IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 January 2024, accepted 24 February 2024, date of publication 29 February 2024, date of current version 8 March 2024.
Digital Object Identifier 10.1109/ACCESS.2024.3371478

RESEARCH ARTICLE

BanSpeech: A Multi-Domain Bangla Speech Recognition Benchmark Toward Robust Performance in Challenging Conditions

AHNAF MOZIB SAMIN[1], M. HUMAYON KOBIR[1], MD. MUSHTAQ SHAHRIYAR RAFEE[1,2], M. FIROZ AHMED[1], MEHEDI HASAN[1], PARTHA GHOSH[1], SHAFKAT KIBRIA[1,3], AND M. SHAHIDUR RAHMAN[1], (Senior Member, IEEE)
[1]Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh
[2]Department of Computer Science and Engineering, Metropolitan University, Sylhet 3109, Bangladesh
[3]Department of Computer Science and Engineering, Leading University, Sylhet 3112, Bangladesh

Corresponding author: M. Shahidur Rahman (rahmanms@sust.edu)

ABSTRACT Despite huge improvements in automatic speech recognition (ASR) employing neural networks, ASR systems still suffer from a lack of robustness and generalizability issues due to domain shifting. This is mainly because principal corpus design criteria are often not identified and examined adequately while compiling ASR datasets. In this study, we investigate the robustness of the fully supervised convolutional neural networks (CNNs), and the state-of-the-art transfer learning approaches, namely self-supervised wav2vec 2.0 and weakly supervised Whisper for multi-domain ASR. We also demonstrate the significance of domain selection while building a corpus by assessing these models on a novel multi-domain Bangladeshi Bangla ASR evaluation benchmark—BanSpeech, which contains approximately 6.52 hours of human-annotated speech, totaling 8085 utterances, across 13 distinct domains. SUBAK.KO, a mostly read speech corpus for the morphologically rich language Bangla, has been used to train the ASR systems. Experimental evaluation reveals that self-supervised cross-lingual pre-training with wav2vec 2.0 is the best strategy compared to weak supervision and full supervision to tackle the multi-domain ASR task. Moreover, the ASR models trained on SUBAK.KO face difficulty recognizing speech from domains with mostly spontaneous speech. The BanSpeech is publicly available to meet the need for a challenging evaluation benchmark for Bangla ASR.[1]

INDEX TERMS Automatic speech recognition, Bangla, domain shifting, read speech, spontaneous speech, transfer learning.

I. INTRODUCTION

Transfer learning has become the state-of-the-art approach in natural language and speech processing [1], [2]. By pre-training a model with a huge amount of unlabelled speech data in a self-supervised fashion and then fine-tuning it with a much smaller dataset, the performance of the ASR models can be significantly improved. Despite the improvements brought by self-supervised training and wav2vec 2.0 [1], some studies indicate that speech models pre-trained in a supervised way for numerous domains are more capable of ensuring robust performance [3], [4], [5]. However, collecting thousands of hours of high-quality supervised speech data is a cumbersome and expensive task. To this end, scaling weakly supervised pre-training is investigated for speech processing and found to be the state-of-the-art for English ASR even in zero-shot transfer learning without additional fine-tuning [6]. While creating a weakly supervised speech dataset, the human-validated annotation is not performed and data is curated using an automated pipeline. However, the benefit of weak supervision is that it is comparatively easier to scale the dataset and as for Whisper, approximately 680K hours of multilingual data is curated in this way [6]. Yet, the zero-shot performance of Whisper for low-resource

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.
[1]BanSpeech can be downloaded from https://huggingface.co/datasets/SUST-CSE-Speech/banspeech

languages is not up-to-the-mark. The authors hypothesize that fine-tuning might improve the results which is yet to be investigated for many languages. It still remains an open question how robust the state-of-the-art fully supervised, self-supervised, and weakly supervised speech models are for multiple-domain low-resource speech recognition.

Bangla is a morphologically rich language from the Indo-Aryan language sub-group. Kibria et al. developed SUBAK.KO, an annotated speech corpus for speech recognition research comprising 241 hours of Bangladeshi Bangla speech data, to address the dearth of annotated speech datasets in Bangla [7]. SUBAK.KO contains 229 hours of clean read speech and 12 hours of broadcast speech utterances. The recording scripts are collected from 40 text domains, including conversations, sports, news, poetry, letters, etc., following the reception and production criteria for the text domains to build the corpus [8]. According to a cross-dataset evaluation, SUBAK.KO is a more balanced corpus with respect to regional accents and other types of speaker variability compared to another large-scale Bangladeshi Bangla speech corpus, LB-ASRTD, when evaluated on clean read speech test sets [7], [9]. The read speech test sets from standard datasets often cannot evaluate the robustness of ASR applications since the same domains are also included in the training set [10]. Thus, it is yet uncertain how well an ASR model based on SUBAK.KO performs in a range of domains, particularly those that include the majority of spontaneous speech. Human verbal communication consists primarily of spontaneous speech with some background noise, and past research indicates significant acoustical and linguistic variations between read speech and spontaneous speech [11].

Motivated by the above-mentioned issues with regard to multi-domain ASR, this paper focuses on the following research questions: 1) In terms of robustness and generalizability, how do the state-of-the-art fully supervised, self-supervised, and weakly supervised speech recognition models perform for multiple domains, especially those that contain mostly spontaneous speech? 2) Do the 40 text domains considered while developing the SUBAK.KO corpus meet all of the reception and production criteria for developing an ASR corpus that can ultimately lead to improved performance in challenging real-world conditions, such as noisy, spontaneous, and multi-talker environments?

The contributions of this paper are as follows:

- We present BanSpeech, a novel multi-domain Bangladeshi Bangla ASR evaluation benchmark consisting of 6.52 hours of human-validated speech data, totaling 8085 utterances, from 13 domains collected from YouTube and manually transcribed by human annotators. This speech dataset consists primarily of spontaneous speech from all domains, with the exception of *audiobooks* and *biography* domains, which comprise read speech. Moreover, we collect approx. 80 minutes of additional data containing dialectal speeches from the seven major parts of Bangladesh, making BanSpeech 7.7 hours long. The latter part, however, has not been validated under human-supervision.
- We explore a fully supervised deep CNN, a self-supervised wav2vec 2.0 XLS-R [12], and a weakly supervised Whisper model [6] to investigate their robustness on multiple domains using BanSpeech evaluation set. We train the CNN model from scratch using SUBAK.KO. For the wav2vec 2.0 and Whisper, we choose the cross-lingual pre-trained models and the same SUBAK.KO dataset for fine-tuning. All of the models are evaluated on the multi-domain BanSpeech and SUBAK.KO test set.
- Through a comprehensive evaluation of ASR models trained or fine-tuned on SUBAK.KO, we shed light on the significance of domain selection for the development of a speech dataset. In this regard, both read and spontaneous speeches from BanSpeech are used to evaluate them.
- We report our findings on deep CNN by experimenting with the number of convolutional layers, applying several normalization techniques, comparing three input feature extraction methods from audio signals, and observing the effect of the number of Mel Frequency Cepstral Coefficients (MFCCs).

The rest of the paper is structured as follows: background is provided in Section II. In Section III, the preparation of the BanSpeech is described. We discuss the experimental setup in Section IV. Results are shown and discussed in Section V. We present the limitations of our work in Section VI. Section VII concludes the article and provides the future direction.

