Khulna University Studies Special Issue ICSTEM4IR: 566-583, 2022

ORIGINAL ARTICLE

SCIENCE & ENGINEERING

AUTOMATED SPEECH-TO-TEXT CONVERSION SYSTEMS IN BANGLA LANGUAGE: A SYSTEMATIC LITERATURE REVIEW

Aysha Akther* and Rameswar Debnath

Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh

KUS: ICSTEM4IR-22/0107

Manuscript submitted: July 27, 2022 Accepted: September 27, 2022

Abstract

The 4th Industrial Revolution (4IR) is creating a new way of working and impacting all disciplines, industries, and economies. In future days, there will be needed seamless communication with machines and has to deal with an enormous amount of information. As speech is the most natural way of communication for humans, research in Natural Language Processing (NLP) is increasing with time. To make human-computer interaction effortless Speech-to-Text (STT) conversion is particularly important. A lot of research works have been carried out to allow machines to interact with humans naturally in many languages like English, Spanish, Japanese, etc. Bangla is the primary language of Bangladesh and West Bengal of India and is spoken by over 250 million people worldwide. Speech processing in Bangla language is still an open research field. This literature review studies the recent advancements in automated speech to text conversion in Bangla language. In this paper, we present a comprehensive comparative study on the state-of-the-art Bangla speech to text conversion systems in accordance with dataset size, feature extraction techniques, methodologies used, toolkits, and accuracies. Furthermore, challenges associated with Bangla speech processing research, applications of automatic speech to text conversion in different fields of Bangla language along possible future research indications are elaborated in this paper.

*Corresponding author: <aysha@cseku.ac.bd>

DOI: https://doi.org/10.53808/KUS.2022.ICSTEM4IR.0107-se

Akhter, A. et al. (2022). Automated speech-to-text conversion systems in Bangla language: A systematic literature review. *Khulna University Studies*, Special Issue (ICSTEM4IR): 566-583.

Keywords: Natural Language Processing, Speech-to-Text conversion, Bangla Speech Recognition, Human-computer interaction

Introduction

The 4th industrial revolution (4IR) is going to change the way of our living and the way of communication. It is going to impact tremendously fields such as artificial intelligence, internet of things, biotechnology, quantum computing along with many more fields. Billions of people are going to connect through smart mobile devices with unlimited storage, access to knowledge, and very fast processing speed. A totally new level of automation in both our daily life and industry level using Internet of Things (IoT) and human-machine and machine-machine communication is emerging at the door. Getting ourselves ready for this coming era of connectivity, speed and an enormous amount of information around us require a seamless connection between human and machines in any conventional natural language.

Speech is the most common and natural form of communication for human being and text or symbols is the most common form of transmission and processing of machines. To make human-machine interactions effortless Speech To Text (STT) conversion is particularly important. A lot of research work has been carried out to allow machines to interact with human in a natural way involving different methodologies for English, Spanish, French, German, Japanese, and some other languages (Rudnicky et al., n.d.). The most promising applications of research in this domain include voice command recognition in car system, medical documentation in healthcare system, transcribe meetings, effortlessly translating and subtitling, voice-powered virtual assistant, in smartphones, in education, and voice command systems for people with disabilities (Sultana et al., 2012). These applications contribute to make life fast, easy, and effortless. Due to the diverse nature of spoken languages, methodology which is suitable for one language may not produce better result for other languages. Specific systematic analysis for language processing is required for each language. Bangla is the seventh most spoken language in the world with over 250 million native speakers (Wikipedia, 2022). Researchers have been working on Bangla speech and language processing for almost two decades. Though it is still an open field of research, some really promising advancements have been done in Bangla speech and language processing.

All modern descriptions of speech are to some extent probabilistic and speech to text conversion is an application of speech so it can not be 100% accurate. In a continuous audio stream or a sequence of states, phones can be defined as similar states. And because words are a combination of phones so theoretically words should be recognized certainly. But practically waveform of phones varies due to environmental factors and speaker-related issues. So certainty of recognizing words decreases. To increase the rate of performance of speech to text conversion systems deep learning and neural network based models are being tested in places of traditional SVM, HMM, DTW based models. In this article, we present a comprehensive comparative study of the state-of-the

art research on Bangla speech to text conversion systems, challenges associated with Bangla speech processing, and applications and future research indications. We hope this study will help Bangla speech and language researchers by showing the current state of advancements and the areas of research that need to be focused on. A brief description of recent advancements in Bangla speech to text conversion are discussed in the following paragraphs.

Hasnat et al., 2007 proposed an HMM-based Bangla isolated and continuous speech to text conversion system. They used an adaptive filter on wav speech signal for noise elimination. By subtracting the surrounding environment's signal from the speech signal noise is eliminated from the speech signal. For converting continuous speech, starting and endpoints are detected and for converting isolated speech pauses and silences are detected too. Then MFCC feature extraction technique was applied to noise eliminated data. For isolated speech recognition, no language model was required as only isolated words are recognized in this case, For continuous speech recognition, a language model was required to recognize a sequence of words. The Viterbi algorithm was used by the model for decoding purpose. The authors used the HTK toolkit for recognition purpose and achieved 70% accuracy for isolated data and 60% accuracy for converting continuous data into text.

