Recent Advancement in Speech Recognition for Bangla: A Survey

Sadia Sultana, M. Shahidur Rahman, M. Zafar Iqbal
Dept. of Computer Science and Engineering
Shahjalal University of Science and Technology Sylhet, Bangladesh

***Abstract*—This paper presents a brief study of remarkable works done for the development of Automatic Speech Recognition (ASR) system for Bangla language. It discusses the information of available speech corpora for this language and reports major contributions made in this research paradigm in the last decade. Some important design issues to develop a speech recognizer are: levels of recognition, vocabulary size, speaker dependency, and approaches for classifications; these have been defined in this paper in the order of complexity of speech recognition. It also highlights some challenges which are very important to resolve in this exciting research field. Different studies carried out on the last decade for Bangla speech recognition have been shortly reviewed in chronological order. It was found that the selection of classification model and training dataset play important roles in speech recognition.**

***Keywords*—Bangla ASR; Bangla speech corpora; speaker dependency; vocabulary size; classification approaches; challenges**

I. Introduction

There are several important applications of a speech recognition system. It is used to develop chat-bots in smartphones and gadgets. For customer service in call centers, speech recognition systems are used for automated replies. ASR systems are widely used in automated machines to detect voice commands. Speech recognizer also can be used in detecting crime planned over phone calls and also for detection of hate speech delivery. A study shows that for the English language more than 10% of searches are made by voice and most of them are done using smartphones [2]. This number will increase day by day. The first paper on speech recognition was published in 1950. Since then researches on Speech Technology have achieved remarkable advancement over the last few decades. Major advancement was started in the 1980s with the introduction of Hidden Markov Model (HMM) for speech recognition. The main objective of all research is to build an ASR system which can operate for large vocabulary continuous speech for different languages.

Bangla language is spoken by more than 228 million people all over the world [1]. People from West Bengal, Tripura, Assam, Barak Valley, Andaman, Nicobar Islands, and diaspora living in various countries speak in Bangla. It is the national language of Bangladesh and the official language of the states of West Bengal. As Bangla language has a large number of speaker groups, a successful ASR system for this language will benefit lots of people. Research on Bangla ASR came into focus in the 90s. Recognition of Bangla speech has been started since around 2000. In 2002, A. Karim et al. presented a method for Spoken Letters Recognition in Bangla [3]. In the same year, K. Roy et al. presented the Bangla speech recognition system using Artificial neural networks (ANN) [4]. In 2003, M. R. Hassan presented a phoneme recognition system using ANN [5], and K. J. Rahman presented a continuous speech recognition system using ANN in 2003 [6]. Recently, Google presented a functional speech recognizer and voice search service (SpeechTexter and Google Assistant) for Bangla and other languages. But, these are available for only android devices. The aim of this paper is to summarise all important works done recently on the development of Bangla ASR to facilitate the researchers working in this filed. Fig 1 shows the diagram of a common ASR system. The system takes voice signal $x(n)$ as input. Then, after preprocessing, feature extraction is done to reduce dimensionality of the input vector while preserving the discriminating attributes for recognition. In the decoder, there are mainly three parts: acoustic model, pronunciation dictionary and language model. Acoustic model calculates the probability of observed acoustic signal $(x_1...x_N)$ for the given word sequence $(w_1...w_N)$. Language model provides the probability of proposed word sequence which is $Pr(w_1...w_N)$. Pronunciation dictionary contains list of words with their phonetic transcriptions, and it propose valid words for a given context. The decoder combines inputs from all three parts and applies classification models to deliver the recognized text as output $y(n)$.

The rest of this paper's contents are organized as follows: related works are discussed in Section II, issues to be considered for developing ASR are presented in Section III, and challenges in developing a successful Bangla ASR are explained in Section IV. A list of available Bangla natural speech corpora is reported in Section V, whereas, recent advancement in the last decade is discussed in Section VI. Discussion and conclusion of this study are presented in Sections VII and VIII, respectively.

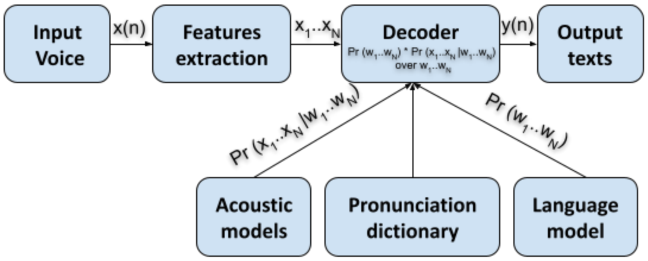

Fig. 1. Automatic Speech Recognition System.

II. Related Works

In 2014, Sultana and Palit surveyed some common speech recognition techniques for Bangla language [7]. In 2020, Badhon et al. reviewed 15 research papers that worked on Bangla ASR. The study represented the datasets and detailed methodologies involved in those researches [8]. A few language-specific surveys on ASR have been conducted for other languages by the researchers. For example, a group of researchers has represented speech recognition techniques for Chinese language [9]. A few studies have been done on speech recognition for Indian Languages [10][11][12]. Lima et al. have studied the speech recognition components for Portuguese language [13]. A literature review on Arabic speech recognition was done by Al-Anzi and AbuZeina [14]. In 2006, Ronzhin et al. studied all methods and models used for Russian speech recognition [15]. The target of such reviews is to represent a useful summary of overall works done on a language-specific ASR.

