BanglaKontho: Closing the Long-Form Gap in Bangla Text-to-Speech

Anonymous ACL submission

Abstract

Bangla, the seventh most spoken language in the world with over 230 million speakers, remains under-resourced for modern neural text-to-speech (TTS). Existing publicly accessible Bangla speech corpora are dominated by short read-prompt utterances aimed at automatic speech recognition, leaving long-form prosody and consistent single-speaker audiobook narration largely uncovered. We present *BanglaKontho*, a single-speaker Bangla TTS corpus of 20 hours derived from professional audiobook recordings, comprising 7,050 segmented utterances aligned with verified transcripts at 24 kHz. Alongside the corpus we release a reusable Bangla text normalizer that handles Bangladeshi-style digit grouping, currency and date expressions, Danda/Daari punctuation, and Unicode normalization, together with the full preprocessing pipeline used to build the dataset. We train an MB-iSTFT-VITS baseline from scratch and report MCD, WER, and naturalness MOS, comparing against the same architecture retrained on the 12-hour IndicTTS-Bn corpus and against an ablation that disables the normalizer. The full 20-hour corpus is openly released on HuggingFace under CC BY-NC 4.0.

1 Introduction

Bangla (Bengali) is the seventh most spoken language in the world, yet text-to-speech (TTS) research for Bangla materially lags English and Mandarin. The principal barrier is data: high-quality TTS training corpora for Bangla are scarce, fragmented across multiple small releases, and dominated by short read-prompt recordings originally collected for automatic speech recognition (ASR) (Kjartansson et al., 2018; Ardila et al., 2020). Modern end-to-end neural TTS architectures such as VITS (Kim et al., 2021) and its lightweight variants (Kawamura et al., 2023) require tens of hours of clean speech aligned with verified text in order to produce natural, intelligible output. Existing Bangla resources are insufficient for this regime, particularly when prosody, sentence-level intonation, and long-form rhythm are required for downstream applications such as audiobook synthesis, screen readers, and conversational agents.

In this paper we describe *BanglaKontho*, a single-speaker Bangla audiobook corpus of 20 hours with verified transcripts and released openly on HuggingFace under CC BY-NC 4.0. The audiobook domain is deliberate: studio recordings by a single trained narrator over long-form material yield acoustic and prosodic patterns that complement the read-prompt corpora that dominate existing Bangla TTS releases. Together with the audio-text pairs, we release the full preprocessing pipeline and an open-source Bangla text normalizer.

Contributions.

1. The first dedicated single-speaker Bangla audiobook TTS corpus with verified transcripts: 20 hours / 7,050 utterances at 24 kHz, openly released under CC BY-NC 4.0 (§3).
2. An open-source Bangla text normalizer with Bangladeshi-style digit grouping, lakh / koti scale, decimal and date handling, and a strict Bangla-Unicode-block mode (§3.5).
3. A reproducible end-to-end preprocessing pipeline and an MB-iSTFT-VITS baseline with MCD, WER, and naturalness MOS, benchmarked against the same architecture retrained on IndicTTS-Bn and against a normalizer-disabled ablation (§4).

2 Related Work

Bangla speech resources. The most widely cited Bangla speech corpus is the crowd-sourced read-prompt collection of Kjartansson et al. (2018), hosted on OpenSLR, containing roughly 200 hours from over 500 speakers. Its many-speaker design

makes it well-suited for ASR but less so for single-voice TTS, and utterance durations are typically short. The IndicTTS project (AI4Bharat, 2023) released paired TTS recordings for major Indian languages including Bangla, totalling around 12 hours per speaker with limited speaker diversity. The Bangla subset of Mozilla Common Voice (Ardila et al., 2020) consists of validated short-form crowd-sourced utterances. Shrutilipi (Bhogale et al., 2023) mines 6,400+ hours of weakly aligned Indic speech (including Bangla) from broadcast audio, but transcript-alignment quality is variable and the corpus targets ASR rather than TTS. None of these resources simultaneously offers (i) a single trained speaker, (ii) long-form audiobook prosody, (iii) studio-clean audio, and (iv) verified transcripts.

