International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.146, October 2026

ConvBangla-115: A Diverse Conversational Speech Corpus and Fine-Tuned Whisper Model for Low-Resource Bangla Speech Recognition

Md. Shadmim Hasan Sifat
Department of Computer Science and Engineering,
Shahjalal University of Science and Technology,
Sylhet, Bangladesh

Mehedi Hasan
Department of Computer Science and Engineering,
Shahjalal University of Science and Technology,
Sylhet, Bangladesh

Md. Ataur Rahman
Department of Computer Science and Engineering,
Shahjalal University of Science and Technology,
Sylhet, Bangladesh

Mahruba Sharmin Chowdhury
Department of Computer Science and Engineering,
Shahjalal University of Science and Technology,
Sylhet, Bangladesh

ABSTRACT

Bangla is used by over 300 million people, but its automated speech-to-text (STT) technology is still one of the least mature. Factors such as lack of high-quality conversational datasets, usage of models trained mostly on scripted speech, and insufficient coverage of dialectal diversity and demographics hinder development. Consequently, current Bangla STT models are only highly accurate under certain conditions and become less accurate in natural conversational setting. In order to overcome this limitation, this work presents ConvBangla-115, a 115.9 hours Bangla conversational speech dataset that consists of spontaneous utterances made by speakers aged from 10 to 50 years old in regions like Dhaka, Rajshahi, Rangpur, and Khulna under diverse acoustic conditions with recorded transcription verifications. Using this dataset, five variations of OpenAI's Whisper model are fine-tuned on training subsets of 5, 10, 50, and 115.9 hours and the results are investigated across three angles: the effect of fine-tuning, amount of training data, and the model size. The fine-tuning yields positive results with all configurations; Whisper large-v3 (Turbo) achieves WER of 19.60% $\pm$ 0.3 with a relative improvement of 60.8% from zero-shot WER of 50.00%, where all observed differences are statistically significant ($p < 0.001$). Besides the main indicator, the results show that WER follows the power law of dependence on the number of training hours; the value of the exponent rises as the model becomes larger (from 0.08 in Tiny to 0.25 in Medium); the first 5 hours cover 49% of the total gains for Tiny but only 17% for Medium; and the model which is of one size larger and fine-tuned for 50 hours performs not worse or better than the model of smaller size trained on 115.9 hours. The paper offers a speech corpus in Bangla, a fine-tuning study in various model sizes and amounts of training data, and recommendations for speech recognition in low-resource scenarios.

General Terms

Speech Recognition, Machine Learning, Human Language Technology

Keywords

Bangla speech-to-text, model fine-tuning, OpenAI Whisper model, dialectal variability, conversational speech recognition

1. INTRODUCTION

Bangla is a language spoken by over 300 million people, but it falls behind in the creation of precise speech-to-text (STT) systems [15]. This lack of technology makes digital content less accessible, makes educational tools difficult to develop, and does not allow many people to use voice-activated systems [29]. The primary issue here is the unavailability of large and diverse speech corpora; current corpora include mostly scripted speech. Therefore, diversity of the Bangla conversational corpus and training of the models on such data solves issues of dataset availability, model robustness, and practical relevance. The novelty of this research is that the authors developed both the dataset and the model evaluation simultaneously. Moving from scripted speech, the researchers collected 115.9 hours of conversational Bangla speech and fine-tuned OpenAI's Whisper at different scales, discovering the relationship between the scale of the model, the amount of training data, and the accuracy [23]. The best model had a word error rate (WER) of 19.60% $\pm$ 0.3, which is 60.8% better in comparison with the zero-shot Turbo baseline ($p < 0.001$). The contributions are:

(1) **Conversational dataset:** The 115.9-hour Bangla conversational dataset for public distribution targets speakers

aged 10-50 years, consists of four divisions, and contains recordings from different settings.

(2) **High-performance model with statistical validation:** The performance of the fine-tuned Whisper large-v3 model is shown to be 19.6% $\pm$ 0.3 WER, verified by using bootstrapped confidence intervals, pair testing, and effect sizes.

(3) **Systematic evaluation:** Five sizes of Whisper model (Tiny, Base, Small, Medium, Turbo) have been tested using four amounts of training data (5, 10, 50, and 115.9 hours).

(4) **Quantitative scaling analysis:** Power-law scaling of WER against training hours is modeled and analyzed, including gains per stage and per doubling and also the data vs. capacity trade-off for other low-resource languages.

A low-error Bangla speech-to-text (STT) system can assist the visually impaired and illiterate communities and enable access to commercial, healthcare, and education services through voice interfaces; the method and dataset are generalizable to other low-resource languages.

2. LITERATURE REVIEW

Bangla STT has moved from rule-based isolated-word systems to deep learning methods and now to large pre-trained transformer models.

2.1 Early Approaches to Bangla STT

Early Bangla speech recognition focused on isolated words using handcrafted features with Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) [24]. Hasnat et al. built an HTK-based recognizer reaching 70% (isolated) and 60% (continuous) word accuracy [9]; Sultana et al. reached 78% on a Latinized (Banglish) newspaper dataset with the SAPI engine [32]; Muhammad et al. exceeded 95% for digits 0–5 with MFCC–HMM but fell below 90% for 6–9 [21]; and Hassan et al. reached 91.64% word accuracy on 3,000 sentences from 60 speakers with gender-independent classifiers [10]. These models were limited by small non-standardized datasets, manual feature engineering, poor handling of continuous speech, and the absence of dialect diversity [19].

