A study on the challenges and opportunities of speech recognition for Bengali language

M. F. Mridha1 · Abu Quwsar Ohi1 · Md Abdul Hamid2 · Muhammad Mostafa Monowar2

Published online: 5 November 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract
Speech recognition is a fascinating process that offers the opportunity to interact and command the machine in the field of human-computer interactions. Speech recognition is a language-dependent system constructed directly based on the linguistic and textual properties of any language. Automatic speech recognition (ASR) systems are currently being used to translate speech to text flawlessly. Although ASR systems are being strongly executed in international languages, ASR systems’ implementation in the Bengali language has not reached an acceptable state. In this research work, we sedulously disclose the current status of the Bengali ASR system’s research endeavors. In what follows, we acquaint the challenges that are mostly encountered while constructing a Bengali ASR system. We split the challenges into language-dependent and language-independent challenges and guide how the particular complications may be overhauled. Following a rigorous investigation and highlighting the challenges, we conclude that Bengali ASR systems require specific construction of ASR architectures based on the Bengali language’s grammatical and phonetic structure.

Keywords Automatic speech recognition · Bengali · Phoneme · Speech to text · Language-dependent challenges · Language-independent challenges

M. F. Mridha
firoz@bubt.edu.bd
Abu Quwsar Ohi
quwsarohi@bubt.edu.bd
Md Abdul Hamid
mabdulhamid1@kau.edu.sa
Muhammad Mostafa Monowar
mmonowar@kau.edu.sa

1 Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh
2 Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Kingdom of Saudi Arabia
1 Introduction

Undoubtedly, speech is the most fascinating and competent form of interaction among one another. Moreover, it is additionally conceivable to utilize speech as an outstanding medium to interact with machines. Wherefore, speech recognition investigation has advanced from research center exhibitions to genuine applications. Hence, speech recognition systems are frequently observed and accepted in daily use applications (Rabiner 1990). This daily usage and dependence on ASR systems require the architecture to be accurate to its best. A user may feel interrupted if the ASR-based search system outputs scrambled or wrong words while he/she is using a voice search feature on the web or calling the wrong person while using ASR-based automated calling functions. Hence, implementing an accurate ASR system requires an in-depth analysis of the speech-to-text translation systems, including grammar and word-level knowledge.

Language dependency is one of the greatest obstructions of a speech recognition system. Thus a speech recognition system has to target a specific language base. Due to language dependency, a system that better recognizes English speech may not correctly recognize other linguistic speech. Moreover, language dependency is solely due to the grammatical properties of specific languages. A similar condition also applies to the Bengali language, which has wider structural and grammatical variations than the English language. However, language dependency has not been frequently investigated by researchers. Apart from discussing existing literature of ASR systems and datasets, most research has been conducted in algorithm selections, speech variation challenges, architectural investigation, etc. Table 1 represents a comparison of some of the notable and recent analyses conducted in the ASR literature. Consequently, in this research endeavor, we deeply investigate the grammatical aspects of speech recognition, along with the challenges of algorithms w.r.t. grammar and phones.

In this paper, we ground our discussion on the challenges and opportunities that a Bengali ASR system pose. The core contribution of the paper includes:

- We conduct a comprehensive investigation of most of the works undertaken in the Bengali ASR system, including speech corpora and architecture. To the best of our knowledge, no comprehensive survey has been made discussing the grammatical and architectural relation of ASR systems.
- We point out various challenges encountered while implementing Bengali ASR systems. Moreover, we provide an anatomy of the challenges and discuss linguistic and grammatical differences between English and Bengali language.
- Finally, we provide future directions that should be recollected while building architectures. Further, we provide an optimal structure that may resolve the issues of Bengali ASR systems.

The rest of the paper is segmented as follows. Section 2 acquaints the generic architectures that are investigated in the ASR domain. Section 3 introduces the attempts which are executed in the track of the Bengali ASR system. Section 4 contains a detailed investigation of the challenges Bengali ASR system poses. Section 5 summarizes the overall challenges introduced in the paper and proposes an optimal architecture to solve the challenges. Finally, Sect. 6 concludes the paper.
Table 1  The table exhibits some of the surveys conducted in the domain of ASR systems. Most of the surveys cover a specific event of the ASR systems

| Study                                      | Reviewed feature extraction strategies | Reviewed deep learning strategies | Discussed existing ASR methods | Reviewed datasets | Discussed grammatical variation | Core contribution                                                                 |
|--------------------------------------------|----------------------------------------|-----------------------------------|--------------------------------|------------------|---------------------------------|----------------------------------------------------------------------------------|
| Trentin and Gori (2001)                    | ✓                                      | ✗                                 | ✓                              | ✗                | ✗                               | Comparing HMM and ANN architectures. Pointing towards the performance improvement of hybrid architectures. |
| Benzeghiba et al. (2007)                   | ✓                                      | ✗                                 | ✓                              | ✗                | ✗                               | Discussion in speech recognition based on speech variations, such as emotion, phycology, speech rate, accent, etc. |
| Mattys et al. (2012)                       | ✗                                      | ✓                                 | ✓                              | ✗                | ✗                               | Through discussion of speech recognition in adverse conditions.                  |
| Besacier et al. (2014)                     | ✓                                      | ✗                                 | ✓                              | ✓                | ✗                               | Focused on under-resourced languages. Discussed extinction, challenges, and resources of such languages. |
| Vadwala et al. (2017)                      | ✓                                      | ✗                                 | ✓                              | ✗                | ✗                               | Addressed the advantages and disadvantages of classic ASR techniques.             |
| ARSLAN and BARIŞÇI (2020)                  | ✓                                      | ✓                                 | ✓                              | ✓                | ✗                               | Brief in Turkish speech recognition.                                             |
| Ours                                       | ✓                                      | ✓                                 | ✓                              | ✓                | ✓                               | Brief in Bengali speech recognition along with architectural strategies concerning grammatical properties. |
2 Attempts in ASR system

The first speech recognition system was introduced in 1920, which was the first machine to recognize speech (Reddy 1976). Later, the journey of speech recognition technology continued to be improved by the independent works of researchers all around the globe. The researchers interested in speech recognition systems introduced and adopted many state-of-the-art techniques that have improved and are still improving the precision of speech recognition systems. Pattern matching approaches such as brute-force techniques, phonetic segmentation, and hybrid systems were first introduced in speech recognition systems. However, the vast improvement is often concerned after Hidden Markov Models (HMM) adaption, which appeared in late 1970. HMM has become popular in ASR systems due to more immeasurable pattern analysis expertise over large vocabularies (Tebelskis 1995; Gales et al. 2008) and being feasible to practice (Rashmi 2014).