III. Issues to Consider for ASR

There are some important research concerns that need to be considered when developing a speech recognition system. The application, development complexity, recognition efficiency of the system depend on a few things like utterance type, vocabulary size, speaker dependency, and pattern matching approaches [16]. These factors are discussed briefly in the following sections in increasing order of recognition complexity of the system.

A. Levels of Speech Recognition

The early stage of Speech recognition started with phoneme recognition from recorded speech [17]. ASR systems can be developed to recognize isolated words, connected words, continuous speech, or spontaneous speech. **Isolated words** are uttered separately with sufficient pause between them. For **connected words**, single words are recorded together but still there are pauses between them. In the **continuous speech**, words are connected and there are overlappings between the words. This means deliberate pauses are not added after each word while recording. The **spontaneous speech** recognition system processes natural speech which is characterized by pauses, silence, disfluencies etc. This type of recognizer is most difficult to develop as it requires additional methods to process the speech.

B. Vocabulary Size

The requirements for the vocabulary size of the training dataset depend on the target applications of recognition systems. Some applications require as small vocabulary as a few words, whereas, some other requires millions of words to train the system. **Small-size vocabulary** dataset comprises only a few to hundreds of words. It is used only when the system needs to recognize a fixed small number of digits or other spoken words. For example, automated digit dialing and access control system. It usually contains 2 to 10 hours of recorded speech. Dataset for **medium-size vocabulary** consists of thousands of words. This may contain 10 to 100 hours of recorded speech. These kinds of datasets are used to recognize under-resourced languages. **Large-size vocabulary** contains millions of words. Large vocabulary recognition systems are used in real-life speech recognition e.g. class lecture transcription. In this case, the speech corpora contains more than 100 hours of recordings involving a large number of speakers.

C. Speaker Dependency

For speech recognition, features are collected from speakers' voice and the classification model is trained for these features. The system can be classified depending on the number of speakers they are able to identify successfully. **Speaker-dependant** voice recognition technique identifies different acoustic features of a single voice. These kinds of systems are easier to develop but they do not perform well for unknown speakers. **Speaker-independent** speech recognition system comprises a large collection of speech from several speakers. Features are calculated for this large size data and recognition is performed by searching the best matching for existing data. **Speaker-adaptive** systems collect features from user samples to enrich the training data. The system adapts to the best suited features for speech recognition collected from users. In this way, the error rate is reduced and the system also performs independent of speakers.

D. Different Approaches of Speech Recognition

The approaches used to classify speech are categorized as follows [18]:

Acoustic-phonetic approach: This type of approach focuses on the nature of the speech. Speech features of phonetic units are detected with help of spectral analysis. For example, analysing the accent features of vowels and diphthongs analysis, investigating the formants and energy of the signals, etc. The target is to discover the acoustic features of the sounds and apply those features to recognize continuous speech. Prior to the recognition, this involves a few steps which are features extraction, segmenting the feature contours, and labeling the segments. **Pattern recognition approach:** involves two steps - pattern training and pattern matching. By applying appropriate statistical methods patterns are extracted from speech units which are probably smaller than a word or a single word are stored in the database. A training algorithm is applied for this stored dataset and direct comparison is done between unknown speech segments and trained patterns during the recognition. **Artificial intelligence approach:** This approach is considered as a mimic of our human brain which actually solves the problems based on its previous learning experiences. The problem solving strategy always follows the steps: learning, reasoning, and perception. Typically this type of speech recognition system is based on neural networks (NN). Actually, it combines the ideas taken from both the acoustic-phonetic approach and pattern recognition approach. Input signals are segmented and the acoustic parameters for these segments are calculated. The system is trained for these parameters and pattern matching is done for recognition. The pattern recognition task can be supervised or unsupervised. For supervised pattern recognition, example input patterns are provided to the system as a predefined class. For unsupervised systems, there are no example patterns and these systems are learn-based. Recent researches focus on speech recognition based on Deep neural network (DNN), Recurrent neural network (RNN), hybrid of HMM-DNN approaches.

E. Performance Analysis

For word recognition systems, raw accuracy rate was used in many studies. For continuous speech recognition, word error rate (WER) and word recognition rate (WRR) are most commonly used performance measures. Word error rate can be computed as [19]: $WER = S+D+I/N$; where, S = number of substitutions, D = number of the deletions, I = number of the insertions, N = number of words in the reference. Word recognition rate (WRR) is defined as: $WRR = 1 - WER$. Reference word sequence and recognized word sequence may have different lengths and orders. To skip this problem, the recognized words are first aligned with the reference word then the error rate is calculated.