Chowdhury et al., 2009 proposed an algorithm for word separation in real-time isolated and continuous speech processing. Sultana et al., 2012, proposed a Bangla Speech to text conversion system using Microsoft developed speech engine SAPI. They tested their system only for an article from a specific newspaper. The authors manually wrote every Bangla word in Banglish format that is With English characters in the XML grammar file. For identifying words with different pronunciations, they included all possible English word combinations of a particular Bangla word in the grammar file. During the testing phase, every word's Banglish form is mapped with words stored in an XML file for Banglish to Bangla word transform. They achieved 78% accuracy for converting the speech data of a paragraph to a transcription. How many speakers were involved in the experiment is not mentioned in the paper. Though as an early attempt at Bangla speech to text conversion, this paper showed a direction, their approach of manually generating the grammar file is not feasible for practical use. SAPI is relatively slower than other available speech engines.

Nasib et al., 2018, used CMU Sphinx4 as an acoustic model. In the research, the transcription was generated in Banglish format(Bengali written in English) first. The Banglish formatted result of the acoustic model was again replaced in Bengali. The acoustic model was trained with 503 unique words, spoken by 5 speakers. The total length of training data was 1.99 hours. The best accuracy achieved was 71.7%. The model lacks enough training also the authors didn't clearly state anything about what type of language model they used and the size of the text corpus for language modeling. The authors used a digital audio workstation, Audacity to provide the environment for speech to text conversion and for manipulation of speech data.

Tausif et al., 2018 applied the speech engine Deepspeech developed by Mozilla for acoustic modeling. Deepspeech is a Neural Network based acoustic modeling approach. Deepspeech per-

forms audio signal to character/phoneme level mapping. The paper dealt with several Bangla language-specific issues such as joint letters and implicit vowels and silent letters. To keep the search space small the authors used only one form of vowels instead of using vowel characters and their modified form(vowels as kar). But this process adds extra overhead to the system. The transcript generated by the acoustic model is in Broken Language Format(BLF), so the text corpus needs to be converted in BLF format too. The final transcript is generated by combining the output of the acoustic model and language model using the Prefix Beam Search algorithm. The model also needs another post-processing by converting the BLF formatted output back to the normal format. The post-processing is done using the Levenshtein distance. The speech dataset used in the study has 1400 audio clips of 170 sentences.

Syfullah et al., 2018 proposed a model for converting Bangla voice characters into equivalent text form. In the proposed model, initially for noise cancellation RLS was used and zero-crossing rate and short time energy were used for noise removal. Then after extracting MFCC features from voice samples the extracted features set is given as input to the back-propagation ANN to formulate the vector codebook. Then a combination of the K-means algorithm and Linde Buzo Gray (LBG) algorithm is used for optimal quantization of data to generate the codebook. Finally, for recognition of input voice, short Euclidian distance value was used to map with generated code-book entries. In this study, they worked on identifying 45 Bangla vowel and consonant detection and used 900 samples from 20 persons as the total dataset for training and testing. Their recognition accuracy rate was 81.61%.

Sumon et al., 2018 proposed Bangla short command word detection based on CNN architecture. They worked on 10 Bangla command words. The authors used the MFCC feature extraction technique and trained the CNN model with 100 samples for each command word. They achieved 74% accuracy for MFCC feature extracted CNN model and 71.44% accuracy for without MFCC and CNN model. For a pre-trained CNN model on English command words, authors achieved 73% accuracy.

Bristy et al., 2019 proposed an HMM-based Bangla speech to text conversion system. They used CMUSphinx 4 speech engine for acoustic modeling. The experimental dataset comprised of 102 words and also they used a language model generated from these 102 words' text corpus. there were 8 speakers for training the model and for unknown speakers achieved accuracy was 59.01% and for known speakers achieved accuracy was 78.57%. The dataset used for the experiment was not sufficient to train the model.

Al Amin et al., 2019 proposed a deep learning based acoustic modeling for Bangla speech to text conversion. The authors used the popular speech engine Kaldi for acoustic modeling. They emphasized on feature extraction for efficient speech to text conversion in the study. Using the MFCC feature extraction technique 143 dimensional features were extracted then LDA (Latent Discriminant Analysis) was applied to reduce dimensionality and de-correlation. Then to extract more precise features they applied MLLT over it and finally applied fMLLR for normalization of

speaker variability. In this paper, GMM-HMM and DNN-HMM based acoustic modeling were applied. The version of speech corpus SHRUTI used in the study contained 21.6 hours of speech data from 34 male and female speakers. DNN-HMM models achieved 0.92% WER and GMM-HMM based model achieved 2.02% WER.