Lately, after the improvement of Artificial Neural Network (ANN) architectures, neural network-based speech recognition systems are also proved and considered to be better. Popular architectures of Deep Neural Networks (DNN) such as Convolutional Neural Networks (CNN) (Haque et al. 2020), Residual Networks (Zoughi et al. 2020) are being implemented in ASR systems, and they are proving to be effective. DNN based architectures are also proven to be more effective than any other architectures implemented in the Bengali ASR system (Khan et al. 2018). Popular feature extraction techniques like Principal Component Analysis (PCA) (Takiguchi and Ariki 2007), Linear Discriminate Analysis (LDA) (Haeb-Umbach and Ney 1992), Independent Component Analysis (ICA) (Kwon and Lee 2004), Wavelet Analysis (Ziółko et al. 2011) has been implemented to extract speech features from acoustic waveforms. Among the aforementioned feature extraction strategies, PCA is used to extract structure from input data. However, the drawback of PCA is that it can only recognize the linearity of data. In contrast to PCA, a deep learning-based strategy, named AutoEncoder, can acknowledge data’s non-linearity. Hence, currently, AutoEncoders are being implemented to embed the non-linearity of data. In the case of LDA, a probabilistic LDA (PLDA) is used mostly to recognize features from speech embeddings. Both LDA and PDA are intensely studied in speaker recognition tasks (Dehak et al. 2010; Variani et al. 2014). Specialized feature extraction systems like Mel-frequency Cepstrum Coefficient (MFCC) (Zheng et al. 2001; Ittichaichareon et al. 2012), Cepstral Mean Subtraction (Westphal 1997), RASTA filtering (Hermansky and Morgan 1994; Hermansky and Fousek 2005) is also observed to be used to extract features from waveforms. MFCC has been deeply investigated in the domain of speech and speaker recognition. Currently, MFCC is fused with various CNN architectures and mostly generating better accuracy in speech recognition frameworks. The reason for achieving better accuracy lies behind the mel-scales of the MFCC. A low-scale MFCC excludes unwanted features and greatly focuses on the phones of speech (Molla and Hirose 2004).

An ASR system has two principal processing architectures, which are observed mostly to exist, a) Feature Extraction and b) Pattern Matching. Feature extraction is the process of extracting speech parameters having acoustic correlation from an acoustic waveform (Dave 2013), and pattern matching is the process of matching the extracted speech features with the correct output from the template database (Gaikwad et al. 2010). The pattern matching can be either speech to phoneme matching (Bird et al. 2019), or speech to word matching (Audhkhasi et al. 2018), although we define a hybrid method that can perform both. Generally, the term hybrid is mostly used to identify such ASR architectures that combine HMM and Multi-Layer Perceptron (MLP) method (Graves et al. 2013; Bourlard
A study on the challenges and opportunities of speech recognition…

However, in this paper, we define the term hybrid on the basis of the combination of speech to text and speech to phoneme scheme. The proper combination and tweaks applied to the two principal architectures (feature extraction and pattern matching) may significantly increase the performance of the system. However, some attachments such as word segmentation (segmenting speech frames from continuous speech), noise reduction, and phoneme to word transformation can be observed in ASR systems to enhance the usage and robustness of ASR systems. Figure 1 demonstrates the overall course of processes which are performed in an ASR system. Furthermore, Fig. 2 demonstrates the overall course of processes that are performed in a hybrid ASR system.

Apart from the general strategies of speech recognition, the current improvement of or recurrent neural networks (RNN) has led the speech recognition system to a new strategy named end-to-end ASR (Graves and Jaitly 2014). A single RNN based architecture performs feature extraction and speech to pattern matching simultaneously in an end-to-end method. The advantage of the end-to-end strategy is that the whole network is always trained using a single loss function. Connectionist temporal classification (CTC) loss is broadly implemented as a loss function in an end-to-end framework. However, the limitation of these methods is that they require a considerably large amount of

---

**Fig. 1** The figure illustrates the common architecture of automated speech recognition systems. The red dashed boxes frame the two basic pattern matching or classification schemes (phoneme and text matching) that are frequently practiced in speech recognition architectures. (Color figure online)

**Fig. 2** The figure illustrates the standard architecture of hybrid automated speech recognition systems. The red-dashed boxes frame the two essential pattern matching, or classification schemes (phoneme and text matching) frequently practiced in speech recognition architectures. The final text output is obtained based on the confidence evaluation of the word matching scheme. (Color figure online)
data to work precisely (Hannun et al. 2014). Moreover, it also requires a considerable amount of time to attain optimal features from the input stream. Figure 3 visualizes the basic structure of an end-to-end framework.

Some modifications of the end-to-end architectures have proven to be remarkably suitable for continuous speech and text processing. Amongst them, sequence-to-sequence (seq2seq) and attention-based models are well-considered. Seq2seq models contain an encoder and a decoder, both having a stack of RNN layers. The encoder generates meaningful embeddings from the input and encourages the decoder towards correct predictions. Figure 4 illustrates a common scenario of seq2seq framework. On the contrary, attention-based architectures (Chorowski et al. 2015) perform similarly as a seq2seq model (Dong et al. 2018). Specifically, the attention mechanism is attached to a seq2seq model that extends the knowledge of previous inputs and outputs, resulting in a superior assumption of the network.
RNN has significantly been investigated in end-to-end architectures. As a result, two sophisticated strategies have been introduced, Long-short term memory (LSTM) (Hochreiter and Schmidhuber 1997) and Gated recurrent units (GRU) (Chung et al. 2014). General RNN based architectures are prone to vanishing gradient problems, whereas LSTM and GRU networks evade such issues. LSTM and GRU both contain a memory of the previous states and have been preferred over general RNNs. GRU network requires fewer parameters in comparison to LSTM. However, LSTM has been proven to perform better in language modeling for speech recognition (Irie et al. 2016). Recurrent architectures are still a region of interest to ASR researchers due to recognizing complex sequences from speech inputs.

3 Attempts in Bengali speech recognition

3.1 Attempts in generating Bengali speech corpora

Efforts have been made in the Bengali speech recognition system, although there is still plenty to explore. Most of the works carried out in Bengali ASR systems are dispersed due to the absence of dataset availability. The scarcity of Bengali speech data caused the individual researchers to create their speech corpora, which has also not been made public. Therefore, most works were incomparable to each other, and it was impossible to prove the authenticity and quality of the corpora as well as research works. Currently, to the best of our knowledge, nine corpora are available for Bengali ASR systems. One is a real number speech corpus, one voice command corpus, and the others are full Bengali speech corpora. A complete analysis of the speech corpora is presented in Table 2.