IV. Challenges in Developing Bangla ASR

Inherently Bangla language has some distinct features like different phonemic systems, presence of long and short vowels, frequent use of consonant clusters, variation in stress and intonation etc. Building an efficient and successful speech recognizer for continuous speech for the Bangla language is a challenging task for the researchers. There are some well-established APIs available for the English language like SAPI, SIRI, IBM Watson API etc. Researchers face few challenges when developing a speech recognizer for Bangla based on a successful API developed for other languages, e.g. English. Some challenges are discussed below:

A. Different Phonemes

Bangla language consists of 14 vowels (7 natural, 7 nasal) and 29 consonants [20]. The numbers of phonemes and phonemic features differ from language to language. For example, Bangla and English language have their distinct phonemic systems [21]. One specialty of Bangla phonemes is that it has 7 nasalized vowels. There also exist two more long vowels /i:/ and /u:/.

B. Speech Patterns

There are some basic differences in the speech pattern of Bangla with that of English and other languages. Bangla is said to be bound stress as for Bangla language stress is high at initial and becomes low at the end of speech [22]. Whereas, English is said to be stress-timed for different stress patterns.

C. Difference in Accents

There are a noticeable differences in accents from region to region, which is true especially for different districts of Bangladesh. Sylhet, Dhaka, Comilla, and other districts have their own dialects. The same word may be pronounced differently for different areas. This is a big challenge to build a common ASR system for all. For example, পাতা -/pata/ (English:leaf) is pronounced as ফাতা- /fata/ by many people from Sylhet.

D. Insufficient Dataset

Still, Bangla is considered to be a low-resource language [23]. That means for the Bangla language, very low resources are publicly available for research purposes. Nowadays almost all the research of Computational Linguistics are concentrating on deep learning-based models. A large training dataset is a key point of getting a highly accurate model for DNN. We only have a few annotated datasets available for Bangla speech recognition.

E. Homophones

There are several words that sound alike but have different spellings and meanings. For example, the words 'শব'- (English: dead body) and 'সব' (English: all) both are pronounced as /ʃob/. Another example is, 'বিশ' (English: twenty) vs 'বিষ' (English: poison); both pronounced as /biʃ/. For the same phonetic representation, these problems cannot be resolved at the acoustic or phonetic levels. A higher level of language analysis is required to do this. To solve these kinds of problems we need a language model and pronunciation dictionary. Unfortunately, still there is a lack of such well-defined models for Bangla continuous speech.

F. Spoken vs Written Words

Sometimes spoken language is not the same as the written text. For example, /bol/ (Englis: ball) and /bolo/ (English: speak) both have the same spelling 'বল'। Another example is /ʃabd̪ʰan/ 'সাবধান'where 'স'is pronounced as 'শ'. Again, a well-defined language model is required to solve this kind of problem.

G. Consonant Clusters

Frequent use of consonant clusters in Bangla speech has made it difficult for word boundary detection. One example is the Bangla word ' চক্কর'(/t̠ɕokːor/). In such case, most of the time the boundary is detected wrongly before the word ends, e.g. 'চক' and 'কর'are identified as two separate words. This degrades the performance accuracy of the overall systems.

H. Mismatched Environment

The background sound in many circumstances is an uncontrollable variable. For example, the level of background noise in the streets of Bangladesh is not as same as that in other developed countries. It is a challenge to build an ASR system that will work efficiently in noisy environments.

I. Unit Selection

Bangla words are pronounced syllable-wise and said to be rhythmic. We know there are different units of speech i.e. syllables, demi-syllables, diphones, phonemes. Bangla and English have different syllable structures [24]. The same pronunciation model can not be applied to both languages. Sometimes it also becomes hard to decide which unit to select to implement an efficient boundary detection method. Syllable segmentation is a challenging task, still, researchers are working on it.

V. Available Speech Corpora for Bangla

There are two standard forms of the Bangla language. One is spoken in West Bengal of India and another form is the official language of Bangladesh. Researchers of both Bangladesh and West Bengal are contributing equally to enrich available language resources in this field. For Bangla, there are a limited number of publicly available speech corpora.

For Kolkata standard, (West Bengal), the first Bangla speech corpus was developed by the Center for Development of Advanced Computing (CDAC), Kolkata [25] in 2005. It is a collection of speech corpora for three East Indian Languages: Bangla, Assamese, and Manipuri. It consists of 32 hours of continuous Bangla speech data. A read speech corpus named SHRUTI was released by IIT, Kharagpur in 2011 which contains 7383 unique Bangla sentences with 49 phonemes [26][27][28]. IARPA (Intelligence Advanced Research Projects Activity) Babel program developed a speech corpus which contains approximately 215 hours of Bengla conversational and scripted telephone speech collected in 2011 and 2012 along with corresponding transcripts [29]. According to a report published in 2014 [30] the Linguistic Data Consortium for Indian Languages (LDC-IL) collected 138 hours of Bangla continuous speech recorded over microphone and telephone line [31]. Technology Development for Indian Languages Program (TDIL) released a speech corpus that contains more than 43000 audio files of Bangla words spoken by 1000 native speakers of West Bengal [32]. Recently European Language Resources Association (ELRA) published a Bangla speech corpus which contains a total of 70 hours of continuous speech recordings [33].