II. BACKGROUND

A. RELATED WORK IN BANGLA

While prior to 2018, Bangla ASR research was limited to only isolated words and digit recognition using small datasets, some work has been done on Bangla large vocabulary continuous speech recognition (LVCSR) later on [13]. Amin et al. examined Deep Neural Network-Hidden Markov Model (DNN-HMM) and Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) based techniques on a relatively small and speaker-independent Shruti corpus (21.64 hours) [14], [15]. According to their findings, a larger data set is necessary for the DNN-HMM method to outperform the GMM-HMM method. Sumit et al. implemented the Recurrent Neural Network-based Deep Speech 2 architecture on the non-public 300-hour-long "Socian" Bangla telephone conversation-based dataset [16], [17]. In addition to Socian, they also added 50 hours from the Bangla Babel corpus. Developed in 2016, Bangla Babel is likewise a telephone conversation-based speech corpus, comprising 215 hours of speech [18]. However, Babel has West Bengal accented speech that is distinct from the Bangladeshi Bangla accent [7]. Ahmed et al. prepared 960 hours of broadcast Bangla speech corpus by transcribing speech data in an automated way with pre-trained

ASR models [19]. The corpus is not publicly accessible, and the authors focused on developing an algorithm to iteratively construct speech corpora in their work. However, the objective of our work is to empirically evaluate the ASR models trained with three distinct approaches and the SUBAK.KO dataset on diverse domains to find out whether performance can vary across different domains.

Samin et al. evaluated the quality of a large-scale publicly available LB-ASRTD corpus (229 hours) using deep learning-based approaches by conducting character-wise error analysis [20]. They also found a deep CNN-based acoustic model and a 5-gram Markov Language Model (LM) to be capable of achieving a lower word error rate (WER) on LB-ASRTD. In this study, we also use a deep CNN-based model while utilizing a higher number of MFCCs during the input feature extraction and introducing layer normalization in each convolution layer. Based on an acoustic study on a regional accented speech and the character-wise error analysis on LB-ASRTD, the requirement of a new corpus with more speaker variability and character-wise well-balancedness was recommended [20], [21]. Therefore, Kibria et al. developed the 241-hour-long publicly available Bangladeshi Bangla SUBAK.KO corpus with the aim of addressing the above-mentioned issues of LB-ASRTD [7].

The Bengali Common Voice Speech dataset with over 400 hours of crowd-sourced data has been made available on the Mozilla Common Voice Platform, and the campaign to address the scarcity of Bangla speech datasets is ongoing [22]. Although there are now some large annotated ASR corpora for Bangla, there is no comprehensive ASR evaluation benchmark in this language that can categorically investigate a model on a variety of domains, dialects, and speech types from multiple speakers.

Shahgir et al. fine-tuned a wav2vec 2.0 model on the Bengali Common Voice Speech dataset and presented a promising performance for Bangla ASR [23]. They choose the wav2vec 2.0 model pre-trained on 53k hours of cross-lingual speech data in a self-supervised way [24]. However, there also exists a wav2vec 2.0 XLS-R model pre-trained on a much larger amount of data (436K hours) and Bengali is among the languages that were considered during pre-training [12]. To the best of our knowledge, no previous work has investigated the Whisper [6] for Bangla ASR.

B. WAV2VEC 2.0

is pre-trained in a self-supervised way by masking input frames in the latent space and using a contrastive loss function, the model learns the inherent representations from speech [1]. These representations have now been used in numerous speech processing tasks including ASR, language identification, keyword spotting, speaker verification, speech translation, etc [25], [26], [27]. Wav2vec 2.0 takes raw audio as input and passes it to a feature encoder consisting of convolution blocks. A quantization module is used to transform the output of the feature encoder from the continuous space into discrete space and a context network with Transformer blocks is trained with a several contrastive objective. Each Transformer block contains multi-head self-attention, feed-forward layer, and add &norm sub-module. In addition to contrastive loss, diversity loss and L2 panelty are used as loss functions. Wav2vec 2.0 contains only encoder and provides contextual representation as output, which can fed to a linear classifier (also referred to as language modeling head) to get the output probability distribution.

Following the success of monolingual wav2vec 2.0, a larger cross-lingual wav2vec 2.0 model named XLS-R is released [12]. XLS-R has been pre-trained on 436K hours of data from 128 languages. There are three versions of this model with 300 million, 1 billion, and 2 billion parameters. During the pre-training stage, several Indo-Aryan languages have been included in XLS-R. More precisely, there are 100 hours of Bengali training data from the common voice (CV) dataset which is used for pre-training the XLS-R. It has been shown that when a model is pre-trained on some closely related languages, then it substantially improves the ASR performance after fine-tuning [12]. For this reason, XLS-R is found to be extremely useful to build ASR applications for many low-resource and mid-resource languages.

C. WHISPER

Contrary to the state-of-the-art method of self-supervised learning, Whisper is pre-trained on 680K hours of speech data adopting a weakly supervised pre-training method [6]. Instead of ensuring gold-standard human-validated transcripts, they collect audio data and corresponding transcripts from the internet. Since these transcripts are often noisy (e.g. ASR generated transcripts, etc.), they utilize several automated filtering strategies. The training datasets for Whisper are prepared without human supervision. Whisper is pre-trained with a standard sequence-to-sequence (seq2seq) encoder-decoder Transformer-based architecture [28]. Each Transformer block in encoder and decoder contains self-attention and feed-forward submodules. Cross attention is performed between the encoder and decoder. Whisper is pre-trained on several speech processing tasks namely multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. The detailed architecture is described in their paper [6]. Similar to wav2vec 2.0 XLS-R [12], Whisper is also pre-trained cross-lingually in 97 languages and contains several versions including tiny, base, small, medium, and large with 39 million, 74 million, 244 million, 769 million, 1550 million of parameters, respectively.

While Whisper sets a state-of-the-art result on the English LibriSpeech dataset, the zero-shot performance of other languages using Whisper is substantially low. The reason lies in the fact that Whisper is pre-trained mostly on English data and the other 96 languages have less than 1000 hours of data. In the case of Bengali, there are only 1.3 hours of annotated speech data used for pre-training and the WER is more than

TABLE 1. Key statistics of BanSpeech. BanSpeech consists of 13 diverse domains along with 6 dialectal domains. The audio length, number of samples and vocabulary size (number of unique words) are shown per domain.

Domain Type	Domain	Length	Number of samples	Vocabulary size
General domains	Television news	30.3 min	571	1786
	Parliament speech	30.0 min	585	1585
	Audiobooks	30.3 min	955	2339
	Drama series	30.5 min	514	1628
	Class lecture	30.2 min	397	1288
	Political talk show	30.0 min	813	1798
	Celebrity Interview	30.5 min	561	1752
	Documentary	30.2 min	615	1773
	Kids' voice	30.3 min	321	1497
	Medicine	31.2 min	704	1294
	Biography	28.2 min	657	1739
	Kids' Cartoon	29.6 min	660	1439
	Sports	30.3 min	732	1784
Dialectal domains	Barisal	10.8 min	129	688
	Chittagong	10.5 min	154	589
	Dhaka	10.5 min	189	690
	Mymensingh	8.6 min	141	810
	Noakhali	10.2 min	93	725
	Rajshahi	10.1 min	229	543
	Sylhet	10.0 min	199	503
	Total validated	391.5 min or 6.52 hours	8085	11630
	Total (including dialects)	462.2 min or 7.7 hours	9219	12807

80% in a zero-shot setting. However, the authors hypothesize that the Whisper performance for low-resource languages can be improved if fine-tuned on a specific language with more data instead of performing zero-shot transfer.