Islam et al., 2019, proposed two models for Bangla STT conversion. One model uses CNN for Bangla STT conversion and the other model uses RNN as the deep learning technique and Connectionist Temporal Classification (CTC) for decoding. Firstly the WAV formatted speech signal was converted to Numpy array format for the sake of memory and processing efficiency then the MFCC feature extraction technique was used. Then for their first model, they trained the acoustic model with a CNN of five hidden layers. The other model consists of five neural network layers: three fully connected layers then a bidirectional RNN layer and then another fully connected layer. CTC was used to find the transcription from RNN model and finally, the sequence is calculated based on the probability of language models. The CNN model was trained on 100 words and 30 samples of each word. The same words but from different samples were used for testing the model. The average accuracy achieved by the CNN model was 86.058%. The RNN based model was trained with 30,000 words but the rate of accuracy achieved was not mentioned in the paper.

Mandal et al., 2020 proposed a combined CNN-RNN model for end-to-end Bangla speech recognition. They have checked the model for different combinations of CNN-RNN by changing the number of CNN layers and the number of RNN layers. The best result was achieved for the combination of 4 initial CNN layers followed by 5 GRU layers having 800 hidden units per GRU. The network was trained end-to-end using the CTC loss function. The model was trained on the largest available speech data on Bangla language of duration 214.6 hours and best achieved performance achieved was WER 13.67%. The convolutional layer also comprised of reduced a kernel size. The first CNN layer had kernel size of 7×3 and the subsequent layer had kernel size of 3×3.

Sharmin et al., 2020 proposed a CNN-based model for Bangla digits classification and achieved 98.37% accuracy. Their dataset consists of 1230 audio files collected from people of different age groups, different genders, and having different dialects. The authors used the MFCC feature extraction method then a combination of 2D-Convolutional layer and Batch Normalization layer was for feature learning and classification purpose.

Materials and Methods

Speech to text conversion is the process of a machine identifying words and phrases in spoken language and converting them to a textual format of that language. The textual conversion of speech signal can not be done accurately using solely the speech data, the conversion system must have prior knowledge about the language in question. The general architecture of speech to text conversion system and the step by step procedure of the conversion is illustrated in this section.

General architecture of speech to text transformation

STT conversion is the method of converting an acoustic signal to a set of words. For speech recognition experimentation one needs to record a voice sample and then convert the voice sample to wav format. Then the speech signal is analyzed to extract important features. Mel Frequency Cepstrum Coefficients(MFCC), Linear Predicted Coefficients (LPC), and Perceptual Linear Prediction (PLP) are some of the popular feature extraction techniques used for speech signal processing (Dave, 2013). A feature extraction method extracts the most important attributes of the speech signal. By using such techniques new spectrum-based parameters are obtained for each word. Each word is recognized by classifying attribute values with the pre-recorded datasets. So speech recognition is a multileveled process of pattern recognition that examines acoustic signals and results in a structured hierarchy of phonemes, words, and sentences (Dave, 2013). Intermediate levels may provide additional temporal constraints such as N-gram probabilities for checking accepted word sequences (Language Model) to compensate uncertainties. Statistical or word mapping based speech recognition systems are not suitable for our-of-vocabulary words. If the target word was not present in the training corpus then the system fails to recognize it. Character or phoneme mapping based systems overcome this problem by identifying each character of the speech signal from utterances. Deep Neural Network based architecture of speech conversion works by mapping audio frames of speech to characters. Character or phoneme mapping performance depends on the amount of training data given.

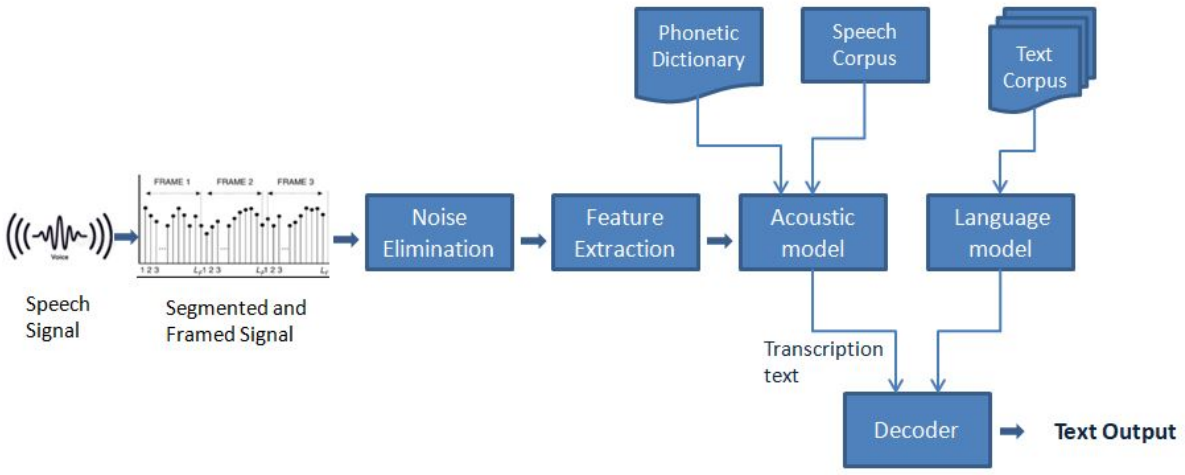

Figure 1: General Architecture of Speech to Text conversion system

The whole speech to text conversion system can be partitioned into several steps. The steps of speech to text conversion system is discussed below in a brief.