For Bangladeshi Bangla, in 2010, Bangladeshi researchers Firoj Alam et al. developed three speech corpora CRBLP for three different purposes [34]. The corpus for acoustic analysis contains 262 sentences, the diphone corpus contains 4335 sentences and the continuous speech corpus contains 10895 sentences collected from nine categories of data. In the next year, Murtoza et al. developed a phonetically balanced Bangla speech corpus which has 2 million sentences with 47 million biphones [35]. Khan et al. created a connected word speech corpus in 2018 containing 62 hours of recordings collected from more than 100 speakers [36]. Khan and Sobhan constructed another speech corpus for isolated words in the same year which has total of 375 hours of recordings collected from 150 speakers [37]. OpenSLR's 'Large Bengali ASR training dataset' was recently published by Google in 2018, the dataset contains 229 hours of continuous speech for Bangladeshi Bangla [38]. There were 323 males and 182 females in a total of 505 speakers who participated in the recording of 217902 utterances. In 2020 Ahmed et al. developed an annotated speech corpus of 960 hours of speech collected from publicly available audio and text data [24]. The authors of that corpus also proposed an algorithm to automatically generate transcription from existing audio sources. At Shahjalal University, the Natural Language Processing (NLP) research team has developed a speech corpus subak.ko which contains 241 hours of recorded speech with 38,470 unique words which is yet to be published [39]. Table I represents the summary of these mentioned corpora.

VI. Recent Advancement in Last Decade

Researches done for developing Bangla ASR have made a moderate progress since 2009. In 2009, Ghulam Muhammad et al. developed an HMM-based speaker independent Bangla digit recognizer [40] which used their own dataset of 10000 words recorded from 50 males and 50 females. The correction rate is above 80% for the system. An ANN and Linear predictive coding (LPC) based ASR has been proposed by Anup Kumar et al. [41] in the same year. Multilayer perceptron (MLP) approach was followed to design the ANN model and LPC coder was used to extract the coefficients. It was able to discriminate four different words uttered by 2 males and 2 females. In the next year, a Bangla phoneme classifier was built by Kotwal et al. [42]. It used hybrid features of Mel-frequency cepstral coefficients (MFCCs), and the phoneme probability was derived from the MFCCs and acoustic features using Multi-layer neural network (MLN). It obtained an accuracy rate of 68.90% using HMM classifier. The dataset contained 4000 sentences uttered by 40 male speakers. In the study [43] carried out by Mahedi Hasan et al., the researchers focused on triphone HMM-based classifier for word recognition. The system could recognize continuous speech using a speech corpus of 4000 sentences spoken by 40 males and the obtained accuracy was above 80%. Mel-frequency cepstral coefficients MFCC38 and MFCC39 were extracted as features for classifications. In 2011, Firoze et al.[44] proposed a word recognition system that used spectral features and a fuzzy logic classifier. The system was trained for a small dataset of 50 words spoken by a male and a female. The reported accuracy was 80%. An ASR method based on context-sensitive triphone acoustic models was represented by Hassan et al. for continuous speech recognition in 2011 [45]. It applied Multilayer neural network (MLN) to extract phoneme probabilities and triphone HMM for classification. It obtained an accuracy of 93.71% using the same dataset from previous work [42]. At about the same time a study was carried out by Sultana et al. [46] that applied a rule-based approach using Microsoft Speech API (SAPI). The achieved accuracy was 74.81% for 270 Bangla unique words for this system. Akkas Ali et al. [47] presented a Bangla word recognizer in 2013 which used MFCC and LPC features for a hybrid of Gaussian mixture model (GMM) and Dynamic time warping (DTW) for classification. A group of researchers applied Back-propagation neural network (BPNN) for Bangla digit recognition [48]. Perceived recognition accuracy for the speaker-dependent system was 96.33% and for speaker-independent system it was 92%. The sample size of the dataset was limited to 300 words taken from 10 male speakers. A speaker-dependent neural network-based speech recognizer for this language was built in 2014 using MFCC features [49]. It employed feed-forward with back-propagation algorithm for classification and the obtained accuracy was 60%. A study carried out by Mahtab et al. in 2015 claimed a high accuracy of 94% which employed Deep Belief Network (DBN) to classify recorded Bangla digits [50]. Seven layers of Restricted Boltzmann machines (RBM) were considered for designing DBN and speech features were collected from MFCCs. Another study [51] applied semantic Modular time-delay neural network (MTDNN) for Bangla isolated word recognition. They conducted recurrent time delay structure to obtain dynamic long-term memory. In total of 525 words were used to obtain an accuracy of 82%. In 2016, Nahid et al. [52] developed a Bangla real-number recognizer using the API CMU Sphinx 4 which was designed based on HMM. They used their own dataset of 3207 sentences that were taken from male speakers. Feature extraction for this system was done using MFCCs and the accuracy of the system was 85%. In 2016, Mukherjee et al. [53] developed a Bangla character recognition system REARC (Record Extract Approximate Reduce Classify). Their database consisted of 3150 Bangla vowel phonemes retrieved from the voices of 18 females and 27 males. They considered