Bangla TTS. Early work on Bangla TTS adopted statistical-parametric methods (Gutkin et al., 2016) on small read-prompt collections such as SHRUTI (Das et al., 2011). More recent neural approaches have explored zero-shot voice cloning (Ahmed et al., 2024), achieving high speaker similarity but relying on centralized fine-tuning that aggregates raw speech on a server, with attendant privacy and licensing risks. To our knowledge, no prior work has released a single-speaker, long-form Bangla audiobook corpus. Outside Bangla, recent end-to-end systems including VITS2 (Kong et al., 2023), Matcha-TTS (Mehta et al., 2024), and StyleTTS 2 (Li et al., 2023) have set new quality bars; we adopt a VITS-based architecture for direct comparability with prior Bangla work, but the corpus is architecture-agnostic.

Text normalization for Indic scripts. Bangla text normalization is non-trivial because of (i) Bangladeshi-style digit grouping and non-compositional number words in the range 20–99, (ii) conjunct characters and decomposed vowel signs that are visually identical but Unicode-distinct, and (iii) mixed Bangla / ASCII digits in modern web sources. Gutkin et al. (2016) note the need for language-specific normalization but do not release a reusable tool; we address this gap.

Table 1 summarizes the resource landscape.

3 Dataset Construction

Figure 1 summarizes the end-to-end pipeline. We describe each stage below.

3.1 Audio Acquisition and Preprocessing

Source audio was acquired in its native delivery format and converted to a uniform working representation of 16 kHz, 16-bit PCM, mono, with peak normalization. Music intros/outros, dynamic-range compression artefacts, and other post-production processing present on the source delivery were manually identified and excised by the authors before segmentation; every removed region was verified by ear. After segmentation, audio is resampled to **24 kHz** for release, which is the operating rate of most modern neural vocoders (Kong et al., 2020; Kawamura et al., 2023).

3.2 Speaker Verification

Because the corpus is single-speaker by construction, the preprocessing challenge is not multi-speaker diarization but reliable removal of non-narrator audio (music intros / outros, chapter-announcement voices distinct from the narrator, occasional guest readings). We use `pyannote.audio` 3.1 (Plaquet and Bredin, 2023) for voice activity detection and segment-level speaker embeddings (ECAPA-TDNN; Desplanques et al., 2020), then verify each speech segment against a manually enrolled narrator reference via cosine similarity. Segments below a calibrated threshold are discarded.

3.3 Segmentation

Segmentation is guided by VAD-derived silence boundaries combined with end-of-sentence punctuation alignment in the time-aligned draft transcript. Clips shorter than 1.0 s or longer than 20 s are dropped; the operating target range is 2–15 s, the standard window for most TTS.

3.4 Transcription and Manual Verification

A first-pass transcript is produced by Google Gemini 2.0 Flash Lite (Gemini Team, Google, 2024) for each clip, exploiting its broad Indic-language coverage. On a randomly sampled 500-clip subset hand-transcribed by the authors, Gemini achieved 2.7% WER against the gold reference, which we treat as a strong but imperfect starting point rather than ground truth. Two volunteer native Bangla speakers then independently reviewed 100% of clips for word-level lexical accuracy, phoneme-relevant punctuation, and audio defects missed by the automated stage; word-level inter-annotator agreement was Cohen's $\kappa = 0.91$ on a 500-clip overlap sub-

Corpus	Hours	Speakers	Domain	SR (kHz)	Access
OpenSLR Bengali (Kjartansson et al., 2018)	∼200	505	Read prompts	16	Open
IndicTTS-Bn (AI4Bharat, 2023)	12	2	Read prompts (studio)	48	Open
Common Voice Bn (Ardila et al., 2020)	60+	many	Read prompts	48	Open
Shrutilipi (Bn portion) (Bhogale et al., 2023)	437	many	Broadcast	16	Open
BanglaKontho (this work)	**20**	**1**	**Audiobook**	**24**	Open (CC BY-NC 4.0)

Table 1: Bangla speech corpora most commonly used as TTS data sources. *BanglaKontho* fills the long-form, single-speaker audiobook slot.