Preprocessing

In the preprocessing step, the recorded audio data in wav format needs to be sampled, trimmed(when necessary) and noise reduced or eliminated. Speech signals are located more or less below 8 kHz, so normally audio data needs to be sampled at 16KHZ for limiting the loss of information and fast processing. In some cases, the silence between words is trimmed in preprocessing step. Noise reduction is an important preprocessing step in real-time or practical speech to text conversion systems. In a practical environment performance of STT conversion systems largely depends on proper noise reduction. Many methods are proposed and applied by researchers for noise estimation and elimination. Hirsch and Ehrlicher, 1995, proposed two noise reduction methods that requires about 400 ms past noisy segments. The first method estimates noise by calculating the weighted sum of past spectral magnitude values for each subband. The second method is based on histograms of past spectrum values of each subband. Both methods work well with HMM-based speech recognition systems. In Sharma and Sardana, 2016, aplied bidirectional KALMAN filter to reduce noise from real-time speech data. A four-layered feed-forward back propagation neural network is proposed by Tamura and Waibel, 1988 for noise estimation and reduction from speech data.

Feature Extraction

To extract characteristic features from speech signals usually each frame of 16-32 ms is taken and updated every 8-16 ms (Dave, 2013). Feature extraction techniques usually applied to extract features from speech signals are Linear Predictive Coding (LPC), Mel-frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), RASTA-PLP (Relative Spectral Filtering of log domain coefficients), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), and Discrete Wavelet Transform (DWT). Perceptual Prediction based methods work well under noisy conditions. MFCC is the widely used and most popular feature extraction method in speech processing. Spectral characteristics of PLP and MFCC are transformed to match with human auditory system. Mahmud and Munni, 2020 present a comparison of performances of PLP, LPC and MFCC feature extraction techniques for Bangla speech recognition on a short dataset consisting of 100 common Bangla sentences. The feature vectors were tarined on LSTM neural network. The performance was evaluated by measuring the Bhattacharyya distance between similar-sounding phonemes. Their study shows that PLP and MFCC performed better than LPC and PLP performed better than MFCC.

Phonetic dictionary

A phonetic dictionary provides a mapping from words to phones. Though dictionary is not the only way of mapping words to phones, it can be also done by using some machine learning algorithms.

Acoustic model

The Acoustic model is used in an automatic speech to text conversion system to represent the relationship between speech signals and linguistic units of speech such as phonemes or words. It can be also considered as a pattern recognition system that matches extracted features of the speech signal with the linguistic unit of speech. The model is trained with speech data and transcription of speech data in the training phase. The Performance of automatic speech to text conversion largely depends on the amount of speech data and transcription text in the training phase. Techniques like Support Vector Machine (SVM), Hidden Markov Model (HMM), Vector Quantization (VQ), Dynamic Time Warping (DTW) are widely used for the classification and recognition of speech data. The Hidden Markov Models (HMM) were the most popular statistically based acoustic models. Each phoneme is modeled using an HMM. As shown in Figure 2, an HMM consists of a set of states, transitions, and output distributions. Hidden Markov Models(HMM) were very popular and widely used technique for acoustic modeling before the emergence of neural network based acoustic models. In the past few years, automatic speech to text conversion achieved a level in terms of robustness for the inclusion of neural networks and deep learning technologies in acoustic modeling. Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and various hybrid neural network models are popular for training acoustic models.

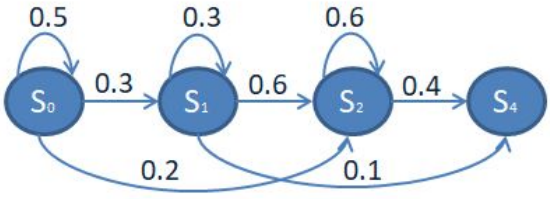

Figure 2: An example HMM phone model

Language model

The language models are used to restrict word search. A language model also compensates the uncertainties generated by the transcription of acoustic models. The process of matching consecutive words is sequential. A language model helps to find the words which can follow previously recognized words by stripping words that are not probable. A language model can be a simple N-gram based statistical model. Statistics based language model stores the probabilities of N-grams based on their appearance in the training corpus. The Performance of a statistics-based language model depends on the size of the text corpus used for training the language model. A major drawback of

statistical language modeling is the inability of dealing with out of vocabulary (OOV) words and synonymous words. The introduction of neural networks helps to overcome this problem. Recently Memory networks and attention mechanisms are used for language modeling (Daniluk et al., 2017). Residual Memory Networks(RMN) and LSTMs are also applied for language modeling. Different types of CNNs used in (Lin et al., 2013) for language modeling. Some studies take character-level input for language modeling instead of word embedding as input (Botha and Blunsom, 2014).