TABLE I. Available Speech Corpora for Bangla

Year	Corpus Name	Size of dataset	No. of Speakers
2005	CDAC, Kolkata	32 hours of continuous speech	Not known
2010	CRBLP	13.50 hours of continuous speech	4 males, 4 females
2011	Phonetically balanced corpus	1.18 hours of continuous speech	One female
2011	SHRUTI	21.64 hours of continuous speech	26 males, 8 females
2012	IARPA-babel103b-v0.4b	215 hours of continuous speech	Not known
2014	LDC-IL	138 hours of continuous speech	240 males, 236 females
2014	TDIL	43000 audio files of isolated words	1000 native speakers
2018	OpenSLR's 'Large Bengali ASR training dataset'	229 hours of continuous speech	323 males, 182 females
2018	Bangla connected word speech corpus	62 hours continuous speech	50 males, 50 females
2018	Bangla isolated word speech corpus	375 hours of connected words	50 males, 50 females
2019	ELRA-U-S 0031	70 hours of continuous speech	Not known
2020	Bangla Speech Corpus from Publicly Available Audio & Text	960 hours of continuous speech	268 males, 251 females
2020	Subak.ko	241 hours of continuous speech	33 males, 28 females

TABLE II. Recent Works Done on Bangla Speech Recognition

Year	Author	Dataset	Unit	Input features/method	Approaches	Accuracy
2009	Muhammad et al.	10K digits	Digits	MFCC	HMM	> 84%
2009	Paul et al.	4 words	Isolated words	LPC	ANN	not mentioned
2010	Kotwal et al.	4K sentences	Phonemes	MFCC,Energy	HMM	>47%
2010	Hasan et al.	4K sentences	Connected words	MFCC38,MFCC39	Triphone HMM	>86%
2011	Firoze et al.	50 words	Isolated words	Energy, frequency	Fuzzy logic	80%
2011	Hassan et al.	4K sentences	Continuous speech	Phoneme probabilities	HMM	93.71%
2012	Sultana et al.	396 words	Connected words	xml grammar	SAPI	78%
2013	Akkas Ali et al.	1K words	Isolated words	MFCC, LPC, and DTW	GMM	>50%
2013	Hossain et al.	300 samples	Digits	F1, F2, MFCC	BPN	>92%
2014	Barua et al.	8 utterances	Continuous speech	MFCC	ANN	60%
2015	Ahmed et al.	840 words	Isolated words	MFCC	DBN	94%
2015	Ali Khan et al.	525 words	Isolated words	MFCC	MTDNN	82.66%
2016	Nahid et al.	3207 sentences	Real numbers	MFCC	CMU Sphinx-HMM	85%
2016	Mukherjee et al.	3150 phonemes	Continuous speech	MFCC	MLP	98.22%
2016	Ahammad et al.	300 digits	Digits	MFCC	BPN	98.46%
2017	Google Voice Search	217902 utterances	Continuous speech	LAS model	LSTM	WER 5.6%
2017	Nahid et al.	2000 words	Real numbers	MFCC	RNN, LSTM	WER 13.2%
2017	Mukherjee et al.	1400 phonemes	Phonemes	MFCC	MLP	98.35%
2018	Saurav et al.	500 words	Isolated words	MFCC	GMM-HMM(Kaldi)	WER 3.96%
2018	Rahman et al.	260 words	Isolated words	DTW	SVM	86.08%
2018	Mukherjee et al.	3710 vowels	Phonemes	LPCC-2	Ensemble Learning	99.06%
2019	Hasan et al.	OpenSLR's dataset	Continuous speech	Improved MFCC	CTC	WER 27.89%
2019	Gupta et al.	240 voice commands	Continuous speech	Energy	Cross-correlation	>75%
2020	Sadeq et al.	28973 sentences	Continuous speech	Labeled LDA	Hybrid CTC-Attention mechanism	>WER 12.8%

MFCCs as features and the recognition rate was reported as 98.22%. Another study was published in the same year [54] which utilized a BPNN to classify Bangla digits using a dataset of 300 connected digits. The system was tested against three different setups and the perceived accuracies were 88.84%, 98.46%, 82.31% for the experiments. It also represents some contrastive analysis of different digit recognition rates. In 2017, the famous “Google Voice Search” [55] added Bangla language to their system which was a turning point for Bangla ASR. The system employed attention-based encoder-decoder architectures such as Listen, Attend, and Spell (LAS) and n-gram model for context detection [56]. Nahid et al. presented their study [57] for Bangla real number recognition using the dataset of their previous experiment [52]. The new experiment employed double-layered Long short-term memory (LSTM) of RNN approach to recognize individual Bangla words and achieved a word detection error rate of 13.2% and phoneme detection error rate of 28.7%. A Bangla phoneme recognition system READ (Record Extract Approximate Distinguish) was developed by the researchers of West Bengal [58] in 2017. A group of 12 males and 8 females volunteered to develop the dataset of 1400 phonemes for the system. The overall recognition accuracy of the system was 98.36%. In 2018, a voice search system[59] for Bangla search engine Pipilika[60] was proposed. The system experimented with two approaches and obtained word error rates of 3.96% and 5.30% for (GMM-HMM) based model and (DNN-HMM) based model, respectively. The dataset consisted of 500 words that was recorded from 43 male speakers and 7 female speakers. Rahman et al. developed a Bangla speech classifier [61] based on DTW-assisted Support vector machine (SVM) which can detect words with an accuracy of 86.08%. MFCC features were obtained from a dataset of 260 words recorded from 52 speakers for this study. Mukherjee et al. presented an ensemble learning-based Bangla phoneme recognition system which used LPCC-2 features for classification [62]. The system was tested on a dataset of 3710 Bangla vowel phonemes and obtained 99.06% recognition accuracy. 32 males and 21 females volunteered to build the dataset. A lexicon-free Bangla speech recognition system [63] was proposed in 2019 by Hasan et al. The model was trained for open-source Bangla speech corpus published by Google Inc. [38]. Two Connectionist temporal classification

