

Lipi-Ghor: A Large-Scale, Diarized, Multi-Domain Bengali Speech Corpus and a Scalable Pipeline for Building It

Sanjid Hasan¹ A H M Fuad² Risalat Labib² Bayazid Hasan²

¹Department of Electrical and Electronic Engineering, KUET, Khulna, Bangladesh

²Department of Computer Science and Engineering, BUET, Dhaka, Bangladesh

{sanjid9657, ahmfuad9, risalatlabib, bayazid741854}@gmail.com

Abstract

Bengali, spoken by over 230 million people, remains severely under-resourced in spoken language research. Existing open Bengali ASR corpora are confined to short, read or crowd-sourced speech and lack speaker diarization and long-form audio. We present **Lipi-Ghor**, a large-scale weakly-supervised Bengali speech corpus comprising 885 hours of audio sourced from 998 YouTube videos across 596 unique channels, spanning 17 diverse domains including audiobooks, drama, religious sermons, news, and regional dialects. Each recording is annotated in SSTT format (Speaker, Speech, Transcription, Timestamp), which combines automated diarization with Bengali YouTube caption tracks. To our knowledge, Lipi-Ghor is the first Bengali speech corpus to pair large-scale coverage with speaker-diarized, long-form audio annotations, both of which are absent from existing open Bengali resources. We also present the construction methodology as a *replicable pipeline* for any language with a sufficient captioned YouTube presence. Targeted channel curation across a handful of rich domains yields corpus scales that were previously out of reach for low-resource languages. We benchmark four widely-used ASR models under zero-shot and fine-tuned conditions. All four degrade consistently on emotive, dialectal, and long-form content. Because caption targets are noisy, we also release a human-annotated gold evaluation split, verified by 17 volunteer annotators over 30 hours of human-in-the-loop correction and stratified across all 17 domains. We release our original contributions (video-ID lists, diarization output, pipeline code, and the gold annotations) under CC-BY-4.0, and distribute caption-derived transcripts for research use only; no raw audio is redistributed.¹

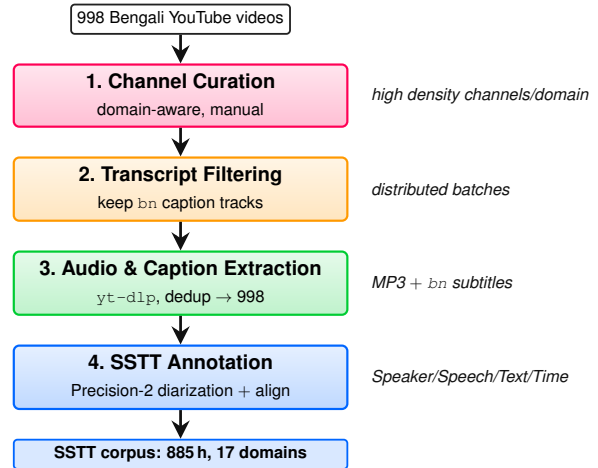


Figure 1: The Lipi-Ghor construction pipeline. Four stages turn a curated set of Bengali YouTube channels into a speaker-diarized, SSTT-annotated corpus. The methodology is language-agnostic; only Stage 1 curation is language-specific.

1 Introduction

Automatic Speech Recognition (ASR) for low-resource languages is bottlenecked by data. English ASR now approaches near-human performance, but languages such as Bengali, the seventh most spoken in the world with over 230 million native speakers, remain severely underserved. Existing open Bengali ASR resources are narrow in domain and confined to short read or crowd-sourced utterances. They lack diarization, speaker annotations, and long-form audio, which limits their use for training modern speech-language models.

Real Bengali speech makes the problem harder. Unlike English or Mandarin, Bengali carries substantial dialectal variation, code-switching with English, register shifts across formal news broadcast, informal Natok drama, and Islamic religious sermons, and high speaker density in long-form audio. Corpora built from read speech or studio recordings miss this diversity, and the models trained on them degrade sharply in the wild.

¹

DATASET

CODE

We address this gap with **Lipi-Ghor** (“House of Scripts”), a large-scale Bengali speech corpus built from YouTube. To our knowledge, it is the first Bengali corpus to combine large-scale coverage with speaker-diarized, long-form audio in a single resource. We make three contributions. First, we describe a scalable, replicable data construction pipeline that uses domain-aware channel curation, caption-track filtering, distributed transcript fetching, and automated diarization to produce SSTT-annotated corpora at scale. Second, we instantiate this pipeline for Bengali, yielding 885 hours of speech across 17 domains from 596 unique channels. A small number of well-chosen channels per domain yields hundreds of hours of training-ready data, which bears directly on corpus construction for other under-resourced languages. Third, we release a human-annotated gold evaluation split and benchmark four widely-used ASR systems in zero-shot and fine-tuned settings, with a per-domain and duration-based analysis of where they degrade.

Figure 1 provides an overview of the Lipi-Ghor construction process.

