Pseudo-Labeling for Domain-Agnostic Bangla Automatic Speech Recognition

Rabindra Nath Nandi[1], **Mehadi Hasan Menon**[1], **Tareq Al Muntasir**[1], **Sagor Sarker**[1], **Quazi Sarwar Muhtaseem**[1], **Md. Tariqul Islam**[1], **Shammur Absar Chowdhury**[2], **Firoj Alam**[2]
[1]Hishab Singapore Pte. Ltd, Singapore
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
rabindra.nandi@hishab.co, shchowdhury@hbku.edu.qa

Abstract

One of the major challenges for developing automatic speech recognition (ASR) for low-resource languages is the limited access to labeled data with domain-specific variations. In this study, we propose a pseudo-labeling approach to develop a large-scale domain-agnostic ASR dataset. With the proposed methodology, we developed a 20k+ hours labeled Bangla speech dataset covering diverse topics, speaking styles, dialects, noisy environments, and conversational scenarios. We then exploited the developed corpus to design a conformer-based ASR system. We benchmarked the trained ASR with publicly available datasets and compared it with other available models. To investigate the efficacy, we designed and developed a human-annotated domain-agnostic test set composed of news, telephony, and conversational data among others. Our results demonstrate the efficacy of the model trained on psuedo-label data for the designed test-set along with publicly-available Bangla datasets. The experimental resources will be publicly available.[1]

1 Introduction

Modern end-to-end automatic speech recognition (E2E-ASR) systems have made remarkable strides, performing well across various types of data (Li et al., 2020; Gulati et al., 2020; Chowdhury et al., 2021; Prabhavalkar et al., 2023). This success can be attributed to the advancement of deep learning techniques relying on different training strategies, highly dependent on large datasets. However, acquiring and maintaining these high-quality human transcriptions is both expensive and time-consuming, and hence hinders further progress for ASR especially in low-resource languages like Bangla.

To overcome these challenges, two dominant methods, leveraging unlabeled audio, are gaining popularity. These methods include: *(i)* pre-training via Self-supervised learning (SSL) (Baevski et al., 2020, 2022; Chung et al., 2021; Hsu et al., 2021); *(ii)* pseudo-labeling (PL) (Kahn et al., 2020; Xu et al., 2020b; Manohar et al., 2021; Zhu et al., 2023; Xu et al., 2020a; Higuchi et al., 2022). In the pre-training approach, the model is initially trained on raw unlabeled data and then fine-tuned using limited labeled data for some downstream ASR tasks. In pseudo-labeling, a pre-trained model generates labels for unlabeled data, which are then used alongside real labels for supervised ASR training. This paradigm is widely adopted due to its simplicity and effectiveness. Both SSL and PL have been shown to achieve competitive results with minimal labeled data, hence making these paradigms, especially PL, suitable for low-resource languages.

Despite being the 6^{th} most widely spoken language globally, Bangla still falls under low resource language family mainly due to the lack of accessible open datasets. To reduce this gap, we introduce a pseudo-labeling approach to develop an extensive, large-scale, and high-quality speech dataset of $\approx 20,000$ hours for developing domain-agnostic Bangla ASR. First, we curated and cleaned the largest collection of Bangla audio-video data from various Bangla TV channels on YouTube (YT) – varying domains, speaking styles, dialects, and communication channels among others. We then leverage the alignments from two ASR systems, to segment and automatically annotate the audio segments. We enrich the quality of pseudo-labels with our confidence and duration-based filtering method. We utilize the created dataset to design an end-to-end state-of-the-art Bangla ASR. Finally, we benchmark the ASR with widely used, domain-agnostic test sets and compare it with both publicly and commercially available Bangla ASR systems. To test domain-generalization capability, we also

[1]https://github.com/hishab-nlp/Pseudo-Labeling-for-Domain-Agnostic-Bangla-ASR

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 152–162

developed manually annotated test sets that include domain-diverse speech segments.
Our contributions are as follows:

- We develop and release **MegaBNSpeech** – the largest Bangla speech ($\approx 20,000$ hours) training corpus, alongside with its metadata;
- We introduce a robust data collection pipeline that systematically extracted audio segments from listed channels, ensuring wide coverage of speech samples;
- We developed and publicly released a domain-agnostic state-of-the-art Bangla ASR model;
- We developed two test sets comprising (a) diversified domain data from YT; and (b) real-life telephony conversational data, to test model generalizability across domains;
- We benchmark the proposed domain-agnostic Bangla ASR with publicly available test data and ASR models.

The rest of the paper is organized as follows: Section 2 presents previous work, Section 3 describes the dataset, Section 4 formulates our experiments, Section 5 discusses the evaluation results. Finally, Section 6 concludes and points to possible directions for future work.