(CTC)-based experiments were carried out and the obtained WERs are 39.61% and 27.89% for the setups. A Bangla voice command detector [64] was developed by Gupta et al. in 2019. This digital personal assistant could execute a task by recognizing a Bangla command. It employed the cross-correlation technique to compare the energy of a given command with a pre-recorded signal. Five speakers delivered speech to build a dataset of 240 audio files for 12 voice commands recorded in Bangla. The system perceived accuarcies of 83%, 83% and 75% for noiseless, moderate, and noisy environments, respectively. Another voice command recognition system was developed in 2020 [65] by Sadeq et al.. The model used a hybrid of CTC and Attention mechanism in the end-to-end architecture with an RNN-based language model. The system was trained for the Bangla corpus that was released by Google [38]. For testing, a separate corpus containing 28973 sentences recorded from 34 male and 22 female speakers was created. Overall WERs of the system for two different setups were 27.2% and 26.9%.

VII. Discussion

The study reveals that a good number of researches have been carried out in the field of developing Bangla ASR in the last decade. A large number of studies concentrated on developing a successful word recognition system for Bangla. Since nowadays researchers are more interested in doing NLP researches using end-to-end systems, there is growing attention to developing a continuous speech recognition system for Bangla based on this type of model. From Table II, it is seen that the most commonly used features are MFC coefficients and the recent trend of using classifiers is focused on ANN-based models. Considering the size of the corpus the largest speech corpus for Bangla is the 'Bangla Speech Corpus from Publicly Available Audio & Text', though publicly available largest natural corpus for this language is Google's 'Large Bengali ASR training dataset'. Considering, the training dataset and accuracy level Google's voice API is performing the best for Bangla speech recognition to date. The system uses n-gram language model which has the problem with synonyms and rigidness. It is evident that using a larger dataset in newer ML-based models improving the overall recognition rate of the system. Though, lots of works have been done related to Bangla ASR still we need to develop an efficient language model and pronunciation model to be used for this purpose.

VIII. Conclusion

In this paper, a concise study has been presented covering all the relevant researches done for Bangla ASR. A brief summary of 24 research papers has been reported to address major advancements in this field. A detailed list of available speech corpora for this language also has been presented. Some challenges regarding the development of a successful and efficient Bangla ASR also have been discussed. The study found that the recent trend focuses on the deep learning-based approaches for classification. Introducing larger datasets for Bangla natural speech is improving the performance of ASR systems. This study may provide some important research insight for the researchers in this field.

References

[1] Wikipedia. Bangali Language; 2021. Available from: https://en.wikipedia.org/wiki/Bengali_language.

[2] TheNewStack. Bangali Language; 2021. Available from: https://thenewstack.io/speech-recognition-getting-smarterstate-art-speech-recognition.

[3] Rezaul Karim AHM, Rahman MS, Iqbal MZ. Recognition of Spoken Letters in Bangla. In: In proceedings of the 5th ICCIT conference. ICCIT; 2002. p. 1–5.

[4] Roy K, Das D, Ali MG. Development of the speech recognition system using artificial neural network. In: Proc. 5th international conference on computer and information technology (ICCIT02); 2002. p. 118–122.

[5] Hassan MR, Nath B, Bhuiyan MA. Bengali phoneme recognition: a new approach. In: Proc. 6th international conference on computer and information technology (ICCIT03); 2003.

[6] Rahman KJ, Hossain MA, Das D, Islam AZMT, Ali DMG. Continuous Bangla Speech Recognition System. In: Proc. 6th Int. Conf. on Computer and Information Technology (ICCIT03); 2003. p. 1–5.

[7] Sultana R, Palit R. A survey on Bengali speech-to-text recognition techniques. In: 2014 9th International Forum on Strategic Technology (IFOST); 2014.

[8] Badhon SSI, Rahaman MH, Rupon FR, Abujar S. State of art Research in Bengali Speech Recognition. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE; 2020. p. 1–6.

[9] Fu SW, Lee C, Clubb OL. A survey on Chinese speech recognition. Communications of COLIPS. 1996;6(1):1–17.

[10] Besacier L, Barnard E, Karpov A, Schultz T. Automatic speech recognition for under-resourced languages: A survey. Speech communication. 2014;56:85–100.