The scarcity of Bengali speech datasets can only be resolved by producing massive, publicly available quality datasets. A quality speech recognition dataset has various usability domains, including speech to text processing, text to speech processing, speaker recognition, far-field speech recognition, etc. (Ohi et al. 2021). While creating a Bengali speech dataset, the following cases should be considered:

- Currently, quality speech datasets target specific environments: clean environment, telephony environment, broadcast (tv/radio) environment, meeting environment, far-field environment, in-the-wild environment. The most challenging environments are telephony, far-field, and in-the-wild environments. Most of the state-of-the-art speech recognition systems target these types of datasets.
- A Bengali speech recognition dataset should contain an accurate transcript of the speech. Also, it may contain speaker information (gender/emotion), environmental information as well.
- Diverse features in the speech dataset are required to make the dataset more challenging and practical. Diversity can be achieved in various constraints: input device, dialect, age, environment, noise constraint, speech disability, etc.
- Most famous datasets maintain clean and noisy datasets separate (Panayotov et al. 2015). Separating clean and noisy datasets help researchers to implement Bengali speech recognition prototypes based on a particular scenario.
- A Bengali speech dataset should target Bengali-specific features, such as collecting speech from different dialects, collecting speech for critical and similar words, and especially handling the letter utterance similarity.
Table 2  The table contains an insight into the currently available corpora suitable for Bengali speech recognition. The column ‘Type’ defines the category of the corpora. The ’Source’ column explains the source from which the data was collected. ‘Speech Length’ column refers to an approximate length of the speech corpora in hours. ‘Unique Utterances’ column generates an approximate unique number of materials available in the corpora. The ’Repository Reference’ column contains the reference link where the dataset can be found. The “Availability” column perceives whether the datasets are publicly or privately available.

| Collector            | Type                  | Source                     | Speech length | Unique utterances | Repository reference                  | Availability |
|----------------------|-----------------------|----------------------------|---------------|-------------------|---------------------------------------|--------------|
| Srivastava et al. (2020) | Speech Phoneme       | Voluntary Contribution    | –             | 47 phonemes       | Srivastava et al.                     | Public       |
| Nahid et al. (2018)   | Spoken Number Corpus  | Voluntary Contribution    | 3.8 hours     | 115 numbers       | Nahid, Md Mahadi Hasan (2018)         | Public       |
| Alam et al. (2010)    | General Speech Corpus | Voluntary Contribution    | 24 h          | –                 | Alam (2018)                           | Public       |
| Das et al. (2011)     | General Speech Corpus | Voluntary Contribution    | 25 h          | 11,000 words      | Das et al. (2021)                     | Public       |
| Reza et al. (2017)    | Speech Command Corpus | Voluntary Contribution    | –             | 30 words          | –                                     | Private      |
| Mandal et al. (2011)  | General Speech Corpus | Voluntary Contribution    | 26 h          | 19,640 words      | –                                     | Private      |
| Gales et al. (2014)   | General Speech Corpus | Telephone Conversation    | 215 h         | –                 | Gales et al. (2021)                   | Private      |
| Google                | General Speech Corpus | Crowd Sourced              | 229 h         | 200,000 words     | Google                                | Public       |
| Ahmed et al. (2020)   | General Speech Corpus | TV News, Audiobooks       | 960 h         | 1,600,000 words   | –                                     | Private      |
3439

• A Bengali speech dataset must cover a large volume of word database and adequately present the statistics of the dataset variations.

Creating a Bengali speech dataset specifically for deep learning architectures is challenging, as training the current deep learning strategies requires vast data. Figure 5 illustrates the general stages of the data collection procedure. The collection of speech datasets may include crowd-sourcing or an especially selected population. However, big datasets are often crowd-sourced. Further, a Bengali speech dataset may require additional statistical analysis to balance variation in numerous domains. It may require pruning and selection process as well. Moreover, the speech dataset requires some pre-processing, such as noise cancellation (optional), sound normalization, reducing silent intervals, and so on. Moreover, a manual or semi-automated process is to be conducted to generate speaker diarization and speech transcription. Finally, after validating the overall process, a quality Bengali speech dataset can be produced.

3.2 Attempts in designing Bengali ASR systems

Research works in the scope of the Bengali ASR system began in late 2000. Recognition of Bengali spoken letters (Karim et al. 2002) was introduced in an earlier stage. The pioneer works in the sector mainly were based on self-made short datasets and used statistical approaches (Houque 2006; Islam et al. 2005; Hassan et al. 2003; Rahman et al. 2003; Khan and Debnath 2002). The first work using Neural Networks, which was witnessed in 2009 (Paul et al. 2009). The authors first pre-processed the input speech using pre-emphasis and hamming window. Then a 12 dimensional Linear Predictive Coding (LPC) is used to produce speech features. Finally, the speech features are fed to Artificial Neural Network to identify speech. However, the research work was conducted using a limited dataset of four persons, and no evidence of performance measurement was included.

In the following year, a continuous Bengali speech-to-text system was introduced. The work was carried out using CMUSphinx (Lee et al. 1990) (a speech recognition system), and a custom dataset was used to train the speech recognition system (Mandal et al. 2010). The system was designed using a phoneme pattern matching scheme and performed a phoneme to text translation system using tri-gram. CMUSphinx implements a three-state (tri-gram) statistical HMM and it uses GMM for probability distribution function. The approach generated 13% word error rate (WER) on 100 sentences.

In the same year, a speech segmenting method was also introduced that could segment Bengali speech from a continuous waveform (Rahman et al. 2010). The authors implemented mean windows to segment each of the words from continuous speech. Then, each segmented word was further referred among three clusters, belonging to mono, di, and tri

Fig. 5 The image illustrates the general steps of preparing a Bengali speech dataset

 Springer
syllable, based on the gaps in each segmented word. With six speakers of a 120 sentence dataset, the authors gained 98.48% accuracy.

The course of study continued, and in 2012, two new methods were introduced, among which, the first method was implemented using Microsoft Speech Application Programming Interface (SAPI) (Sultana et al. 2012). Due to the dependency on SAPI, the research work had a limitation. The architecture had to translate SAPI outputted English words to Bengali, and it was done through a direct English-to-Bengali word matching scheme. Therefore, the method fails to construct a Bengali word if it was not present in the English-to-Bengali word dataset. The second research work claimed that speech recognition systems might have an adverse effect depending on the gender of the speaker (Hassan et al. 2012). The research work was conducted using a self-made speech corpus and introduced an MFCC and HMM-based ASR architecture. It finally concluded that the ASR system performs better if both male and female speeches are present in the training samples.

A continuous speech to word pattern matching method was introduced in 2013, which was implemented using MFCC, Linear Predictive Coding (LPC), Gaussian Mixture Models, and Dynamic Time Wrapping (DTW) (Ali et al. 2013). The authors implemented four different models each with different feature extraction and pattern matching scheme: a) MFCC + DTW, b) LPC + DTW, c) MFCC + GMM, d) MFCC + DTW. Among the four different setups, MFCC+GMM performed best by achieving 84% accuracy. However, the research work was conducted on a self-made dataset, and no comparison is performed. Further, due to the speech to word matching policy, the method may fail to recognize unknown meaningful and meaningless words.

The usage of DNN was first observed in 2017 that was a phoneme classification architecture (Bhowmik et al. 2017). The authors compared DNN and HMM architectures and proved DNN to be the most accurate. The DNN implementation contained stacked denoising autoencoders that took MFCC as input, which is pre-trained. Further, after pre-training the autoencoders, a multi-layer perceptron of three layers has been used to predict the phoneme probabilities. The baseline achieved 82.5% phoneme classification accuracy in a self-made dataset, which is unavailable. The authors also introduced a similar approach for classifying the place of speech sound articulation using DNN and AutoEncoders (Bhowmik et al. 2018).