Decoder

The purpose of a decoder is to produce an accurate textual format of the speech data by combining the output of the acoustic model and language model. The probability of occurrence of N-grams in transcription generated by the acoustic model is checked by the N-gram probability of the language model. The idea is to find a transcription of the speech signal that maximizes the probability of appearing from both the acoustic and language models. Viterbi algorithm, HMM based decoding technique, Prefix beam search algorithms, Connectionist Temporal Classification (CTC) are usually used for designing the decoder.

Evaluation Process of Speech to Text Conversion System

After training the acoustic model and language model with adequate training speech and text data the system is tested on test speech data to evaluate the performance of the system. It is assumed during the evaluation period that we have the system-generated transcription of the speech data and the original text. Usually, the following characteristics are used for evaluating speech to text conversion systems:

- **Accuracy:** Accuracy is a common measure for STT conversion performance. When the total number of words is N, I is the number of inserted words, D is the number of deleted words and S represent the number of substituted words then the accuracy of the system can be measured as:
  Accuracy= (N-D-S)/N

- **Word Error Rate(WER):** WER is another common measure used for evaluating STT conversion systems. WER can be measured as:
  WER= (I+D+S)/N

- **Speed:** Speed is another measure usually used with WER or accuracy. If recording time of speech data of a certain STT system is 2 hours and decoding time is 4 hours then speed is measured as 2×RT.

- **ROC Curve:** ROC curves are the graphical representation of a binary classifier. In the case of conversion of speech to text, there can be true conversion and false conversion. ROC

curves can be created by plotting these true positive rates and false positive rates. It tries to find the optimal point where the true positive is highest and the false positive is the lowest.

Results

In the previous section, we discussed the general procedure of speech to text conversion. Working with any specific language such as Bangla may require changes in some of the methodologies to capture special characteristics of the language. Some techniques might work better for some other languages that may not perform better for Bangla. We discussed the recent advancements in Bangla speech to text conversion in the Introduction. In the following subsection, a comparative study of the recent studies in terms of dataset size, methodology, number of speakers, and accuracy is presented in Table 1. We also present a comparison of the Bangla speech dataset and speech recognition toolkits in the subsequent subsections.

Comparison of Bangla speech to text conversion approaches

In this subsection, a comparison of Bangla speech to text conversion systems is presented in Table 1 in terms of dataset size, number of speakers, and methodology of training, and accuracy rate. Different systems are not directly comparable. Because no specific benchmark speech dataset was used by all of the systems. Most of the systems were tested on the author-created limited dataset and in controlled environment. Because accuracy or WER depends on the complexity and size of the dataset used for experimentation. So we included dataset characteristics in Table 1 to present a detailed idea about the actual performances of the comparative systems.

From Table 1, we observe that Sharmin et al., 2020 achieved very promising result but only for converting Bangla spoken digits. Islam et al., 2019 also achieved very good accuracy. The best result was achieved in Bangla speech to text conversion by Al Amin et al., 2019 using toolkit Kaldi and the system was trained on publicly available speech dataset SHRUTI. Though we can not conclude on a specific acoustic model as best performing model, at least from Table 1 we observe that the deep neural network and traditional HMM based hybrid models are performing better than other models. Research have shown that DNN based acoustic models outperformed on most of the corpora having speech data more than 100 hours (Hinton et al., 2012). Bangla speech to text conversion approaches need to be tested on same training and test dataset or at least on some benchmark speech dataset.

In Table 2, we present a detailed description of the available Bangla speech corpora. All the corpora are not publicly available to download. M. F. Khan and Sobhan, 2018 present a speech corpus only consisting of isolated words, and M. Khan and Sobhan, 2018 present a speech corpus only for continuous words. SUBESCO (Sultana et al., 2021) presents an emotional speech corpus. They captured seven types of emotions in the corpus of 7000 utterances of 10 sentences.

Table 1: Comparison of Bangla speech to text conversion approaches

Approach	Speech Dataset size	No. of Speakers	Model Training Method	Accuracy
Hasnat et al., 2007	100 Words (Created by authors)	5	HMM Toolkit (HTK)	70%(Isolated speech) 60% (Continuous Speech)
Sultana et al., 2012	A newspaper article (Created by authors)	1	SAPI	78%
Nasib et al., 2018	2 Hours(503 words) (Created by authors)	5	CMU Sphinx4	71.7%
Tausif et al., 2018	80 Mins, 170 Sentences (Created by authors)	10(5M, 5F)	DeepSpeech KenLM	50%(Test data) 90%(Training data)
Syfullah et al., 2018	900 Bangla characters (Created by authors)	Unknown	ANN	81.61%
Saurav et al., 2018	500 Unique words (Created by authors)	Unknown	Kaldi	WER 3.96%(GMM-HMM) WER 5.30%(DNN-HMM)
Sumon et al., 2018	10 Command words (Created by authors)	Unknown	CNN	74%
Bristy et al., 2019	101 Words (Created by authors)	8(M and F)	CMU Sphinx4	59.01%(Unknown speaker) 78.57%(known speakers)
Al Amin et al., 2019	SHRUTI(21.64 hours)	26 M and 8 F	Kaldi	WER 0.92%(DNN-HMM) WER 2.02%(GMM-HMM)
Islam et al., 2019	33 Hours (Fraction dataset of Kjartansson et al., 2018)	Unknown	DRNN	86.058%
Sharmin et al., 2020	Bengali spoken digit, 1230 audio files	10	CNN	98%
Mandal et al., 2020	214.6 Hours (Kjartansson et al., 2018)	508	CTC based CNN-RNN	WER 13.67%
Saha et al., 2021	215.53 Hours	Unknown	CTC based CNN-RNN	81.61%

