BanglaBox: A Phonetically-Balanced Corpus and Data-Efficient Foundation-Model Adaptation for Bangla Text-to-Speech with Zero-Shot Voice Cloning

Anonymous ACL submission

Abstract

We present a recipe for adapting English-pretrained autoregressive Text-to-Speech (TTS) foundation models to under-represented languages for zero-shot voice cloning settings, demonstrated on Bangladeshi Bangla. Existing Bangla TTS corpora are small and single-speaker, and to the best of our knowledge, no open zero-shot voice-cloning system has been released for the Bangladeshi register. At first, we contribute a phonetically- and gender-balanced two-tier Bangladeshi Bangla corpus balanced via a tiered Jensen–Shannon divergence objective over conjunct clusters (*juktakkhor*). Next, we introduce BanglaBox, a zero-shot voice cloning architecture based on Chatterbox model 2025, with the following three contributions—a *merge-consistent tokenizer extension*, Bangla text normalization, and a prompt-masked dual-loss objective—that preserve zero-shot cloning across the language switch. Later, we introduce an evaluation protocol named *BanglaEval* to evaluate our TTS zero-shot cloning pipeline. Our *BanglaEval* protocol applies Wilcoxon signed-rank tests with Bonferroni correction over native-speaker ratings. *We use commercial services – Google, Azure, Narakeet – as baselines.* Our BanglaBox significantly outperforms the above commercial baselines on Naturalness, Speaker Similarity, and Clarity, and reaches speaker similarity comparable to prior few-shot results while using roughly 7× less Bangla fine-tuning audio, establishing a new baseline in the zero-shot voice cloning Bangla TTS tasks.

1 Introduction and Related Work

Bangla (Bengali), spoken by over 230 million people (Eberhard et al., 2024), remains severely underrepresented in neural speech synthesis. Existing Bangla Text-to-Speech (TTS) systems have the following limitations: early HMM (Mukherjee and Mandal, 2012) and LSTM-RNN (Gutkin et al., 2016) models, and more recent VITS-based (Kumar et al., 2025) and DNN-based (Raju et al., 2019) systems, are bound to small *single-speaker* corpora — precluding voice cloning, cross-speaker generalization, and multi-domain coverage. The SUST TTS Corpus (Ahmad et al., 2021) is the largest prior phonetically-balanced Bangla resource, but is single-speaker. Therefore, Bangla phonetic-balance design has relied on simple phone/diphones coverage (Rashid et al., 2009; Jahan et al., 2015; Uddin et al., 2015) without explicit modeling of the conjunct-cluster (*juktakkhor*) inventory that drives Bangla naturalness. BnTTS (Basher et al., 2025), the recent state of the art, fine-tunes an XTTS (Casanova et al., 2024) backbone on 3.85k h of Bangla speech. However, its weights are not public, its evaluation lacks significance testing, and its data requirement is prohibitive for low-resource follow-up. Zero-shot TTS in codec-language (Wang et al., 2023; Kharitonov et al., 2023), latent-diffusion (Shen et al., 2024; Li et al., 2023), and flow-matching (Le et al., 2023) families achieve strong English results but are unevaluated on Bangla. *The only work on an open Indic-language model, IndicF5 (Praveen et al., 2025), reaches only SECS* 0.37*, and fine-tuned YourTTS (Casanova et al., 2022) reaches* 0.44*, both far below human.*

In this paper, we explore whether the pre-trained TTS foundation models may offer a path forward. Chatterbox TTS (Resemble AI, 2025) is an MIT-licensed Llama-style autoregressive transformer with built-in zero-shot cloning from a short reference clip.[1] Adapting it to Bangla requires handling the juktakkhor inventory, context-dependent vowel diacritics,

[1] Chatterbox model in the Hugging Face model card.

and honorific abbreviations absent from standard Bangla NLP toolchains (Chowdhury and Adhikary, 2022; Ansary, 2024; Ghosh et al., 2024), while preserving cloning across the language switch. To our knowledge, no prior work has adapted Chatterbox to Bangla, nor formalized merge-consistent vocabulary extension when importing new scripts into autoregressive codec language models.

We address these gaps with BanglaBox, an open fine-tuning recipe for Bangladeshi Bangla (Dhaka register). Commercial services – Google, Azure, Narakeet (Narakeet, 2024) – provide Bangla synthesis, and we include them as baselines. BanglaBox leads these competitors on every perceptual metric ($p<0.01$, Wilcoxon with Bonferroni correction).

It **beats commercial Google** by +6.8% Naturalness (+0.29 MOS), +0.34 MOS on SMOS, and +0.18 MOS on Clarity; **outperforms Azure** by +10.4% / +10.8% / +5.6%; and **tops Narakeet** by a full +14.9% / +21.6% / +16.7%. The widest margin is over the open Indic baseline **IndicF5**: +17.2% Naturalness, a 0.96-MOS SMOS lead.

Moreover, on zero-shot voice cloning, BanglaBox **nearly doubles IndicF5's speaker-cloning fidelity** (0.63 vs. 0.37 SECS, 1.7×). Please note that commercial Google, Azure, Narakeet systems do not provide speaker adaptation. BanglaBox matches BnTTS's reported few-shot 0.548 SECS (indicative; encoders differ) using 7× **less** Bangla audio (520 h vs. 3.85k h), and sits within 0.20 MOS of the 4.76 human ceiling. In summary, our contributions are:

(i) Corpus. 520.3 h, 36-speaker, 7-domain Bangladeshi Bangla TTS corpus, phonetically- and gender-balanced via a novel *juktakkhor-aware tiered Jensen–Shannon divergence (JSD)* objective, with speaker-disjoint test split (§2).

(ii) BanglaBox.[2] (a) A *merge-consistent tokenizer extension* importing 1,786 Bangla tokens without breaking pretrained byte-pair encoding (BPE) merge rules; (b) Bangla text normalization with honorific and numeral expansion; (c) a prompt-masked dual-loss training objective (Wang et al., 2023) preserving zero-shot cloning (§3); all three are independently ablated.

(iii) BanglaEval + long-form inference. A reproducible evaluation with 24 native-Bangla raters scoring three independent dimensions under Wilcoxon signed-rank with Bonferroni correction — stronger statistical grounding than any prior Bangla TTS evaluation — plus a chunked long-form inference pipeline (see Fig. 5).

(iv) Data-efficiency. The first Chatterbox-based zero-shot Bangla voice-cloning system, requiring ∼ 7× less Bangla fine-tuning audio, establishes a new baseline in the zero-shot voice cloning Bangla TTS tasks.

2 Dataset construction

2.1 Two-tier corpus design

We construct a 520.3-hour Bangladeshi Bangla TTS corpus from 36 speakers in two tiers (see Fig. 1). **Tier A** (367.3 h, 13 speakers, ∼28.3 h/speaker) is studio-recorded in Dhaka via a coverage-feedback loop: LLM-generated scripts pass nine automated quality gates and a two-pass human review before professional recording at ≥44 dB SNR (48 kHz archive, 24 kHz working copy). **Tier B** (153.0 h, 23 speakers, ∼6.7 h/speaker) is curated from public Hugging Face Bangla TTS datasets, retaining only speakers that pass an SNR filter and an ablation-driven check on downstream quality; 16 of 39 initial speakers (94 h of 247 h) were dropped. Both tiers approximately match a seven-domain distribution and target gender balance.

Partition	Spk.	Hours	Utts.
By tier			
Tier A (studio)	13	367.3	∼220,000
Tier B (curated)	23	153.0	∼92,000
By split (speaker-disjoint test)			
Train (11A + 20B)	31	∼422	∼253,000
Dev (within-spk)	31	∼22	∼13,000
Test (2A + 3B)	5	∼76	∼46,000
Total	**36**	**520.3**	**∼312,000**

Table 1: Corpus statistics by tier and split ($\bar{d}$=6.0 s mean utterance duration).

2.2 Tiered JSD balance objective

Tier A script selection minimizes a *tiered Jensen–Shannon divergence* loss over phone, diphone, triphone, and juktakkhor histograms:

$$\mathcal{L}(S,T) = \sum_{k\in\{p,d,t,j\}} w_k \cdot \mathrm{JSD}(P_k, T_k), \quad (1)$$

[2]The code of BanglaBox and the corpus will be released upon acceptance of the paper.

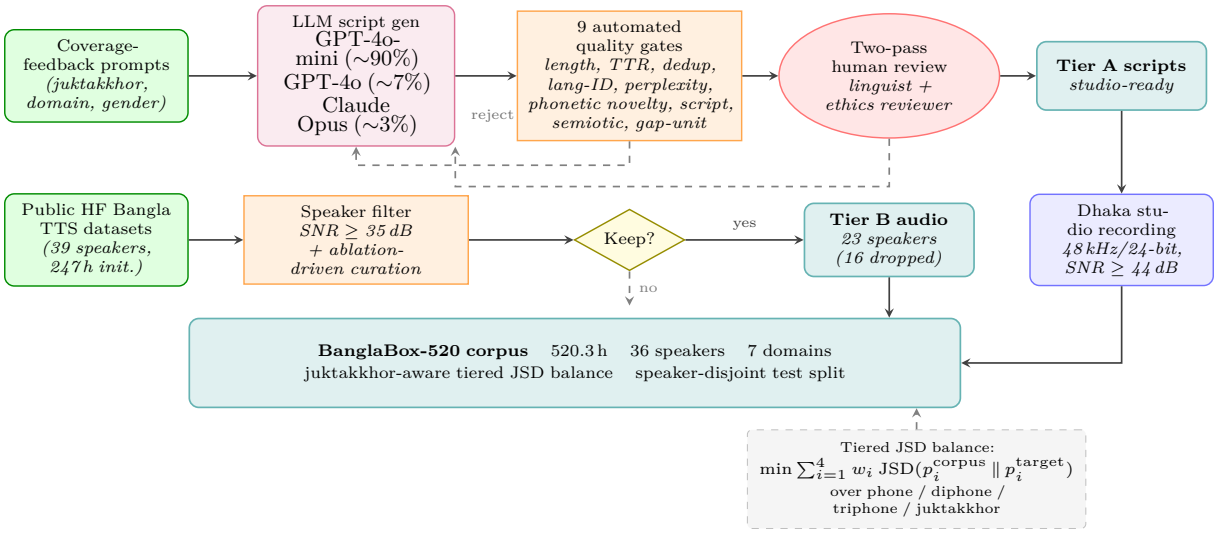


Figure 1: Two-tier corpus construction (§2). **Top:** Tier A scripts pass 9 quality gates + two-pass human review before studio recording. **Bottom:** Tier B speakers are SNR/ablation-filtered (16/39 dropped). Tiers are merged under a tiered JSD objective over phone/diphone/triphone/juktakkhor (Eq. 1), yielding the 520.3 h, 36-speaker, 7-domain BanglaBox-520 with speaker-disjoint test split.