2 Related Work

2.1 Speech Datasets Development

In the realm of speech corpus development, a variety of methods and techniques have been employed across multiple languages. For example, Wang et al. (2005) focused on Mandarin Chinese, creating a speech corpus from broadcast news and aligning the transcriptions. Similarly, Radeck-Arneth et al. (2015) curated data from diverse sources like audiobooks and web recordings to create a comprehensive speech corpus for German. In terms of automatic speech recognition datasets, Chui and Lai (2008) employed a method that constructs a Mandarin Chinese speech corpus using online videos and automated transcription. In a similar vein, Cho et al. (2021) harnessed web data and automatic alignment techniques to develop a Korean speech corpus geared toward speech recognition research.

Furthermore, current literature has also focused on specialized domains or applications. For instance, in the medical field, Cho et al. (2021) crafted a targeted speech corpus designed for medical dictation tasks, featuring recordings from healthcare professionals. Similarly, in the context of voice assistants, Gale et al. (2019) developed a corpus explicitly aimed at training and evaluating voice-controlled systems.

2.2 Speech datasets for Bangla

There have been several recent works for Bangla Speech Recognition. Sumit et al. (2018) proposed a deep learning based on approach and evaluated model on clean (Alam et al., 2010) and noisy speech datasets (Bills et al., 2016). Ahmed et al. (2020) developed a large annotated speech corpus comprising 960 hours, which are automatically curated from publicly accessible audio and text data. The data annotation primarily relies on publicly available audiobooks and TV news recordings from YouTube. It applies automated techniques such as format conversion, noise reduction, speaker diarization, and automatic gender detection. Transcriptions are generated iteratively using two speech recognition systems, with consensus determining accurate transcriptions. The resulting corpus, referred to as the 'Transcribed corpus', encompasses approximately 510 hours of data.

Similarly, Rakib et al. (2023a) created another extensive dataset with a focus on out-of-domain distribution generalization. The dataset is collected via crowdsourcing campaigns on the duration between Feb 2022 and Nov 2022 on the Mozilla Common Voice (MCV) platform. They followed two collection strategies: (i) scripted and (ii)spontaneous. The dataset contains 11.8k hours of training data curated from 22, 645 native Bangla speakers from South Asia. So far, this is the largest dataset available online for Bangla ASR Recognition. Kibria et al. (2022) also developed a speech corpus that includes 241 hours of both recorded and broadcast speech, featuring contributions from over 60 speakers.

Fleur's datasets are derived from the FLoRes-101 collection[2], which comprises 3,001 Wikipedia sentences. The authors translated development and training sentences from FLoRes-101 into 102 languages and annotated them for ASR applications. We extracted the Bangla test dataset, which includes 920 audio files totaling 3.43 hours. Fleur's dataset consists of 3,010 training, 920 testing, and

[2] https://huggingface.co/datasets/gsarti/flores_101

Datasets	Duration (Hours)	Source	Annotation
Fleurs (Conneau et al., 2023)	15.61	Wikipedia	Human
Common Voice13 (Ardila et al., 2020)	65.71	Open domain	Human
OpenSLR (Kjartansson et al., 2018)	229	Open domain	Human
Bengali Speech Corpus (Ahmed et al., 2020)	960	Youtube	Pseudo
OOD-Speech (Rakib et al., 2023a)	12K	Open domain	Human
MegaBNSpeech (Ours)	**19.8K**	YouTube	Pseudo

Table 1: A comparison of commonly used Bangla ASR datasets

402 validation audio files. We isolated the test files to evaluate them using our chosen models.

Common Voice is a comprehensive, multilingual ASR dataset. As of now, the dataset features 17,689 validated hours across 108 languages, with continual additions of new voices and languages (Ardila et al., 2020). The Common Voice 13 dataset includes 20.7k training, 9.23k testing, and 9.23k validation audio files. We also segregated the test files from this dataset for evaluation with our selected models.

The OpenSLR Bangla dataset, identified as OpenSLR-53, is a substantial ASR corpus sponsored by Google. It consists of a total of 232,537 recordings, amounting to 229 hours of audio data. For our evaluation purposes, we downloaded specific portions of this dataset and randomly selected 10,142 files, amounting to 10 hours of audio data.

Our introduced dataset surpasses all other available Bangla ASR datasets in terms of dataset size and annotation strategy, as outlined in Table 1. Compared to other methodologies, our data annotation pipeline is specialized in several crucial aspects. First, we focus on the manual curation of channels, allowing us to select content from reputable sources, thus enhancing both relevance and diversity. Second, our pipeline leverages both Hybrid ASR and Conformer ASR Models, which are potentially fine-tuned for Bangla, resulting in more accurate transcriptions. Finally, we have implemented a duplicate removal system to remove redundant content. These features make our data annotation process an excellent fit for applications that demand high-quality, domain-specific Bangla language resources.