D. MULTI-DOMAIN ASR

Hsu et al. conducted extensive experiments on domain shift in self-supervised pre-training [29]. They found that adding diverse domains in pre-training data helps improve the robustness of the model and the model can recognize data not seen during training. Speech enhancement, data augmentation, and autoencoders are investigated for domain adaptation in distant speech recognition [30]. Domain adaptation for robust ASR is also studied by adopting self-training [31] and multi-task learning [32]. Kawakami et al. pre-trained 8000 hours of noisy multilingual data from diverse domains in an unsupervised way and showed that the learned representations are more robust to domain shifts [33]. To the best of our knowledge, no work has been done to investigate the robustness of self-supervised, weakly supervised, and fully supervised models by dissecting speech data into multiple domains.

III. PREPARATION OF BANSPEECH DATASET

We collect speech data from the open-source platform YouTube. We consider 20 domains, namely *television news, parliament speech, audiobooks, drama, class lectures, political talk shows, interviews, documentaries, kids' voice, medical talk shows, biography, sports, cartoons,* and 7 regional dialects from the major parts in Bangladesh such as *Barisal, Chittagong, Dhaka, Mymensingh, Noakhali, Rajshahi,* and *Sylhet*. All the domains are annotated except for the 7 dialectal domains. The regional dialects are mainly obtained from the region-specific dramas. There are essentially two types of talk shows we consider in this work: political and medical related. On political talk shows, politicians argue about political matters, while on medical talk shows, medical practitioners discuss medical-related topics and frequently employ scientific jargon. Some domains, such as *television news, parliamentary speeches,* and *documentaries*, incorporate both read and spontaneous speech, while others, such as *drama, class lectures, political talk shows, interviews, kid's voices,* and *medical talk shows*, consist predominantly of spontaneous speech. The *audiobook* and *biographical* domains only contain read speech. For each of the domains, there are approximately 30 minutes of speech, except for the dialect domain, which has around 10 minutes of speech per dialect.

We download broadcast speech as waveform audio file format (WAV). We remove commas and brackets from the original audio file names and replace spaces with underscores to avoid potential errors. To make the corpus consistent, each audio file is then converted to a bitrate of 256 kilobits per second (kbps) and a 16 kiloHertz (kHz) mono-channel WAV file. After that, each audio file is automatically divided into smaller chunks based on silence intervals because they are initially too long to pass to a neural network. We use a silence threshold of −40 decibels relative to full scale (dBFS) (consider it silent if quieter than −40 dBFS) and 0.2 seconds as the minimum length of silence (the shortest period of silence before a split can happen). Our dataset contains audio files of varying lengths. The lengths of the audio files range from 0.7 to 35 seconds.

The annotation process is conducted in three stages. After preparing the audio files, we first use the Google Speech-to-Text (STT) system to get the transcripts of those

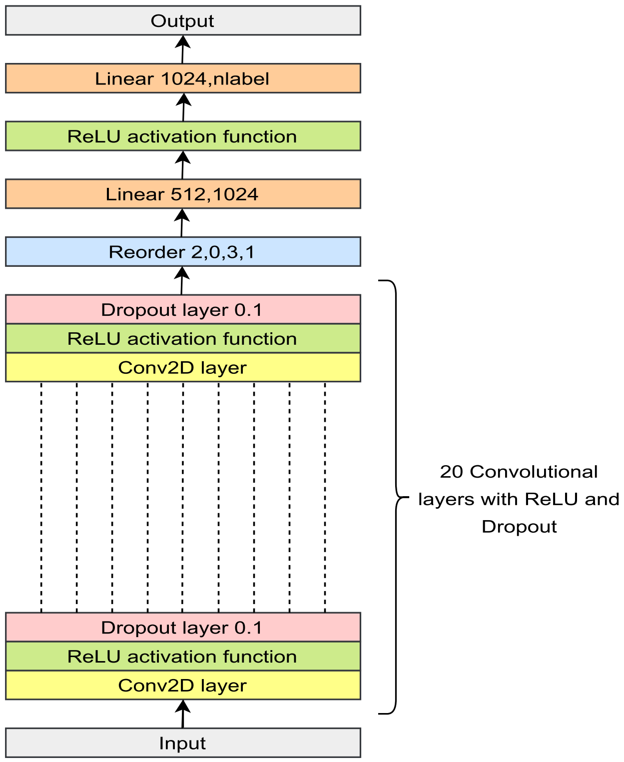

FIGURE 1. **We use the same deep CNN architecture like [20] as our baseline in this study. The architecture consists of 20 convolutional layers, each followed by ReLU activation and a dropout layer. Two linear layers are added to provide the output tokens.**

speeches [34]. Then, two professional native Bangladeshi human annotators with linguistic expertise manually correct the transcriptions. The transcripts are again corrected and finally verified by a new annotator. The corresponding text transcriptions are stored in plain text file format (.txt).

The verified annotated dataset contains 8085 utterances and 6.52 hours of speech with transcriptions. The dialectal domain is not annotated under human supervision since those utterances are extremely deviant from the standard Bangla, possess multiple issues related to annotation, and require native annotators from the corresponding regions. Although the dialectal domains are not suitable to be included in an ASR evaluation set, they can be used for language/dialect identification tasks. The detailed statistics of BanSpeech are shown in Table 1. The verified annotated dataset, containing 8085 audio files, represents substantial representation of both genders, featuring a distribution ratio of 54.5 to 45.5 for male and female speakers, respectively.

IV. EXPERIMENT SETUP

A. DATASET

We train and evaluate our ASR models on SUBAK.KO which contains 241 hours of transcribed Bangladeshi Bangla speech data [7]. This corpus has 229 hours of clean-read speech and only 12 hours of broadcast speech. Clean-read speeches are recorded from 33 native Bangladeshi Bangla male speakers, 28 native female speakers, and 2 second language (L2) speakers. A detailed description of this corpus can be found in the work of [7]. The same train, development (dev), and test splits used in the original paper have been used in our study. The train, dev, and test set contain 200.28 hours, 20.54 hours, and 20.30 hours of speech data, respectively. SUBAK.KO train set is used to train our baseline CNN model from scratch and to fine-tune both wav2vec 2.0 and Whisper. For evaluation, we use the SUBAK.KO test set and BanSpeech.

B. BASELINE

We use a deep CNN model as our baseline in this study. For CNN, three types of features from the frequency domain are used as input to the acoustic model in different experiments: mel-frequency cepstral coefficients (MFCC), mel-frequency spectral coefficients (MFSC), and power spectrum. Frequency domain features exhibit robustness to variations such as speaker differences and noise, and they align with the perceptual characteristics of the human auditory system, in the case of MFCC and MFSC, which involve mel-scale application. Additionally, contrary to raw time-domain features that lead to a substantial increase in input space dimensionality, frequency domain features offer enhanced computational efficiency.