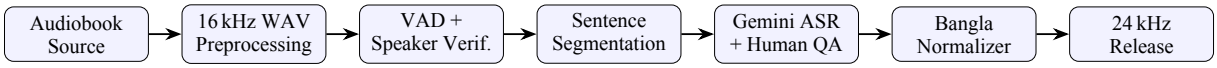

Figure 1: End-to-end *BanglaKontho* construction pipeline.

set, and remaining disagreements were resolved by discussion.

3.5 Bangla Text Normalization

We release `bangla_normalize`, a stand-alone Python module used to clean every transcript in the corpus. The pipeline is intended to be reusable across Bangla TTS / ASR projects and is byte-equivalent to the cleaner shipped in the training filelists.

1. **Unicode NFC.** Decomposed vowel signs (e.g., ে + া → ো) are composed so that each visible grapheme corresponds to a single code-point.
2. **Symbol-to-word.** Currency symbols (৳, £, €, ₹), percent, ampersand, and other ASCII symbols are rewritten to their Bangla word equivalents.
3. **Numeric expansion** over both Bangla (০–৯) and ASCII (0–9) digits. The normalizer hard-codes the 80 unique Bangla compound words for 20–99, since Bangla number words in this range are not productively compositional (26 is *chhabbish*, not *bish-chhoy*).
4. **Bangladeshi digit grouping** on the lakh / koti scale to arbitrary magnitude (e.g., ১,২৩,৪৫,৬৭৮ → *ek koti teish lakh poyetallish hajar chhayshô attattôr*).
5. **Decimals and dates**: ১২.৫ → *baro doshomik pānch*; ২০২৪-০৪-২৩ → *dui hajar chobbish chār teish*.
6. **Danda/Daari punctuation**: repeated Danda/Daari are collapsed (। । । → ।); double-Danda/Daari is folded (॥ → ।); spacing around the Danda/Daari is normalized.

Property	Value
Total duration	20 hours
Utterances	7,050
Speakers	1
Source domain	Audiobook
Sampling rate (release)	24,000 Hz
Sampling rate (baseline)	22,050 Hz
Mean utterance duration	∼10.2 s
Train / val / test split	6,078 / 500 / 472
Split policy	chapter-disjoint
Storage (parquet shards)	∼3.5 GB

Table 2: *BanglaKontho* dataset statistics. Chapter-disjoint split (no chapter contributes utterances to more than one of train / val / test) to prevent same-chapter prosodic leakage between training and evaluation.

7. **Strict mode (optional)** restricts output to the Bangla Unicode block U+0980–U+09EF plus a safe-punctuation set, suitable for downstream G2P with a fixed symbol table.

Example: আজ ১২.৫ টাকা ও 50% ছাড় → আজ বারো দশমিক পাঁচ টাকা ও পঞ্চাশ শতাংশ ছাড়.

3.6 Quality Filtering and Statistics

After transcription and normalization we drop clips with residual non-speech tails, anomalous duration-to-text ratios, or verifier flags from §3.4. Final statistics are reported in Table 2.

4 Baseline TTS and Evaluation

To benchmark the corpus we train MB-iSTFT-VITS (Kawamura et al., 2023) from scratch. MB-iSTFT-VITS replaces the HiFi-GAN-style decoder of standard VITS (Kim et al., 2021) with a multi-band iSTFT head and PQMF subband synthesis, yielding a lightweight, real-time-friendly model directly comparable to prior single-speaker VITS work. The released audio at 24 kHz is resampled

to 22,050 Hz to match the canonical reference configuration of MB-iSTFT-VITS; all evaluation is reported at 22,050 Hz. The architecture combines a Transformer text encoder, a posterior encoder, a stochastic duration predictor, a normalizing-flow prior, and a multi-band iSTFT decoder, with 34.7 M parameters at our configuration (hidden / inter 192, filter 768, two attention heads, six layers; ResBlock-1 kernels $[3, 7, 11]$; upsample rates $[4, 4]$; initial channel 512; four PQMF subbands). Training uses batch size 128 with AdamW (β_1=0.8, β_2=0.99), learning rate 2×10^{-4} with decay 0.999875, segment size 8192 samples, and FP16 mixed precision. All training and preprocessing was conducted on a single NVIDIA H100 SXM5 80 GB GPU.