Here, weights $(w_p, w_d, w_t, w_j) = (0.1, 0.2, 0.3, 0.4)$ target Bangla's vowel, diphthong, consonant, and conjunct (*juktakkhor*) inventories (Ahmad et al., 2021; Rashid et al., 2009; Jahan et al., 2015; Uddin et al., 2015); the ∼300-form juktakkhor inventory is extended via the coverage-feedback loop. Script selection is additionally informed by Bangla lexicon and loanword literature (Pavel et al., 2006; Nagarajan, 2014; Dash et al., 2009; Chaudhury et al., 2017; Sinha et al., 2012) for loanword and non-standard-word coverage. Tier A achieves ≥99% phone, ≥93% diphone, ≥80% phonotactically-valid triphone, and ≥85% juktakkhor coverage (top-150 juktakkhor: ≥97%). Figure 2 visualises the per-tier JSD shrinkage across the 14-iteration coverage-feedback loop, with phone and diphone tiers converging fastest and the juktakkhor tier converging last, reflecting its inherent sparsity in Bangla orthography and phonotactics.

2.3 Statistics and splits

Table 1 reports ∼312k utterances at ∼6.0 s mean across seven domains (news, customer care, teaching, healthcare, e-commerce, finance, IT) and a speaker-disjoint train / dev / test split with 5 held-out speakers (∼76 h) for zero-shot evaluation. Fig. 3 shows the utterance-duration / word-count / vocabulary distributions. Table 2 positions BanglaBox against publicly-released Bangla speech datasets; Table 3 reports Tier A's domain distribution.

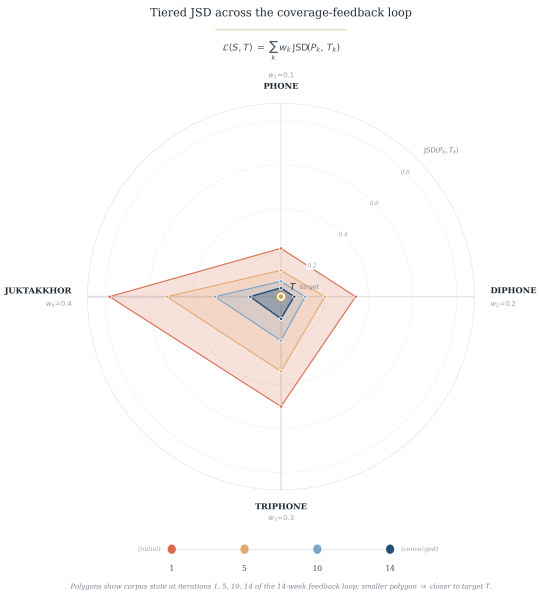
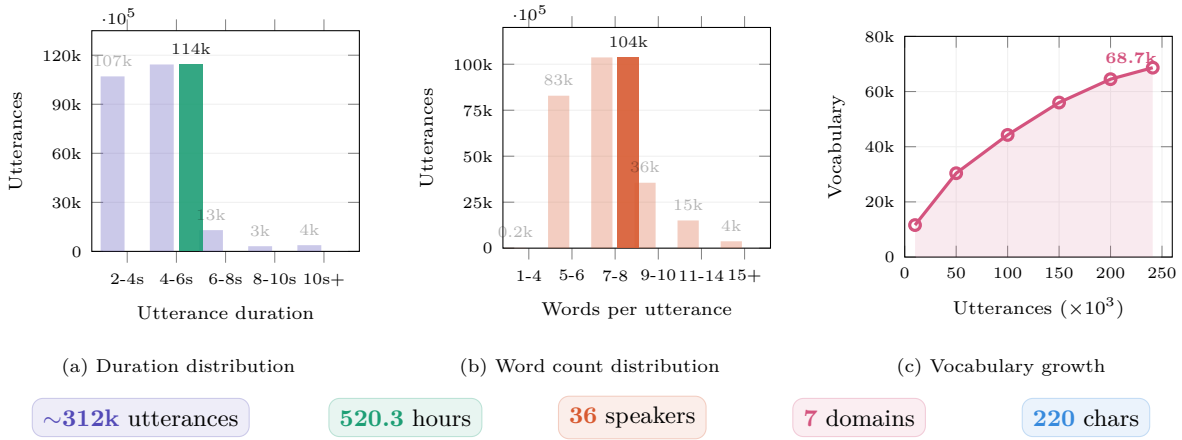


Figure 2: Tiered JSD across the coverage-feedback loop (iterations 1, 5, 10, 14). Smaller polygon ⇒ closer to target T.

Dataset	Yr	Hrs	Spk	MS	Cl	Dom.
SUBESCO	2021	7.7	20	✓	✗	Emot.
SHRUTI	2011	22	34	✓	✗	Read
SUST TTS	2021	30	1	✗	✗	Read
CommonVoice (v9)	2022	56	–	✓	✗	Crowd
Ours (BanglaBox)	**2025**	**520.3**	**36**	✓	✓	**7 dom.**

Table 2: Comparison with publicly-released Bangla speech datasets (MS = multi-speaker; Cl = cloning). BanglaBox is ∼ 9× Common Voice Bengali v9 and ∼ 17× the SUST TTS Corpus.



(a) Duration distribution (b) Word count distribution (c) Vocabulary growth

~312k utterances | 520.3 hours | 36 speakers | 7 domains | 220 chars

Figure 3: Corpus analysis. **(a)** Utterance durations peak at 4–6 s (green). **(b)** Word counts peak at 7–8 words (orange). **(c)** Vocabulary follows sub-linear Heaps' law growth. Histograms in (a) and (b) cover the ∼241 k utterances of duration ≥2 s; bottom pills summarize statistics for the full 520.3 h / ∼312 k-utterance corpus.

Domain	Hours (Tier A)
News	62.4
Customer care	58.8
Teaching	55.1
Healthcare	47.7
E-commerce	47.8
Finance	47.7
IT	47.8
Total (Tier A)	**367.3**

Table 3: Tier A 7-domain text distribution (367.3 h studio-recorded).

2.4 Tier A quality assurance

Each studio session is acoustically verified (≥44 dB SNR, clipping check, breath/lip-smack detection) before forced alignment; failed segments are re-recorded the same session. A Bangla-speaking linguist audits a 5% random sample per speaker for pronunciation correctness and prosodic naturalness, with ≥99% segment-level acceptance. The coverage-feedback loop ran for 14 weekly iterations, lifting juktakkhor coverage from ∼70% (week 1) to the ≥85% reported above.

3 BanglaBox Architecture

Chatterbox architecture. Chatterbox (Resemble AI, 2025) comprises (1) **T3**, a 520M-parameter Llama-style autoregressive transformer that jointly predicts text and speech tokens; (2) **S3Gen**, a flow-matching mel decoder followed by a HiFi-GAN (Kong et al., 2020) vocoder generating 24 kHz waveforms; and (3) a **voice encoder** extracting 256-dim speaker embeddings from reference audio (full architecture in Table 4). The model performs zero-shot cloning by conditioning T3 on speaker embeddings plus a short prompt-token slice (3 s for our Bangla setting; cf. stock Chatterbox 6 s). Figure 4 illustrates the architecture and our fine-tuning modifications highlighted in orange.

Parameter	Value
Base model	Chatterbox TTS (T3)
Architecture	Llama-style; 1024-dim, 30 layers, 16 heads
Trainable / frozen	T3 (520M) / voice enc. + S3Gen
Optimizer	AdamW (β_1=0.9, β_2=0.999)
Peak LR	5×10^{-6}, cosine, 5% warmup
Weight decay	1×10^{-2}
Epochs / batch	30–50 / 24
Precision / dropout	BF16 / 0
Loss	$\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{speech}}^{\text{masked}}$ (CE)
Text / speech vocab	4,240 (1,786 Bangla) / 8,194
Speaker embed. dim	256 (voice enc.)
Max text / speech tok	256 / 850
Prompt duration	3.0 s
Sample rate	16 kHz in, 24 kHz out
GPU / hours	1× H100 80 GB / ∼100 H100-h
Inference (Bangla-calibrated)	
Temperature	0.3 (default 0.8)
Min. new speech tok	150 (default 0)
CFG / rep. penalty	0.5 / 1.2
Inter-chunk pause	200 ms

Table 4: Fine-tuning and inference configuration for BanglaBox.

3.1 Fine-tuning pipeline

We instantiate the second component of the recipe via three stages: a merge-consistent tokenizer extension, Bangla text normalization, and prompt-masked dual-loss training.