Table 2: Comparison of Bangla speech corpora

Speech corpus	Collected in	Recordings	Hours	Speakes
Kjartansson et al., 2018	Bangladesh	232,537 Audio files	229	508 (Male and Female)
Ahmed et al., 2020	Bangladesh	297,065 Audio files	960	268 Male and 251 Female
SHRUTI(Das et al., 2011)	India	7383 Sentences 22012 Words	21.64	26 Male and 8 Female
BdNC01 Speech Corpus of Isolated Words (M. F. Khan and Sobhan, 2018)	Bangladesh	1081 Words	292	150
BdNC01 Speech Corpus of Connected Words (M. Khan and Sobhan, 2018)	Bangladesh	52 Sentences	62	150
Phonetically Balanced Bangla Speech Corpus (Murtoza et al., 2011)	Bangladesh	2,023,162 sentences	1 Hr 11 min	1 Female
SUBESCO (Sultana et al., 2021)	Bangladesh	7000	7	10 Male and 10 Female

A benchmark speech dataset should consists of several hundreds of hours of speech data recorded in practical environment from both male and female speakers of different age group. It should also be publicly available for research. From Table 2 only the first , third and the last speech dataset are publicly available for researchers right now. From speech dataset point of view, the speech dataset collected by Kjartansson et al., 2018 is the best choice for speech researchers. Recently many researchers are using this dataset as it contains a fairly good amount of speech data collected in practical environment from both male and female speakers and is publicly available. Choice of speech dataset also depends on purpose of use. Speech dataset SUBESCO by Sultana et al., 2021, is the only sentimental speech dataset available in Bangla language and it is publicly available.

There are many speech recognition toolkits available online such as CMU Sphinx, Kaldi, SAPI, Julius, HTK, etc. Among them, some of the toolkits were used in Bangla Speech to Text conversion task. Table 3 presents a comparative study of various characteristics of the most popular toolkits used in Bangla Speech to Text conversion.

Discussions

Speech to text conversion has achieved remarkable performance in many languages and many speech to text application programs are available online. Also, there are some dedicated companies for speech to text conversion such as rev.ai, temi.com managed to achieve higher accuracy than Google, Microsoft, and Amazon provided STT applications. Rev.ai had the lowest error rate

Table 3: Comparison of popular toolkits used in Bangla STT conversion

Toolkit	Basic Technology	Open Source	Operating System	Programming Language
CMU Sphinx	HMM	Yes	Cross-platform	Java
Kaldi	Neural Network	Yes	Cross-platform	C++
HTK	HMM and Neural Network	No	Cross-platform	C
Julius	HMM Trigrams	Yes	Cross-platform	C

among speech-to-text technology companies as of 2020. With an accuracy rate of 86%. Rev.ai outperformed companies like Google, Microsoft, and Amazon (Liu, 2021). Still, speech to text conversion applications in any language has to face some challenges. The following two subsections provide the general challenges and Bangla language-specific challenges that speech to text conversion system faces correspondingly.

Challenges associated with STT conversion

In this era of connectivity and speed, human-machine interaction become essential and inevitable. Connectivity through the Internet of things (IoT), Ubiquitous computing, enhances the necessity of end-to-end transcription of speech data. Though the research on speech to text conversion began last century, still it does not achieve 100% accuracy in any language. There are some environmental and speaker-related issues that affect the performance of STT conversion. The different types of variability that affects the performance are discussed below:

- **Environmental noise:** In the case of speech to text conversion in a practical environment there might add some unwanted sound from surroundings with the speech signal. The performance of the system can be degraded by the background sounds like a car horn, chattering, echo, etc. Even in a controlled environment, the performance of the system changes due to different microphone types.

- **Speaker variability:** Performance of speech recognition changes for speaker variability. It may vary on the gender of the speaker and age of the speaker. The training speech data needs to contain speech data from different kinds of speakers of different sex and from people of different ages. The same person can speak differently depending on his/her mood. The same person speaks in different tone in a sad and angry mood.

- **Speaking Style:** Pronunciation of the same word may vary due to the accent style of the word. Every person has his/her own way of pronouncing a word. While speaking the vocal tract of a person can vary in terms of roughness, nasality, volume, speed, and pitch. Pronunciation of a word also depends on the locality of the speaker belong to.