A renowned Bengali search engine Pipilika (SUST(Accessed April 1, 2020)) developed a Bengali ASR system that used a larger vocabulary and performed better than previous DNN based methods (Saurav et al. 2018). Hybrid models combining DNN-HMM and GMM-HMM were also introduced and proved to perform better than previously applied architectures (Al Amin et al. 2019). The DNN-GMM model firstly performed GMM, and the outputs of the GMM states were passed to DNN fully connected layers. An error pattern analysis of HMM is also analyzed for Bengali speech (Aura et al. 2020). Similar efforts are given in speech to word ASR (Ali et al. 2013; Sumon et al. 2018; Sumit et al. 2018), phoneme-based ASR (Chowdhury and Khan 2019), spoken digit recognition systems (Hossain et al. 2013; Ahmed et al. 2015; Nahid et al. 2016, 2017; Sharmin et al. 2020), and word segmentation system (Bhowmik and Mandal 2019).

Table 3 gives a detailed insight into the various architectures that are implemented in the scope of the Bengali Speech Recognition system. Although the paper focuses on the speech-to-text procedures, all methods that only operate in speech recognition (does not perform text translation) are evaluated. From the presented data, it can be observed that most methods are implemented using self-made datasets, which in most cases are inadequate in size. Therefore, the results presented in most works remain incompetent in performance standards. Further, Fig. 6 illustrates a taxonomy of the implemented system toward
A study on the challenges and opportunities of speech recognition...

| Author               | Domain                  | Matching scheme | Features  | Recognition method        | Dataset              | Accuracy (%) |
|----------------------|-------------------------|-----------------|-----------|---------------------------|----------------------|--------------|
| Hossain et al. (2013)| Digit Recognition       | Word            | MFCC      | Neural Network            | Self-made            | 92           |
| Al Amin et al. (2019)| Speech to Text          | Word            | MFCC      | DNN                       | Mandal et al. (2011) | 99.08        |
| Nahid et al. (2016)  | Digit Recognition       | Word            | MFCC      | Sphinx-4 (Walker et al. 2004) | Nahid et al. (2018)  | 85           |
| Kotwal et al. (2010) | Speech to Text          | Phoneme         | MFCC      | DNN & HMM                 | Self-made            | 54.7         |
| Ahmed et al. (2015)  | Digit Recognition       | Word            | MFCC      | Deep Belief Network (Hinton 2009) | Self-made            | 94           |
| Sumon et al. (2018)  | Speech Command Recogni- | Word            | MFCC      | CNN                       | Self-made            | 74           |
|                     | tion                   |                 |           |                           |                      |              |
| Nahid et al. (2017)  | Digit Recognition       | Word            | MFCC      | LSTM                      | Nahid et al. (2018)  | 86.8         |
| Sharmin et al. (2020)| Digit Recognition       | Word            | MFCC      | CNN                       | Self-made            | 98           |
| Bhowmik et al. (2017)| Speech to Phoneme       | Phoneme         | MFCC      | AutoEncoder               | Self-made            | 82.5         |
| Zinnat et al. (2014)| Speech to Text          | Word            | MFCC & Local Features | HMM                      | Self-made            | 93.7         |
| Chowdhury and Khan (2019) | Speech to Text | Word | Spectral Analysis | Feed-Forward Network | Self-made            | 60           |
| Ali et al. (2013)    | Speech to Text          | Word            | MFCC & LPC | HMM & DTW                | Self-made            | 84           |
| Hassan et al. (2012) | Speech to Text          | Phoneme         | MFCC      | HMM                       | Self-made            | 88.6         |
| Sumit et al. (2018)  | Speech to Text          | Word            | Raw wave  | End-to-End Recurrent Network | Gales et al. (2014); Alam et al. (2010) | 59.7         |
| Mandal et al. (2010) | Speech to Text          | Phoneme         | MFCC      | Sphinx-3 (Placeway et al. 1997) | Self-made            | 87           |
| Khan et al. (2018)   | Speech to Text          | Phoneme         | MFCC, LDA (Bal-akrishnama and Ganapathiraju 1998), & MLLT (Gales 1998) | Kaldi (Povey et al. 2011) | Self-made            | 94.6         |
| Saurav et al. (2018) | Speech to Text          | Phoneme         | MFCC, LDA, MLLT | GMM, DNN, HMM | Self-made            | 96.04        |
Bengali ASR. Among the various methods, most of them implement phoneme-level recognition, which is the smallest and simplistic possible speech recognition level.

In Bengali ASR, no research works have performed an in-depth analysis of language-dependent challenges of ASR systems. Therefore, research works introduce their strength by showing better Word Error Rate (WER) and solving language-independent tasks. Table 4 shows the generated report based on our analysis of solving

![Fig. 6](image-url) The figure illustrates a taxonomy of the papers in the domain of Bengali ASR. The taxonomy separates existing strategies firstly, based on the vocabulary limit and secondly, based on recognition level

| Method          | Authors            | Method | Authors          |
|-----------------|--------------------|--------|------------------|
| Neural Network  | Hussain et al. (2013) | DNN    | Al Amin et al. (2019) |
| Sphinx-4        | Nahid et al. (2016) | HMM    | Zinnat et al. (2014) |
| Deep Belief Network | Ahmed et al. (2015) | Feed Forward Network | Chowdhury and Khan (2019) |
| LSTM            | Nahid et al. (2017) | LFC, HMM, DTW | Ali et al. (2013) |
| CNN             | Sharmin et al. (2020) | End-to-End | Sumit et al. (2018) |
| CNN             | Sumon et al. (2018) | CNN    | Sharmin et al. (2020) |

Table 4 The analysis report is conducted based on the language-independent challenges. The tick mark (✓) ensures the specific challenge is solved, the cross mark (✗) defines that the specific challenge is unsolved, and the unreported information is marked as null

| Method          | Noise reduction | Speaker independence | Speech variability | Speech segmentation | Recording             |
|-----------------|-----------------|----------------------|-------------------|---------------------|----------------------|
| Ahmed et al. (2015) | ✗               | ✓                    | ✓                  | ✓                   | 8192 Hz              |
| Nahid et al. (2016) | ✗               | ✓                    | ✓                  | ✓                   | Null                 |
| Bhowmik et al. (2017) | ✓               | ✓                    | ✓                  | ✓                   | 16,000 Hz            |
| Nahid et al. (2017) | ✓               | ✓                    | ✓                  | ✓                   | Null                 |
| Saurav et al. (2018) | ✗               | ✓                    | ✓                  | ✓                   | Null                 |
| Sumit et al. (2018)   | ✓               | ✓                    | ✓                  | ✓                   | 16,000 Hz, single channel |
| Sumon et al. (2018)   | ✗               | ✓                    | ✓                  | ✗                   | Null                 |
| Chowdhury and Khan (2019) | ✗               | ✓                    | ✗                  | ✓                   | Null                 |
| Al Amin et al. (2019) | ✗               | ✓                    | ✓                  | ✓                   | 16,000 Hz, 16 bit, mono channel |
| Sharmin et al. (2020)  | ✗               | ✓                    | ✓                  | ✗                   | Null                 |
language-independent challenges. However, the investigation lacks all papers described in Table 3 as they do not disclose any parameters properly in the reports. The investigation shows that Bhowmik et al. (2017) and Sumit et al. (2018) have solved all the language-independent challenges. However, Bhowmik et al. (2017) only performed phoneme classification, and Sumit et al. (2018) concluded that the implemented architecture gave unsatisfactory WER. Hence, this study proves that the Bengali ASR systems have not converged to the acceptance level.