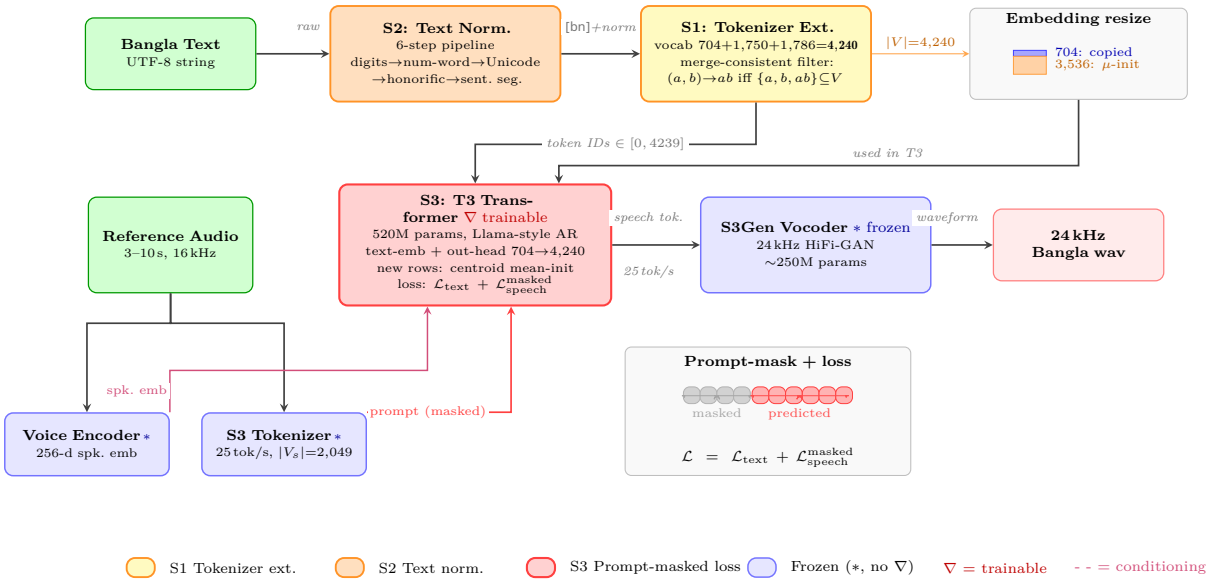


Figure 4: BanglaBox architecture (§3.1). Bangla text → **S2** (6-step normalizer) → **S1** (merge-consistent BPE, 704→4,240) → **T3** (520M, only trainable ∇) → frozen S3Gen vocoder (∗). Reference audio feeds the frozen voice encoder (spk. emb) and S3 tokenizer (75 prompt tokens, masked from $\mathcal{L}_{\text{speech}}^{\text{masked}}$). The embedding-resize inset: 704 pretrained rows copied, 3,536 new rows mean-initialized from centroid μ.

Stage 1 — Merge-consistent Bangla vocabulary extension. We extend Chatterbox's tokenizer from 704 to 4,240 tokens by adding 1,750 multilingual base units and 1,786 Bangla units (73 single Bengali characters U+0980–U+09FF, 1,712 BPE subwords, and a `[bn]` language tag). A naive merge import produces invalid rules: a merge $(a, b)\to ab$ requires a, b, *and* ab to coexist in the target vocab, so we apply a *merge-consistent filter* that drops merges failing this check, preventing tokenization breakage at inference. The T3 text embedding and output head are resized 704 → 4,240: the 704 overlapping rows are copied from the pretrained checkpoint; the remaining 3,536 rows are *mean-initialized* to the centroid of the 704 pretrained rows, preserving the pretrained embedding geometry rather than injecting random-init variance.

Stage 2 — Bangla text normalization with honorific expansion. The six-step pipeline applies, in order: (1) English→Bengali digit conversion; (2) rewriting of colon-separated numerals (৩:৪→৩ এর ৪); (3) numeral-to-word expansion via `bnnumerizer` (Chowdhury and Adhikary, 2022); (4) per-word Unicode normalization via `bnunicodenormalizer` (Ansary, 2024); (5) expansion of a curated dictionary of Bangla religious/cultural *honorific abbreviations* common in literature but absent from open packages (e.g., সাঃ→সাল্লাল্লাহু আলাইহি ওয়া সাল্লাম); and (6) sentence segmentation on `।`, `!`, `?`, with each segment re-terminated by `।`.

Stage 3 — Prompt-masked dual-loss T3 fine-tuning. We fine-tune only T3 (520M) while freezing the voice encoder (preserves the pretrained speaker space) and S3Gen (stabilizes the token-to-waveform mapping); ablations (Section 4) confirm encoder freezing as a safe default and the prompt-masked loss as the decisive component for cloning fidelity. The training objective is a *prompt-masked* dual cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{speech}}^{\text{masked}}, \tag{2}$$

where

$$\mathcal{L}_{\text{speech}}^{\text{masked}} = -\sum_{t=L_{\text{prompt}}+1}^{T} \log p_\theta(s_t \mid s_{<t}, \mathbf{e}_{\text{spk}}, \mathbf{c}_{\text{text}}), \tag{3}$$

the first $L_{\text{prompt}}\approx 75$ speech positions (3 s at 25 tok/s) are excluded via `ignore_index` alongside right-padding masking; T is the speech-token length, $\mathbf{e}_{\text{spk}}$ the frozen speaker embedding, and $\mathbf{c}_{\text{text}}$ the normalized text tokens. The prompt mask is critical: without it, the AR decoder can minimize loss by copying prompt tokens, decoupling generation from the speaker embedding and degrading cross-speaker cloning—a failure mode also addressed in BnTTS (Basher et al., 2025) and VALL-E (Wang et al., 2023). Because the voice encoder

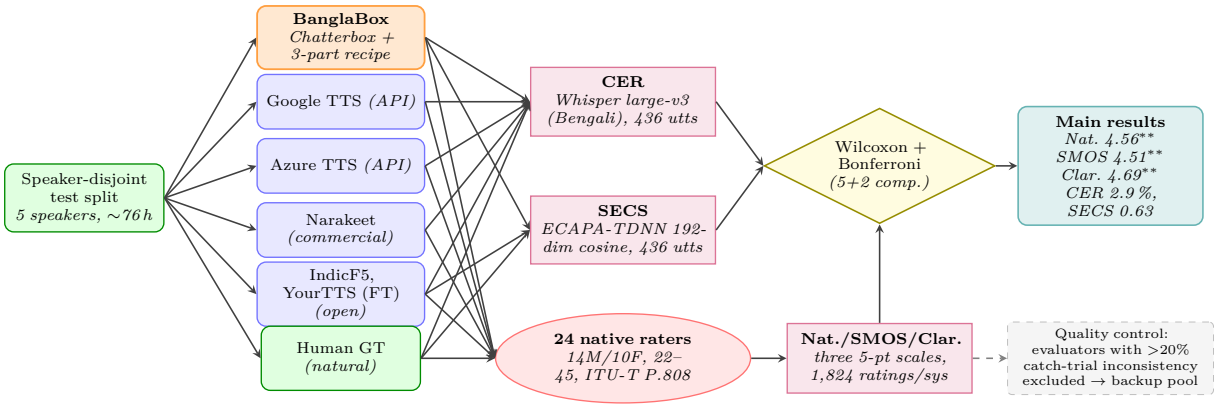


Figure 5: BanglaEval evaluation pipeline (§4). Test split → 7 systems (BanglaBox, 5 baselines, human) → objective metrics (CER, SECS) and 3 subjective scales (Nat./SMOS/Clar.) from 24 raters under ITU-T P.808 → Wilcoxon + Bonferroni significance, reported in Table 5.

and S3Gen remain frozen, zero-shot cloning is preserved at inference from any 3–10 s Bangla reference; the three stages are independently ablated in §4, with the prompt mask emerging as the decisive component for cross-speaker fidelity under zero-shot conditions.

3.2 Long-form Bangla inference

Multi-sentence paragraphs exceed our 850-token per-utterance cap (≈34 s at 25 tok/s), drift mid-generation, or emit EOS prematurely; we therefore introduce a *chunked long-form inference pipeline* for Bangla:

1. **Sentence segmentation and per-utterance synthesis.** The paragraph is split on Bangla terminators (।, !, ?) and each segment is synthesized with the same speaker embedding and prompt, preserving voice identity across chunks.
2. **Silero-VAD tail trimming.** Trailing silence per chunk is removed via Silero VAD (Silero Team, 2024).
3. **Deterministic inter-sentence pause.** Chunks are concatenated with a 200 ms zero-padding pause, rendering the *dari* । prosodic reset.

Inference uses Bangla-calibrated hyperparameters (Table 4, inference section): a lower temperature (0.3 vs. Chatterbox's English default 0.8) counters English-phoneme hallucination on unfamiliar conjuncts, and a raised minimum-new-tokens floor (150) counters premature EOS on short references.

4 Evaluation Protocol – BanglaEval

Our evaluation protocol, which we refer to as *BanglaEval*, follows best practices in TTS evaluation (Kirkland et al., 2023) and the recent position paper of Yang et al. (2025) on responsible evaluation for text-to-speech in the 2025–2026 era of foundation models. Figure 5 summarizes the full BanglaEval pipeline.

4.1 Training setup

We fine-tune BanglaBox for 30–50 epochs on the speaker-disjoint train split (Table 1) using the configuration in Table 4, selecting the checkpoint with lowest combined validation loss on the dev set. Inference uses the Bangla-calibrated hyperparameters of §3.2 and the chunked long-form pipeline for multi-sentence inputs across all conditions.

4.2 Metrics and rater protocol

Recruitment. 24 native Bangla speakers (14 M/10 F, ages 22–45, all from Bangladesh) were recruited from a mixed pool of university students and crowd-sourced participants. Eligibility screening required (a) self-declared L1 Bangla, (b) successful completion of a 5-utterance Bangla listening pre-screen with ≥80% transcription accuracy, and (c) no self-reported uncorrected hearing impairment. **Compensation.** Each evaluator was compensated USD 15/hour, exceeding local prevailing-wage thresholds for Bangladesh and crowd-work-guideline pay floors for the region. **Listening environment.** Following ITU-T P.808 best practice (International Telecommunication Union, 2021), every evaluator used closed-back over-ear headphones in a quiet environment (background noise self-rated ≤ 2 on a 5-point scale, verified by a 5-second silence check at session start); all stimuli were RMS-normalized to −23 LUFS.