2 Related Work

Bengali ASR Datasets. Prior open Bengali ASR resources are limited in scale, register, and annotation richness, and none pair large-scale training audio with speaker diarization and long-form recordings. Early corpora such as SHRUTI (Das et al., 2011) contain only tens of hours of read speech from a handful of speakers. Common Voice Bengali (Alam et al., 2022; Ardila et al., 2020) provides several hundred hours of crowd-sourced read speech but covers only a narrow, sentence-level acoustic register. The BengaliAI Speech Corpus (Howard et al., 2023) offers broader, larger-scale coverage but was collected under controlled crowd-sourced conditions without speaker-attributed long-form audio. Large Indic multilingual efforts include the Bengali portions of IndicSUPERB/Kathbath (Javed et al., 2023), IndicWay2Vec (Javed et al., 2022), and the Vistaar benchmark suite (Bhogale et al., 2023), but these remain sentence-level and lack conversational diarization. OOD-Speech (Rakib et al., 2023) introduced the first out-of-distribution benchmark for Bengali, with 23 hours of manually annotated test data across 17 sources, but it is a test set, not a diarized training corpus. BengaliLoop (Tabib et al., 2026) released a long-form diarization benchmark

Corpus	Hrs	Dom.	Diar.	Long
Common Voice bn	~40	1	✗	✗
BengaliAI Speech	~1200*	few	✗	✗
OOD-Speech (test)	23	17	✗	✗
BengaliLoop (eval)	small	few	✓	✓
Lipi-Ghor	885	17	✓	✓

Table 1: Open Bengali speech resources by scale, domain count, speaker diarization, and long-form audio.

*Predominantly short, read/controlled speech.

with human-in-the-loop transcripts, but at evaluation scale. Table 1 situates Lipi-Ghor against prior Bengali resources.

Web- and YouTube-Based Corpus Construction.

The People’s Speech (Galvez et al., 2021) and GigaSpeech (Chen et al., 2021) assembled tens of thousands of hours of English from web sources, while VoxPopuli (Wang et al., 2021) scaled a similar approach to 23 European languages. Most directly related, YODAS (Li et al., 2023) harvested over 500k hours from YouTube subtitles across more than 100 languages, only 27 hours for bangla language. These efforts are either English-centric or provide only segment-level transcripts without speaker diarization, and the Bengali material they contain is sparse. Two problems recur in such pipelines: filtering for target-language caption quality, and managing API rate constraints at scale. We address both in our pipeline design (§3).

Speaker Diarization Corpora. Diarization benchmarks such as AMI (Carletta et al., 2005), DIHARD III (Ryant et al., 2021), and VoxConverse (Chung et al., 2020) establish the difficulty of speaker attribution on naturalistic audio, but none cover Bengali. Lipi-Ghor brings diarized annotation to Bengali at training scale.

Weakly Supervised ASR. Weakly supervised learning from noisy web transcripts is well established (Radford et al., 2022), as are massively multilingual models such as MMS (Pratap et al., 2024) and SeamlessM4T (Seamless Communication et al., 2023). Auto-generated captions exhibit systematic error patterns distinct from human transcription (Kafle and Huenerfauth, 2017), a limitation we address through our gold evaluation split.

3 The Lipi-Ghor Pipeline

Our construction pipeline consists of four stages: domain-aware channel curation, transcript avail-

ability filtering, audio and caption extraction, and SSTT annotation.

3.1 Stage 1: Domain-Aware Channel Curation

Rather than crawling YouTube indiscriminately, we perform targeted manual curation of Bengali YouTube channels across predefined domain categories.

The initial collection process focused on identifying high-volume channels within each domain to efficiently bootstrap data acquisition. These anchor channels provided substantial amounts of domain-relevant speech, after which additional channels were incorporated to improve speaker, topic, and recording diversity. The final corpus therefore combines contributions from both high-volume sources and a broader long-tail distribution of channels. The bottleneck is not data collection but *domain-aware channel discovery*, that is, finding which channels have enough Bengali-language caption coverage.

3.2 Stage 2: Transcript Availability Filtering

Only videos with an existing Bengali caption track (language code bn) are retained. This filtering uses our `ytranscript` toolkit,² which queries caption availability without downloading audio. To manage YouTube API rate constraints at scale, video IDs are split into batches and processed in parallel across multiple compute environments (Kaggle notebooks, Google Colab sessions, GitHub Codespaces), with results merged after each batch.

3.3 Stage 3: Audio and Caption Extraction

Audio (MP3) and Bengali subtitle tracks are extracted using `yt-dlp`. Only Bengali-track captions are retained; auto-translated tracks are discarded. The extraction stage produced 1,019 audio-caption recordings. Cross-domain deduplication removed duplicate content shared across channels, yielding 998 unique videos that constitute the released corpus.

3.4 Stage 4: SSTT Annotation

Every audio file is diarized with Precision-2, a speaker-diarization model from pyannoteAI ([pyannoteAI, 2025](#)), built on the `pyannote.audio` framework ([Bredin et al., 2020](#)). Speaker count is inferred per recording; labels are local to each video and not linked across videos. Precision-2

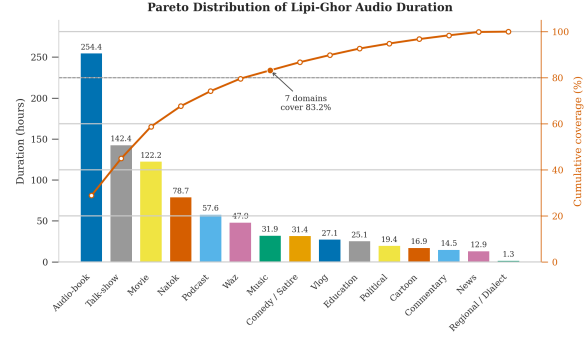


Figure 2: Pareto view of audio duration by domain. Seven of the seventeen domains account for roughly 83% of the corpus, reflecting where Bengali caption coverage is densest on YouTube.

output is aligned with Bengali caption transcripts to produce SSTT (Speaker, Speech, Transcription, Timestamp) annotations at the segment level. All 1,019 files passed integrity checks: no invalid intervals, no duplicate segments, no annotation extending past recording end. Because Precision-2 is a commercial, API-gated model, we release its output so that the corpus can be used without re-running diarization. Final data is packaged into three splits: (i) raw diarization output, (ii) diarization aligned with transcription (885-hour SSTT corpus), and (iii) segment-chunked ASR-ready data (364 videos pre-chunked for direct training use).