3 Dataset

3.1 Data Collection

To develop a large-scale dataset focused on diverse domains, we selected YouTube as our data source due to its extensive coverage of Bangla speech. We gathered content from popular news channels such as ATN News, Banglavision News, ZEE 24 Ghanta, News18bangla, Republic Bangla, DD Bangla News, ABP Ananda, NTV News, DBC News, BBC News Bangla, Channel 24, mytvbd news, News24, and Channel I News, among others. Additionally, we included talk shows like RTV Talkshow and ATN Bangla Talk Show. We have also incorporated travel VLOGs into our dataset.

Crawler: To facilitate the collection of data from YouTube, we developed a web crawler that periodically collects videos using youtube-dl.[3] This crawler operates on a list of YouTube channels that we manually pre-select to ensure domain diversity. The crawler then lists all available videos from each channel and proceeds to download them. The download module within the crawler stores the downloaded videos in a Google Cloud Storage (GCS) bucket. The resulting collection consists of $\sim$53K hours with 42K number of videos.

Audio Extraction: We extracted audio from the videos, which were originally in Opus format. To ensure compatibility and standardization, we converted these Opus files to WAV format with a sampling rate of 16 kHz. The conversion process demanded the use of both high CPU and low memory resources. In Figure 1, we provide the data collection pipeline.

3.2 Pseudo Labeling

In Figure 2, we report the architecture of our proposed pseudo labeling approach for the ***MegaBNSpeech*** corpus development. The system takes audio files extracted from videos and passes them into two distinct in-house developed ASR systems:

- **Hybrid ASR (E_1):** Kaldi (Povey et al., 2011) based Factorized Time Delayed Neu-

[3] https://github.com/ytdl-org/youtube-dl

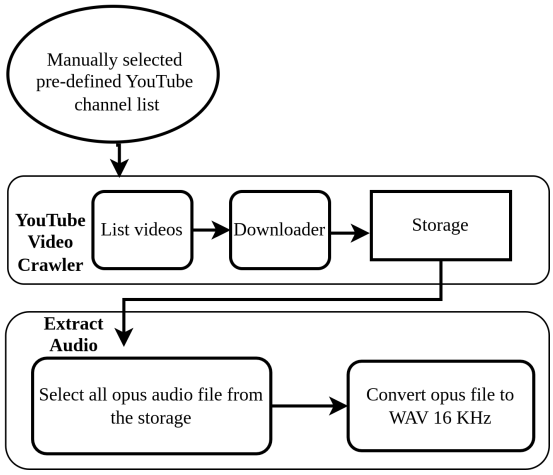

Figure 1: Data collection pipeline.

ral Network (TDNN) (Povey et al., 2018) model is used for training on 1.2K hours transcribed YouTube audio dataset which is manually collected. The model is called hybrid because firstly a Gaussian Mixture Model (GMM) is trained on speech acoustic features for phoneme level alignment and then DNN model is trained on the aligned features. During training, we use 15 factorized TDNN layers in model architecture and 4 epochs. The training recipe is available in the Kaldi Website.[4]

- **End-to-End Conformer ASR (E_2):** Nemo Tooklit based Conformer-CTC model (Gulati et al., 2020) is trained on 4k hours of transcribed YouTube data. A byte-pair encoding (BPE) tokenizer (Wang et al., 2005) is first built using the transcripts of the train set. At training time, pretrained weights of Nemo English ASR [5] are used for initializing weights of the encoder part only. The training parameters are epochs 16, batch size 32, sampling rate 16kHz, use_start_end_token TRUE, pin memory TRUE, number_of_workers 48, trim_silence False, max duration 18.5 and min duration 0.2. The training script is customized from the following the script. [6]

The objective was to leverage the capabilities of these ASR systems to generate transcription based on their decisions. We use the term expert to refer to these systems.

As part of our proposed pseudo-labeling approach, we consider them as expert systems. Based on the transcripts they generate, we take their decisions on segments that match, as depicted in Figure 2. To formally define this, we have two expert systems E_1 and E_2, each of which generates transcripts $T1$ and $T2$, respectively. We use a matching algorithm, Algorithm 1, that employs exact string matching to align the text of segments from the experts E_1 and E_2 ASR systems. The next step involves segmenting the audio based on matching text and removing the segments that do not match. For example, the words highlighted in red in Figure 2 indicate mismatched segments. We therefore remove these segments. The subsequent step is to filter out segments based on predefined criteria. These include: *(i)* confidence score of the ASR systems, *(ii)* minimum and maximum duration of the segments, *(iii)* the ratio of segment duration to the number of words, and *(iv)* the minimum number of words required in a segment. These steps resulted in the final MegaBNSpeech corpus.

Algorithm 1 Transcription matching algorithm.