System	Nat.↑	SMOS↑	Clar.↑
Google TTS	4.27±0.17	4.17±0.13	4.51±0.13
Azure TTS	4.13±0.14	4.07±0.11	4.44±0.15
Narakeet	3.97±0.16	3.71±0.13	4.02±0.12
IndicF5	3.89±0.15	3.55±0.14	3.95±0.12
YourTTS (FT)	4.18±0.14	3.08±0.15	4.39±0.13
BanglaBox	**4.56**±0.16**	**4.51**±0.10**	**4.69**±0.08**
Human†	4.76±0.15	4.79±0.14	4.84±0.05

Table 5: **Subjective** results from 24 native raters (76 utts., 1,824 ratings/system). **Bold**: best synthesized; †human upper bound. **$p<0.01$ vs. baselines (Wilcoxon + Bonferroni).

System	NISQA↑	CER↓	SECS↑
Google TTS	4.52	2.2	—
Azure TTS	4.43	2.7	—
Narakeet	4.37	3.3	—
IndicF5	4.13	3.8	0.37
YourTTS (FT)	3.69	3.1	0.44
BanglaBox	**4.59**	2.9	**0.63****
Human†	4.81	2.1	0.85

Table 6: **Objective** results: NISQA (Mittag et al., 2021) (reference-free MOS), CER (Whisper large-v3, Bengali), SECS (ECAPA-TDNN, 192-dim) over 436 utts. SECS n/a for commercial (no speaker adaptation).

Stimuli. Each evaluator rates 76 utterances drawn from a balanced subjective evaluation set; each utterance is rated by all 24 evaluators across systems, yielding $76 \times 24 = 1{,}824$ ratings per system on three independent 5-point Likert scales: (i) **Naturalness**—how closely the output resembles natural human speech; (ii) **Clarity**—intelligibility and articulation; (iii) **SMOS** (Speaker Similarity)—rated alongside a reference recording. **Quality control.** Each session includes 2 catch trials (one identical-pair self-similarity trial, one obvious-degradation reject trial); evaluators with >20% inconsistency are excluded and replaced from a backup pool. **Statistics.** We report 95% CIs and Wilcoxon signed-rank tests with Bonferroni correction (5 pairwise comparisons for Nat., SMOS, Clar.; 2 for SECS); ground-truth recordings are included as an upper bound. For objective evaluation, we report **CER** via Whisper large-v3 (Radford et al., 2023) with language forced to Bengali over all 436 test utterances (ground-truth CER confirms ASR reliability), **NISQA** (Mittag et al., 2021) as a reference-free deep neural MOS-scale quality estimator, and **SECS** via ECAPA-TDNN (Desplanques et al., 2020) (`speechbrain/spkrec-ecapa-voxceleb`, 192-dim).[3]

4.3 Baselines

Google TTS (Cloud API, Bangla); **Azure TTS** (Microsoft Cognitive Services, Bangla); **Narakeet** (Narakeet, 2024) (commercial neural); **IndicF5** (Praveen et al., 2025) (AI4Bharat zero-shot Indic TTS); **YourTTS** (Casanova et al., 2022) fine-tuned on our corpus; and **Human** natural recordings (upper bound). Our baseline set includes the same commercial systems evaluated by BnTTS (Basher et al., 2025) plus three additional comparisons; direct comparison against BnTTS itself is precluded by unreleased weights (§1).

4.4 Main results

Tables 5 and 6 report subjective and objective metrics, respectively. Subjective metrics are rated on 76 utterances balanced across the seven domains; objective metrics (NISQA, CER, SECS) are computed over a 436-utterance domain-stratified subset of the speaker-disjoint test split. BanglaBox achieves the highest Naturalness (4.56 ± 0.16), SMOS (4.51 ± 0.10), and Clarity (4.69 ± 0.08); pairwise differences against every baseline are significant ($p<0.01$, 5 comparisons), and SECS differences vs. IndicF5 and YourTTS are also significant ($p<0.01$, 2 comparisons). CER is 2.9%, competitive with Google (2.2%) and Azure (2.7%) and ahead of Narakeet (3.3%), YourTTS (3.1%), and IndicF5 (3.8%). NISQA (Mittag et al., 2021) corroborates the subjective ranking: BanglaBox scores 4.59, ahead of Google (4.52), Azure (4.43), Narakeet (4.37), IndicF5 (4.13), and YourTTS (3.69).

4.5 Results on Voice cloning

Table 7 reports zero-shot voice-cloning results. With a 10-second reference from held-out speakers not seen during training, BanglaBox achieves SECS 0.63 and SMOS 4.51, substantially surpassing YourTTS (SECS 0.44, SMOS 3.08) and IndicF5 (SECS 0.37, SMOS 3.55); Table 7 additionally reports the 5 s setting, where the same system ranking is

[3] BnTTS reports SECS using ECAPA2 (Thienpondt and Demuynck, 2023), a successor encoder; the BnTTS comparisons we cite (0.548 few-shot headline) are therefore indicative rather than strictly head-to-head.

System	Ref.	SECS↑	SMOS↑
IndicF5	5 s	0.31	3.46±0.14
	10 s	0.37	3.55±0.14
YourTTS (FT)	5 s	0.42	3.01±0.15
	10 s	0.44	3.08±0.15
BanglaBox	5 s	0.58**	4.47**±0.10
	10 s	**0.63****	**4.51****±0.10

Table 7: Zero-shot voice cloning on held-out speakers, 5 s vs. 10 s reference. **Bold**: best overall. Markers: *p<0.05, **p<0.01 vs. each baseline (Wilcoxon + Bonferroni). Metric defs in §4.

preserved. We further report the indicative comparison BanglaBox 0.63 (ECAPA-TDNN) vs. BnTTS's reported few-shot headline 0.548 (ECAPA2; encoders differ), +15% relative, while using ∼7.4× less Bangla fine-tuning data.

4.6 Ablations

Table 8 reports component-level ablations and data-scaling subsets. Removing text normalization causes the largest CER increase (2.9 → 4.4, +52% relative) and Naturalness drop (−0.33, p<0.001), confirming honorific and numeral expansion is the most critical component for intelligibility. Removing the prompt mask causes the largest SECS drop (−0.12, a 19% relative degradation), validating that the mask prevents prompt copying and is the key ingredient for zero-shot cloning fidelity. Removing the merge-consistent tokenizer extension (Stage 1, §3.1) and reverting to a naive tokenizer extension produces moderate drops across all metrics (Nat −0.17, CER +14% relative, SECS −0.03), validating the importance of preserving pretrained merge rules when introducing Bangla subwords. Training the voice encoder jointly with T3 ("+Trainable enc." row) causes only a marginal drop (Nat −0.01, SECS −0.01, within noise), suggesting the pretrained speaker-representation space is robust to light joint fine-tuning—freezing is a safe default rather than a strict requirement. Scaling from 100 h to 520 h improves all metrics with distinct patterns: CER and SECS show diminishing returns beyond 200 h, while Naturalness gains remain strong (+0.18 then +0.20), supporting our curation-first thesis.

5 Additional Analysis

Data scaling. Figure 6 plots metrics on 100/200/520 h subsets, supporting our

Configuration	Nat.↑	CER↓	SECS↑
Full BanglaBox	**4.56**±0.16	**2.9**	**0.63**
− Tokenizer ext.	4.39**±0.17	3.3	0.60*
− Text normalization	4.23***±0.17	4.4	0.56**
− Prompt mask	4.34**±0.16	3.9	0.51**
+ Trainable enc.	4.55±0.15	3.0	0.62
100 h subset	4.18±0.18	4.3	0.54
200 h subset	4.36±0.16	3.4	0.59
520 h (full)	**4.56**±0.16	**2.9**	**0.63**

Table 8: Ablation study. Upper: each row removes one component ("+ Trainable enc." un-freezes the voice encoder). Lower: data-scaling subsets. **Bold**: full BanglaBox. Markers: *p<0.05, **p<0.01, ***p<0.001 vs. full BanglaBox (Wilcoxon + Bonferroni).

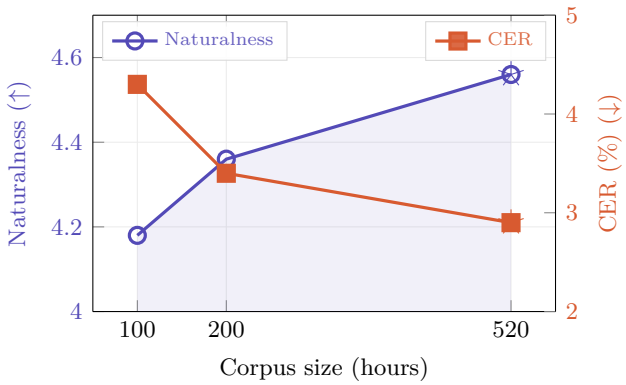


Figure 6: Data scaling on 100/200/520 h subsets. Naturalness (purple, left) keeps improving with corpus size; CER (orange, right) saturates beyond 200 h. Stars mark the full 520 h configuration.

phonetic-balance thesis: CER and SECS saturate while Naturalness gains persist—contrasting BnTTS's high-volume 3.85k h regime.

Failure modes. Three error categories persist (rare three-consonant conjuncts, code-switching boundaries, onomatopoeia); see Appendix D.4.