2.2 Deep Learning and Pre-trained Models

Deep learning enabled automatic feature extraction and better handling of complex linguistic structure [11]. Nasib et al. used CMU Sphinx4 on a 1.99-hour dataset, reaching 71% accuracy with Banglish-only output [22, 16], and a Kaldi HMM–GMM system on 512 hours reaching 23% WER [22, 7]. Mandal et al. combined CNN–RNN for end-to-end recognition on 214.6 hours, achieving 13.67% WER [19]. Saurav et al. proposed a DNN–HMM voice-search system reporting 96.04% accuracy on a self-built corpus with limited dialect diversity [31], and Mridha et al. surveyed hybrid HMM, GMM, CNN, RNN, and DNN systems, finding dataset limitations a recurring challenge [20]. More recent work relies on pre-trained multilingual models: Hasan released a transformer-based Whisper-Bangla model [8]; Ridoy et al. found Wav2Vec-BERT (14.42% WER) better than Whisper on 70,000 samples [28]; Rakib et al. improved wav2vec2 with n-gram language models [27]; Karim et al. combined Whisper and HuBERT for Bangla–English input (15.2% WER) [14]; Rahman et al. achieved a 10% decrease in word error rate (WER) using synthetic audio augmentation [25]. Bhadra et al. obtained an accuracy of 93.42% in isolated words using a feed-forward network utilizing MFCC features [4]. Issues remain with respect to accents, punctuation, and noise [6, 34].

2.3 Challenges, Datasets, and Positioning

Open issues include the problem of regional accent, which is poorly reflected in such resources as SUBAK.KO (241 hours) [15], dialects [7], restoration of punctuation [23], resistance to noise [33], as well as problems of homograph and proper names [19, 22]. Available datasets are also limited: Bengali.AI (196k words) [3] and Mozilla Common Voice (400 hours) [2] suffer from demographic bias, limited dialect coverage, and noise; SHRUTI (21.64 hours) [18] and SUBAK.KO [15] lack conversational data; BanSpeech consists of TV and audiobook recordings rather than natural speech [30]; and OOD-Speech (1,177 hours) [26], Common Voice-Bengali [13], BRADS [1], and BanglaDial [17] widen dialect coverage but lack demographic balance or quality annotation. KBES adds 900 emotion-labeled samples from 35 speakers [5]. Table 1 summarizes existing Bangla ASR systems and Table 2 compares ConvBangla-115 with popular Bangla speech datasets. They show an imbalance that exists in terms of dialects, demographics, and conversations, which is what ConvBangla-115 and the fine-tuned Whisper models try to resolve.

3. METHODOLOGY

The end-to-end data collection, preprocessing, training, and evaluation process is illustrated in Figure 1.

3.1 Choice of Whisper

Whisper was chosen since it provides speech-to-text transcription in one shot, without a need for a pipeline, pre-trained for multiple languages, shows outstanding performance in low-resource languages with minimal linguistic engineering, can perform transcription directly into Bangla script, and has encoder–decoder Transformer architecture, noise resistant.

3.2 Dataset Collection and Composition

Recordings were collected using an online collection system; all of them were manually annotated and verified. Types of recordings include informal conversations, interviews, and voice notes recorded both in the controlled (studios, quiet rooms) and uncontrolled (in public places, at homes) environment, including single and paired dialogue utterances. Metadata associated with every sample contain the speaker age, gender, region background, and recording setting.

The breakdown of the corpus is presented in Figure 2. Speakers of the corpus include 63.34% men and 36.66% women. Ages vary from 10 to 50 years old, where 63% of the speakers are aged from 15 to 29 (maximum concentration is 20–24 years-old, accounting for 30%), 29% is aged 30 or older and 8% is aged from 10 to 14. This distribution correlates with the digitally engaged population easy to gather data from. A total of 115.9 h is distributed non-uniformly among four divisions, which are Rangpur (45.6 h, 39%), Rajshahi (41.2 h, 36%), Dhaka (22.7 h, 20%) and Khulna (6.5 h, 6%), so Rangpur and Rajshahi provide approximately 75% of the total audio samples (see Section 5.2); Khulna and the speakers over the age of 40 are underrepresented. Chittagong, Sylhet, and Barishal were left out of the voices because there was difference in the dialects causing phonetic consistency to be lost. Comparison between the corpora is presented in Table 2.

Table 1. Comprehensive Analysis of Bangla Automatic Speech Recognition Research (chronological; every entry corresponds to a work discussed in Section 2).