$T_1 \leftarrow$ Kaldi model (E1)
$T_2 \leftarrow$ Conformer CTC model (E2)
for each (t_1, t_2) in zip (T_1, T_2) **do**
  $\mathcal{M} \leftarrow f(t_1, t_2)$
  for each m in $\mathcal{M}$ **do**
    $r_w \leftarrow$ word rate of m
    $d_a \leftarrow$ segment duration of m
    $c_t \leftarrow$ total characters in m
    $w_t \leftarrow$ total words in m
    if $r_w < r_{w,\text{min}}$ **or** $r_w > r_{w,\text{max}}$ **or** $d_a < d_{a,\text{min}}$ **or** $d_a > d_{a,\text{max}}$ **or** $c_t < c_{t,\text{min}}$ **or** $w_t < w_{t,\text{min}}$ **then**
      continue
    end if
    Write matched transcript and segment
  end for
end for

where $r_{w,\text{min}}$ and $r_{w,\text{max}}$ refers to minimum and maximum word rate; $d_{a,\text{min}}$ and $d_{a,\text{max}}$ refers to minimum and maximum segment duration; $c_{t,\text{min}}$ refers to minimum number of characters, and $w_{t,\text{min}}$ refers to minimum number of total words; $f(t_1, t_2)$ is the longest substring matching function.

[4] https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/local/chain/tuning/run_tdnn_1d.sh

[5] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_medium

[6] https://github.com/NVIDIA/NeMo/blob/main/examples/asr/conf/conformer/conformer_ctc_bpe.yaml

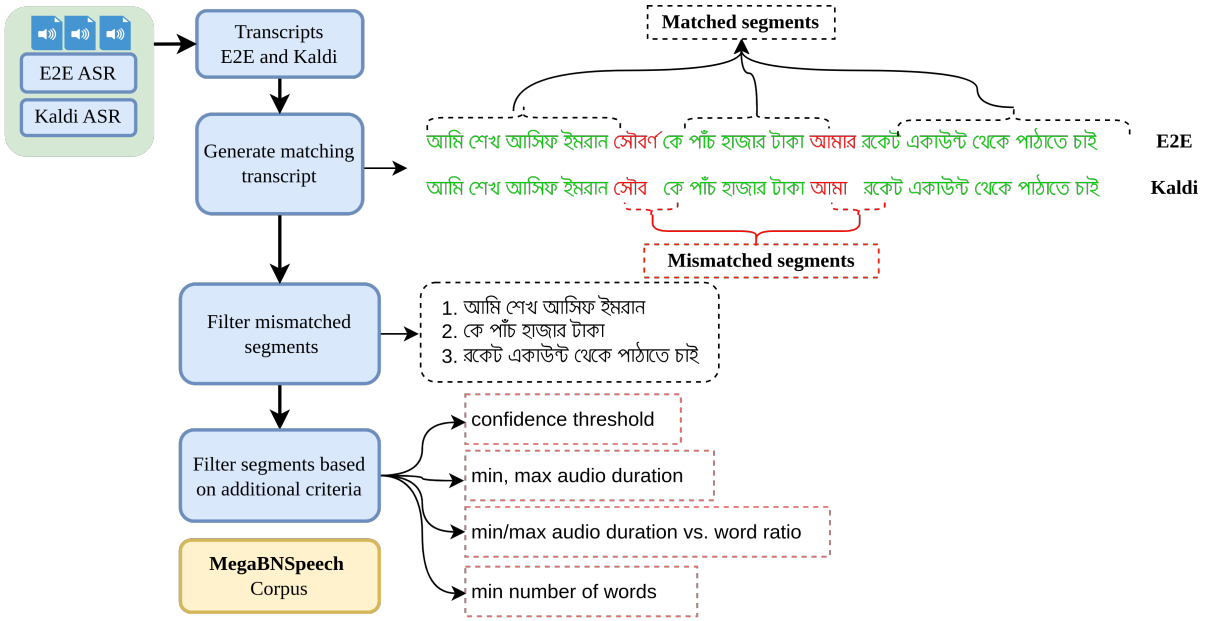

Figure 2: Architecture of the proposed pseudo labeling approach.

3.3 Metadata

To ensure both reproducibility and transferability, we store the metadata in JSON format. This metadata includes the following key elements: *(i)* audio_filepath, *(ii)* text, and *(iii)* duration. The audio_filepath field specifies the path to the audio file, with channel information embedded in the filename. The text field contains the data generated by the pseudo-labeling pipeline, while the duration field indicates the length of the audio in seconds. The audio files have a sampling rate of 16 kHz.

4 Experiments

4.1 Data splits

Training set For training the model, the dataset we selected comprises 17.64k hours of news channel content, 688.82 hours of talk shows, 0.02 hours of vlogs, and 4.08 hours of crime shows. Table 2 provides detailed information about each category and its corresponding duration in hours.