- **Homonyms words:** When two words have different meaning and different spelling but pronounce similarly they are called homonyms words. The acoustic model can not differentiate between homonyms words. N-gram based language model may compensate for uncertainties aroused by homonyms words.
- **Proper Nouns:** Handling proper nouns is still a challenge in STT conversion. One solution proposed was to include syllable N-grams in the language model. But this decreases overall STT conversion performance.
- **Accurate Use of Punctuation:** Correct use of punctuation mark is an open challenge for speech to text conversion in any language.
- **Use of Jargon:** To be usable on technical field or medical field the speech corpus and text corpus must contain data of technical or medical sector. For identifying jargon's of specific domain the language model and acoustic model needs to be trained on recordings of that specific domain.

Challenges associated with Bangla language

Along with general challenges associated with speech to text conversion, Bangla speech to text conversion has some additional challenges that need to be handled carefully. Special challenges associated with Bangla speech to text conversion are pointed out below:

- **Resource unavailability:** In the research of NLP, Bangla is labeled as a low resource language. Most of the research done on Bangla speech to text conversion worked on limited vocabulary speech and text datasets. But for practical application, the system should be trained on a very large speech dataset consisting of speakers from different age groups, different gender, from different localities, and speech data should be collected at different times of the day. This should help to achieve speaker independence and include different dialects.
- **Language Speciality:** Bangla is a very special language and word formation in Bangla is different than in English. Modified forms of vowels are used with consonants to form words. Also, clustered consonant letters are used in words. Also, modified forms of consonants are used in word formations. These specialties of Bangla language add extra difficulties in transforming the acoustic model generated transcriptions to the correct textual format. Some recent works tried to deal with language specialty (Tausif et al., 2018). Still, there are many language issues that need to be addressed.

Future research directions

The leading companies that are providing professional quality speech to text translation are trained on several thousands of hours of speech data whereas Bangla speech to text translation systems are

trained on several hours of speech data only (very few systems worked on more than one hundred hour speech data). So developing speech resource is the most important task for doing research in this field. For example, the promising application rev.ai is trained on 50,000+ hours of human-transcribed content from a wide range of topics, industries, and accents. There are some application areas such as medical documentation and technology-related fields that require domain-specific data. To get the benefit of such applications we need to develop domain-specific speech and text corpus. Some recent research has indicated that low-resource languages can be benefited by transferring knowledge learning from rich-resource languages (Hou et al., 2021 and Dalmia et al., 2018). Bangla speech to text researchers can use their idea to tackle Bangla dialect. Hou et al., 2021 presents experimental results with an improvement of WER by using knowledge transfer algorithm *SimAdapter* on very little training data. There is very little speech corpus available on dialectal Bangla speech. To tackle this data sparsity problem a cross-lingual approach can be followed for Bangla by applying transfer learning on standard Bangla spoken dataset and dialectal Bangla speech data. In this case, acoustic differences between standard Bangla and dialectal Bangla is smaller than language level. More research should be carry on recognizing out-of-vocabulary words that are not present in the training corpus. Statistical models fail in this case. More research with neural network models on huge speech corpus may help. Special attention should be given to recognize named entities or proper nouns in speech. Bangla language has many specialities including the use of clustered letters in words. This specialities results in more spelling errors in transcripts unless special care taken. A context aware spell checker can be used to solve this problem. A context aware spell checker with the availability of a very large scale text corpus that contains maximum vocabulary of Bangla language should help in this context.

Conclusion

Speech to text conversion did not achieve full accuracy and reliability in any language yet. Despite its limitations, present speech recognition technology can be a very useful tool for a variety of applications. Speech recognition is already used for live subtitling on television, as dictation tools in the healthcare system, in car system, in military applications, and for off-line speech-to-text conversion for the education system for several languages. For all these applications, human editing of the output is needed to achieve really good levels of accuracy. Bangla is a very rich language with great historical value and a huge number of native speakers. Yet we can not take advantage of these types of applications in our native language due to available resource scarcity and lack of enough research in this domain. Some recent research has shown promising results in controlled environment with limited vocabulary. Professional speech to text conversion quality is yet to be achieved in Bangla language. This article provides a detailed study of recent advancements in Bangla speech to text conversion along with a brief discussion of general speech to text conversion architecture. We believe, the comprehensive comparative study of this literature review of Bangla

speech to text conversion systems and Bangla speech datasets will help researchers to select better approaches to overcome both the general challenges and language-specific challenges. We also hope that future research directions will provide important guidelines for research in this field.

References

Ahmed, S., Sadeq, N., Shubha, S. S., Islam, M. N., Adnan, M. A., & Islam, M. Z. (2020). Preparation of bangla speech corpus from publicly available audio & text. *Proceedings of The 12th language resources and evaluation conference*, 6586–6592.

Al Amin, M. A., Islam, M. T., Kibria, S., & Rahman, M. S. (2019). Continuous bengali speech recognition based on deep neural network. *2019 international conference on electrical, computer and communication engineering (ECCE)*, 1–6.

Botha, J., & Blunsom, P. (2014). Compositional morphology for word representations and language modelling. *International Conference on Machine Learning*, 1899–1907.

Bristy, I. J., Shakil, N. I., Musavee, T., & Choton, A. R. (2019). *Bangla speech to text conversion using cmu sphinx* (Doctoral dissertation). Brac University.