MFCCs are a widely used input features in ASR. To get MFCC features, discrete raw waveform in the time-domain go through pre-emphasis filter to boost the higher frequencies at first, and then, short-time Fourier Transform (STFT) is computed after windowing to convert from the signal from time-domain to frequency domain. Then, mel filterbank is applied to obtain mel-scale, which is similar to how human perceive speech of different frequencies. After taking the log of the mel-spectrum, discrete cosine transform (DCT) is performed to get the MFCCs. Although 13 MFCCs are commonly taken from each input frame, we perform an experiment with different number of MFCCs corresponding to each frame to see the impact (See Figure 4).

MFSC refers to the log energy directly computed from the mel-frequency spectral coefficients without applying DCT. Power-spectrum features have been used in acoustic modeling in speech recognition [17] and they are computed by converting the time-domain signal to frequency domain by applying STFT. Therefore, for power spectrum, mel-filterbanks are not applied unlike MFCC and MFSC.

The non-stationary speech signal is split into smaller windows or frames where it can be assumed as stationary. Different frame sizes and frame strides can affect the WER which we also investigate in this work (See Table 4).

There are several variations in the architecture for conducting numerous experiments. In the case of our baseline CNN model, MFCC feature vectors are fed into the CNN. Our baseline architecture consists of mainly convolutional layers and fully connected linear layers (See Figure 1). Our architecture includes 20 convolutional layers each with a kernel size of 8. In the first convolutional layer, 18 MFCCs are mapped to an embedding space of size 256, and stride size

TABLE 2. List of some of the hyper-parameters and experimental choices used in this work.

Hyper-parameter	CNN	wav2vec 2.0 XLS-R	Whisper
Learning rate	0.05	3e-5	1e-5
Warm up steps	-	300	500
Max grad norm	0.2	-	-
Momentum	0.8	-	-
Learning criterion	0.006	-	-
Cosine scheduler	False	True	True
Batchsize	16	1	4
Gradient accumulation	1	2	1
Max epochs	200	30	15
Early stopping patience	-	10	10
Seed	100	42	42
Token type	char	char	char
Pre-train model	-	XLS-R 300M	Medium 769M

2 is used only for this layer. In the rest of the convolutional layers, there are 256 input channels and 256 output channels (feature maps), and feature extraction through convolution operation is done by 256 filters with stride size 1. Each of the convolutional layers is followed by a rectified linear unit (ReLU) activation function to add non-linearity to the network and dropout to address the over-fitting problem. The dropout probability is 0.1. Lastly, two fully connected linear layers are attached for the classification of character-level tokens.

Connectionist temporal classification (CTC) loss function is used while training the network [35]. The CTC loss is computed using a probability distribution over all possible alignments between the input and output sequences. The probability distribution is obtained from the linear layer at each time step in the output sequence. The CTC loss function can be represented as follows:

$$P(Y|X) = \sum_{A \epsilon A_{X,Y}} \prod_{t=1}^{T} p_t(a_t|X) \tag{1}$$

Here, $P(Y|X)$ is the CTC loss for the output alignment $Y = [y_1, y_2, y_3, \ldots, y_M]$ given the input sequence $X = [x_1, x_2, x_3, \ldots, x_T]$. $p_t(a_t|X)$ is the probability distribution for x_t over the possible characters. Considering all probability vectors $p_t(a_t|X)$ for timesteps, $t = 1, \ldots, T$, we can calculate the probability associated with a particular alignment (representing a specific output sequence). Subsequently, we perform marginalization over the set of alignments.

Apart from the baseline architecture, we train several acoustic models incorporating the batch norm [36], layer norm [37], and weight norm techniques [38]. After running numerous experiments, we obtain our best-performing CNN. We only use our best-performing CNN model to compare its performance with wav2vec 2.0 and Whisper on the BanSpeech. For training the acoustic model, we utilize the wav2letter++ speech processing toolkit [39].

C. TRANSFER LEARNING

We fine-tune the wav2vec 2.0 XLS-R cross-lingual pre-trained model that has 300 million trainable parameters [12]. Raw waveform in the discrete time-domain is passed as input to XLS-R. During fine-tuning XLS-R, the CNN encoder is kept frozen while the contextual network based on Transformer along with a linear layer are updated.

As for the Whisper model, we choose the medium-sized model with 769 million parameters. Log-mel spectrograms are extracted from speech signal and used as input to the CNN blocks, according to the Whisper architecture [6]. Both encoder and decoder blocks of Whisper are fine-tuned. We use SUBAK.KO train set to fine-tune both XLS-R and Whisper using CTC loss function. There exist standard HuggingFace implementations for wav2vec 2.0 and Whisper, which we make use of to run our experiments [40]. The hyper-parameter choices during training of self-supervised XLS-R, weakly supervised Whisper and fully supervised CNN are shown in Table 2

D. GREEDY DECODING

Instead of beam search decoding with language model rescoring, we only perform greedy decoding using connectionist temporal classification (CTC) to get the output e.g. character-level tokens [35]. This is done for fully supervised CNN, weakly supervised Whisper, and self-supervised wav2vec 2.0. While rescoring with the help of a language model achieves better performance [20], we would like to observe the WERs and CERs of the acoustic models without any interference from the language model. The goal of this study is to investigate the robustness of several ASR systems across multiple domains and evaluate SUBAK.KO on both read and spontaneous speech. For this, we do not require language model rescoring.

All the experiments are conducted using a single GPU with 24 gigabytes (GB) of VRAM and a central processing unit (CPU) with 24 cores.

V. RESULTS AND DISCUSSION

A. EVALUATING ROBUSTNESS USING BANSPEECH

We present the results of deep CNN, wav2vec 2.0, and Whisper models, either trained from scratch (CNN) or fine-tuned (wav2vec 2.0 and Whisper) on SUBAK.KO in

TABLE 3. CERs/WERs calculated for CNN, wav2vec 2.0 XLS-R (300M parameters), and Whisper (769M parameters) trained/fine-tuned on SUBAK.KO. 13 distinct domains from BanSpeech as well as the SUBAK.KO test set are used for evaluation. Out-of-vocabulary (OOV) rate is also provided for each domain. No external LM is used for rescoring. Wav2vec 2.0 outperforms the CNN and Whisper in recognizing the OOV words. All models demonstrate comparatively superior performance for *Audio books* and *biography* domains, which are primarily characterized by read speech. In contrast, other domains encompass more spontaneous speech and domain-specific vocabulary, hampering the ASR performance.

Domain/Dataset	Compared to SUBAK.KO train+val sets		CNN		wav2vec 2.0 XLS-R		Whisper (multilingual)	
	# of OOV words	OOV rate (%)	CER	WER	CER	WER	CER	WER
Television news	240	13.44	25.57	63.04	11.93	37.97	13.15	52.21
Parliament speech	228	14.38	40.44	75.97	14.06	40.55	14.77	56.37
Audio books	406	17.36	17.93	53.76	8.74	32.11	15.65	51.45
Drama series	288	17.69	41.72	79.79	19.07	48.56	17.31	55.00
Class lecture	213	16.54	28.82	65.52	12.46	37.72	17.64	51.65
Political talk show	238	13.24	27.99	64.82	13.50	40.41	12.41	57.82
Celebrity Interview	275	15.70	25.15	60.09	11.54	34.66	12.92	49.06
Documentary	262	14.78	24.93	61.20	10.19	32.48	14.05	51.93
Kids' voice	247	16.50	59.62	86.11	19.63	50.75	21.56	54.77
Medicine	446	34.47	34.85	74.96	14.51	42.40	18.72	63.86
Biography	318	18.29	12.79	43.07	7.78	28.25	10.76	56.18
Kids' Cartoon	247	17.16	48.55	81.46	23.65	54.17	28.26	66.86
Sports	377	21.13	32.06	69.40	17.30	47.37	13.30	61.67
BanSpeech	3585	30.83	31.84	67.65	**14.02**	**40.62**	15.94	55.92
SUBAK.KO test	918	5.24	4.02	13.94	**2.02**	**7.56**	4.96	22.15