Channels Category	Hours
News	17,640.00
Talkshow	688.82
Vlog	0.02
Crime Show	4.08
Total	**18,332.92**

Table 2: Training data distribution according to channel category and hours

Development set To investigate the robustness of the pseudo-labeling approach, we randomly selected 10 hours of speech to create a development set.

Test set To evaluate the performance of the models, we used four test sets. Two of these were developed as part of the MegaBNSpeech corpus, while the remaining two (Fleurs and Common Voice) are commonly used test sets that are widely recognized by the speech community.

- **MegaBNSpeech-YT Test Set :** The test set has been prepared from a recent collection of YouTube videos, resulting in 8 hours of data. This set is manually transcribed for evaluation purposes. The domains of this set include News, Talkshow, Courses, Drama, Science, Waz (Islamic preaching), etc.
- **MegaBNSpeech-Tele Test Set:** To assess the model's generalization capabilities, we also included 1.9 hours of telephony conversations from our in-house dataset collection, which were subsequently manually transcribed. It involves telephone conversations covering various discussion topics, including online food orders, health services, online ticket bookings, and online banking. The calls were originally recorded using 8kHz sampling rate, which we then upsampled to 16kHz to match the ASR input.[7]

[7]The curated telephony dataset is open-ended conversa-

- **Fleurs:** Fleur's (Conneau et al., 2023) datasets are from FLoRes-101 datasets[8] which contain 3001 Wikipedia sentences. The authors translated dev and train sentences from FLoRes-101 to 102 languages and annotated them to develop ASR. We have separated the Bangla test datasets which contain 920 audio files with 3.43 hours of data. Fleurs contains a total of 3,010 train, 920 test, and 402 validation audio files. We separated the test datasets and evaluated them with our selected models.

- **Common Voice:** Common voice (Ardila et al., 2020) is a massively multilingual ASR dataset. The dataset currently consists of 17,689 validated hours in 108 languages, but more voices and languages are always added. Common Voice 13 contains a total of 20.7k train, 9.23k of test, and 9.23k of validation audio files. We separated the test datasets and evaluated them with our selected models.

4.2 Contemporary ASR Models

Google: Google speech-to-text[9] is a cloud-based ASR service that provides transcription from input Audio for several languages. It provides different domain-specific models for task-specific ASR services. We used the default model and settings and set the language to Bangla.

MMS: Massively Multilingual Speech (MMS (Pratap et al., 2023)) is a fine-tuned model developed by Meta. This model is based on the Wav2Vec2 (Baevski et al., 2020) architecture and makes use of adapter models to transcribe 1000+ languages. The model consists of 1 billion parameters and has been fine-tuned in 1,162 languages. The model checkpoint is published in the HuggingFace model hub.[10]

OOD-speech ASR: OOD-speech ASR is a Conformer-CTC-based model trained on ODD speech datasets (Rakib et al., 2023b). The model consists of 121 million parameters and is trained on 1,100+ hours of audio data which is crowdsourced from native Bangla speakers. The model was trained using NVIDIA NeMo[11] framework and published in Huggingface model hub.[12]

Parameter	Value
epoch	15
global_step	90,911
learning_rate	0.000073287
train_backward_timing	0.1630282
train_loss	11.203718
training_batch_wer	0.149231
val_loss	27.58967
val_wer	0.203385
validation_step_timing	0.089399

Table 3: Details of the hyperparameter settings.

4.3 MegaBNSpeech ASR

We trained the FastConformer model (Rekesh et al., 2023) using the full 18k MegaBNSpeech training sets. During the training phase, we employed a set of predefined parameters: a learning rate of 0.5, a weight decay of 0.001, a batch size of 32, AdamW optimizer, and a maximum audio duration of 15 seconds. We provide details of the hyperparameter settings in Table 3.

To optimize the performance of our model, we conducted experiments with various NVIDIA NeMo architectures and assessed their training accuracy. Specifically, we evaluated the Conformer-CTC, Conformer-Transducer, and Fast-Conformer models. Among these, the Conformer-CTC model exhibited the best performance, achieving a training loss of approximately 11.2%.

To accelerate the training process, we deployed a total of 16 $A100 - 40G$ GPUs to handle the entire dataset. Despite leveraging significant computational resources, the training still took approximately 112 hours to complete.

The model underwent training for 15 epochs, completing approximately 90,911 global steps. The chosen learning rate was relatively low, contributing to stable and incremental updates of the model's parameters. Although the training loss suggests potential for further improvement, it does indicate a narrowing gap between predicted and actual values during the training phase.

As for the WER the value indicates that our

tions with pre-defined topics and includes consent from the interlocutors.

