Dataset	Unit	Input features/method	Approaches	Accuracy
2016	Nahid et al. [14]	3207 sentences	Real numbers	MFCC	CMU Sphinx-HMM	85%
2016	Mukherjee et al. [15]	3150 phonemes	Continuous speech	MFCC	MLP	98.22%
2016	Ahammad et al. [16]	300 digits	Digits	MFCC	BPN	98.46%
2017	Google Voice Search [17]	217902 utterances	Continuous speech	LAS model	LSTM	WER 5.6%
2017	Nahid et al. [18]	2000 words	Real numbers	MFCC	RNN, LSTM	WER 13.2%
2017	Mukherjee et al. [19]	1400 phonemes	Phonemes	MFCC	MLP	98.35%
2018	Saurav et al. [20]	500 words	Isolated words	MFCC	GMM-HMM(Kaldi)	WER 3.96%
2018	Rahman et al. [21]	260 words	Isolated words	DTW	SVM	86.08%
2018	Mukherjee et al. [22]	3710 vowels	Phonemes	LPCC-2	Ensemble Learning	99.06%
2019	Hasan et al. [23]	OpenSLR's dataset	Continuous speech	Improved MFCC	CTC	WER 27.89%
2019	Gupta et al. [24]	240 voice commands	Continuous speech	Energy	Cross-correlation	>75%
2020	Sadeq et al. [25]	28973 sentences	Continuous speech	Labeled LDA	Hybrid CTC-Attention mechanism	>WER 12.8%

of this method in Bengali speech recognition. Most likely, Firoze et al. [29] made the first speech recognition system in Bengali that used fuzzy logic and an ANN. Fuzzy logic was employed in the development of a speech recognition system. Fuzzy logic was proposed as the foundation for all Bengali ambiguity issues, and as a result, the system was created. They were able to show that fuzzy logic helps people respond better to more ambiguous linguistic entities in Bengali speech. This is the first time that cepstral analysis has been used in an artificial neural network (ANN) to recognize Bengali speech. A commercial Hidden Markov Model (HMM) based speech recognizer has an average accuracy of 73%. The fuzzy logic system had an accuracy rate of 86% and the ANN system had an accuracy rate of 90%. However, despite the fact that the fuzzy logic-based system is more accurate when it comes to "difficult" or "polysyllabic" words, the ANN system is more accurate overall.

VI. Analysis

The verge of developing Bengali ASR has been started by many research groups since 2009. It was a great advancement for Bengali ASR when Google added Bengali to their algorithm interface. Receiving an accuracy rate of 85% using the CMU Sphinx-HMM model [14], Nahid et al. used double-layered long short-term memory (LSTM) of the RNN approach and achieved an accuracy rate of 13.2% for word error detection. Record Extract Approximate Distinguish, which is also known as READ, was developed in West Bengal [19], where word mapping accuracy was 98.36%. For the (GMM-HMM) based model and the (DNN-HMM) based model, Pipilika [30] found word error rates of 3.96 percent and 5.30 percent, respectively, for the two models in 2018. The dataset consisted of 500 words recorded by 50 speakers. With a dataset of 3.7K Bengali Vowel Phonemes, Mukherjee et al. obtained a recognition accuracy of 99.06% using LPCC 2 features for classification [22]. A support vector machine was used by Rahman et al. in a Bengali speech classifier [21], which yielded an accuracy of 86.08%. Using an open-source Bengali speech corpus from Google Inc. [31], Hasan et al. proposed a lexicon-free Bengali speech recognition system [23] in 2019.

Use of Connectionist temporal classification (CTC) obtained a word error rate of 39.61% and 27.89% for the two different systems. In 2019, Gupta et al. developed a Bengali voice command detector [24], where a cross-correlation technique was used. In both noiseless and moderate environments, the accuracy was 83%. But in a noisy environment, it went down to 75%. Sadeq et al. [25] also developed a voice command detector in the next year, using the Attention Mechanism and CTC with an RNN-based model. A total of around 29 thousand sentences were recorded from 56 speakers to test the system, where the word error rate was 27.2% and 26.9%, respectively, for the two systems.

VII. Conclusion

There has been a lots of work done for Bengali language ASR. Our paper has listed studies from 2009 to 2020. We have also discussed about the result of these studies and the methods they used for Automated speech recognition. We reviewed these paper to understand the state of ASR of Bengali. The main problems we saw in these researchers faced was lack of proper dataset which is a crucial part for any automated system. The richness in Consonant clusters of Bengali language make it harder for the ML system to understand the word properly. The different sentence structure, the need for preposition and merged between noun and preposition brings more challenge to make the system work properly and efficiently. There are also many words where the pronunciation is very similar between two words. These are some of the challenges the papers faced which we reviewed in our paper. Some of the techniques these papers used are : Linear prediction coding, Mel Frequency Cepstral Coefficient, Hidden Markov Model, Neural Network and Fuzzy logic. Bengali ASR needs more work and research in future. Even though recent works are promising they are nowhere near close to the perfection and efficiency of ASR of other languages like English.