Study	Year	Methodology	Dataset	Result	Key Finding	Limitation
Hasnat et al. [9]	2007	Viterbi + HTK, isolated/continuous	Not specified	70% isolated, 60% continuous	Needs noise/silence detection	Struggles in real-time noisy conditions
Muhammad et al. [21]	2009	MFCC + HMM digit recognition	Medium, quiet/noisy	>95% (0–5), <90% (6–9)	Effective for simple digits	Phoneme confusion (6–9)
Hassan et al. [10]	2011	Gender-independent ASR	3,000 sent., 60 spk.	91.64% word, 91.30% sent.	Gender-independent > gender-specific	No dialect diversity
Sultana et al. [32]	2012	SAPI-based STT (Banglish)	Newspaper text (Latinized)	78%	Works on Banglish text	Reduced generalizability to native script
Nasib et al. [22, 16]	2018	CMU Sphinx4, real-time	503 words, 5 spk. (1.99 hrs)	71%	Real-time feasible at small scale	Banglish-only output
Nasib et al. [22, 7]	2018	Kaldi HMM–GMM	512 hrs	WER 23%	Viable for low-resource setups	Weak on dialects, proper nouns
Saurav et al. [31]	2018	DNN–HMM hybrid, voice search	Large-vocab, self-built	96.04%	Large-vocabulary voice-search system	Limited dialect diversity
Mandal et al. [19]	2020	CNN–RNN end-to-end	214.6 hrs	WER 13.67%	Outperforms traditional models	Homonym/dialect issues remain
Mridha et al. [20]	2022	Survey of HMM / GMM / CNN / RNN / DNN systems	Multiple corpora (survey)	– (survey, no new system)	Dataset scarcity is a recurring bottleneck	Review only
Hasan [8]	2022	Transformer-based Whisper-Bangla	Not reported	Not reported	Early Whisper adaptation for Bangla	Repository release only
Rakib et al. [27]	2023	wav2vec2 + n-gram LM (Bangla-Wave)	Common Voice	Improved over wav2vec2 baseline	n-gram LM helps wav2vec2	Read-speech corpus
Karim et al. [14]	2024	Whisper + HuBERT	Mixed Bangla–English	WER 15.2%	Robust to code-switching	Limited pure-Bangla evaluation
Ridoy et al. [28]	2025	Whisper vs. Wav2Vec-BERT	70k samples (Common Voice, OpenSLR)	WER 14.42% (Wav2Vec-BERT)	Wav2Vec-BERT > Whisper	Read-speech corpora only
Bhadra et al. [4]	2025	MFCC + feed-forward NN	Isolated-word speech	93.42% accuracy	Strong on isolated words	Isolated words only
Rahman et al. [25]	2025	Synthetic-audio augmentation	Existing Bangla corpora	WER reduced by 10%	Augmentation helps existing models	Synthetic rather than natural speech
This Work	**2026**	**Fine-tuned Whisper large-v3**	**115.9 hrs, conversational**	**WER 19.60% ± 0.3**	**Strong on real conversational speech**	**Excludes Sylheti/Chittagonian dialects**

3.3 Data Processing and Partitioning

All audio was converted to 16 kHz, as required by Whisper [23], silent segments were trimmed, volume was normalized, and long recordings were divided into single-utterance segments. To increase acoustic variability, noise and pitch-variation augmentation was applied (Figure 1, stage 4). Bangla-speaking linguists transcribed the audio, and a second linguist verified every transcript. To study scaling, training subsets of approximately 5, 10, and 50 hours were drawn from the corpus, with the complete 115.9-hour configuration used for the final model. For every configuration, the corpus was split into a Training Set (70%, ≈ 81.1 h), a Validation Set (15%, ≈ 17.4 h) used to monitor overfitting and tune hyperparameters, and a held-out Testing Set (15%, ≈ 17.4 h) for final evaluation.

3.4 Model Architecture and Training Setup

Whisper is an end-to-end Transformer encoder-decoder [35]: the encoder consumes a log-Mel spectrogram and the decoder predicts the output text. The public checkpoints were pre-trained on 680,000 hours of multilingual audio [23]. The following five models were tuned to examine their cost effectiveness vs. accuracy (Table 3): Tiny (39M params), Base (74M), Small (244M), Medium (769M), and Turbo/large-v3 (809M). Fine-tuning used PyTorch with Hugging Face Transformers. Tiny, Base, Small, and Medium

Table 2. Comparison Between ConvBangla-115 and Other Existing Bangla Speech Datasets.

Aspect	ConvBangla-115	Other Datasets	Distinguishing Feature
Gender balance	Male 63.34%, Female 36.66%	Bengali Sentence Aligned corpus near-balanced (223F/227M); Common Voice–Bengali 84.3% M / 15.7% F	More balanced than Common Voice–Bengali; reduces gender bias in the trained model
Age range coverage	10–50 yrs across eight brackets (heaviest at 20–24, 30%)	Most datasets lack age metadata or skew adult (e.g., BRADS mainly 23–27)	Broader, documented age spread supports age-robust ASR
Regional audio hours	Rangpur 45.6, Rajshahi 41.2, Dhaka 22.7, Khulna 6.5 hrs	BRADS: 8 divisions, fewer hrs/region; Common Voice–Bengali: 400+ hrs, little regional breakdown	More hours per region, especially underrepresented ones, improving accent robustness
Dialect/regional diversity	Explicit coverage of four key divisions, incl. underrepresented Rangpur	BRADS/BanglaDial include dialects but with small, imbalanced samples	Larger, better-balanced regional representation supports dialect-aware training
Data volume and quality	∼116 hrs with controlled demographics and detailed metadata	OOD-Speech: 1,177 hrs, less controlled; BRADS: 2,439 samples, small scale	Balances scale with demographic control, addressing gaps at both extremes
Use-case suitability	Dialect-, gender-, and age-aware ASR with fairness metadata	Existing sets target one dimension: scale, dialect, or a narrow age band	Supports multidimensional, inclusive ASR modeling

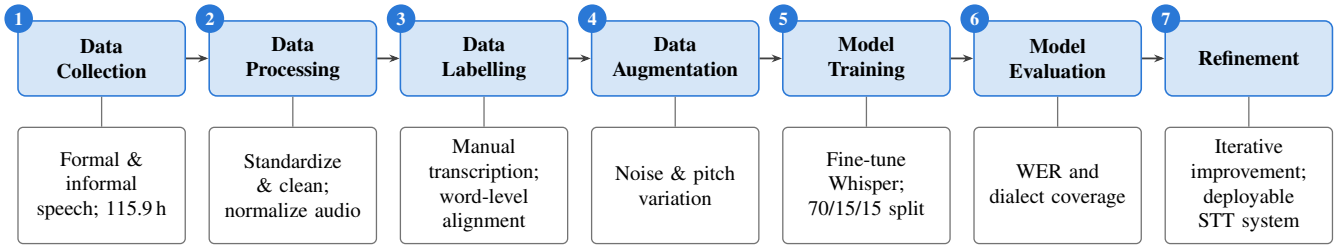


Fig. 1. Overview of the methodology: from data collection and processing (stages 1–4) to model training, evaluation, and refinement (stages 5–7).

were trained on Kaggle NVIDIA T4 GPUs (16 GB VRAM); Turbo was trained on Google Colab Pro with an NVIDIA L4 GPU (22.5 GB VRAM, 53 GB system RAM).