In the next section, the explicit challenges that the Bengali ASR system casts are addressed. To investigate the challenges of the Bengali ASR system, we assume the system contains the following properties:

- The end-to-end system must take speech as input, and the output must be in Bengali text.
- The Bengali ASR system will process continuous speech. The system will continuously get user voice input and segment the speech from the voice input and perform recognition.
- Speech may represent meaningful or meaningless words as human names are often out of the scope of Bengali vocabulary. However, word-matching ASR systems would fail to recognize meaningless words.

## 4 Challenges of speech recognition for Bengali

The difficulties of speech recognition can be split into two sections, a) language-dependent challenges and b) language-independent challenges. The principal processing architectures must be designed considering these challenges, and resolving these issues will cause a Bengali ASR system to perform better. The language-independent challenges are, a) noise, b) speaker dependency, c) speech variability, d) speech segmentation, and e) recording device. Contrastively, the language-dependent challenges are, a) structural properties, b) consonant conjuncts, c) diacritics, d) word database, e) dialects, f) silent letters, g) word utterance similarity, h) letter utterance similarity. Figure 7 illustrates the dependency of the hurdles.

The researchers adequately perceive the language-independent speech-recognition challenges, and there exist state-of-the-art methods to suppress the difficulties. Also, efforts have been made to demonstrate the language-independent difficulties of speech recognition techniques (Vadwala et al. 2017; Sahu et al. 2018) and feature extraction procedures (Nivetha 2020; Singh et al. 2019). Therefore, in the following subsections, we manifest the language-dependent difficulties that are overlooked concerning a Bengali speech recognition system and report some possible solutions. Nevertheless, we shortly define the language-independent challenges in Table 5.

### 4.1 Structural properties

Every language has its structural properties that differ from language to language. Structural properties define the construction criteria of a meaningful sentence, which is set by grammar. However, languages composed of the same states or continents hold similar grammatical structures, linguistic patterns, and writing patterns. In this case, the Bengali
language has a significant share of relation in grammatical structure to the Hindi language. To reveal the structural properties of Bengali language, some of the fundamental structural differences between English and Bengali sentence are reported as follows,

Table 5 A summary of the language-independent challenges

| Challenge                  | Description                                                                 |
|----------------------------|-----------------------------------------------------------------------------|
| Noise                      | The environmental sound mixed with speech. Noise distorts the speech features and may cause incorrect word outputs. Therefore, noise reduction/elimination is an important factor in preprocessing |
| Speaker Dependency         | Speaker dependency targets speakers for an ASR system. If an ASR system is designed for a particular individual, it is considered as a speaker-dependent ASR system, otherwise a speaker-independent ASR system. Modern ASR systems are speaker-independent, and therefore, they are trained with speeches of different individuals |
| Speech Variability         | Speech variability describes the change of utterance depending on human emotion, environmental, and age factors. Proper ASR architectures trained with variable speech datasets can overcome this challenge |
| Recording Device           | The recording device fixes the audio type used for the ASR system. The input audio can be a single channel (mono), dual-channel, stereo, or even intensity stereo. Every input type has its advantages and disadvantages depending on circumstances which is also a challenge |
| Speech Segmentation        | Speech segmentation can be classified into two types: a) word segmentation, b) phoneme segmentation. Word and phoneme segmentation are required for continuous speech recognition. Error in segmentation causes misleads in speech pattern matching. However, some present end-to-end ASR systems do not require speech segmentation (Hsu et al. 2020) |
A study on the challenges and opportunities of speech recognition…

4.2 Consonant conjuncts

Consonant conjuncts are characters that hold two or more joined consonants represented as a single character. In the Bengali language, 118 consonant conjuncts are mostly used. Consonant conjuncts have been derived from the ancient Brahmi script, and it is also being used in many other scripts (Tuṅga and S. Śekhara 1995). Figure 8 derives the difference between the utterance of consonant and consonant conjuncts. The utterance of consonants contains two portions, a consonant utterance followed by a vowel utterance. On the contrary, a consonant conjunct contains three portions, a consonant utterance followed by another consonant utterance, and finally, a vowel utterance. Consonant conjuncts may cause great difficulty to the phoneme-based speech recognition architecture. The precision of recognizing correct phoneme patterns must be ensured to recognize consonant conjuncts from the extracted speech features correctly.
4.3 Diacritics

In English, diacritics are practiced to express the correct accent of a word. Whereas, in the Bengali language, diacritics are greatly utilized to express words. The Bengali letters may contain at most two types of diacritics, vowel diacritics and consonant diacritics. The main difference between diacritics and consonant conjuncts is that diacritics are mostly considered an extension of a particular letter. On the contrary, consonant conjuncts are often considered as a single letter, and they can contain a diacritic as well. The diacritics are limited to 11 vowels and 7 consonants. In contrast, consonant conjuncts can be constructed with any pair of consonants. However, the second consonant can not be used as diacritics. The usage of diacritics introduces obstructions in phoneme matching and phoneme transformation processes.

4.4 Word database

A rich word database is one of the vastest language-dependent challenges of an ASR system. Word database is mainly required for an ASR system that uses speech-to-word identification. However, phoneme-based ASR systems are also trained using a word database, but they mostly learn to classify phonemes. The Bengali language has a complex structure of words due to the diacritics and consonant conjuncts. Figure 10 illustrates an example of the construction of Bengali words. Also, Fig. 11 explains the construction of a grapheme,
which is considered to be the smallest unit in a word writing system. A grapheme root can be obtained by excluding the diacritics from a grapheme. Diacritics and consonant conjunctions are the most critical challenge for a phoneme-based ASR system. Therefore, speech-to-word identification systems may be considered as a better choice. Nevertheless, due to the centuries of contact with the Europeans, Persians, Arabians, and Mughals, the Bengali vocabulary has a larger subset of adopted words. A linguistic difference is also considered in the Bengali and west Bengali continent. Therefore, generating a reliable speech-to-word database is also a significant challenge. A more extensive word database increases the probability of pattern mismatch. An incomplete word database will cause database-excluded words to be faultily recognized, mostly in word pattern matching ASR systems.

4.5 Dialects

Dialects refer to the linguistic variances that may differ in accent, vocabulary, spelling, and grammar of a language. Dialects are observed in almost every widely spoken language of the globe. According to the phonology and pronunciation of different dialects, the dialects of the Bengali language can be divided into six classes (Parishad 2001), a) Bengali, b) Rarhi, c) Varendri, d) Manbhumi, e) Rangpuri, and f) Sundarbani. Also, Bengali has more than 33 regional dialects. The dialects introduce more phoneme patterns and more words in vocabularies in an ASR system. These dialects should also be considered to implement a flawless Bengali ASR system.