Chowdhury, N., Sattar, M. A., & Bishwas, A. K. (2009). Separating words from continuous bangla speech. *Global Journal of Computer Science and Technology*, *9*(4).

Dalmia, S., Sanabria, R., Metze, F., & Black, A. W. (2018). Sequence-based multi-lingual low resource speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4909–4913.

Daniluk, M., Rocktäschel, T., Welbl, J., & Riedel, S. (2017). Frustratingly short attention spans in neural language modeling. *arXiv preprint arXiv:1702.04521*.

Das, B., Mandal, S., & Mitra, P. (2011). Bengali speech corpus for continuous auutomatic speech recognition system. *2011 International conference on speech database and assessments (Oriental COCOSDA)*, 51–55.

Dave, N. (2013). Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, *1*(6), 1–4.

Hasnat, M., Molwa, J., & Khan, M. (2007). Isolated and continuous bangla speech recognition: Implementation. *Performance and application perspective*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82–97.

Hirsch, H.-G., & Ehrlicher, C. (1995). Noise estimation techniques for robust speech recognition. *1995 International conference on acoustics, speech, and signal processing*, *1*, 153–156.

Hou, W., Zhu, H., Wang, Y., Wang, J., Qin, T., Xu, R., & Shinozaki, T. (2021). Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 317–329.

Islam, J., Mubassira, M., Islam, M. R., & Das, A. K. (2019). A speech recognition system for bengali language using recurrent neural network. *2019 IEEE 4th international conference on computer and communication systems (ICCCS)*, 73–76.

Khan, M. F., & Sobhan, M. A. (2018). Construction of large scale isolated word speech corpus in bangla. *Global Journal of Computer Science and Technology*.

Khan, M., & Sobhan, M. (2018). Creation of connected word speech corpus for bangla speech recognition systems. *Asian Journal of Research in Computer Science*, 1–6.

Kjartansson, O., Sarin, S., Pipatsrisawat, K., Jansche, M., & Ha, L. (2018). Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Liu, S. (2021). Speech-to-text transcript accuracy rate among leading companies 2020. https://www.statista.com/statistics/1133833/speech-to-text-transcript-accuracy-rate-among-leading-companies/

Mahmud, N. A., & Munni, S. A. (2020). Qualitative analysis of plp in lstm for bangla speech recognition. *The International Journal of Multimedia & Its Applications (IJMA) Vol*, *12*.

Mandal, S., Yadav, S., & Rai, A. (2020). End-to-end bengali speech recognition. *arXiv preprint arXiv:2009.09615*.

Murtoza, S., Alam, F., Sultana, R., Chowdhur, S., & Khan, M. (2011). Phonetically balanced bangla speech corpus. *Proc. Conference on Human Language Technology for Development*, *2011*, 87–93.

Nasib, A. U., Kabir, H., Ahmed, R., & Uddin, J. (2018). A real time speech to text conversion technique for bengali language. *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, 1–4.

Rudnicky, A. I., Hauptmann, A. G., & Lee, K.-F. (n.d.). Survey of current speech technology. *Communications of the ACM*, *37*.

Saha, S., et al. (2021). Development of a bangla speech to text conversion system using deep learning. *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 1–7.

Saurav, J. R., Amin, S., Kibria, S., & Rahman, M. S. (2018). Bangla speech recognition for voice search. *2018 international conference on Bangla speech and language processing (ICBSLP)*, 1–4.

Sharma, N., & Sardana, S. (2016). A real time speech to text conversion system using bidirectional kalman filter in matlab. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2353–2357.

Sharmin, R., Rahut, S. K., & Huq, M. R. (2020). Bengali spoken digit classification: A deep learning approach using convolutional neural network. *Procedia Computer Science*, *171*, 1381–1388.

Akhter, A. et al. (2022). Automated speech-to-text conversion systems in Bangla language: A systematic literature review. *Khulna University Studies*, Special Issue (ICSTEM4IR): 566-583.

Sultana, S., Rahman, M. S., Selim, M. R., & Iqbal, M. Z. (2021). Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla. *Plos one*, *16*(4), e0250173.

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

Sumon, S. A., Chowdhury, J., Debnath, S., Mohammed, N., & Momen, S. (2018). Bangla short speech commands recognition using convolutional neural networks. *2018 international conference on bangla speech and language processing (ICBSLP)*, 1–6.

Syfullah, S. M., Zakaria, Z. B., Uddin, M. P., Rabbi, M. F., Afjal, M. I., & Nitu, A. M. (2018). Efficient vector code-book generation using k-means and linde-buzo-gray (lbg) algorithm for bengali voice recognition. *2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, 1–4.

Tamura, S., & Waibel, A. (1988). Noise reduction using connectionist models. *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 553–556.

Tausif, M. T., Chowdhury, S., Hawlader, M. S., Hasanuzzaman, M., & Heickal, H. (2018). Deep learning based bangla speech-to-text conversion. *2018 5th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*, 49–54.

Wikipedia. (2022). Bangla language, wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Bengali_language