3.5 Statistical Validation

In order to determine that any improvement observed was due to anything other than random variability, the 95% confidence intervals (CIs) were computed through the bootstrap resampling technique for the predictions of each model from the test set, 1,000 times, by utilizing the 2.5th and 97.5th percentiles of the distribution of the word error rate (WER); the $\pm$ numbers mentioned above correspond to the CI half-widths. Paired t-tests were conducted in order to determine the difference between WERs in each consecutive model on the same test dataset, whereas practical significance was determined using the value of Cohen's d. The results obtained yielded $p < 0.001$ with large or very large effects ($d > 1.5$) after Bonferroni adjustment.

3.6 Evaluation Metrics and Derived Quantities

The performance was evaluated using the Word Error Rate (WER) [12] given by the sum of word substitution (S), deletion (D), and insertion (I) divided by the number of reference words N:

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\% \tag{1}$$

Lower values indicate better recognition. Four quantities are derived from the measured WERs. The *absolute reduction* is $\mathrm{WER}_{\mathrm{untrained}} - \mathrm{WER}_{\mathrm{tuned}}$ in percentage points, and the *relative reduction* is that difference divided by $\mathrm{WER}_{\mathrm{untrained}}$. Data scaling is described by the power law

$$\mathrm{WER}(h) = A\, h^{-b}, \tag{2}$$

where h is the number of fine-tuning hours and the exponent b measures how efficiently a model converts additional data into lower error (A and b by least squares in log–log space). The *gain per doubling* of data normalizes each interval by $\log_2(h_2/h_1)$ so that intervals of different length are comparable. The fits use only four points per model (two for Turbo), so they are descriptive and not used for extrapolation.

4. RESULTS AND ANALYSIS

Table 3 illustrates Word Error Rate (WER) for all the models in accordance with the number of hours in fine-tuning data ranging from 5 to 115.9 hours, whereas Table 4 contains the statistical results of the figures. Fine-tuning results in lower WERs for all model sizes; bigger models perform better than smaller ones at any data volume, and higher data volume leads to lower WER. The best result, $19.60\% \pm 0.3$, is obtained by Turbo with 115.9 hours, a 60.8% relative reduction from its zero-shot WER of 50.00%. Paired t-tests between successive model sizes are all significant ($p < 0.001$) with Cohen's d between 1.97 and 8.91. In this section, the results are broken down in order to analyze the impact of fine tuning, amount of data, and size of the model.

4.1 Impact of Fine-Tuning

Zero-shot Whisper shows poor performance on conversational Bangla for all model sizes except Turbo (Figure 3a): untrained word

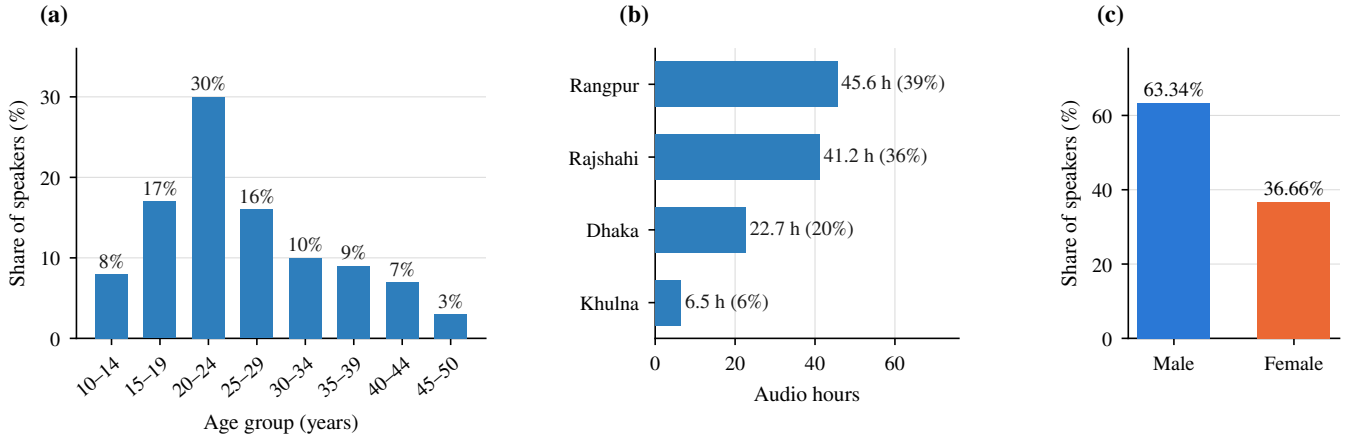


Fig. 2. Composition of ConvBangla-115: (a) speaker age distribution, (b) audio hours per division (share of the 115.9 h in parentheses), and (c) speaker gender distribution.

Table 3. WER (%, with 95% bootstrap CI half-width) by Model Size and Fine-Tuning Data Volume, With Absolute and Relative WER Reduction From the Untrained State to the 115.9-Hour Model.