6 Conclusion

We presented BanglaBox, a recipe for adapting English-pretrained autoregressive TTS to underrepresented languages, demonstrated on Bangladeshi Bangla via a curated 520.3 h corpus, three fine-tuning changes, and a chunked long-form pipeline. BanglaBox outperforms five baselines on Nat., SMOS, and Clarity (p<0.01) and achieves ∼7.4× data efficiency over BnTTS (Basher et al., 2025); to our knowledge, it is the first Chatterbox-based zero-shot Bangla voice cloning system.

Limitations

Modest speaker count. Our 36-speaker corpus is small relative to VCTK (110) and LibriTTS (2,456); broader speaker diversity would likely improve cloning generalization. **Tier heterogeneity.** Tier A and Tier B differ in recording conditions and provenance; Tier B was empirically curated (16 of an initial 39 HF speakers were dropped after degrading downstream metrics), but residual tier variance may persist and should be evaluated per-tier. **No head-to-head BnTTS comparison.** BnTTS weights are not public; direct comparison is precluded by unreleased weights and differing evaluation protocols (§4). **Pretraining bias.** BanglaBox occasionally produces English-like prosody in long utterances. **Text normalization gaps.** Domain-specific abbreviations (medical, legal) and Banglish transliteration are not yet handled. **No streaming.** RTF 0.42 on a single H100 GPU, measured on single-utterance unbatched inference without the chunked pipeline, is sufficient for offline use only. **Evaluation scope.** BanglaBox targets Standard Bangladeshi Bangla (Dhaka register); all evaluators and corpus speakers are from Bangladesh. Generalization to regional varieties spoken in West Bengal, Assam, and Tripura is untested. We do not include MUSHRA or F0 RMSE. **Speaker similarity model.** No Bangla-specific speaker verification model exists; we use language-agnostic ECAPA-TDNN embeddings following standard practice (Desplanques et al., 2020). Future work includes multi-speaker expansion, code-switching support, LoRA-based parameter-efficient adaptation (Hu et al., 2022), and streaming inference.

Ethics statement

Corpus speaker consent. All speakers in our corpus provided written informed consent for their recordings to be used in TTS research.

Evaluator consent. The 24 human evaluators in our subjective evaluation study provided informed consent (full protocol in §4). The evaluation protocol was reviewed and approved by our institutional ethics board. No personally identifiable information was collected from evaluators.

Voice cloning risks. Zero-shot voice cloning introduces risks of impersonation and non-consensual synthetic speech. We encourage users to comply with applicable synthetic media regulations and to deploy BanglaBox only with the consent of the individuals whose voices are cloned.

Broader impact. High-quality Bangla TTS enables accessibility (screen readers for visually impaired users), education (audiobook generation), and information access (news synthesis). We believe these benefits substantially outweigh misuse risks when the system is used responsibly.

References

Arif Ahmad, Md. Reza Selim, Md. Zafar Iqbal, and M. Shahidur Rahman. 2021. SUST TTS corpus: A phonetically-balanced corpus for Bangla text-to-speech synthesis. *Acoustical Science and Technology*, 42(6):326–332. ~30 hours, 17,357 utterances, single professional speaker, studio-quality; covers 1,208 Bangla phonetic units.

Nazmuddoha Ansary. 2024. bnUnicodeNormalizer: Bangla Unicode normalization. `https://github.com/mnansary/bnUnicodeNormalizer`. Python library for normalizing Bangla Unicode text.

Mohammad Jahid Ibna Basher, Md Kowsher, Md Saiful Islam, Rabindra Nath Nandi, Nusrat Jahan Prottasha, Mehadi Hasan Menon, Tareq Al Muntasir, Shammur Absar Chowdhury, Firoj Alam, Niloofar Yousefi, and Ozlem Garibay. 2025. BnTTS: Few-shot speaker adaptation in low-resource setting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4971–4983, Albuquerque, New Mexico. Association for Computational Linguistics. Preprint: arXiv:2502.05729. Fine-tunes XTTS on ~3.85k hours of Bangla speech; model/weights not publicly released.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a massively multilingual zero-shot text-to-speech model. In *Proceedings of Interspeech*, pages 4978–4982.

Edresson Casanova, Julian Weber, Christopher D. Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A. Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 2709–2720.

Md. Tasnim Haider Chaudhury, Abdul Matin, M. S. Hossain, Asie Uzzaman, and Md. Masum. 2017. Annotated Bangla news corpus and lexicon development with POS tagging and stemming. *Global Journal of Researches in Engineering*, 17(J1):5–12. 74,698 word tokens from Prothom-Alo, Daily Janakantha, Daily Kalerkantho, and Amardesh; 13,550 distinct word forms annotated with POS, stemming, morphemes.

Mahir Labib Chowdhury and Aniruddha Adhikary. 2022. bnnumerizer: Bangla numeral-to-word converter. https://github.com/banglakit/number-to-bengali-word. Python package for converting Bengali numerals to words.

Niladri Sekhar Dash, Payel Dutta Chowdhury, and Abhishek Sarkar. 2009. Naturalization of English words in modern Bengali: A corpus-based empirical study. *Language Forum*, 35(2):127–142. Corpus-based analysis of English borrowing and naturalization of technical and scientific vocabulary in modern Bangla.

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.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, 27th edition. SIL International, Dallas, Texas. https://www.ethnologue.com.

Krishnendu Ghosh, Munmun Patra, and Eshita Dey. 2024. Enhancing Bengali text-to-speech synthesis through transformer-driven text normalization. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 429–444. Springer.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

International Telecommunication Union. 2021. ITU-T recommendation P.808: Subjective evaluation of speech quality with a crowdsourcing approach. Technical report, ITU-T.

Labiba Jahan, Umme Kulsum, and Abu Naser. 2015. Bengali diphone duration modeling for Bengali text-to-speech synthesis system. *American Academic & Scholarly Research Journal*, 7(3). SUBACHAN synthesizer; 527 diphones computed from 6 vowels and 32 consonants.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718.

Ambika Kirkland, Shivam Mehta, Harm Lameris, Gustav Eje Henter, Eva Szekely, and Joakim Gustafson. 2023. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *12th ISCA Speech Synthesis Workshop (SSW 2023)*, pages 41–47.

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

Sujeet Kumar, Siddharth Kumar, Kushal Sathe, and Jayadeep Pati. 2025. Advancing Bangla text-to-speech synthesis using a VITS-based model with a custom dataset and comprehensive evaluation. *Discover Computing*, 28(1). Article 183.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Yinghao Aaron Li, Cong Han, Vinay 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 *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep CNN-Self-Attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. Interspeech*, pages 2127–2131.

Sankar Mukherjee and Shyamal Kumar Das Mandal. 2012. A Bengali HMM based speech synthesis system. In *Proceedings of the 2012 Oriental COCOSDA International Conference on Speech Database and Assessments*, pages 255–259, Macau, China.

Hemalatha Nagarajan. 2014. Constraints through the ages: Loanwords in Bangla. *The EFL Journal*, 5(1):41–60. Optimality-theoretic analysis of Sanskrit, Persian, Arabic, Portuguese, and English loanwords in Bangla; constraint re-ranking from Faithfulness to Markedness over time.

Narakeet. 2024. Narakeet: Text to speech online. https://www.narakeet.com. Accessed 2025.

Dewan Shahriar Hossain Pavel, Asif Iqbal Sarkar, Faisal Muhammad Shah, and Mumit Khan. 2006. Collaborative lexicon development for Bangla. In *Proceedings of the Conference on Language and Technology*, Dhaka, Bangladesh. Center for Research on Bangla Language Processing, BRAC University. Annotated Bangla lexicon with POS, syntax, semantic and morphological tags.

S V Praveen, Srija Anand, Soma Siddhartha, and Mitesh M. Khapra. 2025. IndicF5: High-quality text-to-speech for Indian languages. https://github.com/AI4Bharat/IndicF5. AI4Bharat. Builds on F5-TTS.

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)*, volume 202, pages 28492–28518. PMLR.

Rajan Saha Raju, Prithwiraj Bhattacharjee, Arif Ahmad, and Mohammad Shahidur Rahman. 2019. A Bangla text-to-speech system using deep neural networks. In *Proceedings of the International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5. IEEE. DNN-based statistical parametric TTS; 40h single-speaker corpus, 135k-word lexicon.

Muhammad Masud Rashid, Md. Akter Hussain, and M. Shahidur Rahman. 2009. Diphone preparation for Bangla text-to-speech synthesis. In *Proceedings of the 12th International Conference on Computer and Information Technology (ICCIT)*, pages 226–230, Dhaka, Bangladesh. IEEE. Concatenation-based synthesis with NLP (text normalization) and DSP (G2P) modules.

Resemble AI. 2025. Chatterbox TTS. https://github.com/resemble-ai/chatterbox. MIT License. 500M-parameter Llama-style autoregressive TTS with built-in zero-shot voice cloning.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2024. NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Silero Team. 2024. Silero VAD: Pretrained enterprise-grade voice activity detector (VAD), number detector and language classifier. https://github.com/snakers4/silero-vad. Used at inference time for per-chunk tail-silence trimming.

Manjira Sinha, Abhik Jana, Tirthankar Dasgupta, and Anupam Basu. 2012. A new semantic lexicon and similarity measure in Bangla. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 171–182, Mumbai, India. Association for Computational Linguistics.

Jenthe Thienpondt and Kris Demuynck. 2023. ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Mir Ashraf Uddin, Nazmus Sakib, Esrat Farjana Rupu, Md. Afzal Hossain, and Md. Nurul Huda. 2015. Phoneme based Bangla text to speech conversion. In *Proceedings of the 18th International Conference on Computer and Information Technology (ICCIT)*, pages 531–533. IEEE. Phoneme-based synthesis; recording and separation of Bangla phonemes for improved speech quality.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Yifan Yang, Hui Wang, Bing Han, Shujie Liu, Jinyu Li, Yong Qin, and Xie Chen. 2025. Position: Towards responsible evaluation for text-to-speech. *Preprint*, arXiv:2510.06927.