4 Dataset Description

4.1 Scale and Coverage

Table 2 summarizes the released corpus after cross-domain deduplication. Per-domain breakdowns appear in Appendix A. The corpus is dominated by audiobooks (254h), talkshows (142h), and movies (122h), with smaller but linguistically important contributions from regional dialects, religious sermons, and children’s cartoons.

The distribution is heavily skewed, as Figure 2 makes clear: three domains account for 58.6% of audio, and seven of seventeen reach 83.0%. This shape is a direct consequence of the curation strategy: audiobook and talkshow channels upload long, caption-rich episodes, whereas dialect and health content on YouTube is short-form and rarely captioned. The skew means domain balance is a function of what is available with Bengali captions, not of what would be ideal for training.

² [Sanjiddh090/ytranscript](#)

Property	Value
Total Videos	998
Unique Channels	596
Total Duration	885.34 hours
Avg. Duration/Video	~52 minutes
Domains	17
Total Word Tokens	5,832,680
Unique Vocabulary	181,648
Avg. WPM	114.81
License	CC-BY-4.0*

Table 2: Lipi-Ghor corpus summary. *CC-BY-4.0 applies to our original contributions; caption-derived transcripts are for research use only (§9).

4.2 Linguistic Analysis

Speaking Rate. Average WPM varies widely across domains, from 170.17 in Service/Promo down to 59.87 in Regional/Dialect, spanning a broad range of speech styles relevant to ASR robustness.

Vocabulary and OOV. The corpus contains 181,648 unique word types. OOV rates vary sharply by domain: audiobooks reach 50.35% (literary vocabulary), while news and political content fall to 15–16% (constrained formal register).

Code-Switching. 262 utterances (0.026% of segments) contain Latin-script characters from natural Bengali–English code-switching. These are flagged in metadata but retained, as code-switching is a realistic feature of Bengali speech.

Transcription Quality. Transcriptions come from YouTube caption tracks, a mix of manually contributed and auto-generated captions, and are therefore weakly supervised. Lipi-Ghor is a *pre-training corpus* rather than an evaluation benchmark.

4.3 Conversational Structure

Because every recording is diarized, we can characterize the corpus’s conversational structure beyond raw duration. Table 3 reports aggregate statistics over all 1,019 recordings. Of the 885.34 hours of audio in the released corpus, 639 hours (72.2%) are covered by at least one detected speaker segment, while the remaining 246.34 hours contain silence, music, ambient sound, or other non-speech regions.

Speaker density is bimodal. The median recording has 4 speakers, but 25.9% are single-speaker narration such as audiobooks and sermons, while 31% carry ten or more speakers, mostly talkshows

Property	Value
Diarized recordings	1,019
Speaker-attributed segments	984,647
Speaker labels (all recordings)	9,056
Speech / non-speech	639 h / 243 h
Speech ratio (mean)	0.76
Speakers per rec. (median, IQR)	4 (1–12)
Single-speaker recordings	25.9%
Recordings with ≥ 10 speakers	31.0%
Speaker turns per minute (mean)	14.2
Median segment duration	1.54 s
Segments under 2 s	62.6%
Detected overlap (median)	0.31%
Total overlapped speech	11.6 h

Table 3: Conversational structure of the diarized corpus. Speech ratio is the fraction of each recording covered by at least one speaker turn, averaged over recordings.

and debates. Segments are also short, with a median of 1.54 s and 62.6% falling under two seconds, because diarization decomposes long recordings into turn-level chunks before ASR ever sees the audio. This bears on the duration analysis in §6, where segmentation has already removed most of the long-form difficulty. Per-domain conversational statistics and the full speaker-density and segment-duration distributions appear in Appendix A.

Table 4 gives the per-domain conversational statistics summarized in §4.3, obtained by joining the diarization output to the channel metadata on video ID (all 1,019 recordings match). The domains fall into three groups. Narration domains such as audiobook, education, and health are mostly single-speaker, turn slowly, and produce the longest segments. Conversational domains such as talkshow, podcast, vlog, and comedy sit at 20–25 turns per minute with few solo recordings and high speech ratios. Performance domains such as movie, natok, and cartoon show the highest speaker counts and the shortest segments, with speech ratios near 0.56 because music and sound effects occupy much of the audio. Waz is a useful check on the diarizer: sermons are near-monologue, and the model agrees, giving a median of two speakers and a dominant-speaker share of 0.94.

Figure 3 shows the segment-duration distribution underlying the median of 1.54 s reported in Table 3. The curve is steep throughout the sub-second range and flattens well before ten seconds, indicating that the diarization process converts long-form recordings into predominantly short speaker segments before ASR processing, even though the original

Domain	Spk	Solo	Turn	SR	Seg
Audio-book	1	62	3.4	0.74	1.83
Talk-show	5	2	20.1	0.86	2.38
Movie	24	14	20.0	0.58	0.95
Natok	18	2	21.6	0.56	1.14
Podcast	3	10	21.9	0.85	1.73
Waz	2	32	5.8	0.84	2.21
Music	3	14	17.4	0.71	2.12
Comedy/Satire	7	5	22.0	0.73	1.54
Vlog	9	0	25.1	0.70	1.22
Education	1	64	3.8	0.90	3.63
Political	3	17	11.5	0.90	2.56
Cartoon	15	0	28.5	0.75	1.92
Commentary	2	38	7.3	0.66	2.82
News	5	14	18.1	0.84	2.82
Regional/Dialect	3	25	26.5	0.64	1.11
Health	1	50	12.1	0.90	3.39
Service/Promo	2	0	27.3	0.93	2.61

Table 4: Conversational structure by domain. Spk = median speakers per recording; Solo = % of single-speaker recordings; Turn = mean speaker turns per minute; SR = mean speech ratio; Seg = median segment duration (s). Ordered by total duration.