4.6 Silent letters

Silent letters are frequently observed in most languages. Usually, a word containing letters that are not uttered is referred to as silent letters. Example: pneumonia (p silent) and ghost (h silent). Silent letters also occur in the Bengali language such as ব্যাঙ্ক (b silent), থাকাও (t silent). Speech to phoneme matching ASR system fails to recognize the silent letters. In this case, using a pre-defined lexicon rule can be used to auto-correct the words containing silent letters.

4.7 Word utterance similarity

In the Bengali language, some words have similar utterances but have different grapheme construction. As an illustration, the word pair অন্য (else), and অন্ন (food) has similar utterance, although their meaning is different. In such circumstances, humans mostly relate the correct word by the concept of the sentence and some intuition. This problem can essentially be solved using n-gram or recurrent neural network models over previous predictions. Figure 12 contains an illustration explaining the above scenario. From the

Fig. 12 An example of guessing the correct word from a set of similar utterance words

তোমাকে না, আমি __ __ কাওকে ভালোবাসি

It is not you, I love someone __ __.

Similar Utterance in Bengali: অন্য (else), অন্ন (food)
Correct Word in the Blank: অন্য (else)
example, from a set of similar uttered words, we humans pick the correct word by relating each word with the sentence. This explains the requirement of n-grams and recurrent networks over previously predicted words.

### 4.8 Letter utterance similarity

Some Bengali letters also contain mostly similar utterances. Tables 6 and 7 contain a list of vowels and consonants, their phone, Bengali word examples (written in English), and the meanings, respectively. Based on the example of Tables 6 and 7, it can be observed that Bengali language contains some phonetically similar word clusters ({{উ, এল}}, {জ, ঘ}, {ন, ধ}, {র, স, হ}, and {হ, থ}). Also, humans often tend to fail to guess the correct letters from these clusters applied in particular words. For example, a vowel utterance “u” can be constructed using two different letters “ঘ + থ” + “ঘ”. However, for a particular Bengali word, “chul” the correct word construction is “চল” (implicated in Table 7, row 4). Further similar variation is observed for consonant letters as well. An example can be drawn for the letter cluster “চ”. The consonant utterance “n” can be constructed using either “চ” or “চ”.

| Phoneme | Letters | Bengali Word | Meaning |
|---------|---------|--------------|---------|
| k       | ক       | kapor        | cloth   |
| kʰ      | খ       | khabar       | food    |
| g        | গ       | golap        | rose    |
| gʰ      | ঘ       | ghor         | home    |
| ɦ        | হ       | bang         | frog    |
| ɬ        | ছ       | chaka        | tire    |
| ʃ        | চ       | char         | offer   |
| j        | জ, ঝ    | jal          | mesh    |
| ɻ        | ল       | jhal         | hot taste |
| t        | ট       | taka         | money   |
| tʰ      | ঠ       | thela        | push    |
| d        | ড       | dal          | branch  |
| dʰ      | ঢ       | dhaka        | covered |
| f        | ফ       | tala         | lock    |
| fʰ      | ফ       | thana        | police station |
| d        | দ       | dalan        | building |
| dʰ      | ধ       | dhakka       | push    |
| n        | ন, ণ    | nam          | name    |
| p        | প       | poka         | insect  |
| pʰ      | পু       | phol         | fruit   |
| b        | ব       | boka         | fool    |
| bʰ      | ভ       | bharo        | fare    |
| m        | ম       | mash         | month   |
| ʃ, ʃʰ  | শ, স, হ   | shaban       | soap    |
| r        | র       | rod          | sun ray |
| l        | ল       | lathi        | stick   |
| h        | ঙ       | hashi        | smile   |
A study on the challenges and opportunities of speech recognition

However, a specific word “nam” has a fixed word construction “নাম” (shown in Table 6, row 18). The problem can be resolved either by applying a robust dataset that can give the pattern matcher a proper intuition or hard-implementing the Bengali grammatical rules (ঔ, ও, Bangla Academy laws) (Chatterji 1988; Bangladesh 1995).

| Phoneme | Letters | Bengali Word | Meaning |
|---------|---------|--------------|---------|
| \(a\) | অ | olopo | less |
| \(ā\) | আ | aamar | my |
| \(i\) | ই, ঈ | itihash | history |
| \(u\) | উ, ঊ | chul | hair |
| \(e\) | এ | ke | who |
| \(o\) | ও | golap | rose |
| \(ou\) | ঔ | koushol | strategy |

Table 7 The table illustrates phones of vowels in the Bengali language, along with letters, word examples (written in English), and the corresponding meanings. In the example, the words {ঔ, ঔ} and {ঔ, ঔ} contain a similar phone structure “ন + আ + স”. However, a specific word “nam” has a fixed word construction “নাম” (shown in Table 6, row 18). The problem can be resolved either by applying a robust dataset that can give the pattern matcher a proper intuition or hard-implementing the Bengali grammatical rules (ঔ, ও, Bangla Academy laws) (Chatterji 1988; Bangladesh 1995).

5 Future research scope on bengali ASR

This section summarizes the key challenges of a Bengali ASR system required to further extend the existing schemes’ performance. Moreover, we propose an architecture that may solve the challenges. From the overall discussion of Sect. 4, three essential language-dependent challenges can be summarized:

- **Grammatical and literal dependency of words**: The grammatical dependency of words causes filter-out words depending on the grammatical structure of the previous words. Furthermore, literal dependency may help to obtain the proper word from a set of similar words containing a similar utterance pattern. Therefore, the search space for the proper word can be reduced. However, a powerful memory-based architecture is required to extract grammatical and literal dependencies properly. An attempt to implement this scheme may result in solving the challenges discussed in Sects. 4.1, 4.4, 4.5, and 4.7.

- **Grammatical and preceding dependency of characters**: The grammatical and preceding dependency deals with exploring the correct graphemes, vowel diacritics, and consonant diacritics of a word. Every language has grammatical patterns that correctly guess the proper graphemes from a grapheme set of a similar utterance. The extraction of these patterns also requires a memory-based generator. An attempt to implement this scheme may result in solving the challenges discussed in Sects. 4.2, 4.3, and 4.8.

- **Dissimilar uttered words due to silent letters**: Through the discussion in Sect. 4.6, it can be observed that dissimilarity of utterance and text mainly occurs due to silent letters. In the scope of the Bengali language, silent letters mostly do not contain grammatical dependencies. Therefore, a direct word-to-text transition may result in solving the difficulty.
The present researches in Bengali ASR systems often evade the interrelation of the grammatical issues and correct word predictions. Therefore the problems mentioned above are the future research scope of the Bengali ASR. Furthermore, we contribute to the future scope of the Bengali ASR by proposing a theoretical architecture. In Fig. 13, we introduce an architecture that we believe to be optimal based on our research endeavor. To the best of our knowledge, the suggested architecture has not been investigated or implemented in any research endeavors. Also, the proposed architecture includes recurrent hybrid architecture that can create a new architectural perspective in the current research field. Hence, we point out the properties of the suggested ASR system as follows.