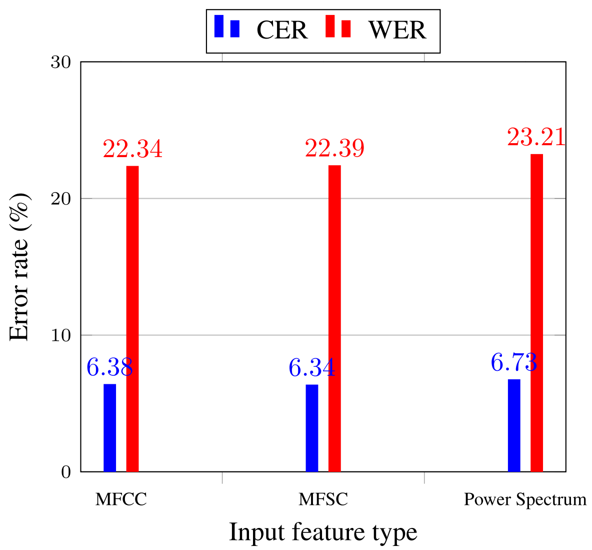

FIGURE 2. ASR performance evaluation in terms of CERs and WERs calculated using the SUBAK.KO test set for the three input feature extraction methodologies, namely MFCC, MFSC, and Power Spectrum. MFCC performs slightly better than the rest of the approaches.

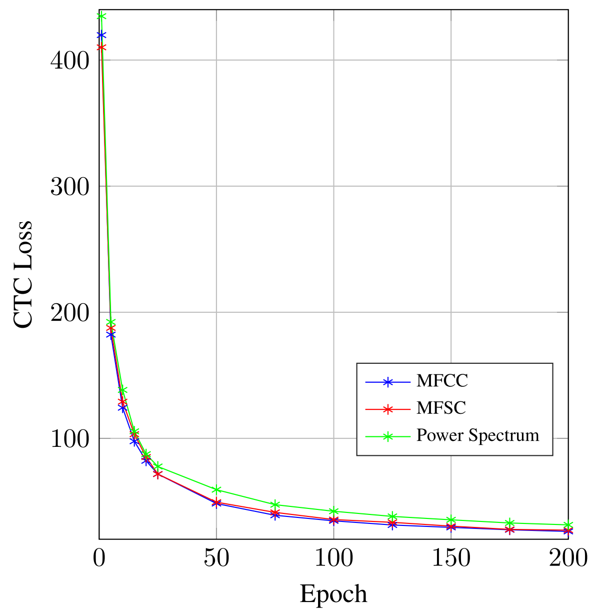

FIGURE 3. The CTC loss vs epoch computed on the SUBAK.KO dev set for MFCC, MFSC, and power spectrum features given input to the CNN. Similar trend in CTC loss decrease is observed for the three feature types.

Table 3. We use our best-performing CNN with 20 convolutional layers, layer normalization in each convolutional block, and 21 MFCCs extracted from each frame as input features. We show the performance across 13 domains from BanSpeech as well as using the SUBAK.KO test set. To evaluate the models, we use the most standard ASR metrics such as word error rate and character error rate. Moreover, we calculate the out-of-vocabulary (OOV) rates for each domain in reference to SUBAK.KO train and val sets. The OOV rate of a domain is calculated by dividing the vocabulary size of that domain by the total vocabulary of the SUBAK.KO train and dev sets. For example, the *audiobook* domain has a vocabulary of 2339, and the vocabulary of the SUBAK.KO train and dev sets are 37301. There are 406 words in that domain which are unavailable in SUBAK.KO. Dividing 406 (number of OOV words) by 2339 (vocabulary size) provides us with a 17.36% OOV rate.

For each of the domains, the OOV rate is greater than 10% which implies the difficulty of building a robust ASR model on the morphologically rich language Bangla. The *medicine* domain has the highest OOV rate because of the academic jargon used in this domain. BanSpeech has a much higher OOV rate (30.83%) than SUBAK.KO test set (5.24%), which

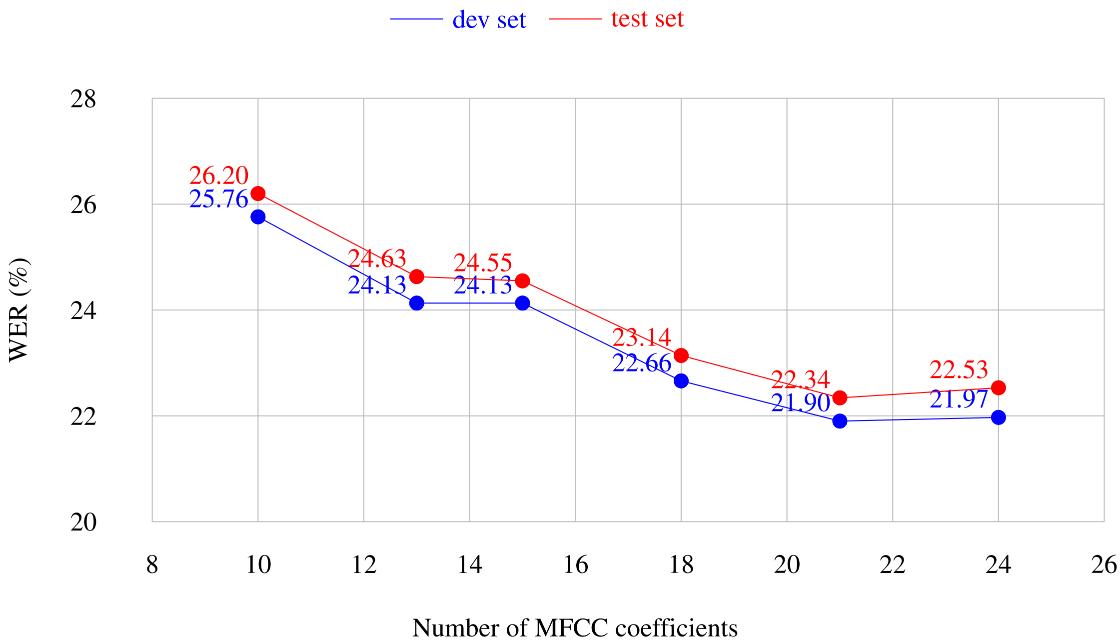

FIGURE 4. Impact of the number of MFCC coefficients on WER (%). SUBAK.KO dev and test sets are used for evaluation. CNN with 20 convolutional layers is used as the acoustic model. 21 MFCCs are found to obtain the lowest WER.

indicates that domain shifting introduces new vocabularies that are unseen during training.