Model	Params	Untrained	5 h	10 h	50 h	115.9 h	Abs. red. (pts)	Rel. red.
Tiny	39M	93.57±1.2	75.52±0.9	64.82±0.7	60.10±0.6	56.52±0.5	37.05	39.6%
Base	74M	86.34±1.1	70.80±0.8	63.10±0.7	54.40±0.6	45.87±0.5	40.47	46.9%
Small	244M	80.77±1.0	68.55±0.8	51.92±0.6	44.32±0.5	35.10±0.4	45.67	56.5%
Medium	769M	72.64±0.9	65.10±0.8	48.30±0.6	34.50±0.4	28.80±0.4	43.84	60.4%
Turbo	**809M**	**50.00±0.8**	–	–	**24.40±0.3**	**19.60±0.3**	**30.40**	**60.8%**

error rate (WER) drops from 93.57% (Tiny) to 72.64% (Medium); only the Turbo model starts close to 50%. Fine-tuning lowers WER for all model sizes. The relative reduction increases monotonically as the model size increases, from 39.6% (Tiny) to 46.9%, 56.5%, 60.4%, and 60.8% (Turbo) (Figure 3b), which indicates that the larger models gain relatively more from Bangla data than the others depending on their sizes. However, the absolute reduction is highest for Small (45.67 points) and lowest for Turbo (30.40 points; Figure 3c), because the Turbo model gains the least absolute benefit due to its significantly low starting zero-shot error while still maintaining the lowest ending error. This is important to interpret the results, because a lower starting error leaves less scope for absolute improvement. Furthermore, the rate at which the benefit is gained depends on the model size. The first 5 hours make up 48.7% of Tiny's total benefit (from untrained to 115.9 h), 38.4% of Base's, 26.8% of Small's, and 17.2% of Medium's (Table 5). While Medium gains only 7.5 points (10.4% relative) in the first 5 hours, it gains another 16.8 points between 5 and 10 hours. This indicates that larger models need more in-domain data before benefiting from the adaptation process, but intermediate points are based on a single run and further investigation is required. A practical corollary is that the zero-shot Turbo model (50.00%) is already better than every Tiny configuration (best: 56.52%), than Base up to 50 hours (54.40%), than Small up to 10 hours (51.92%), and than Medium at 5 hours (65.10%). It might be more beneficial to choose a better pre-trained checkpoint than spend dozens of hours collecting more data for a poorer model.

4.2 Effect of Training-Data Volume

The WER drops monotonically with increasing data for each model (Figure 4a). Each successive data increment is significantly greater than the uncertainty of the measurement, and the difference in WER between successive data volumes is 4.6 (Tiny, 50 to 115.9 h) to 19.1 times greater (Medium, 10 to 50 h) than the root-sum-square of the CI half-widths.

Whether the returns are "diminishing" depends on the unit of measurement. Per additional hour, the gains shrink sharply: Small improves by 3.33 points per hour between 5 and 10 hours, 0.19 between 10 and 50 hours, and 0.14 between 50 and 115.9 hours. As regards data doubling, improvements remain sizable even in the studied range (Figure 5): the WER decreases by 7.6 points due to doubling for the Small configuration, 7.0 points for Base, 4.7 points for the Medium and 4.0 for Turbo in the range from 50 up to 115.9 hours. For Base and Small configurations this number is greater than the improvement during the previous 10-to-50 hours period (3.7 and 3.3 points of improvement due to doubling respectively). Hence, saturation occurs not at 115.9 hours, and the addition of data is expected to improve WER further.

Power-law fitting (Table 5) characterizes the behavior described above with the R^2 coefficients equal to 0.90 up to 0.98. The value of the exponent b grows along with the configuration size from 0.08 (for Tiny) to 0.13 (Base), 0.19 (Small) and 0.25 (Medium). Two points fit of Turbo gives an exponent of 0.26. Alternatively, each doubling of the data volume allows one to reduce the logarithmically linearized WER value by 3.7 points for Tiny and 7.5 for Medium configurations. Thus, small models reach the saturation quickly and do not gain much from extra data.

Table 4. Statistical Significance of Model Improvements (Full 115.9-Hour Training).

Comparison	WER Difference	p-value	Cohen's d	Interpretation
Turbo vs. Medium	−9.20 pts	<0.001	2.84	Very large effect
Medium vs. Small	−6.30 pts	<0.001	1.97	Large effect
Small vs. Base	−10.77 pts	<0.001	2.42	Very large effect
Base vs. Tiny	−10.65 pts	<0.001	2.31	Very large effect
Fine-tuned vs. Untrained (Turbo)	−30.40 pts	<0.001	8.91	Extremely large effect