[8] https://huggingface.co/datasets/gsarti/flores_101

[9] https://cloud.google.com/speech-to-text

[10] https://huggingface.co/facebook/mms-1b-all

[11] https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/asr/examples/kinyarwanda_asr.html

[12] https://huggingface.co/bengaliAI/BanglaConformer

model performed with commendable accuracy. However, the validation loss remains somewhat elevated. These metrics offer valuable insights into the model's performance and serve as a road map for future optimization efforts.

4.4 Data Post-processing

During the evaluation of the test sets, we apply a set of post-processing on predicted transcription and human annotation to reduce unexpected symbols, confused words, and misleading alignment. We find that there are some typing issues during manual labeling. To resolve this, a typing error minimization function is applied. In addition, we added two common normalization rules including: (i) number-to-word conversation and (ii) punctuation removal.

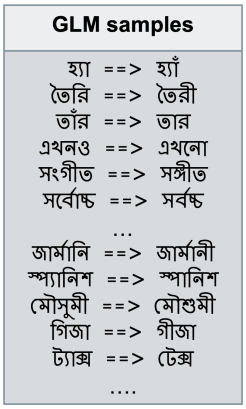

Figure 3: Sample of GLM entries.

Minimizing the confusion due to writing style An extensive analysis of transcriptions indicates many words have different forms of writing (as shown in Figure 3) based on different character combinations. In some cases, both words of confused pairs are acceptable as people annotated in different ways, especially for country names, along with borrowed or code-mixed words.

To minimize these differences, we created a simple Global Mapping File (GLM) that allows different variations of the word to be accepted during evaluation. The GLM file contains entries for different homophones, primarily those with spelling variations. We employed the most frequently occurring confusion patterns for the task, although this approach may not cover all possible variations.

4.5 Evaluation Metrics

To evaluate the performance of the models, we used widely accepted metrics such as Word Error Rate (WER) and Character Error Rate (CER). The reported WER values are presented using the GLM and postprocessing of the hypothesis and references.

5 Results and Discussion

5.1 Robustness of Pseudo-labelling

We first evaluate the robustness of our annotation process for unlabeled audio by utilizing our proposed pseudo-labeling approach. To investigate the quality of these annotations, we used the development set mentioned earlier. This set was subsequently annotated by a human annotator who had no prior knowledge of the ASR-generated pseudo-labels. We then computed the Word Error Rate (WER) and Character Error Rate (CER) between these pseudo-labels (serving as predictions) and the human annotations (acting as ground truth). We observed WER and CER rates of less than 3% (specifically, 2.89% for WER and 2.27% for CER), thereby increasing our confidence in the reliability of the pseudo-labeled datasets.

5.2 Effectiveness of MegaBNSpeech ASR

We initially assess the performance of MegaBNSpeech ASR, which is fully trained on a pseudo-labeled dataset, and compare its ASR performance against other systems such as Google, MMS, and OOD-speech ASRs. Utilizing our in-domain test set (MegaBNSpeech-YT), we noticed a significant performance gap; MegaBNSpeech ASR outperformed the commercial Google ASR, which itself was notably better than the rest (see Table 4).

One plausible explanation for MegaBNSpeech's high performance could be the nature of its training data, which is predominantly sourced from News and Talkshow segments, followed by Science content. These sources typically feature formal speaking styles and limited linguistic diversity, thereby contributing to improved performance. This hypothesis is further supported by the category-level performance data, especially within the 'News' category, as indicated in Table 5.

Across different categories: In Table 5, we report the WER for each category within the MegaBNSpeech-YT test set. From the table, it is evident that all the ASRs (except MMS) perform exceptionally well in the broadcast domain, specifically in News, with MegaBNSpeech achieving nearly 98% accuracy. In the case of talk shows – a

Category	Duration(hr)	MegaBNSpeech	Google	MMS	OOD-speech
MegaBNSpeech-YT	8.1	6.4/3.39	28.3/18.88	51.1/23.49	44.4/33.43
MegaBNSpeech-Tel	1.9	*40.7/24.38	*59/41.26	*76.8/39.36	*69.9/52.93
Fleurs	3.42	*36.1/8.43	24.6/8.54	*39.4/11.58	29.5/13.97
Common Voice	16.5	*42.3/11.44	23.6/ 8.31	*48/14.72	23.6/10.49

Table 4: Reported Word error rate (WER) /character error rate (CER) on four test sets using four ASR systems. * represent the training portion of the corresponding test set **was not** present in the ASR model.