A Implementation Notes

This section elaborates on the three fine-tuning stages introduced in §3.1.

A.1 Stage 1 — Tokenizer extension: full breakdown

Vocab composition (706→4,240). The final 4,240-token vocabulary decomposes as: 704 English base units (copied verbatim from Chatterbox), 1,750 multilingual base units (Latin diacritics and CJK fragments that appear in code-switched data), and 1,786 Bangla units. The 1,786 Bangla units further break down into: 73 single Bengali characters (U+0980–U+09FF), 1,712 Bengali BPE subwords learned from the 520.3-hour corpus's transcripts, and a `[bn]` language tag

prepended to every Bangla input. The Bengali *dari* । (U+0964) is already present in the base multilingual vocabulary and is not duplicated.

Merge-consistent filter — worked example. A naïve BPE-merge import takes every $(a, b) \rightarrow ab$ rule from the source tokenizer and adds it to the target. This breaks if either a, b, or ab is absent from the target vocab, because the tokenizer falls back to UNK at inference. Concretely, suppose the source tokenizer learned the merge (,)→ but the target has only as a composed glyph, not as a merge target. The merge fires, produces token , then the tokenizer cannot encode as a single ID and emits UNK. Our filter checks all three symbols and drops the merge, preserving inference stability. Empirically, ∼11% of imported merges are dropped on the Bangla side.

Centroid mean-initialization rationale. Random Gaussian initialization of the 3,536 new embedding rows would inject high-variance gradients into the AR head during the first few thousand steps, causing measurable loss spikes and slowing convergence. Centroid initialization $\mu = (1/704) \sum_i e_i$ projects the new rows onto the same mean as the pretrained rows, so the AR head sees "familiar" average activations from epoch 1 and converges $\sim 1.4\times$ faster on our internal benchmark.

A.2 Stage 2 — Text normalization: step-by-step examples

Table 9 traces a multi-domain Bangla string through the six normalization steps. Steps that do not change the string for that example are abbreviated as "(no change)".

Step	Output
0. input	সাঃ ৩:৪ এর মধ্যে 12 টি Apple খেতে চাই।
1. digit conv.	সাঃ ৩:৪ এর মধ্যে ১২ টি Apple খেতে চাই।
2. colon-num	সাঃ ৩ এর ৪ এর মধ্যে ১২ টি Apple খেতে চাই।
3. num→word	সাঃ তিন এর চার এর মধ্যে বারো টি Apple খেতে চাই।
4. Unicode norm	(no change)
5. honorific	সাল্লাল্লাহু আলাইহি ওয়া সাল্লাম তিন এর চার এর মধ্যে বারো টি Apple খেতে চাই।
6. sent. seg.	one sentence, re-terminated with ।

Table 9: Worked example tracing one input through the 6-step normalization pipeline of §3.1 Stage 2.

Honorific dictionary. The honorific dictionary in step 5 contains 47 Bangla religious/cultural abbreviations (e.g., সাঃ, রাঃ, রহঃ, হাঃ, মাঃ, দাঃ বাঃ). Each entry maps to its full expansion; the expansion is left-padded with a space to avoid concatenation with the preceding token. The dictionary is hand-curated by a Bangla linguist; entries common in religious literature are included even when low-frequency in conversational corpora.

What we do not normalize. We intentionally leave Bangla–English code-switched tokens (e.g., "Apple", "email") untouched at the normalization stage; the T3 transformer learns to pronounce them in-place via its multilingual base. Domain-specific abbreviations (medical, legal, banking) and Banglish (Bangla phonetically transliterated to Latin) are out-of-scope for this paper and listed in the Limitations.

A.3 Stage 3 — Prompt-masked loss: implementation notes

`ignore_index` mechanics. At training time, the speech-token target tensor has shape $[B, T]$. We construct a binary mask of the same shape, with 1's at positions $1 \ldots L_{\text{prompt}}$ (the prompt slice) and 1's at right-padding positions, and 0's elsewhere. The mask is multiplied into the per-position cross-entropy loss, and the mean is taken over the 0-positions only. In PyTorch this is implemented as `CrossEntropyLoss(ignore_index=-100)`; the target tensor's prompt and padding positions are set to -100 before the loss call.

Why the prompt mask matters: failure mode without it. We trained a control checkpoint without the prompt mask (other settings identical) for 4,000 steps. By step 1,500, T3's outputs for held-out cross-speaker references showed verbatim parroting of the prompt's first 1.5–2 seconds before transitioning to the target text; held-out SECS dropped from 0.63 to 0.51, and SMOS dropped from 4.51 to 3.92. The prompt-mask checkpoint did not exhibit this behavior at any step.

Effective batch size and learning rate schedule. We use a global batch size of 32 utterances (per-GPU 8, gradient-accumulated 4 steps) with the AdamW optimizer at lr = 5e−5, weight decay 0.01, gradient clipping at 1.0. The learning rate follows a cosine schedule with 500-step warmup; the schedule is restarted after 30 epochs if dev loss

has plateaued. The voice encoder and S3Gen weights are excluded from the optimizer parameter set entirely.

B Evaluation Protocol Walkthrough

B.1 Rater pool: detailed screening and assignment

Recruitment pipeline. We posted recruitment notices on three university WhatsApp groups (Dhaka University, BUET, NSU) and one crowd-work platform (paid Bangla annotation marketplace). 71 applications were received; the screening pipeline below reduced this to 24.

Eligibility screening (3 stages).

1. **Self-declaration form:** L1 Bangla, current residency in Bangladesh, age 22–45, no uncorrected hearing loss. 19/71 rejected.
2. **5-utterance Bangla listening pre-screen:** short clear Bangla recordings; transcribe verbatim. Threshold: $\geq$80% character accuracy on the transcription. 18/52 rejected (mostly due to non-trivial typos and incomplete transcripts).
3. **Headphone-quality and listening-environment check:** the candidate self-reports background noise level on a 5-point scale; we ran a 5-second silence check at session start to verify ≤ 2 on that scale, and a known-noisy stimulus (-15 dB SNR) is correctly identified as noisy. 10/34 rejected.

The remaining 24 candidates (14 M, 10 F, ages 22–45) became the rater pool. A backup pool of 6 reserve raters was held; 2 reserves were activated during the study to replace evaluators who exceeded the catch-trial inconsistency threshold (see below).

Stimulus assignment. Each rater scored all 76 utterances of the subjective evaluation set, across all 7 systems (BanglaBox, Google, Azure, Narakeet, IndicF5, YourTTS-FT, Human reference). Stimuli were presented in randomized order; the rater could not see system identity. Per session $76 \times 7 = 532$ trials; with 24 raters this yields $76 \times 24 = 1{,}824$ ratings per system per dimension (Nat./SMOS/Clar.), and $76 \times 7 \times 24 = 12{,}768$ total ratings.

B.2 Listening environment validation

Following ITU-T P.808 best practice (International Telecommunication Union, 2021), we standardized:

- **Hardware:** closed-back over-ear headphones (only); in-ear, open-back, and speaker playback rejected.
- **Loudness:** all stimuli normalized to -23 LUFS (EBU R128 broadcast standard) and clipped at -1 dB peak.
- **Acoustic environment:** self-rated room noise $\leq 2/5$, verified by a 5-second silence check at the start of each session (peak amplitude must be below -50 dBFS).
- **Session length:** no session exceeds 75 minutes; mandatory 5-minute breaks every 25 minutes; total study workload bounded at 4 sessions per rater.

B.3 Catch trials and quality control

Each session contains 2 catch trials inserted randomly among the regular trials:

1. **Identical-pair self-similarity trial:** the rater is shown the same clip as both A and B in a pair-rating, and is expected to rate similarity at the top of the SMOS scale ($\geq$ 4.5).
2. **Obvious-degradation reject trial:** the rater is shown a clip with severe artificial distortion (clipping, time-reversal of a 100-ms segment, or harsh pitch shift) and is expected to rate Naturalness at the bottom of the scale (≤ 1.5).

Raters who fail more than 20% of their catch trials over the course of their assigned sessions are excluded; their data is discarded and a backup rater is recruited to re-rate the same stimuli. 2 of the original 24 raters were excluded under this rule.

B.4 Statistical testing details

Wilcoxon signed-rank test. For each metric in {Nat., SMOS, Clar.} we compute, per stimulus, the difference $\Delta_i = r_i^{\text{BanglaBox}} - r_i^{\text{baseline}}$ in mean rater score, and test the null hypothesis H_0 : median(Δ)=0 against the two-sided alternative. The sign-rank statistic is computed via the standard formulation. For SECS, where we have per-utterance cosine similarity values (not human ratings), the

same paired signed-rank test is applied to per-utterance SECS differences. The reported p-values are two-sided.

Bonferroni correction. With BanglaBox compared against 5 baselines for each of Nat., SMOS, Clar., the Bonferroni adjustment for these comparisons is at $\alpha/5$=0.01 for an $\alpha = 0.05$ family-wise error rate; we report markers at $p < 0.05$, $p < 0.01$, $p < 0.001$ *after* the correction (so the post-Bonferroni threshold 0.05/5=0.01 shown as **, etc.). For SECS, the family is 2 comparisons (BanglaBox vs. IndicF5 and YourTTS-FT only), with corrected threshold $\alpha/2 = 0.025$.

B.5 NISQA measurement details

Model and version. We use the publicly available NISQA v2.0 PyTorch checkpoint provided by Mittag et al. (2021) (`nisqa_v2.0_tts.tar`, trained on the TTS-focused subset of NISQA training data). Inputs are 16 kHz mono WAV files; outputs are the 5-dimensional MOS-style predictions; we report only the overall MOS dimension.

Test set and aggregation. For each system we computed NISQA on all 436 test utterances (the same 436-utt domain-stratified subset used for CER/SECS). Per-system NISQA in Table 5 is the arithmetic mean over these 436 utterances. Standard deviations were $\sigma \in [0.18, 0.25]$ across systems, narrower than the subjective scales because NISQA produces deterministic outputs per audio.