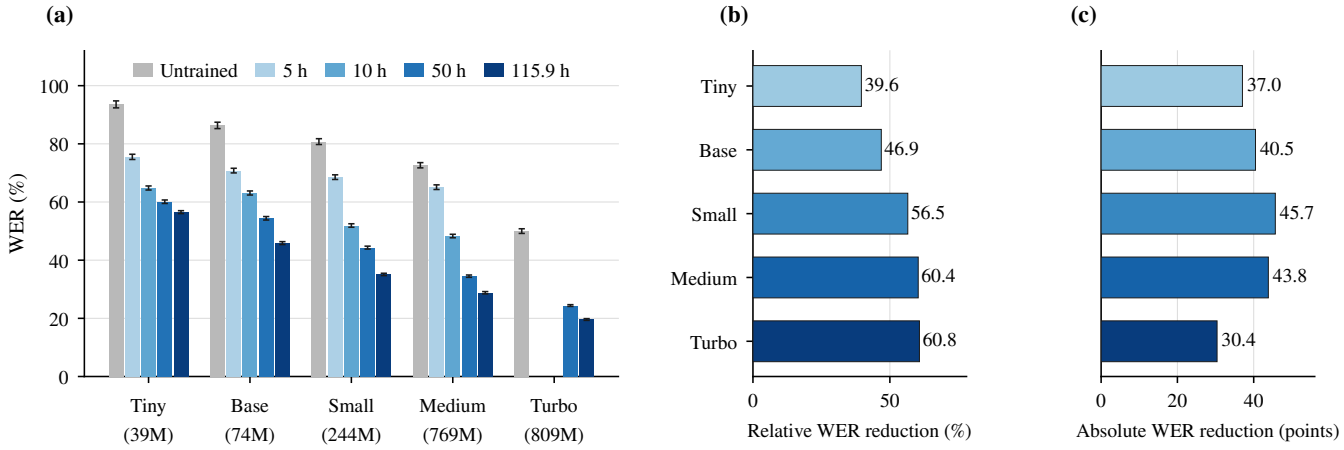


Fig. 3. Effect of fine-tuning: (a) WER for each model before and after fine-tuning on 5, 10, 50, and 115.9 h (error bars: 95% bootstrap CI; Turbo was not trained on 5 or 10 h), (b) relative WER reduction and (c) absolute WER reduction (points) from the untrained state to the 115.9-h model.

Table 5. Power-Law Fit WER $= A\,h^{-b}$ to Fine-Tuned Results, Log-Linear Gain per Doubling of Data, and Share of Total Gain Obtained in the First 5 Hours.

Model	b	R^2	Pts / doubling	First 5 h
Tiny	0.08	0.90	3.7	48.7%
Base	0.13	0.98	5.1	38.4%
Small	0.19	0.94	6.5	26.8%
Medium	0.25	0.98	7.5	17.2%
Turbo†	0.26	–	4.0	–

†Two data points only (50 and 115.9 h); no R^2 is reported.

4.3 Effect of Model Size

At every data volume, WER decreases with model size (Figure 4b). From Tiny to Medium (39M to 769M parameters, about 4.3 doublings), WER falls by roughly 6 points per doubling of parameters (6.3 at 115.9 h and 5.9 at 50 h), and all adjacent differences are significant with large effects (Table 4). Parameter count alone does not explain the whole trend, however. Turbo has only about 5% more parameters than Medium (809M versus 769M) yet reduces WER by a further 9.2 points at 115.9 hours (28.80% to 19.60%) and starts from a zero-shot WER that is 22.6 points lower. Turbo belongs to the large-v3 generation, which keeps the large-v3 encoder but uses a much shallower decoder [23]; its advantage therefore most plausibly reflects the larger encoder and the broader multilingual pre-training of that generation rather than parameter count. This study did not isolate those factors, so the attribution remains a hypothesis.

Table 6. Data–Capacity Trade-off: Smaller Model on 115.9 h vs. Next Larger Model on 50 h (WER %).

Smaller, 115.9 h	WER	Larger, 50 h	WER	Δ (pts)
Tiny	56.52±0.5	Base	54.40±0.6	2.12
Base	45.87±0.5	Small	44.32±0.5	1.55
Small	35.10±0.4	Medium	34.50±0.4	0.60
Medium	28.80±0.4	Turbo	24.40±0.3	4.40

4.4 Trade-off Between Data and Model Capacity

Because both factors reduce WER, a practical question is which is the cheaper lever. Table 6 compares each model fine-tuned on all 115.9 hours with the next larger model fine-tuned on only 50 hours (43% of the data). In all cases where the larger model is trained on less data, it achieves either equal or better results than the smaller model with more data: by 2.12 points (Base at 50 h vs. Tiny), 1.55 (Small vs. Base), 0.60 (Medium vs. Small), and 4.40 (Turbo vs. Medium). Since the margin between Medium and Small is 0.60 points and the CI half-widths are only 0.4 for both, they should be interpreted as parity rather than showing superiority of one model over another. Thus, moving one model size up requires at least a 2.3-fold increase in training data.

5. DISCUSSION

5.1 Best Model and Comparison with Existing Work

The best result comes from Turbo trained on the full dataset. It starts from the strongest zero-shot point (50.00%), and at 50 hours (24.40%) already outperforms Medium trained on all 115.9 hours