For all the 13 domains, wav2vec 2.0 obtains substantially lower WERs compared to Whisper and our baseline CNN. Whisper performs better compared to CNN in each domain except for *biography*. Moreover, the difference in WERs between CNN and Whisper for the *audiobook* domain is marginal. It is worthwhile to mention that both *biography* and *audiobook* contain read speeches. From these findings, we can conclude that a self-supervised wav2vec 2.0 can recognize speeches from both read speech and spontaneous speech domains better than CNN and Whisper. A CNN, trained from scratch, is comparatively better with read speech samples, however, performs poorly on spontaneous speeches. The fact that CNN is trained solely with a read speech SUBAK.KO corpus might be the reason behind the poor performance in spontaneous speeches. This emphasizes the requirement for a spontaneous speech corpus for Bangla ASR. In the case of the Whisper that is pre-trained under weak supervision, the performance is not up to the mark in Bangla even after fine-tuning with 200 hours of high-quality supervised data.

On the BanSpeech benchmark, wav2vec 2.0 achieves 40.62% WER while CNN and Whisper obtain 67.65% and 55.92% WERs, respectively. wav2vec 2.0 also sets a benchmark on the SUBAK.KO test set getting a WER of 7.56%. CNN and Whisper get 13.94% and 22.15% WERs on the SUBAK.KO test set. We argue that since BanSpeech contains a large amount of spontaneous speech, it makes it a challenging evaluation set for ASR models. These results also prove the fact that when ASR models are evaluated on the test set split from the same corpus, the performance cannot be properly evaluated as the same domains/speakers are seen during training.

All the models struggle to recognize speeches from domains like *kids' voices, cartoons, medicine, drama,* and *sports*. These domains containing spontaneous speeches are not included in the SUBAK.KO as this dataset is mostly a read speech corpus.

B. EXPERIMENTS WITH CNN

Although a fully supervised CNN performs poorly on out-of-domain and spontaneous data, training a CNN is still a parameter-efficient way compared to fine-tuning large-scale pre-trained models such as wav2vec 2.0 and Whisper. CNN is applicable, especially for low-resource and mid-resource languages with limited availability of graphics processing units. We conduct several experiments with CNN which is described in this subsection.

Three input feature extraction methods from audio signals such as MFCC, MFSC, and Power Spectrum are implemented to build Bangla ASR models trained with our baseline deep CNN network and CERs/WERs are compared. Figure 2 illustrates the outcome of this experiment. MFCC technique gets a slightly lower WER of 22.34% compared to MFSC which obtains 22.39% WER. Comparing the CERs, however, MFSC gets a slight edge with 6.34% CER than MFCC with 6.38% CER. On the other hand, using Power Spectrum features, the ASR model performs worse, getting a WER of 23.21% and a CER of 6.73%. Figure 3 presents the CTC loss decreasing per epoch while training the CNN with MFCC, MFSC, and power spectrum features. We observe similar trend in the decrease of CTC loss for all the features.

TABLE 4. Impact of windowing on CERs and WERs calculated using an acoustic model trained with deep CNN and evaluated on the SUBAK.KO test set. A frame size of 30 and stride of 20 can reduce the CER/WER as well as the training time for a SUBAK.KO-based acoustic model.

Frame size (in ms)	Frame stride (in ms)	CER (%)	WER (%)	Training Time (in hours)
25	10	6.63	23.15	31.11
25	20	6.21	20.76	17.44
30	10	6.62	23.12	31.11
30	15	6.06	20.87	22.56
30	20	**6.04**	**20.51**	**17.44**
30	25	6.89	22.99	15.22

TABLE 5. The impact of training dataset size and the number of convolutional layers measured on CER/WER using SUBAK.KO test set. Deeper CNN model with 20 convolutional layers achieves lower CER/WER with more training data.

Hours of training data	CNN 10 Layers		CNN 15 Layers		CNN 20 Layers	
	CER	WER	CER	WER	CER	WER
40	18.17	54.30	16.45	47.94	18.67	52.32
80	11.84	38.72	10.81	34.16	11.14	34.07
120	9.43	31.57	8.09	26.13	7.95	25.03
160	8.13	27.66	6.79	22.02	6.33	19.94
200	7.27	25.04	5.86	19.23	5.61	**17.79**

TABLE 6. CER/WERs calculated on the SUBAK.KO test set for several normalization techniques. Here, B.N., W.N., and L.N. refer to batch normalization, weight normalization, and layer normalization, respectively. SUBAK.KO training set is used to develop the acoustic models. Layer normalization applied to CNN outperforms our baseline CNN as well as the rest of the normalization techniques.

Model	Test CER	Test WER
Baseline CNN	6.04	20.51
CNN+batch norm	5.83	19.78
CNN+weight norm	5.95	20.32
CNN+layer norm	**5.41**	**18.89**
CNN+B.N.+W.N.	5.96	20.14
CNN+L.N.+B.N.+W.N.	5.55	19.29

Since we get lower WER using MFCC features, we conduct further experiments with it by extracting different numbers of MFCCs from each frame. The WERs are calculated using SUBAK.KO dev and test sets and shown in Figure 4. We observe substantial variation in WERs while altering the number of MFCCs from 10 to 24 per frame. Initially, by increasing the MFCCs from 10 to 21, we can constantly improve the WERs from 26.20% to 22.34%. Taking 24 MFCCs per frame, however, slightly degrades the ASR performance with 22.53% WER. A similar phenomenon is observed on the dev set as well. These results suggest that a higher number of MFCCs e.g. 21 MFCCs generally ensures lower WER in the case of clean read speech training data.

Table 4 presents the impact of windowing on CERs and WERs for a deep CNN-based ASR model. As seen from Table 4, frame size and frame stride have a strong impact on the ASR performance as well as the computational cost. Setting the frame size to 30 milliseconds (ms) and stride to 20 ms provides the lowest WER of 20.51% and CER of 6.04%. This setting also reduces the training time to only 17.44 hours. Based on this experiment, we suggest experimenting with the frame sizes and strides to find the suitable setting for each dataset while developing an ASR system.

The performance of Neural Networks rests on the amount of data and the complexity of the networks/number of parameters. We split the SUBAK.KO train set and prepared 5 subsets containing 40 hours, 80 hours, 120 hours, 160 hours, and 200 hours of speech data. Using each of the subsets, we train three ASR models with 10, 15, and 20 convolutional blocks. The relation between training data size and the number of convolutional layers is shown in Table 5. In the case of 40 hours of training data, the CNN model with 15 convolutional layers outperforms the models with 10 layers and 20 layers, getting a WER of 47.94% and a CER of 16.45%. The advantage of the deep CNN with 20 convolutional layers becomes more apparent as we increase the amount of data. Using 200 hours of train data, CNNs with 10, 15, and 20 convolutional layers obtain 25.04%, 19.23%, and 17.79% WERs, respectively.

In Table 6, we provide the CERs and WERs on the SUBAK.KO test set for several types of normalization techniques integrated into our CNN model with 20 convolutional blocks. The normalization methods investigated in this study are batch norm [36], layer norm [37], and weight norm [38]. To implement that, the normalization method is applied to each convolutional layer, precisely after the dropout layer in each convolutional layer. Here, baseline CNN refers to the model where no normalization is applied. By adopting normalization, our model achieves lower WERs and CERs compared to our baseline. Using layer norm, we get 18.89% WER whereas applying batch norm and weight norm, 19.78% and 20.32% WERs can be obtained, respectively. While training a CNN with both batch norm and weight norm, our ASR model is able to get 20.14% WER. By building a model incorporating all of these normalization techniques (e.g. laye norm, batch norm, and weight norm), we achieve 19.29% WER. From this experiment, it can be concluded that CNN with layer norm outperforms the rest of the models.