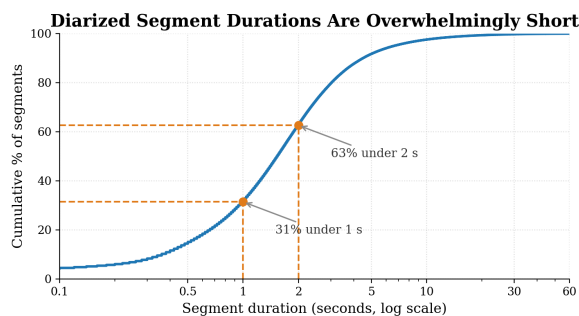


Figure 3: Cumulative distribution of diarized segment durations (log-scaled x-axis). Nearly two-thirds of the 984,647 segments are shorter than two seconds.

source recordings average roughly 52 minutes in duration.

5 Gold Evaluation Split

We release a human-annotated gold split for evaluation. The split comprises 500 clips drawn by stratified sampling across all 17 domains, with per-domain quotas so that even small domains (regional dialect, health) receive enough clips for meaningful per-domain estimates. Seventeen volunteer native Bengali-speaking annotators produced approximately 30 hours of manual annotation: correcting caption-derived transcripts against the audio. The gold split therefore carries both human transcriptions and human speaker labels, held out from training data, removing the main weakness of caption-referenced evaluation.

Domain	WER	CER
Cartoon	28.57	25.85
Vlog	14.17	4.07
Commentary	10.81	5.33
News	10.34	6.57
Movie	7.32	5.28
Waz	6.51	3.88
Music	6.20	4.32
Natok	5.88	3.43
Podcast	3.81	2.81
Talk-show	3.52	1.96
Audio-book	2.57	1.12
Comedy/Satire	0.00	0.00

Table 5: Domain-wise disagreement between the weakly supervised YouTube transcripts and the human-corrected gold references. WER and CER are corpus-level percentages calculated separately for each domain.

The gold split provides human-verified transcript references for ASR evaluation. The released diarization output remains automatically generated and is intended to support future diarization analysis (Section 7).

Table 6 reports global WER and CER on the held-out test split, where all ASR systems are evaluated using the same reference transcripts from youtube captions. These results provide a consistent comparison of ASR performance under the same evaluation setting. Differences across systems therefore reflect model behavior on the held-out test data rather than differences in reference quality. The diarization analysis further highlights the limitations of automatic speaker segmentation discussed in §7. In particular, over-segmentation and false speaker assignments can affect speaker-attributed transcript quality, especially in domains with frequent speaker changes.

Table 5 compares the weakly supervised YouTube transcripts with the human-corrected gold references across different content domains.

6 Baseline Experiments

Figure 4 places all six systems on one axis, the four zero-shot models alongside the two we fine-tuned on Lipi-Ghor. The ordering is the main result: fine-tuning a mid-sized model on this corpus beats every off-the-shelf system we tested, including the Bengali-specific one.

6.1 Zero-Shot Evaluation

Table 6 reports four ASR systems evaluated zero-shot on Lipi-Ghor using YouTube captions as reference. All hypotheses and references are passed

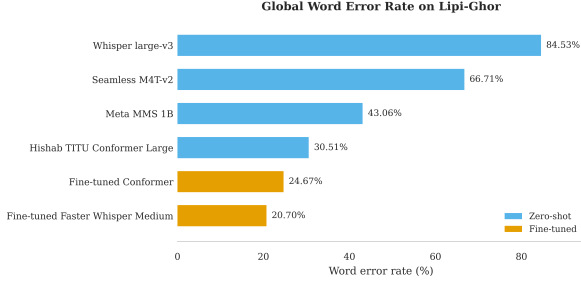


Figure 4: Global WER for the four zero-shot systems (blue) and two fine-tuned models (orange). Fine-tuned Faster Whisper Medium reaches 20.70%, below the best zero-shot system (Hishab TITU, 30.51%).

Model	WER	CER
Hishab TITU Conformer-L	30.51	18.23
Meta MMS 1B	43.06	21.16
Seamless M4T-v2	66.71	45.54
Whisper large-v3	84.53	72.92

Table 6: Zero-shot ASR performance on Lipi-Ghor (%).

through an identical Bengali text-normalization pipeline (Unicode NFC, punctuation and digit standardization) before scoring.

The WER spread (30.51%–84.53%) shows both how hard the corpus is and how much Bengali-specific pre-training matters. Whisper large-v3’s very high error reflects its multilingual but English-centric training distribution and known weakness on Bengali script. These figures overestimate true error against noisy captions; Section 5 quantifies this reference noise using the human-corrected gold split.

Domain-Level Degradation. Table 7 shows domain-specific WER for representative domains; Figure 5 extends this to all seventeen. All models follow a consistent degradation pattern: lowest WER on audiobooks and news, highest on emotive drama, comedy, and regional dialects. Seamless M4T-v2 exceeds 100% WER on Regional/Dialect, indicating near-complete failure. Domains are sorted by TITU’s error rate, and the other three curves rise with it almost monotonically, which suggests the difficulty ordering is a property of the audio rather than of any one model’s training data. Whisper large-v3 is the exception, staying near 85% almost everywhere and failing at a level largely independent of domain.