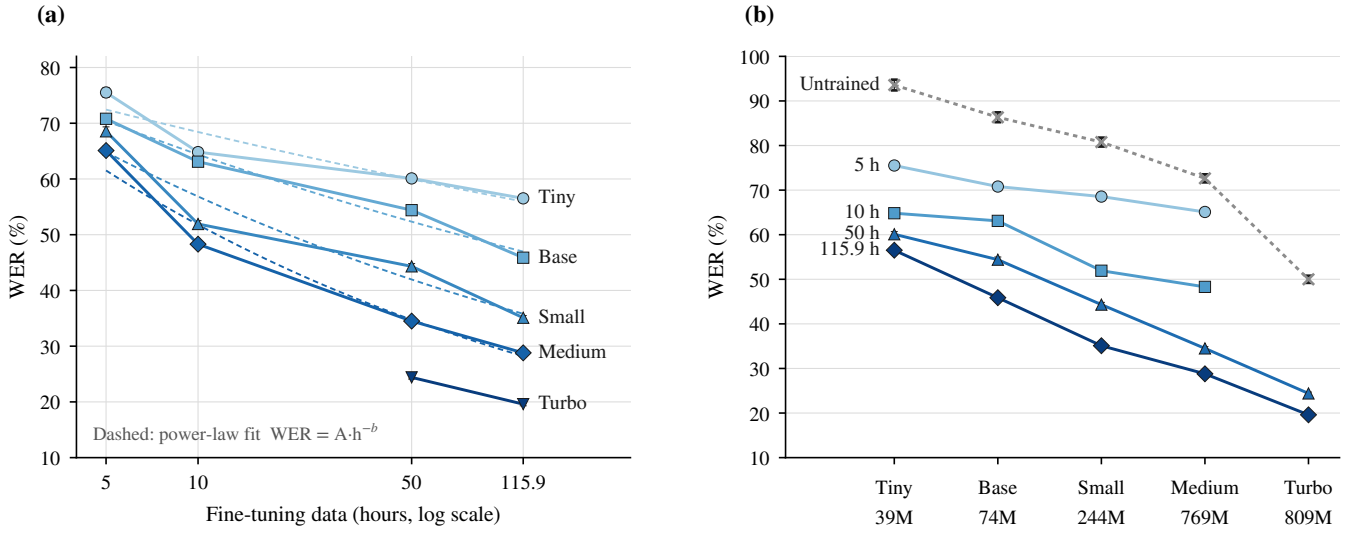


Fig. 4. Scaling behavior: (a) WER versus fine-tuning hours (log axis) with power-law fits (dashed) for Tiny–Medium, and (b) WER versus model size for each data volume. Error bars: 95% bootstrap CI.

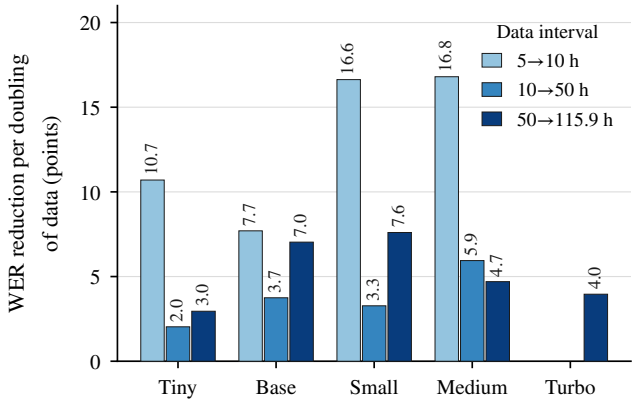


Fig. 5. WER reduction per doubling of fine-tuning data for each data interval. Turbo has no 5-h or 10-h model.

(28.80%); the further 4.8-point reduction (19.7% relative) gives the final 19.60%, with a narrow CI (±0.3, 1.5% of the WER). This WER is also competitive for existing Bangla STT research. It outperforms the system of Nasib et al., based on Kaldi, that achieved 23% WER using a larger 512-hour dataset [22, 7], and comes close to 13.67% WER of Mandal et al. using CNN-RNN on 214.6 hours [19]. Such comparisons need to be done with some care, since the earlier models were trained using different datasets, which consisted of either read or scripted speech, whereas the present dataset consists of spontaneous speech. Power-law trends in performance show that ConvBangla-115 could benefit from growth further, although beyond 115.9 hours extrapolations are unsure.

5.2 Limitations and Future Work

Dataset biases: Despite being varied, ConvBangla-115 is not comprehensive. Barishal, Sylhet, and Chittagong dialects are not included to ensure phonetic consistency, but Rangpur and Rajshahi make up about 75% of the hours, Khulna 6%, people older than 40 years old 10%, and 63% of the participants are men; the performance on underrepresented groups can be inferior to that of the overall WER. **Scope of statistics:** The bootstrap confidence intervals take into account only the test set sampling variance; there is no run-to-run variance because each setup was trained only once, and the power law fits use four points per setup (two for Turbo, which was not trained on 5 and 10 hours). **Computational cost:** Turbo is the most expensive model to train and run, making it difficult to deploy in real time on consumer devices. Potential future work would include the addition of speech from the other divisions, balancing regions, ages, and genders; reporting WER by division, age, and gender, along with a substitution/deletion/insertion split, using multiple seeds; quantization and distillation for real-time use; speaker diarization and punctuation restoration.

6. CONCLUSION

In order to fill the gap created due to the lack of Bangla conversational speech corpus, this research gathered ConvBangla-115, a 115.9-hour diverse conversational corpus in demographic and dialectal terms, and fine-tuned five different sizes of Whisper, OpenAI's speech recognizer, on it. The best performing model is Whisper large-v3 (Turbo) achieving 19.6% ± 0.3 WER, a relative improvement of 60.8% from its zero-shot baseline. Fine-tuning, amount of training data, and model size reduce WER significantly; the efficiency of conversion of data into accuracy increases with model capacity (power-law exponent 0.08 to 0.25); a one-size-larger model trained on 50 hours outperforms or equals the smaller model trained on 115.9 hours; and there is no model that has saturated at 115.9 hours. Thus, this research contributes with a competitive Bangla speech recognition model, a useful conversational corpus, and quantitative insight about trade-off between data collection and model capacity.

7. ACKNOWLEDGMENT

The authors would like to thank all the participants who have made speech recordings, the annotators for their painstaking transcription

efforts, and the open source community for the software used in this study.

8. REFERENCES

[1] U. Aiman, M. N. Islam, H. Sultan Chowdhury, M. S. Rahman, M. T. Habib, and M. Hasan. BRADS: a multipurpose audio dataset for Bangla regional word detection, March 2025.

[2] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. Common Voice: a massively-multilingual speech corpus. In *Proc. 12th Language Resources and Evaluation Conf.*, pages 4218–4222, May 2020.

[3] Bengali.AI. Open source datasets. `https://www.bengali.ai/datasets`, 2023.

[4] D. Bhadra, M. Hosain, and F. Alam. Deep feed-forward neural network for bangla isolated speech recognition. *Journal Of Nazrul University*, 10(1-2):85–96, 2025.

[5] M. M. Billah, M. L. Sarker, and M. A. H. Akhand. KBES: a dataset for realistic Bangla speech emotion recognition with intensity level. *Data in Brief*, 51:109741, December 2023.

[6] S. K. Chatterji. *Origin and Development of the Bengali Language, Pt. 1*. 1926.