- The grammatical dependency of words mostly serves to find optimal literary words by generating some fixed rules. Short-term memory can be combined to correlate these rules. Using the short term memory, the system can optimally learn the grammatical relation only if trained on a large speech corpus.
- The grammatical and preceding dependency of characters can also be determined by combining a short-term memory with a speech character generator. The popular systems (Dong et al. 2018) depend on short-term memory to explore the dependency of character-level prediction.
- Every language, including the Bengali, contains words with irregular letter sequences. This problem can be solved by memorizing some fixed words. Therefore, it is optimal to implement both speech-to-word matching and phoneme-to-word matching. The current architectures implement end-to-end schemes (Baevski et al. 2020; Ravanelli et al. 2018) that only generates characters and receives information from the previous characters only. Therefore it is usual to overlook most of the irregular word representations.
- The current implementations (Dong et al. 2018) only emphasize character recognition schemes. However, a hybrid implementation of a word and character matching scheme can solve the problem of generating irregular words and non-dictionary words. Therefore, our suggested system may search for optimal word matching. Further, the model may extract characters from the speech if the optimal word match is not found.

Fig. 13 The suggested architecture of an optimal Bengali ASR system. Combining short-term memories will enable the architecture to recognize both words and characters’ grammatical and literary dependency. Word matching schemes may help to recognize words that contain dissimilar and silent letters. Confidence evaluation defines if the model is confident that a speech exists in the present speech to word dictionary. Otherwise, the model extracts recurrent characters based on the speech.
The suggested architecture pattern may solve the overall challenges discussed in the paper only if it is trained with speech corpora with a proper variation of speech and grammar variability.

6 Conclusion

In this survey, we begin with the investigation of the current research endeavors conducted in the Bengali ASR system, including speech corpora and recognition methods. Then, we have examined several difficulties that prevail in the domain of the Bengali ASR system. We have explained the structural and linguistic dissimilarities between languages on which an ASR system researcher should concentrate. We have rigorously presented grammatical fundamentals and suggestions on solving challenges. Although the examined challenges are also witnessed for most other languages, we have explained the challenges and opportunities regarding the Bengali language in particular. We have also investigated most of the latest works that implemented Bengali ASR systems, and through onerous exploration, we have shown that they lack perfection. We strongly believe that our gentle excavation on this very topic may expand the research scope of the Bengali as well as universal ASR systems and guide researchers scrupulously to target the exact challenges to be resolved.

Acknowledgements We would like to thank the Advanced Machine Learning (AML) lab and Bangladesh University of Business and Technology (BUBT) for their support.

References

Ahmed M, Shill PC, Islam K, Mollah MAS, Akhand M (2015) Acoustic modeling using deep belief network for bangla speech recognition. In: 2015 18th international conference on computer and information technology (ICCIT), pp 306–311. IEEE
Ahmed S, Sadeq N, Shubha SS, Islam MN, Adnan MA, Islam MZ (2020) Preparation of bangla speech corpus from publicly available audio & text. In: Proceedings of The 12th language resources and evaluation conference, pp 6586–6592
Al Amin MA, Islam MT, Kibria S, Rahman MS (2019) Continuous bengali speech recognition based on deep neural network. In: 2019 international conference on electrical, computer and communication engineering (ECCE), pp 1–6. IEEE
Alam F (2018) Development of annotated bangla speech corpora. https://data.mendeley.com/datasets/c79z6gz9rm/1
Alam F, Habib S, Sultana DA, Khan M (2010) Development of annotated bangla speech corpora
Ali MA, Hussain M, Bhuiyan MN (2013) Automatic speech recognition technique for bangla words. Int J Adv Sci Technol 50:51–60
Arslan RS, Barişçi N (2020) A detailed survey of turkish automatic speech recognition. Turk J Electr Eng Comput Sci 28(6):3253–3269
Audhkhasi K, Kingsbury B, Ramabhadrar B, Saon G, Picheny M (2018) Building competitive direct acoustics-to-word models for english conversational speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4759–4763. IEEE
Aura SR, Rahimi MJ, Baroi OL (2020) Analysis of the error pattern of hmm based bangla asr. Int J Image Graph Signal Process 12(1):1
Baevski A, Zhou H, Mohamed A, Auli M (2020) wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems. Curran Associates, Inc., vol 134, pp 12449–12460
Balakrishnan S, Ganapathiraju A (1998) Linear discriminant analysis-a brief tutorial. Inst Signal Inf Process 18:1–8
Benzeghiba M, De Mori R, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C et al. (2007) Automatic speech recognition and speech variability: a review. Speech Commun 49(10–11):763–786

Besacier L, Barnard E, Karpov A, Schultz T. (2014) Automatic speech recognition for under-resourced languages: a survey. Speech Commun 56:85–100

Bhowmik T, Mandal SKD. (2019) Prosodic word boundary detection from bengali continuous speech. Lang Resour Eval pp 1–19

Bhowmik T, Choudhury A, Mandal SKD. (2017) Deep neural network based recognition and classification of bengali phonemes: a case study of bengali unconstrained speech. In: International conference on next generation computing technologies, pp 750–760. Springer

Bhowmik T, Chowdhury A, Mandal SKD. (2018) Deep neural network based recognition and classification of bengali phonemes: a case study of bengali unconstrained speech. In: Proceedings of the genetic and evolutionary computation conference companion, pp 362–363

Bourlard HA, Morgan N. (2012) Connectionist speech recognition: a hybrid approach, volume 247. Springer Science & Business Media

Chowdhury MSA, Khan MF. (2019) Linear predictor coefficient, power spectral analysis and two-layer feed forward network for bangla speech recognition. In: 2019 IEEE international conference on signal processing, pp 1–6. IEEE

Chung J, Gulcehre C, Cho K, Bengio Y. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555

Das B, Mandal S, Mitra P. (2011) Bengali speech corpus for continuous automatic speech recognition system. In: 2011 international conference on speech database and assessments (Oriental COCOSDA), pp 51–55. IEEE

Das B, Mandal S, Mitra P. (2021) SHRUTI bengali continuous ASR speech corpus. https://cse.iitkgp.ac.in/~pabitra/shruti_corpus.html

Dave N. (2013) Feature extraction methods lpc, plp and mfcc in speech recognition. Int J Adv Res Eng Technol 1(6):1–4

Dekhak N, Kenny PJ, Dekhak R, Dumouchel P, Ouellet P. (2010) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798

Dong L, Xu S, Xu B. (2018) Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP), pp 5884–5888. IEEE

Gaikwad SK, Gawali BW, Yannawar P. (2010) A review on speech recognition technique. Int J Compt Appl 10(3):16–24

Gales M, Young S, et al. (2008) The application of hidden markov models in speech recognition. Found Trends® Signal Process 1(3):195–304

Gales MJ. (1998) Maximum likelihood linear transformations for hmm-based speech recognition. Comput Speech Lang 12(2):75–98