References

[1] Wikipedia contributors, "Bengali language — Wikipedia, the free encyclopedia," 2022, [Online; accessed 30-April-2022]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bengali_language

[2] R. Karim, M. S. Rahman, and M. Z. Iqbal, "Recognition of spoken letters in bangla," in *Proc. 5th international conference on computer and information technology (ICCIT02)*, 2002.

[3] K. Roy, D. Das, and M. G. Ali, "Development of the speech recognition system using artificial neural network," in *Proc. 5th international conference on computer and information technology (ICCIT02)*, 2002, pp. 118–122.

[4] M. R. Hassan, B. Nath, and M. A. Bhuiyan, "Bengali phoneme recognition: a new approach," in *Proc. 6th international conference on computer and information technology (ICCIT03)*, 2003.

[5] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous auutomatic speech recognition system," in *2011 International conference on speech database and assessments (Oriental COCOSDA)*. IEEE, 2011, pp. 51–55.

[6] S. S. I. Badhon, M. H. Rahaman, F. R. Rupon, and S. Abujar, "State of art research in bengali speech recognition," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2020, pp. 1–6.

[7] R. Sultana and R. Palit, "A survey on bengali speech-to-text recognition techniques," in *2014 9th International Forum on Strategic Technology (IFOST)*. IEEE, 2014, pp. 26–29.

[8] R. M. Doha, "Welcome to the machine learning repository of bengali.ai." [Online]. Available: https://bengali.ai/datasets/

[9] S. Mandal, B. Das, P. Mitra, and A. Basu, "Developing bengali speech corpus for phone recognizer using optimum text selection technique," in *2011 international conference on asian language processing*. IEEE, 2011, pp. 268–271.

[10] "Bengali raw speech corpus," https://data.ldcil.org/bengali-raw-speech-corpus, (Accessed on 04/28/2022).

[11] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *Plos one*, vol. 16, no. 4, p. e0250173, 2021.

[12] B. V. Parishad, *Praci Bhasavijnan: Indian Journal of Linguistics*. Bhasa Vidya Parishad., 2001, vol. 20.

[13] M. F. Mridha, A. Q. Ohi, M. A. Hamid, and M. M. Monowar, "Challenges and opportunities of speech recognition for bengali language," 2021.

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

[15] H. Mukherjee, S. Phadikar, P. Rakshit, and K. Roy, "Rearc-a bangla phoneme recognizer," *2016 International Conference on Accessibility to Digital World (ICADW)*, pp. 177–180, 2016.

[16] K. Ahammad and M. M. Rahman, "Connected bangla speech recognition using artificial neural network," *International Journal of Computer Applications*, vol. 149, pp. 38–41, 2016.

[17] G. Inc., "Google voice search," https://voice.google.com/about, 2011.

[18] M. M. H. Nahid, B. Purkaystha, and M. S. Islam, "Bengali speech recognition: A double layered lstm-rnn approach," *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–6, 2017.

[19] H. Mukherjee, C. Halder, S. Phadikar, and K. Roy, "Read—a bangla phoneme recognition system," in *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*. Springer, 2017, pp. 599–607.

[20] J. R. Saurav, S. Amin, S. Kibria, and M. S. Rahman, "Bangla speech recognition for voice search," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–4, 2018.

[21] M. M. Rahman, D. R. Dipta, and M. Hasan, "Dynamic time warping assisted svm classifier for bangla speech recognition," *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–6, 2018.

[22] H. Mukherjee, S. Phadikar, and K. Roy, "An ensemble learning-based bangla phoneme recognition system using lpcc-2 features," 2018.

[23] M. Hasan, M. A. Islam, S. Kibria, and M. S. Rahman, "Towards lexicon-free bangla automatic speech recognition system," *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–6, 2019.

[24] D. Gupta, E. Hossain, M. S. Hossain, K. Andersson, and S. Hossain, "A digital personal assistant using bangla voice command recognition and face detection," *2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON)*, pp. 116–121, 2019.

[25] N. Sadeq, S. Ahmed, S. S. Shubha, M. N. Islam, and M. A. Adnan, "Bangla voice command recognition in end-to-end system using topic modeling based contextual rescoring," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7894–7898, 2020.

[26] M. A. Ali, M. Hossain, M. N. Bhuiyan *et al.*, "Automatic speech recognition technique for bangla words," *International Journal of Advanced Science and Technology*, vol. 50, 2013.

[27] M. F. Khan and D. R. C. Debnath, "Comparative study of feature extraction methods for bangla phoneme recognition," in *5th ICCIT*, 2002, pp. 27–28.

[28] M. Hasnat, J. Mowla, M. Khan *et al.*, "Isolated and continuous bangla speech recognition: implementation, performance and application perspective," 2007.

[29] A. Firoze, M. S. Arifin, and R. M. Rahman, "Bangla user adaptive word speech recognition: approaches and comparisons," *International Journal of Fuzzy System Applications (IJFSA)*, vol. 3, no. 3, pp. 1–36, 2013.

[30] T. of Science SU, "Pipilika," https://pipilika.com/.

[31] O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, "Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 52–55.