[11] Hemakumar G, Punitha P. Speech recognition technology: a survey on Indian languages. International Journal of Information Science and Intelligent System. 2013;2(4):1–38.

[12] Kurian C. A survey on speech recognition in Indian languages. International Journal of Computer Science and Information Technologies. 2014;5(5):6169–6175.

[13] de Lima TA, Da Costa-Abreu M. A survey on automatic speech recognition systems for Portuguese language and its variations. Computer Speech & Language. 2020;62:101055.

[14] Al-Anzi F, AbuZeina D. Literature survey of Arabic speech recognition. In: 2018 International Conference on Computing Sciences and Engineering (ICCSE). IEEE; 2018. p. 1–6.

[15] Ronzhin AL, Yusupov RM, Li IV, Leontieva AB. Survey of russian speech recognition systems. In: Proc. of 11th International Conference SPECOM; 2006. p. 54–60.

[16] Saksamudre SK, Shrishrimal P, Deshmukh R. A review on different approaches for speech recognition system. International Journal of Computer Applications. 2015;115(22).

[17] Ostendorf M, Roukos S. A stochastic segment model for phoneme-based continuous speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1989;37(12):1857–1869.

[18] Singh AP, Nath R, Kumar S. A Survey: Speech Recognition Approaches and Techniques. In: 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE; 2018. p. 1–4.

[19] Gaikwad SK, Gawali BW, Yannawar P. A review on speech recognition technique. International Journal of Computer Applications. 2010;10(3):16–24.

[20] Wikipedia. Bengali phonology; 2011. Available from: https://en.wikipedia.org/wiki/Bengali_phonology.

[21] Barman B. A contrastive analysis of English and Bangla phonemics. Dhaka University Journal of Linguistics. 2009;2(4):19–42.

[22] Mandal SKD, Gupta B, Datta AK. Word boundary detection based on suprasegmental features: A case study on Bangla speech. International Journal of Speech Technology. 2007;9(1-2):17–28.

[23] Bhattacharjee A, Hasan T, Samin K, Rahman MS, Iqbal A, Shahriyar R. BanglaBERT: Combating Embedding Barrier for Low-Resource Language Understanding. arXiv preprint arXiv:210100204. 2021;.

[24] Ahmed S, Sadeq N, Shubha SS, Islam MN, Adnan MA, Islam MZ. Preparation of Bangla Speech Corpus from Publicly Available Audio & Text. In: Proceedings of The 12th Language Resources and Evaluation Conference; 2020. p. 6586–6592.

[25] C-DAC. Annotated Speech Corpora for 3 East Indian Languages viz. Bangla, Assamese and Manipuri; 2005. Available from: https://www.cdac.in/index.aspx?id=mc_ilf_Speech_Corpora.

[26] Das B, Mandal S, Mitra P. SHRUTI Bengali Continuous ASR Speech Corpus; 2011. Available from: https://cse.iitkgp.ac.in/~pabitra/shruti_corpus.html.

[27] Das B, Mandal S, Mitra P. Bengali speech corpus for continuous auutomatic speech recognition system. In: 2011 International conference on speech database and assessments (Oriental COCOSDA). IEEE; 2011. p. 51–55.

[28] Mandal S, Das B, Mitra P, Basu A. Developing Bengali speech corpus for phone recognizer using optimum text selection technique. In: 2011 International Conference on Asian Language Processing. IEEE; 2011. p. 268–271.

[29] IARPA. IARPA Babel Bengali Language Pack IARPA-babel103b-v0.4b; 2016. Available from: https://catalog.ldc.upenn.edu/LDC2016S08.

[30] Kandagal AP, Udayashankara V. Speech Corpus Development for Speaker Independent Speech Recognition for Indian Languages. Grenze International Journal of Computer Theory and Engineering. 2017;3(4).

[31] LDCIL. Bengali Raw Speech Corpus; 2014. Available from: https://data.ldcil.org/bengali-raw-speech-corpus.

[32] TDIL. Bengali Speech Data – ASR; 2018. Available from: http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=2000&lang=en.

[33] ELRA. ELRA-U-S 0031; 2018. Available from: http://universal.elra.info/product_info.php?cPath=37_39&products_id=1669.

[34] Alam F, Habib S, Sultana DA, Khan M. Development of annotated Bangla speech corpora. 2010;.

[35] Murtoza S, Alam F, Sultana R, Chowdhur S, Khan M. Phonetically balanced Bangla speech corpus. In: Proc. Conference on Human Language Technology for Development 2011; 2011. p. 87–93.

[36] Khan MF, Sobhan MA. Creation of Connected Word Speech Corpus for Bangla Speech Recognition Systems. Asian Journal of Research in Computer Science. 2018; p. 1–6.

[37] Khan MF, Sobhan MA. Construction of large scale isolated word speech corpus in Bangla. Global Journal of Computer Science and Technology. 2018;.

[38] Kjartansson O, Sarin S, Pipatsrisawat K, Jansche M, Ha L. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. 2018;.

[39] Subakko. Speech to Text; 2019. Available from: https://stt.sustbanglaresearch.org/.