Category	Duration(hr)	MegaBNSpeech ASR	Google ASR	MMS ASR	OOD-speech
News	1.21	2.5/1.21	18.9/10.46	52.2/21.65	32.3/20.71
Talkshow	1.39	6/3.29	28/18.71	48.8/21.5	45.8/34.59
Courses	3.81	6.8/3.79	30.8/21.64	50.2/23.52	46/35.99
Drama	0.03	10.3/7.47	37.3/27.43	64.3/32.74	53.6/45.14
Science	0.26	5/1.92	20.6/11.4	45.3/19.93	33.4/23.11
Vlog	0.18	11.3/6.69	33/22.9	57.9/27.18	49.3/37.22
Recipie	0.58	7.5/3.29	26.4/16.6	53.3/26.89	41.2/29.39
Waz	0.49	9.6/5.45	33.3/23.1	57.3/27.46	59.9/50.38
Movie	0.1	8/4.64	35.2/23.88	64.4/34.96	50.9/42.13

Table 5: Reported Word error rate (WER) /character error rate (CER) on different categories present in MegaBNSpeech - YT test set for four different ASR systems.

synchronized conversational setup – both MegaBNSpeech and Google significantly outperform MMS and OOD-speech. This trend is observed across almost all the categories.

5.3 Generalization Capability to unknown Dataset and Channel

Dataset: To understand how the model performs in unknown domains or datasets, we evaluated the four ASRs using the widely used Fleurs and Common Voice test sets. As seen in Table 4, MegaBNSpeech performs slightly better than MMS ASR on both Fleurs and Common Voice test sets, even though these two datasets are unfamiliar to both MMS and MegaBNSpeech ASR. On the other hand, Google and OOD-speech perform significantly well, with a Word Error Rate (WER) in the range of 23-29%. It should be noted, however, that OOD-speech ASR has been trained on Common Voice data – a crowdsourced dataset where the text prompts are randomly selected from Wikipedia, making it similar to Fleurs. Therefore, the content and style of these datasets are not entirely unknown to these models.

Telephony Channel: To assess how ASR models perform not just in unfamiliar domains but also across different communication channels,[13] we evaluated these four models using telephony conversational data, as shown in MegaBNSpeech-Tel Table 4. Our results indicate that MegaBNSpeech ASR significantly outperforms all other ASRs, with Google coming in second place. This level of performance is consistent with our earlier observations that MegaBNSpeech ASR excels in conversation-style categories like talk shows and vlogs.

5.4 Key Points: Psuedo-labelling based ASR vs Fully-supervised ASR

Traditional ASR training relies heavily on extensive labeled datasets, a requirement that becomes both challenging and expensive to meet for languages, dialects, and domains with limited resources. In contrast, pseudo-labeling not only enriches the training data but also diversifies domain-specific variations, as demonstrated in this study.

From our analysis, we found that MegaBNSpeech performs comparably to supervised out-of-domain (OOD) speech ASR systems, even when exposed to data or domains it has not previously encountered. This shows the efficacy of pseudo-labeling as well as the potential of both the MegaBNSpeech datasets and the model. In this study, we

[13]The collected data was upsampled from an 8K to a 16K sampling rate to match the input sampling rates of the models.

trained MegaBNSpeech exclusively with pseudo-labels to demonstrate the impact of this automated labeling technique. In practical applications, supplementing pseudo-labels with a small amount of manually annotated data can further enhance ASR performance while leveraging the model's strong generalization capabilities.

6 Conclusion and Future Work

This study offers a significant contribution in Bangla speech processing, in addition to the field of ASR particularly for low-resource language. The primary contribution of this paper lies in demonstrating that the model trained with pseudo-labeling only, offers comparable performance with supervised ASR systems. Specifically, the MegaBNSpeech model excels in their ability to generalize across multiple domains and channels as shown in the results.

Additionally, the developed train, development, and two test sets of MegaBNSpeech corpus of $\approx 20,000$ hours of data will serve as a valuable resource for the research community. The MegaBNSpeech corpus, especially the manually annotated YT and telephony test sets, can be used as a benchmark for future studies, enabling other researchers to build upon our work and potentially discover even more effective methods for designing low-resource ASR.

Acknowledgments

We are grateful to HISHAB[14] for providing us with all the necessary working facilities, computational resources, and an appropriate environment throughout our entire work.

7 Limitations

Our data collection originated from YouTube and in-house telephony conversations. Due to restrictions on sharing most of the YouTube content directly, we will instead release links to the YouTube videos along with their transcriptions.

References

Shafayat Ahmed, Nafis Sadeq, Sudipta Saha Shubha, Md Nahidul Islam, Muhammad Abdullah Adnan, and Mohammad Zuberul Islam. 2020. Preparation of bangla speech corpus from publicly available audio & text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6586–6592.

Firoj Alam, S. M. Murtoza Habib, Dil Afroza Sultana, and Mumit Khan. 2010. Development of annotated bangla speech corpora. In *Spoken Language Technologies for Under-resourced languages (SLTU'10)*, volume 1, pages 35–41, Penang, Malaysia.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Aric Bills, Judith Bishop, Anne David, Eyal Dubinski, Jonathan G. Fiscus, Breanna Gillies, Mary Harper, Amy Jarrett, María Encarnación Pérez Molina, Anton Rytting Jessica Ray, Shelley Paget, Wade Shen, Ronnie Silber, Evelyne Tzoukermann, and Jamie Wong. 2016. IARPA babel bengali language pack iarpa-babel103b-v0.4b ldc2016s08.