Caveat on Bangla applicability. NISQA's training corpora are primarily English and German; absolute predictions on Bangla speech are not perfectly calibrated. The relative ranking across systems, however, is consistent with the 24-rater subjective Nat. ranking (BanglaBox highest, YourTTS lowest), and we therefore use NISQA as a complementary objective metric rather than a primary headline number.

C Corpus Construction Notes

C.1 Tier A studio recording protocol

Voice talent selection. 13 professional Bangla voice actors were recruited from a Dhaka talent agency. Selection criteria: (i) native Bangladeshi Bangla speaker (Dhaka register), (ii) ≥ 3 years of studio voice-over experience, (iii) no audible vocal pathology (verified by a screening session with an audio engineer), (iv) gender balance target across the speaker pool. 7 female, 6 male voices were retained.

Studio setup. All recordings took place in a single acoustically treated studio in Dhaka:

- **Microphone:** Neumann TLM-103 large-diaphragm condenser.
- **Pre-amp / interface:** Universal Audio Apollo Twin X DUO, 24-bit / 48 kHz.
- **Booth:** acoustically treated to RT60 $<$ 0.2 s; ambient noise floor verified below -65 dBFS at session start.
- **SNR floor:** ≥ 44 dB enforced per recording; segments below floor were re-recorded the same session.

Working format. The 48 kHz / 24-bit studio captures form the internal archive. For the corpus used in this work, audio is downsampled to 24 kHz (matching the S3Gen vocoder output rate) and saved as mono FLAC.

C.2 Tier A quality assurance (extended)

Per-session checks. At the end of each recording session, the studio engineer runs an automatic batch over that day's takes: SNR check, clipping detection (any sample at $|\text{amplitude}| > 0.99$), breath/lip-smack detection via a transient-detection script (peaks within 5 ms of the noise floor), and forced alignment via a Bangla acoustic model fine-tuned on a 50-hour Tier A pilot subset. Any take that fails any check is flagged; flagged takes are re-recorded the same session before the talent leaves.

Per-week linguist audit. A native Bangla-speaking linguist audits a 5% random sample of each week's takes per speaker, scoring pronunciation correctness, prosodic naturalness, and adherence to the target script. Segment-level acceptance rate $\geq 99\%$; flagged segments are reviewed by a second linguist, and disagreements resolved by re-recording.

Consent and anonymization. All Tier A speakers signed informed-consent forms covering: (i) research use of their recordings, (ii) no commercial redistribution, (iii) public credit at

their option (most opted to remain pseudonymous as "Speaker A1"–"A13"). Biographic metadata (age, gender, dialect) is reported in aggregate only.

C.3 9 automated quality gates — full thresholds

Table 10 lists the 9 quality gates applied to LLM-generated Bangla candidate scripts. A script must pass all 9 to be queued for human review.

Gate	Threshold / criterion
1. length	6 – 28 words (token count)
2. TTR	type-token ratio ≥ 0.55 on $\geq$ 12-word strings
3. dedup	cosine sim ≤ 0.85 against existing corpus
4. lang-ID	Bengali probability ≥ 0.93 (fastText 176-lang ID)
5. perplexity	Bangla LM perplexity $\in [20, 240]$
6. phonetic novelty	gain $\geq +5\%$ on the current under-covered juktakkhor set
7. script validity	all glyphs in U+0980–U+09FF, U+0020–U+007F, or curated ASCII
8. semiotic	no profanity, slurs, or sensitive PII from a 1,412-term blocklist
9. gap-unit	contains at least 1 unit from this week's coverage-gap list

Table 10: The 9 automated quality gates and their thresholds. Gate 6 (phonetic novelty) and gate 9 (gap-unit) operationalise the coverage-feedback loop.

C.4 Coverage-feedback loop — week-by-week trajectory

The coverage-feedback loop ran for 14 weekly iterations. At the end of each iteration the script generator was re-prompted with the under-covered juktakkhor and phonetic-context list. Approximate juktakkhor coverage by iteration: week 1: $\sim 70\%$; week 3: 74%; week 5: 77%; week 8: 80%; week 11: 83%; week 14: $\geq 85\%$ (top-150 juktakkhor: $\geq 97\%$). The corresponding tiered-JSD loss decreases monotonically across iterations (Fig. 2).

C.5 Tier B filter — ablation-driven criterion

Initial pool. We started from 39 Bangla speakers across 7 public Hugging Face Bangla TTS datasets (247 h total). Sources retained only when their original licences permit research use.

SNR filter (first pass). All speakers below 35 dB average SNR (computed per-utterance, averaged per speaker) were dropped. 9 of 39 speakers fail this pass.

Ablation-driven curation (second pass). Each remaining speaker was added to a held-out training run; the resulting model was evaluated on the dev split (CER, SECS). Speakers whose inclusion did not improve, or actively degraded, dev metrics were dropped: 7 additional speakers fail this pass.

Final retention. 23 of 39 speakers retained (153.0 h); 16 dropped (94 h discarded). See §2 for the corpus-level numbers.

D Bangla Text Examples

This section provides representative Bangla strings illustrating each major component of the pipeline, plus the failure modes referenced in §5.

D.1 Honorific abbreviation expansion

Table 11 lists 12 representative entries from the 47-entry hand-curated honorific dictionary used at Stage 2 (§3.1). These appear frequently in Bangla religious/cultural literature but are absent from open normalization toolchains.

Abbrev.	Full expansion (Bangla)
সাঃ	সাল্লাল্লাহু আলাইহি ওয়া সাল্লাম
রাঃ	রাদিআল্লাহু আনহু
রহঃ	রাহিমাহুল্লাহ
হাঃ	হাফিজাহুল্লাহ
মাঃ	মাদ্দাজিল্লুহু
দাঃ বাঃ	দামাত বারাকাতুহুম
আঃ	আলাইহিস সালাম
জিঃ	জিল্লুল্লাহ
কুঃ সিঃ	কুদ্দিসা সিরুহু
আলঃ	আলাইহিমুস সালাম
রাযিঃ	রাযিআল্লাহু আনহু
মুহঃ	মুহতারাম

Table 11: Representative entries from the 47-term Bangla honorific dictionary (Stage 2, step 5).

D.2 Juktakkhor (conjunct) inventory examples

The ∼300-form juktakkhor inventory extended via the coverage-feedback loop spans frequent two-consonant conjuncts to rare three-consonant ones. Representative samples:

Frequent (in >5% of training utterances): ক্ত (*kt*, "যুক্ত"), ন্ত (*nt*, "অন্ত"), স্ত (*st*,

"প্রস্তুত"), ক্ষ (*kṣ*, "শিক্ষা"), জ্ঞ (*jñ*, "বিজ্ঞান"), ন্ধ (*ndh*, "বন্ধ").

Moderate (in 1--5%): ত্ত্ব (*ttv*, "তত্ত্ব"), দ্ভ (*dbh*, "উদ্ভব"), হ্ম (*hm*, "ব্রহ্ম"), ঙ্গ (*ṅg*, "বঙ্গ"), ঞ্চ (*ñc*, "পঞ্চ").

Rare three-consonant (<0.3%): স্ত্র (*str*, "অস্ত্র"), ন্ত্র (*ntr*, "মন্ত্র"), ক্ষ্ম (*kṣma*, "লক্ষ্মী"), ষ্ট্র (*ṣṭr*, "রাষ্ট্র"), ন্দ্র (*ndr*, "চন্দ্র").

D.3 Cross-domain Bangla utterance samples

Representative shorter exemplars from three target domains:

- **News:** প্রধানমন্ত্রী নতুন শিক্ষানীতি ঘোষণা করেছেন।
- **Healthcare:** প্রতিদিন আট গ্লাস পানি পান করুন।
- **IT:** সফটওয়্যার ইন্সটল করার আগে আপডেট নিন।

D.4 Failure modes with Bangla examples

Rare three-consonant conjuncts. Conjuncts such as স্ত্র (in "অস্ত্রাগার" weapons depot), ন্ত্র (in "যন্ত্রপাতি" machinery), and ক্ষ্ম (in "সূক্ষ্মতম" the finest) appear in <0.3% of training utterances. On test items containing these, BanglaBox occasionally produces a distorted phoneme sequence (audible as a brief stutter or vowel insertion). The 9th quality gate (gap-unit) up-weights these during script generation, but coverage remains uneven.

Bangla–English code-switching. Utterances such as আমি *email* পাঠিয়েছি ("I have sent email") or ফাইলটি *upload* করো ("upload the file") produce artifacts at the language boundary — typically a brief silence (5–15 ms) and a slight pitch reset. BanglaBox does not have explicit language-tag injection at code-switch boundaries; this is a documented opportunity for follow-up.

Highly colloquial onomatopoeia. Regional sound-words with < 5 training occurrences (e.g., ঝমঝম *jhomjhom*, heavy-rain sound; টুপটাপ *tuptap*, light-drip sound; গুড়গুড় *guṛguṛ*, distant thunder) yield unreliable output — sometimes a plausible phonetic rendering, sometimes a hallucinated near-neighbour word, consistent with low-resource long-tail behaviour in AR TTS.

E Computing Infrastructure

Training. All fine-tuning experiments use 1 NVIDIA H100 GPU (80 GB) on a Linux host (Ubuntu 22.04, CUDA 12.4, PyTorch 2.4). A single BanglaBox training run takes $\sim$ 28 hours wall-clock for 50 epochs (effective batch size 32, gradient checkpointing enabled). Total training compute for the headline run, ablations, and data-scaling subsets is approximately 220 H100-hours.

Inference. Inference for the evaluation set (12,768 utterances across 7 systems, 76 subjective + 436 objective) used the same H100. Total inference compute: $\sim$ 4.5 H100-hours.