[40] Muhammad G, Alotaibi YA, Huda MN. Automatic speech recognition for Bangla digits. In: 2009 12th International Conference on Computers and Information Technology. IEEE; 2009. p. 379–383.

[41] Paul AK, Das D, Kamal MM. Bangla speech recognition system using LPC and ANN. In: 2009 Seventh International Conference on Advances in Pattern Recognition. IEEE; 2009. p. 171–174.

[42] Kotwal MRA, Hossain MS, Hassan F, Muhammad G, Huda MN, Rahman CM. Bangla phoneme recognition using hybrid features. In: International Conference on Electrical & Computer Engineering (ICECE 2010). IEEE; 2010. p. 718–721.

[43] Hasan MM, Hassan F, Islam GMM, Banik M, Kotwal MRA, Rahman SMM, et al. Bangla triphone hmm based word recognition. In: 2010 IEEE Asia Pacific Conference on Circuits and Systems. IEEE; 2010. p. 883–886.

[44] Firoze A, Arifin MS, Quadir R, Rahman RM. Bangla Isolated Word Speech Recognition. In: ICEIS (2); 2011. p. 73–82.

[45] Hassan F, Kotwal MRA, Muhammad G, Huda MN. MLN-based Bangla ASR using context sensitive triphone HMM. International Journal of Speech Technology. 2011;14(3):183–191.

[46] Sultana S, Akhand M, Das PK, Rahman MH. Bangla Speech-to-Text conversion using SAPI. In: 2012 International Conference on Computer and Communication Engineering (ICCCE). IEEE; 2012. p. 385–390.

[47] Ali MA, Hossain M, Bhuiyan MN, et al. Automatic speech recognition technique for Bangla words. International Journal of Advanced Science and Technology. 2013;50.

[48] Hossain M, Rahman M, Prodhan UK, Khan M, et al. Implementation of back-propagation neural network for isolated Bangla speech recognition. arXiv preprint arXiv:13083785. 2013;.

[49] Barua P, Ahmad K, Khan AAS, Sanaullah M. Neural network based recognition of speech using MFCC features. In: 2014 international conference on informatics, electronics & vision (ICIEV). IEEE; 2014. p. 1–6.

[50] Ahmed M, Shill PC, Islam K, Mollah MAS, Akhand M. Acoustic modeling using deep belief network for Bangla speech recognition. In: 2015 18th International Conference on Computer and Information Technology (ICCIT). IEEE; 2015. p. 306–311.

[51] Khan MYA, Hossain SM, Hoque MM. Isolated Bangla word recognition and speaker detection by semantic modular time delay neural network (MTDNN). In: 2015 18th International Conference on Computer and Information Technology (ICCIT). IEEE; 2015. p. 560–565.

[52] Nahid MMH, Islam MA, Islam MS. A noble approach for recognizing bangla real number automatically using cmu sphinx4. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV). IEEE; 2016. p. 844–849.

[53] Mukherjee H, Phadikar S, Rakshit P, Roy K. REARC-A Bangla phoneme recognizer. In: 2016 International Conference on Accessibility to Digital World (ICADW). IEEE; 2016. p. 177–180.

[54] Ahammad K, Rahman MM. Connected bangla speech recognition using artificial neural network. International Journal of Computer Applications. 2016;149(9):38–41.

[55] Inc G. Google Voice Search; 2011. Available from: https://voice.google.com/about.

[56] Chiu CC, Sainath TN, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, et al. State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 4774–4778.

[57] Nahid MMH, Purkaystha B, Islam MS. Bengali speech recognition: A double layered LSTM-RNN approach. In: 2017 20th International Conference of Computer and Information Technology (ICCIT). IEEE; 2017. p. 1–6.

[58] Mukherjee H, Halder C, Phadikar S, Roy K. READ—a Bangla phoneme recognition system. In: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. Springer; 2017. p. 599–607.

[59] Saurav JR, Amin S, Kibria S, Rahman MS. Bangla speech recognition for voice search. In: 2018 international conference on Bangla speech and language processing (ICBSLP). IEEE; 2018. p. 1–4.

[60] of Science SU, Technology. Pipilika; 2013. Available from: https://pipilika.com/.

[61] Rahman MM, Dipta DR, Hasan MM. Dynamic time warping assisted svm classifier for bangla speech recognition. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). IEEE; 2018. p. 1–6.

[62] Mukherjee H, Phadikar S, Roy K. An ensemble learning-based Bangla phoneme recognition system using LPCC-2 features. In: Intelligent Engineering Informatics. Springer; 2018. p. 61–69.

[63] Hasan MM, Islam MA, Kibria S, Rahman MS. Towards Lexicon-free Bangla Automatic Speech Recognition System. In: 2019 International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE; 2019. p. 1–6.

[64] Gupta D, Hossain E, Hossain MS, Andersson K, Hossain S. A digital personal assistant using bangla voice command recognition and face detection. In: 2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON). IEEE; 2019. p. 116–121.

[65] Sadeq N, Ahmed S, Shubha SS, Islam MN, Adnan MA. Bangla Voice Command Recognition in end-to-end System Using Topic Modeling based Contextual Rescoring. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2020. p. 7894–7898.