Won Ik Cho, Seok Min Kim, Hyunchang Cho, and Nam Soo Kim. 2021. Kosp2e: Korean speech to english translation corpus. *arXiv preprint arXiv:2107.02875*.

S. Chowdhury, A. Hussein, A. Abdelali, and A. Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic Asr. *Interspeech 2021*.

Kawai Chui and Huei-ling Lai. 2008. The nccu corpus of spoken chinese: Mandarin, hakka, and southern min. *Taiwan Journal of Linguistics*, 6(2).

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

[14] https://hishab.co/

Robert Gale, Liu Chen, Jill Dolata, Jan Van Santen, and Meysam Asgari. 2019. Improving asr systems for children with autism and language impairment using domain-focused dnn transfer techniques. In *Interspeech*, volume 2019, page 11. NIH Public Access.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori. 2022. Momentum pseudo-labeling: Semi-supervised asr with continuously improving pseudo-labels. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1424–1438.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jacob Kahn, Ann Lee, and Awni Hannun. 2020. Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7084–7088. IEEE.

Shafkat Kibria, Ahnaf Mozib Samin, M Humayon Kobir, M Shahidur Rahman, M Reza Selim, and M Zafar Iqbal. 2022. Bangladeshi bangla speech corpus for automatic speech recognition research. *Speech Communication*, 136:84–97.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali.

Jinyu Li, Rui Zhao, Zhong Meng, Yanqing Liu, Wenning Wei, Sarangarajan Parthasarathy, Vadim Mazalov, Zhenghao Wang, Lei He, Sheng Zhao, et al. 2020. Developing rnn-t models surpassing high-performance hybrid models with customization capability. *arXiv preprint arXiv:2007.15188*.

Vimal Manohar, Tatiana Likhomanenko, Qiantong Xu, Wei-Ning Hsu, Ronan Collobert, Yatharth Saraf, Geoffrey Zweig, and Abdelrahman Mohamed. 2021. Kaizen: Continuously improving teacher using exponential moving average for semi-supervised speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 518–525. IEEE.

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *arXiv preprint arXiv:2303.03329*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Stephan Radeck-Arneth, Benjamin Milde, Arvid Lange, Evandro Gouvêa, Stefan Radomski, Max Mühlhäuser, and Chris Biemann. 2015. Open source german distant speech recognition: Corpus and acoustic model. In *Text, Speech, and Dialogue: 18th International Conference, TSD 2015, Pilsen, Czech Republic, September 14-17, 2015, Proceedings 18*, pages 480–488. Springer.

Fazle Rabbi Rakib, Souhardya Saha Dip, Samiul Alam, Nazia Tasnim, Md Istiak Hossain Shihab, Md Nazmuddoha Ansary, Syed Mobassir Hossen, Marsia Haque Meghla, Mamunur Mamun, Farig Sadeque, et al. 2023a. Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking. *arXiv preprint arXiv:2305.09688*.

Mohammed Rakib, Md Ismail Hossain, Nabeel Mohammed, and Fuad Rahman. 2023b. Bangla-Wave: Improving bangla automatic speech recognition utilizing n-gram language models. In *Proceedings of the 2023 12th International Conference on Software and Computer Applications*, pages 297–301.

Dima Rekesh, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Juang, Oleksii Hrinchuk, Ankur Kumar, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *arXiv preprint arXiv:2305.05084*.

Sakhawat Hosain Sumit, Tareq Al Muntasir, MM Arefin Zaman, Rabindra Nath Nandi, and Tanvir Sourov. 2018. Noise robust end-to-end speech recognition for bangla language. In *2018 international conference on bangla speech and language processing (ICBSLP)*, pages 1–5. IEEE.

Hsin-Min Wang, Berlin Chen, Jen-Wei Kuo, and Shih-Sian Cheng. 2005. Matbn: A mandarin chinese broadcast news corpus. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 2, June 2005: Special Issue on Annotated Speech Corpora*, pages 219–236.

Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. 2020a. Iterative pseudo-labeling for speech recognition. *arXiv preprint arXiv:2005.09267*.

Tatiana Likhomanenko Qiantong Xu, Jacob Kahn, and Gabriel Synnaeve Ronan Collobert. 2020b. Slimipl: Language-model-free iterative pseudo-labeling. *arXiv preprint arXiv:2010.11524*.

Han Zhu, Dongji Gao, Gaofeng Cheng, Daniel Povey, Pengyuan Zhang, and Yonghong Yan. 2023. Alternative pseudo-labeling for semi-supervised automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.