VI. LIMITATION

The speech utterances for BanSpeech have been collected from publicly available sources using an automated pipeline, so any sort of bias such as gender bias, political bias, etc. cannot be identified and filtered. Yet, we strictly monitor the domains from which we extract our data to build this benchmark.

We collect speech data for the dialectal domain and use Google STT to transcribe them. In our initial evaluation, the quality of automated annotation of dialectal data is quite poor and we are not able to label them under human supervision as these are highly deviant from standard Bangla. These speeches, however, can be used for the language/dialect identification task.

All the domains contain approximately 30 minutes of annotated speech data. Due to resource constraints e.g., time, cost, and human labour, the size of these domains cannot be expanded. This study underscores the need for a multi-domain spontaneous speech training corpus in Bangla. Given the availability of resources, expanding the domains could be a promising future endeavor to create a training corpus covering various domains, predominantly featuring spontaneous speech. Nonetheless, the current size of BanSpeech is sufficient to be used as an evaluation benchmark.

wav2vec 2.0 XLS-R is pre-trained on a large amount of cross-lingual data. Although Whisper is also cross-lingual, it mostly contains English data and most of the other languages have less than 1000 hours of data [6]. If Whisper had more cross-lingual data, it would have been a better comparison between self-supervised and weakly supervised-based pre-training approaches. However, to the best of our knowledge, Whisper is the only large-scale weakly supervised pre-trained speech model as of now that can be compared to XLS-R. Moreover, both Whisper and XLS-R contain Bangla data in their pre-training dataset and so we make use of them in our work. Most importantly, we use the same training dataset from SUBAK.KO to fine-tune both models and evaluate them.

VII. CONCLUSION AND FUTURE SCOPE

In this work, we introduce a multi-domain Bangla ASR evaluation benchmark, named *BanSpeech*, consisting of 6.52 hours of human-annotated speech data and 8085 utterances from 13 distinct domains. This dataset has been utilized to evaluate three state-of-the-art ASR training strategies including full supervision, self-supervision, and weak supervision. We use SUBAK.KO, which was developed by collecting recording scripts from 40 text domains following the reception and production criteria set for text domains for training/fine-tuning. Through a comprehensive evaluation, we find that a self-supervised, pre-trained, cross-lingual wav2vec 2.0 is considerably robust for recognizing out-of-domain speech data. Although Whisper has seen only 1.3 hours of Bengali data during weakly supervised pre-training, the performance does not improve even after fine-tuning it with 200 hours of high-quality annotated speech data. A CNN, trained from scratch, is found to be performing quite poorly in terms of robustness and generalizability. Thus, transfer learning or self-supervised speech models, to be more specific, is the best strategy to handle challenging spontaneous, noisy, and multi-domain speech recognition.

The empirical evaluation also suggests that SUBAK.KO-based CNN can perform comparatively better in domains containing mostly read speech, such as *audiobooks*, and *biography*. On the other hand, this model is not well-trained for perceiving spontaneous speech, as seen by its high WERs in domains such as *drama, talk shows, sports,* etc. Although 40 text domains were considered while building the mostly read speech corpus SUBAK.KO, these are not sufficient to address the complexities of spontaneous speech. We emphasize the requirement of compiling a spontaneous Bangla speech corpus considering distinct domains such as *medicine, sports, children's voices,* etc. for better ASR performance.

This paper also reports the experimental results on feature extraction techniques, normalization methods integrated into the deep CNN architecture, and the impact of convolutional layers with respect to different training data sizes. We find that the MFCC features with a high number of coefficients, the layer normalization strategy, and the stacking of many convolutional blocks can improve the performance of CNN-based ASR.

We would like to build a high-quality spontaneous Bangla speech corpus consisting of diverse domains as part of our future work. Since self-supervised models are proven to be much better for ASR under challenging conditions, we aim to conduct more research into them. More specially, there does not exist any mono-lingual Bangla wav2vec 2.0 and so we would like to pre-train a large-scale Bangla wav2vec 2.0 using both read speech and spontaneous speech utterances and compare it with the XLS-R for numerous speech processing tasks such as ASR, language identification, keyword spotting, etc. We released BanSpeech and the CNN, wav2vec 2.0, and Whisper models for public use to foster research in Bangla.

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[3] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "SpeechStew: Simply mix all available speech recognition data to train one large neural network," 2021, *arXiv:2104.02133*.

[4] T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, "Rethinking evaluation in ASR: Are our models robust enough?" 2020, *arXiv:2010.11745*.

[5] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, ''Toward domain-invariant speech recognition via large scale training,'' in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 441–447.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, ''Robust speech recognition via large-scale weak supervision,'' in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.

[7] S. Kibria, A. M. Samin, M. H. Kobir, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, ''Bangladeshi Bangla speech corpus for automatic speech recognition research,'' *Speech Commun.*, vol. 136, pp. 84–97, Jan. 2022.

[8] S. Atkins, ''Corpus design criteria,'' *Literary Linguistic Comput.*, vol. 7, no. 1, pp. 1–16, Jan. 1992.

[9] O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, ''Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and Bangladeshi Bengali,'' in *Proc. 6th Workshop Spoken Lang. Technol. Under-Resourced Lang. (SLTU)*, sundanese, sinhala, Nepal, Aug. 2018, pp. 1–4.

[10] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, ''Shortcut learning in deep neural networks,'' *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[11] M. Nakamura, K. Iwano, and S. Furui, ''Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance,'' *Comput. Speech Lang.*, vol. 22, no. 2, pp. 171–184, Apr. 2008.

[12] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, ''XLS-R: Self-supervised cross-lingual speech representation learning at scale,'' 2021, *arXiv:2111.09296*.

[13] S. Sultana, M. Shahidur, and M. Zafar, ''Recent advancement in speech recognition for Bangla: A survey,'' *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 3, pp. 546–552, 2021.

[14] Md. A. A. Amin, Md. T. Islam, S. Kibria, and M. S. Rahman, ''Continuous Bengali speech recognition based on deep neural network,'' in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6.

[15] B. Das, S. Mandal, and P. Mitra, ''Bengali speech corpus for continuous auutomatic speech recognition system,'' in *Proc. Int. Conf. Speech Database Assessments*, Oct. 2011, pp. 51–55.

[16] S. Hosain Sumit, T. Al Muntasir, M. M. Arefin Zaman, R. Nath Nandi, and T. Sourov, ''Noise robust end-to-end speech recognition for Bangla language,'' in *Proc. Int. Conf. Bangla Speech Lang. Process. (ICBSLP)*, Sep. 2018, pp. 1–5.

[17] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, and G. Chen, ''Deep speech 2: End-to-end speech recognition in English and Mandarin,'' in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[18] A. Bills and J. Bishop. (2016). *Iarpa Babel Bengali Language Pack Iarpa-babel103b-v0.4b, 2016*. Accessed: Aug. 11, 2020. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016S08

[19] S. Ahmed, N. Sadeq, S. S. Shubha, M. N. Islam, M. A. Adnan, and M. Z. Islam, ''Preparation of Bangla speech corpus from publicly available audio & text,'' in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 6586–6592.

[20] A. M. Samin, M. H. Kobir, S. Kibria, and M. S. Rahman, ''Deep learning based large vocabulary continuous speech recognition of an under-resourced language Bangladeshi Bangla,'' *Acoust. Sci. Technol.*, vol. 42, no. 5, pp. 252–260, 2021.