Gales MJ, Knill KM, Ragni A, Rath SP. (2014) Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In: Fourth international workshop on spoken language technologies for under-resourced languages (SLTU-2014), pp 16–23. International Speech Communication Association (ISCA)

Gales MJ, Knill KM, Ragni A, Rath SP. (2021) IARPA babel bengali language pack. https://catalog.ldc.upenn.edu/LDC2016S08

Google. Large Bengali ASR training data set. http://www.openslr.org/53/

Graves A, Jaitly N. (2014) Towards end-to-end speech recognition with recurrent neural networks. In: International conference on machine learning, pp 1764–1772

Graves A, Jaitly N, Mohamed A-R. (2013) Hybrid speech recognition with deep bidirectional lstm. In: 2013 IEEE workshop on automatic speech recognition and understanding, pp 273–278. IEEE

Haeb-Umbach R, Ney H. (1992) Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proceedings of ICASSP, volume 1, pp 13–16. USA: ICASSP
Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, et al. (2014) Deep speech: scaling up end-to-end speech recognition. arXiv:1412.5567

Haque MA, Verma A, Alex JSR, Venkatesan N (2020) Experimental evaluation of cnn architecture for speech recognition. In: First international conference on sustainable technologies for computational intelligence, pp 507–514. Springer

Hassan F, Khan MSA, Kotwal MRA, Huda MN (2012) Gender independent bangla automatic speech recognition. In: 2012 international conference on informatics, electronics & vision (ICIEV), pp 144–148. IEEE

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

Hermansky H, Fousek P (2005) Multi-resolution rasta filtering for tandem-based asr. Technical report, IDIAP

Hermansky H, Morgan N (1994) Rasta processing of speech. IEEE Trans Speech Audio Process 2(4):578–589

Hinton GE (2009) Deep belief networks. Scholarpedia 4(5):5947

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hossain M, Rahman M, Prodhan UK, Khan M, et al. (2013) Implementation of back-propagation neural network for isolated bangla speech recognition. arXiv:1308.3785

Houque A (2006) Bengali segmented speech recognition system. Undergraduate thesis, BRAC University, Bangladesh

Hsu J-Y, Chen Y-J, Lee H-Y (2020) Meta learning for end-to-end low-resource speech recognition. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7844–7848. IEEE

Irie K, Tüské Z, Alkhouli T, Schütter R, Ney H (2016) Lstm, gru, highway and a bit of attention: an empirical overview for language modeling in speech recognition. In: Interspeech, pp 3519–3523

Islam MR, Sohail ASM, Sadid MWH, Mottalib M (2005) Bangla speech recognition using three layer back-propagation neural network. In: Proceedings of the national conference on computer processing of Bangla (NCCPB), Dhaka

Ittichaichareon C, Suksri S, Yingthawornsuk T (2012) Speech recognition using mfcc. In: International conference on computer graphics, simulation and modeling (ICGSM’2012), pp 28–29

Karim R, Rahman MS, Iqbal MZ (2002) Recognition of spoken letters in bangla. In: Proceedings of 5th international conference on computer and information technology (ICCIT02)

Khan MF, Debnath DRC (2002) Comparative study of feature extraction methods for bangla phoneme recognition. In: 5th ICCIT, pp 27–28

Khan S, Pal M, Basu J, Bepari MS, Roy R (2018) Assessing performance of bangli speech recognizers under real world conditions using gmm-hmm and dnn based methods. In: SLTU, pp 192–196

Kotwal MRA, Banik M, Eity QN, Huda MN, Muhammad G, Alotaibi YA (2010) Bangla phoneme recognition for asr using multilayer neural network. In: 2010 13th international conference on computer and information technology (ICCIT), pp 103–107. IEEE

Kwon O-W, Lee T-W (2004) Phoneme recognition using ica-based feature extraction and transformation. Signal Process 84(6):1005–1019

Lee K-F, Hon H-W, Reddy R (1990) An overview of the sphinx speech recognition system. IEEE Trans Acoust Speech Signal Process 38(1):35–45

Mandal S, Das B, Mitra P (2010) Shruti-ii: a vernacular speech recognition system in bangli and an application for visually impaired community. In: 2010 IEEE students technology symposium (TechSym), pp 229–233. IEEE

Mandal S, Das B, Mitra P, Basu A (2011) Developing bangali speech corpus for phone recognizer using optimum text selection technique. In: 2011 international conference on asian language processing, pp 268–271. IEEE

Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: a review. Lang Cognit Process 27(7–8):953–978

Molla K, Hirose K (2004) On the effectiveness of mfccs and their statistical distribution properties in speaker identification. In: 2004 IEEE symposium on virtual environments, human-computer interfaces and measurement systems, 2004.(VCIMS)., pp 136–141. IEEE

Nahid MMH (2018) Bengali speech recognition—bangla real number audio dataset. https://data.mendeley.com/datasets/t33byr6cpt/6

Nahid MMH, Islam MA, Islam MS (2016) A noble approach for recognizing bangla real number automatically using cmu sphinx4. In: 2016 5th international conference on informatics, electronics and vision (ICIEV), pp 844–849. IEEE
SUST SUoST (2020) Pipilika: (Bengali Search Engine). Accessed April 1, 2020. https://www.pipilika.com/
Takiguchi T, Ariki Y (2007) PCA-based speech enhancement for distorted speech recognition. J Multimed 2(5)
Tebelskis J (1995) Speech recognition using neural networks. PhD thesis, Carnegie Mellon University
Trentin E, Gori M (2001) A survey of hybrid ann/hmm models for automatic speech recognition. Neurocomputing 37(1–4):91–126
Tunga, Sekhara S (1995) Bengali and other related dialects of south Assam. Mittal Publications, 1 edition
Vadwala AY, Suthar KA, Karmakar YA, Pandya N, Patel B (2017) Survey paper on different speech recognition algorithm: challenges and techniques. Int J Comput Appl 175(1):31–36
Variani E, Lei X, McDermott E, Moreno IL, Gonzalez-Dominguez J (2014) Deep neural networks for small footprint text-dependent speaker verification. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4052–4056. IEEE
Walker W, Lamere P, Kwok P, Raj B, Singh R, Gouvea E, Wolf P, Woelfel J (2004) Sphinx-4: a flexible open source framework for speech recognition
Westphal M (1997) The use of cepstral means in conversational speech recognition. In: Fifth European conference on speech communication and technology
Zheng F, Zhang G, Song Z (2001) Comparison of different implementations of mfcc. J Comput Sci Technol 16(6):582–589
Zinnat SB, Siddique RMA, Hossain MI, Abdullah DM, Huda MN (2014) Automatic word recognition for bangla spoken language. In: 2014 international conference on signal propagation and computer technology (ICSPCT 2014), pp 470–475. IEEE
Ziolkoko M, Samborski R, Galka J, Ziolkko B (2011) Wavelet-Fourier analysis for speaker recognition. In: 17th national conference on applications of mathematics in biology and medicine, vol 134, p 129
Zoughi T, Homayounpour MM, Deypir M (2020) Adaptive windows multiple deep residual networks for speech recognition. Expert Syst Appl 139:112840

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.