Structure and Difficulty. Across the seven domains in Table 7, domains with the shortest diarized segments also have the highest error (Spear-

Domain	TITU	MMS	SM4T	Wh-v3
Audio-book	21.02	33.04	57.09	80.87
News	28.03	41.68	57.83	86.23
Talk-show	29.88	43.38	64.81	86.17
Waz	35.75	45.02	74.27	85.15
Natok	44.37	60.14	73.94	89.01
Comedy/Satire	44.56	56.18	85.37	88.64
Regional/Dialect	57.01	70.39	113.70	90.74

Table 7: Domain-specific WER (%) for representative domains. SM4T = Seamless M4T-v2.

Quartile	TITU	MMS	SM4T	Wh-v3
Short	30.54	41.98	66.11	85.01
Medium	29.14	41.79	64.49	84.25
Long	28.46	40.74	63.55	84.00
Very Long	32.60	45.53	70.13	84.93

Table 8: WER (%) by utterance duration quartile.

man $\rho = -0.75$; Figure 6). Audiobooks sit at one extreme with 1.83 s median segments and 21.0% WER, and Natok and Regional/Dialect at the other with roughly 1.1 s segments and 44–57% WER. We want to be careful about how much weight this carries. The relationship is suggestive rather than established, since it rests on seven points and weakens when the smallest domains are dropped, and turn rate on its own does not survive extension to all seventeen domains ($\rho = +0.42$, $p = 0.09$). Cartoons are the clearest counterexample: they turn faster than any other domain at 28.5 turns per minute, yet remain easy to recognize because the speech is clearly enunciated. What the correlation does support is a weaker and more useful claim, that fragmented turn-taking marks the domains where Bengali ASR struggles, and that segment duration is a cheap proxy for finding them without reference transcripts.

Utterance-Length Effects. Table 8 shows WER by utterance duration quartile. The effect is modest, because diarization has already segmented long recordings into short turns and absorbed most of the long-form difficulty. Only the very-long quartile shows a consistent uptick (TITU: 28.5% $\rightarrow$ 32.6%).

6.2 Fine-Tuned Baselines

Table 9 presents a downstream utility evaluation of ASR models fine-tuned using Lipi-Ghor data. The evaluated models and their fine-tuning configurations follow the training setups described in Hasan et al. (2026). Both models were

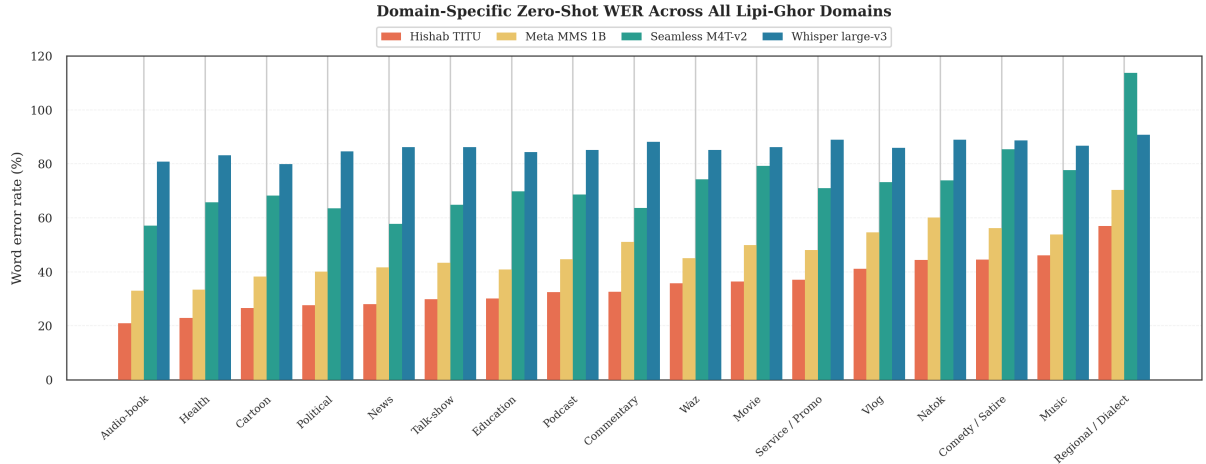


Figure 5: Zero-shot WER across all 17 Lipi-Ghor domains, sorted by Hishab TITU’s error rate. The three Bengali-capable systems degrade together as speech becomes more spontaneous and dialectal; Whisper large-v3 fails near-uniformly across domains. Table 7 lists a representative subset.