Evaluation protocol. We hold out 472 utterances as a test split, drawn from chapters disjoint from those used for training or validation to avoid intra-chapter prosodic leakage that would inflate evaluation scores under a naive utterance-level random split. We report: (i) **MCD** (Kubichek, 1993), DTW-aligned 13-dim MFCC distance; (ii) **WER** computed by transcribing the synthesized audio with IndicWav2Vec (Javed et al., 2022), a Bangla-capable ASR (generic multilingual models such as Whisper-base (Radford et al., 2023) produce uninformative scores on Bangla synthesis); and (iii) **Naturalness MOS** on a 5-point scale, with 15 volunteer native Bangla raters scoring 50 utterances per system (balanced for length) following the ITU-T P.800 protocol (ITU-T, 1996), reported as mean $\pm$ 95% CI. All MOS sessions used headphones in a quiet environment, with system identity hidden and clip order randomized per rater; Krippendorff's $\alpha = 0.74$ across raters. As an ASR-floor calibration, IndicWav2Vec on the *natural* test-split audio (i.e., human ground truth) yields 8% WER, so our reported synthesis WERs should be read relative to this floor rather than to 0%. We further include a **ground-truth MOS** ceiling obtained by submitting the natural test-split audio to the same rating protocol. We compare our model against an MB-iSTFT-VITS baseline that we retrained from scratch on the IndicTTS-Bn dataset (AI4Bharat, 2023) (12 hours) using the same configuration as our model; we do not use any released checkpoint.

Baseline results. Table 3 compares our MB-iSTFT-VITS model trained on *BanglaKontho* against the same architecture retrained on IndicTTS-Bn and against the natural-speech ceiling. Our model achieves substantially lower WER (9.5% vs. 16.0%) and higher naturalness MOS (4.46 vs. 3.16), with comparable MCD (5.85 vs. 5.97), and approaches the natural-speech ceiling of 4.90 MOS / 8% ASR-floor WER. Because the data-size gap between *BanglaKontho* and IndicTTS-Bn is modest (20 h vs. 12 h, a factor of 1.67), the large WER and MOS gaps cannot be explained by scale alone; they are consistent with the domain difference (consistent long-form audiobook narration vs. short read-prompt recordings) being the principal driver.

Metric	Ours	Ours w/o norm.	IndicTTS-Bn	Ground truth
MCD (dB) ↓	5.85	5.81	5.97	–
WER (%) ↓	9.5	14.0	16.0	8.0
Naturalness MOS ↑	4.46 ± 0.06	3.80 ± 0.08	3.16 ± 0.09	4.90 ± 0.03

Table 3: Comparison of our VITS-based model trained on *BanglaKontho* against a VITS baseline trained on IndicTTS-Bn (12 h). WER computed via IndicWav2Vec on synthesized audio. MOS over 15 native Bangla raters following ITU-T (1996).

Normalizer ablation. To isolate the contribution of `bangla_normalize`, we retrained the same MB-iSTFT-VITS architecture on *BanglaKontho* with the normalizer disabled (raw verified transcripts only). The ablated model degrades to WER 14.0% and naturalness MOS 3.80 (Table 3); MCD is comparable (5.81). The gap is concentrated on utterances containing numerals, dates, currency symbols, or mixed Bangla/ASCII digits (18% of the test set), where the ablated model frequently mispronounces symbol-bearing tokens; on the symbol-free remainder, the gap narrows substantially. This pattern confirms that Bangla-specific text normalization contributes a localized but consistent intelligibility gain, a contribution distinct from corpus scale.