[7] S. Guchhait, A. S. A. Hans, and J. Augustine. Automatic speech recognition of Bengali using Kaldi. In *Proc. 2nd Int. Conf. Sustainable Expert Systems*, pages 153–166, Singapore, 2022. Springer Nature.

[8] M. Hasan. Transformer-based Whisper Bangla ASR model. `https://github.com/menon92/BangalASR`, 2022.

[9] A. Hasnat, J. Mowla, and M. Khan. Isolated and continuous Bangla speech recognition: implementation, performance and application perspective. `https://www.semanticscholar.org/paper/e6697a3bb8601ab533012473b6844b2121c7d7bd`, 2007.

[10] F. Hassan, M. R. A. Kotwal, and M. N. Huda. Bangla ASR design by suppressing gender factor with gender-independent and gender-based HMM classifiers. In *Proc. World Congress on Information and Communication Technologies*, pages 1276–1281, December 2011.

[11] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[12] D. Jurafsky and J. H. Martin. Speech and language processing, 2023. 3rd ed. draft.

[13] M. A.-A.-K. Kafi, R. Moni, and D. M. Raza. Bangla voice dataset: simple, complex, and compound structures, September 2025.

[14] M. R. Karim, S. Islam, and M. M. Rahman. Multilingual speech-to-text framework for Bangla-English code-switching scenarios. In *Proc. Int. Conf. Natural Language Processing (ICNLP)*, pages 112–120. IEEE, 2024.

[15] 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 Communication*, 136:84–97, January 2022.

[16] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. K. Warmuth, and P. Wolf. The CMU Sphinx-4 speech recognition system. `https://www.semanticscholar.org/paper/5064c602c3a57f4e6f1e4c8f8fb137384c5d41a7`, 2001.

[17] Mehraj Hossain Mahi, Anzir Rahman Khan, and Mayen Uddin Mojumdar. Bangladial: A merged and imbalanced text dataset for bengali regional dialect analysis. *Data in Brief*, 63:112200, 2025.

[18] S. Mandal, B. Das, and P. Mitra. Shruti-ii: A vernacular speech recognition system in bengali and an application for visually impaired community. In *2010 IEEE Students Technology Symposium (TechSym)*, pages 229–233, 2010.

[19] S. Mandal, S. Yadav, and A. Rai. End-to-end Bengali speech recognition. arXiv, `https://www.semanticscholar.org/paper/463c98d2fd4954923db2366786480339c95f9dfb`, September 2020.

[20] M. F. Mridha, A. Q. Ohi, M. A. Hamid, and M. M. Monowar. A study on the challenges and opportunities of speech recognition for Bengali language. *Artificial Intelligence Review*, 55(4):3431–3455, April 2022.

[21] G. Muhammad, Y. A. Alotaibi, and M. N. Huda. Automatic speech recognition for Bangla digits. In *Proc. 12th Int. Conf. Computers and Information Technology (ICCIT)*, pages 379–383, December 2009.

[22] A. U. Nasib, H. Kabir, R. Ahmed, and J. Uddin. A real-time speech-to-text conversion technique for the Bengali language. In *Proc. Int. Conf. Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pages 1–4. IEEE, 2018.

[23] OpenAI. Robust speech recognition via large-scale weak supervision. `https://github.com/openai/whisper`, 2022.

[24] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[25] T. Rahman, F. Ahmed, and F. Chowdhury. Data augmentation for bangla speech recognition using synthetic audio generation. arXiv preprint arXiv:2501.08765, 2025.

[26] F. R. Rakib, S. S. Dip, S. Alam, N. Tasnim, M. I. H. Shihab, M. N. Ansary, S. M. Hossen, M. H. Meghla, M. Mamun, F. Sadeque, S. S. Chowdhury, T. Reasat, A. Sushmit, and A. I. Humayun. OOD-Speech: a large Bengali speech recognition dataset for out-of-distribution benchmarking. arXiv:2305.09688, May 2023.

[27] M. Rakib, M. I. Hossain, N. Mohammed, and F. Rahman. Bangla-Wave: improving Bangla automatic speech recognition utilizing n-gram language models. In *Proc. 12th Int. Conf. Software and Computer Applications (ICSCA '23)*, pages 297–301, June 2023.

[28] M. S. Islam Ridoy, S. Akter, and M. A. Rahman. Adaptability of ASR models on low-resource language: a comparative study of Whisper and Wav2Vec-BERT on Bangla. In *Proc. 2nd Int. Conf. Next-Generation Computing, IoT and Machine Learning (NCIM)*, pages 1–6, June 2025.

[29] 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. *Acoustical Science and Technology*, 42(5):252–260, 2021.

[30] A. M. Samin, M. H. Kobir, M. M. S. Rafee, M. F. Ahmed, M. Hasan, P. Ghosh, S. Kibria, and M. S. Rahman. BanSpeech: a multi-domain Bangla speech recognition benchmark toward robust performance in challenging conditions. *IEEE Access*, 12:34527–34538, 2024.

[31] J. R. Saurav, S. Amin, S. Kibria, and M. S. Rahman. Bangla speech recognition for voice search. In *Proc. International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE, 2018.

[32] S. Sultana, M. A. H. Akhand, P. K. Das, and M. M. Hafizur Rahman. Bangla speech-to-text conversion using SAPI. In *Proc. Int. Conf. Computer and Communication Engineering (ICCCE)*, pages 385–390, July 2012.

[33] S. Hosain Sumit, T. Al Muntasir, 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 and Language Processing (ICBSLP)*, pages 1–5, September 2018.

[34] O. Tilk and T. Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proc. Interspeech 2016*, pages 3047–3051, September 2016.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.