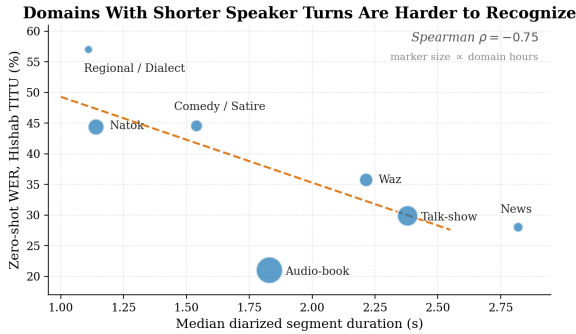


Figure 6: Median diarized segment duration vs. zero-shot WER (Hishab TITU) across the seven domains of Table 7. Marker area proportional to domain hours.

trained using approximately 58 hours of Lipi-Ghor data under comparable data conditions. The corresponding training subset is publicly available at [Train_lighten_check_bangla](#), together with the associated model checkpoints and training resources. **Model A** (Fast Conformer) achieves 24.67% WER and 15.56% CER globally. **Model B** (Faster Whisper Medium) achieves 20.70% WER and 11.90% CER globally. On FLEURS Bengali, Model B scores 21.7% and Model A 20.03%, consistent with prior Bengali Whisper fine-tuning results. For reference, Moonshine-Base fine-tuned with a transplanted BanglaBERT WordPiece vocabulary (Hasan and Rahman, 2026) achieves 21.54% WER and 10.79% CER. These gains over zero-shot baselines indicate that Lipi-Ghor is effective training data despite its weak supervision. However, external benchmark results should be interpreted cautiously, as public

Model	WER	CER
Fast Conformer (A)	24.67	15.56
Faster Whisper Med. (B)	20.70	11.90
Moonshine-Base Bengali [†]	21.54	10.79

Table 9: Fine-tuned model performance on Lipi-Ghor (%). [†]Hasan and Rahman (2026); RTF 0.0053.

models may have prior exposure to related Bengali resources.

6.3 Gold-Split Evaluation

Table 5 provides an analysis of weak supervision quality on the 500-clip human-corrected split. It compares YouTube caption transcripts against human-corrected references across domains using WER and CER. These values measure transcript disagreement and should not be interpreted as ASR model error.

The level of transcript disagreement varies across domains. Some domains show higher differences between YouTube-generated transcripts and human corrections due to variations in caption quality and speech characteristics. Therefore, these values indicate reference transcript noise rather than ASR system performance.

7 Discussion

What the Diarization Gets Wrong. Two patterns in the diarizer’s output are artifacts rather than facts about Bengali speech.

First, the annotation never contains more than two simultaneous speakers: 682 recordings peak at

two concurrent speakers, 337 at one. Genuine talk-show and debate audio contains moments where three or more people talk over each other; this ceiling comes from Precision-2’s output representation. The overlap figures in Table 3 should be read as detected overlap under a two-speaker cap.

Second, detected overlap *falls* as speaker count rises, which inverts what multi-party conversation would predict. The likely cause is over-segmentation: when the diarizer splits a single voice across many labels, those fragments are consecutive rather than concurrent, so overlap drops. Supporting evidence: 21.8% of all 9,056 speaker labels account for under ten seconds of speech each, and the 105-label recording has a dominant-speaker share of 0.11. Speaker counts are therefore reliable at the low end and inflated at the high end; the median of 4 is more trustworthy than the mean of 8.9.

Pipeline Generalizability. The construction methodology is not Bengali-specific by design. Any language with a substantial YouTube presence and native-language caption tracks can serve as a corpus source. We have only validated it on Bengali, so cross-lingual transfer is a hypothesis rather than a demonstrated result. The bottleneck is domain-aware channel discovery: in our collection, 3–5 rich channels per domain collected 20–50 hours, suggesting comparable scales are achievable for other South Asian or low-resource languages with modest curation effort.

8 Conclusion

We presented Lipi-Ghor, an 885-hour weakly supervised Bengali speech corpus spanning 17 domains, together with a scalable YouTube-based construction pipeline. Experiments reveal consistent degradation on emotive, dialectal, and long-form speech, while fine-tuning demonstrates the corpus’s utility for Bengali ASR. We release the corpus metadata, annotations, pipeline code, and human-annotated gold evaluation split.

Limitations

The most significant limitation is weak supervision: transcripts are derived from YouTube caption tracks, have not been fully human-verified, and systematically differ from human transcriptions in punctuation, digit normalization, and code-switched speech handling. Speaker labels over the corpus are produced by Precision-2 and are not

human-verified; therefore, the “Speaker” dimension of SSTT should be regarded as automatically generated corpus metadata rather than a manually validated diarization benchmark. Precision-2 is a commercial, API-gated model, so the diarization stage is not fully open; we release its output so downstream use does not require re-running it. Regional dialect coverage is sparse (4 videos, 1.29 hours), insufficient for dialect-robust training. Exact reproducibility is constrained by YouTube’s evolving API rate policies and the partially manual collection process; we mitigate this by releasing video identifiers and pipeline scripts.

9 Ethical Considerations

Data Source, Copyright, and Licensing. We do not redistribute raw audio. We release video identifiers and pipeline scripts so users reconstruct audio locally, following the video-ID release model of Tabib et al. (2026) and Galvez et al. (2021). Our original contributions (video-ID lists, diarization output, pipeline code, and gold annotations) are released under CC-BY-4.0. Caption-derived transcripts remain subject to the rights of the original uploaders and are distributed for non-commercial research use only.

Data Removal Policy. Rights holders, uploaders, or individuals whose voices appear in source videos may request removal by supplying video identifiers. On a valid request we remove the corresponding entries, publish an updated dataset version, and record removed identifiers in a public changelog.

Speaker Consent and Privacy. Speech comes from public YouTube channels whose creators published the content themselves. We restrict the release to research use and prohibit speaker re-identification and surveillance applications. Speaker labels are local to each video and do not support cross-video re-identification.

Intended Use. Lipi-Ghor is intended for ASR pre-training, spoken language modeling, and multi-lingual NLP research. It is not intended for speaker re-identification, surveillance, or high-stakes production ASR without additional human verification.

Acknowledgments

The authors are grateful to Almighty Allah for His guidance and blessings throughout this work.