5 Conclusion

We release *BanglaKontho*, a 20-hour single-speaker Bangla audiobook corpus sourced from publicly available professional recordings, together with a reusable Bangla text normalizer and an MB-iSTFT-VITS baseline. The resource addresses a documented gap in long-form Bangla speech data and provides ready-to-use tooling for the broader Indic-TTS community. Future work will (i) extend to multiple narrators, (ii) develop a Bangla-specific CMOS / preference-test benchmark, and (iii) evaluate downstream applications such as screen readers and audiobook synthesis.

Limitations

Scope. The corpus is single-speaker by construction; multi-speaker extension is left to future work. The audiobook domain shapes prosody toward narrative reading style, which differs from conversational TTS; for conversational use we recommend combining *BanglaKontho* with a small in-domain adaptation set.

License and release. The full 20-hour *BanglaKontho* corpus is openly released on HuggingFace under CC BY-NC 4.0; the accompanying preprocessing pipeline, including the `bangla_normalize` library, is released under a permissive open-source license.

Data provenance and ethical use. The source audiobook recordings are publicly available online and were obtained through their official distribution channels. The corpus is redistributed under CC BY-NC 4.0 strictly for non-commercial academic research; commercial use, voice cloning for impersonation, fraud, and the synthesis of political content are explicitly prohibited by the accompanying data-use agreement that all downloaders must accept. We maintain a documented point of contact for the original narrator and rights-holders, and operate a standing takedown / opt-out process: any such request will be honoured and the corresponding material removed from the release within a published service window. For additional provenance protection we recommend that downstream systems trained on *BanglaKontho* apply inaudible neural watermarking to synthesized audio.

Speaker anonymity. We omit the narrator's identity from the released metadata, but acknowledge that voice-based identification remains possible for anyone in possession of the audio; we make no cryptographic anonymity claim about the speech signal itself.

ASR-based intelligibility. WER depends on the Bangla ASR's own accuracy, which is improving rapidly but not error-free. We accordingly emphasize MOS as the primary intelligibility signal.

References

Abu Ahmed, P. Roy, D. C. Paul, and M. S. Rahman. 2024. Efficient zero-shot voice cloning for Bengali speech synthesis. In *27th International Conference on Computer and Information Technology (ICCIT)*, pages 1470–1474. IEEE.

AI4Bharat. 2023. IndicTTS: Multilingual text-to-speech for Indian languages. https://github.com/AI4Bharat/Indic-TTS.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. Effectiveness of mining audio and text pairs from public data for improving ASR systems for low-resource languages. In *Proceedings of ICASSP*. IEEE.

Biswajit Das, Sandipan Mandal, and Pabitra Mitra. 2011. Bengali speech corpus for continuous auutomatic speech recognition system. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 51–55. IEEE.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proceedings of Interspeech*, pages 3830–3834.

Gemini Team, Google. 2024. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alexander Gutkin, Linne Ha, Martin Jansche, Knot Pipatsrisawat, and Richard Sproat. 2016. TTS for low resource languages: A Bangla synthesizer. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, pages 2005–2010.

ITU-T. 1996. Methods for subjective determination of transmission quality. Technical Report Recommendation P.800, International Telecommunication Union.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building ASR systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10813–10821.

Masaya Kawamura, Yuma Shirahata, Ryuichi Yamamoto, and Kentaro Tachibana. 2023. Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time Fourier transform. In *Proceedings of ICASSP*, pages 1–5. IEEE.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5530–5540.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-sourced speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 52–55.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 17022–17033.

Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. 2023. VITS2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. In *Proceedings of Interspeech*, pages 4374–4378.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, volume 1, pages 125–128.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proceedings of ICASSP*, pages 11341–11345. IEEE.

Alexis Plaquet and Hervé Bredin. 2023. Powering speaker diarization with multi-scale neural speaker embeddings. In *Proceedings of ICASSP*, pages 1–5. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518.