[21] S. Kibria, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, ''Acoustic analysis of the speakers' variability for regional accent-affected pronunciation in Bangladeshi Bangla: A study on sylheti accent,'' *IEEE Access*, vol. 8, pp. 35200–35221, 2020.

[22] S. Alam, A. Sushmit, Z. Abdullah, S. Nakkhatra, M. Nazmuddoha Ansary, S. M. Hossen, S. M. Mehnaz, T. Reasat, and A. I. Humayun, ''Bengali common voice speech dataset for automatic speech recognition,'' 2022, *arXiv:2206.14053*.

[23] H. A. Z. S. Shahgir, K. S. Sayeed, and T. A. Zaman, ''Applying wav2vec2 for speech recognition on Bengali common voices dataset,'' 2022, *arXiv:2209.06581*.

[24] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, ''Unsupervised cross-lingual representation learning for speech recognition,'' 2020, *arXiv:2006.13979*.

[25] Z. Fan, M. Li, S. Zhou, and B. Xu, ''Exploring wav2vec 2.0 on speaker verification and language identification,'' 2020, *arXiv:2012.06185*.

[26] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, ''Multilingual speech translation with efficient finetuning of pretrained models,'' 2020, *arXiv:2010.12829*.

[27] N. San, M. Bartelds, M. Browne, L. Clifford, F. Gibson, J. Mansfield, D. Nash, J. Simpson, M. Turpin, M. Vollmer, S. Wilmoth, and D. Jurafsky, ''Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages,'' in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 1094–1101.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[29] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, ''Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training,'' 2021, *arXiv:2104.01027*.

[30] H. Tang, W.-N. Hsu, F. Grondin, and J. Glass, ''A study of enhancement, augmentation, and autoencoder methods for domain adaptation in distant speech recognition,'' 2018, *arXiv:1806.04841*.

[31] S. Khurana, N. Moritz, T. Hori, and J. L. Roux, ''Unsupervised domain adaptation for speech recognition via uncertainty driven self-training,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6553–6557.

[32] S. Sun, B. Zhang, L. Xie, and Y. Zhang, ''An unsupervised deep domain adaptation approach for robust speech recognition,'' *Neurocomputing*, vol. 257, pp. 79–87, Sep. 2017.

[33] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. van den Oord, ''Learning robust and multilingual speech representations,'' 2020, *arXiv:2001.11128*.

[34] (2023). *Google Speech-to-Text API A Google Cloud's AI & Machine Learning Product*. Accessed: Oct. 22, 2023. [Online]. Available: https://cloud.google.com/speech-to-text/

[35] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, ''Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,'' in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[36] S. Ioffe and C. Szegedy, ''Batch normalization: Accelerating deep network training by reducing internal covariate shift,'' in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[37] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, ''Layer normalization,'' 2016, *arXiv:1607.06450*.

[38] T. Salimans and D. P. Kingma, ''Weight normalization: A simple reparameterization to accelerate training of deep neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1–9.

[39] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, ''Wav2letter++: The fastest open-source speech recognition system,'' 2018, *arXiv:1812.07625*.

[40] T. Wolf et al., ''HuggingFace's transformers: State-of-the-art natural language processing,'' 2019, *arXiv:1910.03771*.

AHNAF MOZIB SAMIN received the B.Sc. degree in computer science and engineering from the Shahjalal University of Science and Technology, Bangladesh, in 2020, and the joint M.Sc. degree from the Erasmus Mundus Language and Communication Technologies (LCT) Program, University of Groningen, The Netherlands, and the University of Malta, Malta, in 2023.

His research interests include deep learning, speech processing, computational linguistics, and natural language processing. He is particularly interested in zero/low-resource language processing. He actively serves as a peer reviewer for several journals and conferences. He was a recipient of the Erasmus+ Scholarship funded by European Commission for the M.Sc. degree in two European countries. For more information, see https://www.ahnafsamin.com/.

M. HUMAYON KOBIR received the degree in computer science from the Shahjalal University of Science and Technology, in 2020. He is currently a Software Engineer with the Samsung Research and Development Institute. His professional focus encompasses the fascinating world of natural language processing (NLP), where he specializes in speech recognition, particularly for low-resource languages, such as Bengali. In addition to his interest in NLP, he is equally enthusiastic about image processing, deep neural networks, and large language models.

MD. MUSHTAQ SHAHRIYAR RAFEE was born in Thakurgaon, Bangladesh, in 1999. He received the B.Sc. degree in computer science and engineering from the Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh, in 2021. Since September 2021, he has been a Faculty Member with the Department of Computer Science and Engineering, Metropolitan University, Sylhet. He is currently researching on Bangla noisy speech dataset and Bangla multi-domain automatic speech recognition (ASR). His research interests include speech processing, computational linguistics, natural language processing, and machine learning.

M. FIROZ AHMED was born in Brahmanbaria, Bangladesh, in 1998. He received the B.Sc. degree in computer science and engineering from the Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh, in 2021. Since November 2021, he has been a Software Engineer at Vivasoft Ltd., a reputed Bangladeshi Company. He has also been working on advanced web development alongside researching the Bangla noisy speech dataset and Bangla multi-domain automatic speech recognition (ASR). His research interests include speech processing, computer vision, machine learning, and artificial intelligence.

MEHEDI HASAN received the bachelor's degree in computer science and engineering from the Shahjalal University of Science and Technology, Bangladesh. He is a dedicated IT professional. He is currently a Machine Learning Engineer with a prominent private firm (REVE Systems), Bangladesh, leveraging a robust foundation. With a passion for innovation and technology, he has amassed practical expertise in machine learning, deep learning, natural language processing, and computer vision.

PARTHA GHOSH was born in Dhaka, Bangladesh. He received the B.Sc. degree in computer science and engineering from the Shahjalal University of Science and Technology, Bangladesh. After completing the B.Sc. studies, he is currently a Machine Learning Engineer at REVE Systems, Dhaka.

SHAFKAT KIBRIA was born in Hawapara, Sylhet, Bangladesh, in 1978. He received the B.Sc. degree in computer science from East West University, Dhaka, in 2001, the M.Sc. degree in computer science from Umeå University, Umeå, Sweden, in 2005, and the Ph.D. degree in computer science and engineering from the Shahjalal University of Science and Technology (SUST), Sylhet, in 2023.

Since March 2020, he has been an Assistant Professor with the Department of Computer Science and Engineering, Leading University, Sylhet. His research interests include speech recognition, accent analysis, Bangla speech-to-text, Bangla natural language processing, human–robot interaction (HRI), machine learning systems (behavior-based robotics), mobile robotics (path planning, outdoor robotics), human–computer interaction (HCI), financial computation and analysis for decision making, and embedded system-based smart and ubiquitous environment. He has reviewed several journal articles as a peer reviewer.

M. SHAHIDUR RAHMAN (Senior Member, IEEE) was born in Jamalpur, Bangladesh, in 1975. He received the B.Sc. and M.Sc. degrees in electronics and computer science from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 1995 and 1997, respectively, and the Ph.D. degree in mathematical information systems from Saitama University, Saitama, Japan, in 2006.

In 1997, he joined the Shahjalal University of Science and Technology as a Junior Faculty, where he is currently a Professor. From 2009 to 2011, he was a JSPS Postdoctoral Research Fellow with Saitama University. His current research interests include speech analysis, speech synthesis, speech recognition, enhancement of bone-conducted speech, and digital signal processing.