References

- Samiul Alam, Asif Sushmit, Zaowad Abdullah, Shahrin Nakkhatra, Md. Nazmuddoha Ansary, Syed Mobassir Hossen, Sazia Morshed Mehnaz, Tahsin Reasat, and Ahmed Imtiaz Humayun. 2022. [Bengali common voice speech dataset for automatic speech recognition](#). *arXiv preprint arXiv:2206.14053*.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, 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. European Language Resources Association.
- Kaushal Santosh Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M. Khapra, and Pratyush Kumar. 2023. [Vistaar: Diverse benchmarks and training sets for indian language ASR](#). In *Proc. Interspeech 2023*, pages 4384–4388.
- Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. [pyannote.audio: neural building blocks for speaker diarization](#). In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain McCowan, Wilfried M. Post, Dennis Reidsma, and Pierre D. Wellner. 2005. [The ami meeting corpus: A pre-announcement](#). In *Machine Learning for Multimodal Interaction*.
- Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. [GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio](#). In *Proc. Interspeech 2021*, pages 3670–3674.
- Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Senior. 2020. [Spot the conversation: Speaker diarisation in the wild](#). In *Proc. Interspeech 2020*, pages 299–303.
- Biswajit Das, Sandipan Mandal, and Pabitra Mitra. 2011. [Bengali speech corpus for continuous automatic speech recognition system](#). In *2011 International Conference on Speech Database and Assessments (Oriental COCOSA)*, pages 51–55. IEEE.
- Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. [The people’s speech: A large-scale diverse english speech recognition dataset for commercial usage](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Sanjid Hasan, Risalat Labib, A H M Fuad, and Bayazid Hasan. 2026. [Make it hard to hear, easy to learn: Long-form bengali asr and speaker diarization via extreme augmentation and perfect alignment](#). *Preprint*, arXiv:2602.23070.
- Sanjid Hasan and Md. Abdur Rahman. 2026. [Tokenizer transplantation: Mitigating autoregressive collapse in edge-efficient bengali ASR](#). *Preprint*, arXiv:2607.09598.
- Addison Howard, Ahmed Imtiaz Humayun, Ashley Chow, Ryan Holbrook, Sushmit, and Tahsin. 2023. [Bengali.ai speech recognition](#). <https://www.kaggle.com/competitions/bengaliai-speech>. Kaggle.
- Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. 2023. [IndicSUPERB: A speech processing universal performance benchmark for indian languages](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12942–12950.
- 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.
- Sushant Kafle and Matt Huenerfauth. 2017. [Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing](#). In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, pages 165–174.
- Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. [YODAS: Youtube-oriented dataset for audio and speech](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoeng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. [Scaling speech technology to 1,000+ languages](#). *Journal of Machine Learning Research*, 25(97):1–52.
- pyannoteAI. 2025. [Precision-2: Premium speaker diarization](#). Model card, <https://huggingface.co/pyannote/speaker-diarization-precision-2>. Commercial speaker-diarization model accessed via the pyannoteAI API.

- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *arXiv preprint arXiv:2212.04356*.
- Fazle Rabbi Rakib, Souhardya Saha Dip, Samiul Alam, Nazia Tasnim, Md. Istiak Hossain Shihab, Md. Nazmuddoha Ansary, Syed Mobassir Hossen, Marisia Haque Meghla, Mamunur Mamun, Farig Sadique, Sayma Sultana Chowdhury, Tahsin Reasat, Asif Sushmit, and Ahmed Imtiaz Humayun. 2023. [Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking](#). *arXiv preprint arXiv:2305.09688*. Presented at Interspeech 2023.
- Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2021. [The third DIHARD diarization challenge](#). In *Proc. Interspeech 2021*, pages 3570–3574.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, et al. 2023. [SeamlessM4T: Massively multilingual and multimodal machine translation](#). *arXiv preprint arXiv:2308.11596*.
- H. M. Shadman Tabib, Istiak Ahmmed Rifti, Abdullah Muhammed Amimul Ehsan, Somik Dasgupta, Md Zim Mim Siddiquee Sowdha, Abrar Jahin Sarker, Md. Rafiul Islam Nijamy, Tanvir Hossain, Mst. Metaly Khatun, Munzer Mahmood, Rakesh Debnath, Gourab Biswas, Asif Karim, Wahid Al Azad Navid, Masnoon Muztahid, Fuad Ahmed Udoy, Shahad Shahriar Rahman, Md. Tashdiqur Rahman Shifat, Most. Sonia Khatun, and 8 others. 2026. [Bengali-Loop: Community benchmarks for long-form bangla ASR and speaker diarization](#). *Preprint*, arXiv:2602.14291.
- Changhan Wang, Morgane Rivi re, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. [VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 993–1003.

A Extended Dataset Statistics

Domain	Videos*	Hours
Audio-book	277	254.37
Talk-show	176	142.40
Movie	74	122.18
Natok (Drama)	95	78.66
Podcast	62	57.59
Waz (Sermons)	66	47.89
Music	37	31.91
Comedy/Satire	43	31.35
Vlog	42	27.13
Education	39	25.14
Political	29	19.35
Cartoon	23	16.88
Commentary	13	14.51
News	35	12.88
Regional/Dialect	4	1.29
Health	2	1.14
Service/Promo	2	0.67
Total	1,019	885.34

Table 10: Domain distribution of Lipi-Ghor. *Per-domain counts are raw video counts before cross-domain deduplication (1,019 total); deduplication yields 998 unique videos. Hours are on the deduplicated corpus.