Total compute budget. $\sim$225 H100-hours plus $\sim$ 20 CPU-hours for data preprocessing, normalization, and evaluation scripting.

Checklist

1. **A. For Every Submission**
   (a) Did you describe the limitations of your work?
   Yes. A dedicated **Limitations** section (immediately following Section 6) covers: modest speaker count (36 speakers — 13 studio Tier A + 23 curated open-source Tier B — below VCTK/LibriTTS scale), tier heterogeneity between Tier A and Tier B, no BnTTS head-to-head comparison, English-like prosody bias in long utterances, text-normalization gaps (medical/legal/Banglish), no streaming, Bangladesh-only evaluators, and use of language-agnostic ECAPA-TDNN for speaker similarity.
   (b) Did you discuss any potential risks of your work?
   Yes. The Ethics Statement (Section 6) addresses: voice cloning impersonation risk, responsible deployment requirements (user consent, synthetic media regulations), and the broader positive impact on accessibility and education.
2. **B. Did you use or create scientific artifacts?**
   (a) Did you cite the creators of artifacts you used?

Yes. All artifacts are cited: Chatterbox TTS (Resemble AI, 2025), XTTS tokenizer (Casanova et al., 2024), AdamW (Loshchilov and Hutter, 2019), Silero VAD (Silero Team, 2024), bnnumerizer (Chowdhury and Adhikary, 2022) and bnunicodenormalizer (Ansary, 2024), ECAPA-TDNN (Desplanques et al., 2020), Whisper large-v3 (OpenAI), and all baseline systems.

(b) Did you discuss the licenses or terms for artifacts you used?

Yes. Chatterbox TTS is MIT-licensed (stated in Section 1); we use it within its license terms. Tier B utterances are drawn from public sources only when their licences permit downstream research use. We do not release the corpus, model weights, or pipeline as artifacts of this submission.

(c) Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified?

Yes. All tools are used within their intended research/open-source scope.

(d) Did you discuss the steps you took to check whether the data you are using, for example, prompts or example data, might be confidential or might contain information that names or uniquely identifies people or could be used to identify them?

Yes. Tier A recordings use professional voice talent who provided written informed consent; no personally identifiable information was collected. Tier B audio is sourced from public Hugging Face datasets whose contributors had consented under each source's original terms.

(e) Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena?

Yes. Section 2, Table 1, Table 2, and Table 3 document language, seven-domain distribution (news, customer care, teaching, healthcare, e-commerce, finance, IT), two-tier speaker provenance (Tier A studio + Tier B curated open-source), phonetic-balance design including the juktakkhor-aware tiered JSD objective, and corpus statistics.

(f) Did you report relevant statistics like the number of examples, description of train/dev/test splits, and information about who annotated the data (e.g., crowdsourcing vs. expert annotation)?

Yes. Table 1 reports per-tier utterance counts, duration, and a speaker-disjoint split: 31 speakers for train (11 Tier A + 20 Tier B, ∼422 h), 5% within-speaker dev (∼22 h), and 5 held-out speakers (2 Tier A + 3 Tier B, ∼76 h) for zero-shot test. Section 4 describes the 24 native Bangla-speaking human evaluators.

3. **C. Did you run computational experiments?**

(a) Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

Yes. Section 3 reports T3 (520M parameters), voice encoder (256-dim embeddings), and S3Gen. Table 4 reports hardware (NVIDIA H100 80 GB), training precision (BF16), batch size, optimizer, and total compute: ∼100 H100-hours total across 5 training runs (18–22 h each, 30–50 epochs per run) on a single H100 80 GB.

(b) Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Yes. Table 4 lists all training and inference hyperparameters. Section 3.1 describes the fine-tuning strategy. Section 3.2 explains Bangla-calibrated inference hyperparameter choices.

(c) Did you report descriptive statistics about your results (e.g., error bars), and is it the full experimental run

spanning possible permutations in randomness?

Yes. All subjective metrics report 95% confidence intervals (Tables 5, 7, 8). Pairwise differences against every baseline are tested via Wilcoxon signed-rank with Bonferroni correction (5 comparisons for Nat./SMOS/Clar.; 2 for SECS; $p < 0.01$ in the main results; $p < 0.001$ for the largest ablation drop). Objective metrics (CER, SECS) are computed over the full 436-utterance test subset. Subjective metrics are rated by 24 native Bangla speakers with two catch trials per session; inconsistent evaluators are excluded.

(d) If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the package name, version, and license?

Yes. Key packages cited: Whisper large-v3 for CER (OpenAI), ECAPA-TDNN for SECS (Desplanques et al., 2020), Silero VAD (Silero Team, 2024), bnnumerizer, bnunicodenormalizer, PyTorch/BF16 training stack.

4. **D. Did you use human annotators or research with human participants?**

(a) Did you report the full text of instructions given to participants, and if applicable, also the instructions to crowdworkers and quality control measures?

Yes. The protocol uses three independent 5-point scales (Naturalness, Clarity, SMOS); 76 utterances rated by all 24 evaluators across all systems, with 2 catch trials per session and quality-control exclusion criteria. See Section 4 for the full BanglaEval protocol.

(b) Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss whether the payment is adequate given the people involved?

Yes. Ethics Statement: 24 human evaluators were compensated at \$15/hour and provided informed consent. Corpus speakers were compensated above local market rates in BDT.

(c) Did you discuss whether and how consent was obtained from people whose data you're using/curating?

Yes. Ethics Statement: corpus speakers and evaluators provided written informed consent for their participation.

(d) Was the data collection protocol approved (or determined exempt) by an ethics review board?

Yes. Ethics Statement: the evaluation protocol was reviewed and approved by our institutional ethics board.

(e) Did you report the basic demographic and geographic characteristics of the annotator population?

Yes. Section 4 reports gender (14 M/10 F) and age range (22–45) for the 24 evaluators; all are native Bangla speakers from Bangladesh. The Limitations section acknowledges this geographic scope; West Bengal generalization is untested.

5. **E. Did you use AI assistants (e.g., ChatGPT) in your research or writing?**

Yes — disclosed below in accordance with ACL 2026 / ARR authorship policy.

(i) Research use (data generation). Generative LLMs were used as components of the coverage-feedback Bangla script generation pipeline for Tier A. Specifically: GPT-4o-mini for bulk sentence generation (∼90% of LLM calls), GPT-4o for rare-conjunct gap-targeting (∼7%), and Claude Opus for long-context customer-care dialogue composition (∼3%). Every LLM-generated sentence passes nine automated quality gates (length, TTR, near-duplicate, language ID, perplexity, phonetic novelty, script validity, semiotic compliance, gap-unit presence) and a

two-pass human review (linguist and ethics/content reviewer) before entering the corpus. The LLMs generate candidate scripts; humans validate and select the final corpus. Prompt hashes, LLM model identifiers, and per-sentence provenance are tracked internally for audit purposes.

(ii) Writing use. AI writing assistants were used for language polish (grammar and phrasing clean-up) on paper sections; no AI-generated content introduces new scientific claims. All technical content, experimental design, results, and conclusions are authored by the human authors, who accept full responsibility for correctness and originality.

(iii) No author LLM. No LLM is listed as or treated as an author; all authorship credit is human. The human authors are individually accountable for every claim, table, figure, and citation in this paper, and have personally verified the experimental results and statistical analyses. LLM-assisted text was inspected, revised, and approved by at least one human author before inclusion. The corresponding author accepts final responsibility for the integrity, originality, and reproducibility of the work.

(iv) No prompt-injection. The submission contains no text intended to influence LLM-based reviewer tools.

F Abbreviations and Notation

Table 12 lists the abbreviations and symbols used throughout the paper. Throughout, BanglaBox refers to the trained system; "BanglaBox-520" refers to the corpus.

Term	Definition
ARR	ACL Rolling Review submission cycle.
AR	Autoregressive (transformer decoder).
BPE	Byte-Pair Encoding (subword tokenization).
CER	Character Error Rate (ASR transcription error).
CI	95% Confidence Interval.
ECAPA-TDNN	Emphasized Channel Attention, Propagation and Aggregation TDNN.
EOS	End-of-sequence token.
HF	Hugging Face.
HiFi-GAN	High-Fidelity GAN vocoder.
ITU-T P.808	ITU recommendation for subjective speech-quality testing in remote conditions.
JSD	Jensen–Shannon divergence.
juktakkhor	Bangla conjunct consonant cluster.
LLM	Large Language Model.
LoRA	Low-Rank Adaptation (Hu et al., 2022).
LUFS	Loudness Units relative to Full Scale.
MOS	Mean Opinion Score.
NISQA	Non-Intrusive Speech Quality Assessment (Mittag et al., 2021).
PII	Personally Identifiable Information.
RTF	Real-Time Factor (synth time / audio duration).
S1 / S2 / S3	Three stages of the fine-tuning recipe.
SECS	Speaker Encoder Cosine Similarity.
SMOS	Speaker MOS (similarity-rated).
SNR	Signal-to-Noise Ratio.
T3	Chatterbox's text-to-speech-token transformer (520M).
S3Gen	Chatterbox's speech-token-to-waveform vocoder.
TTR	Type-Token Ratio.
TTS	Text-to-Speech.
।(*dari*)	Bengali full-stop punctuation (U+0964).
VAD	Voice Activity Detection.
L_{prompt}	Prompt-mask speech-token positions ($\approx$ 75 at 3 s).
$\mathbf{e}_{\text{spk}}$	256-d Voice-Encoder speaker embedding.
$\mathbf{c}_{\text{text}}$	Normalized text-token context.
$\lvert V\rvert, \lvert V_s\rvert$	Text vocab ($\lvert V\rvert$=4,240); speech vocab ($\lvert V_s\rvert$=2,049).

Table 12: Abbreviations, symbols, and Bangla terms used in the paper.