Table 11: Distribution of YouTube Caption Types in Lipi-Ghor

Caption Type	Videos	Percentage
Manually created captions	102	10.22%
Auto-generated captions	896	89.78%
Total	998	100.00%

B The Annotation Toolkit

The 500-clip gold split (§5) was produced with a purpose-built web annotation tool rather than a general-purpose labeling platform. We describe it here because the design determines what the gold labels can and cannot be trusted to contain, and because the corpus release includes the tool’s export scripts. Figure 7 shows the annotation screen as a volunteer sees it.

Interface. Each annotation job loads one recording together with its automatic annotations, so annotators correct existing output rather than transcribing from scratch. The two “Load Original” controls re-fetch the caption-derived transcript and the Precision-2 diarization, which lets an annotator revert to the machine baseline if an edit goes wrong.

The waveform supports variable-speed playback from $0.5\times$ to $1.5\times$ and 2 s or 5 s jumps, and the coloured ribbon under it shows the current speaker assignment across the whole recording, so an annotator can see turn structure at a glance instead of inferring it row by row. The segment list below is the editing surface: every row exposes the speaker label, the start and end timestamps, and the Bengali transcript, and rows can be split at the playhead, merged with a neighbour, or deleted outright. Keyboard shortcuts cover the frequent operations, including play/pause, ± 2 s seeking, split, merge, delete, undo, and direct speaker assignment via the number keys. Work is saved as a draft continuously, and a job is submitted only once the annotator has reviewed every segment.

Why the Video Is Retained. The tool deliberately keeps the source video visible, which is a departure from audio-only transcription interfaces. The reason is that our annotators verify speaker turns as well as transcripts, and speaker attribution from audio alone is unreliable in exactly the domains where the corpus is densest. Talkshow panels often place several speakers of the same gender in similar acoustic conditions, and natok scenes may involve several actors with deliberately similar delivery; in both cases voice quality is a weak cue. Seeing who is on screen, whose lips are moving, and how the camera frames a turn resolves attributions that the waveform leaves genuinely ambiguous. Removing the player would have made the gold speaker labels less reliable, which would defeat their purpose as a reference for measuring diarization error. The trade-off is that visual cues can also mislead, since an off-screen speaker may be talking while the camera holds on someone else, so annotators were instructed to treat the audio as authoritative and the video as a disambiguation aid rather than the reverse.

Provenance and Export. Each submitted job records the annotator, the timestamps of the edits, and the machine baseline the annotator started from, so a corrected segment can always be compared against the automatic output it replaced. This provenance enables future fine-grained analysis of diarization errors at the segment level and supports reproducible evaluation against human-corrected annotations. Completed jobs are exported to the same SSTT schema used by the bulk corpus, so the

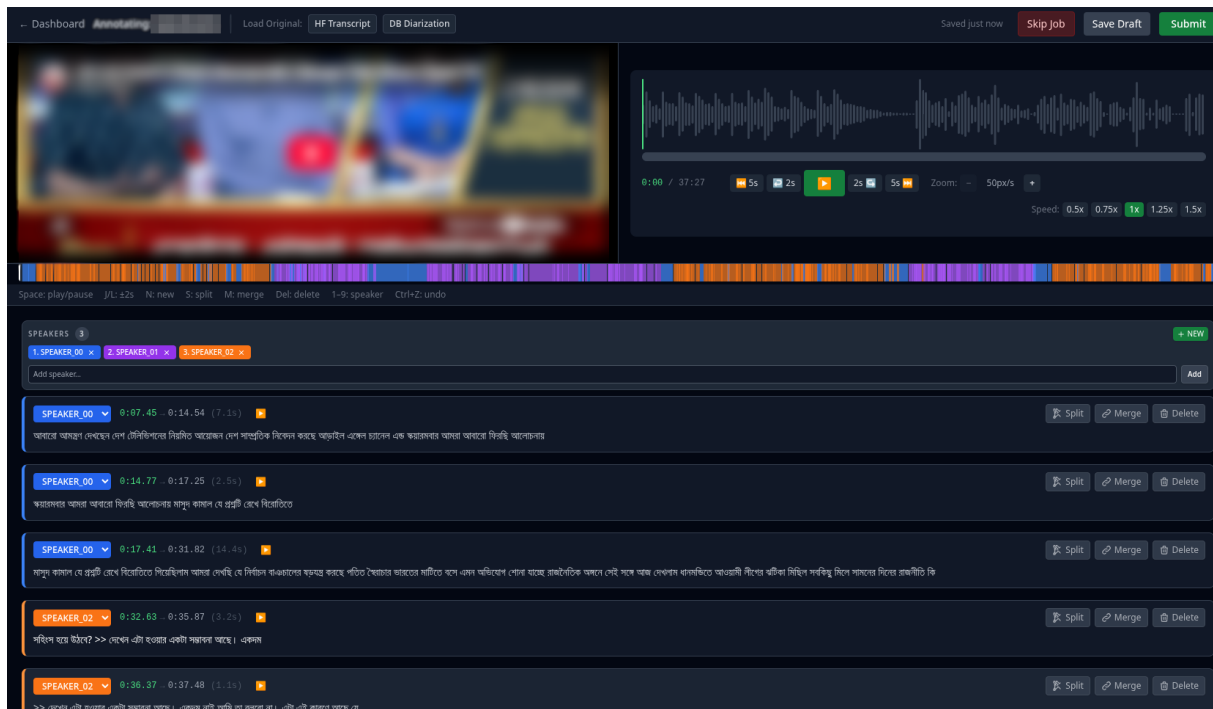


Figure 7: The Lipi-Ghor annotation interface. Clockwise from top left: the embedded video player, the waveform with a per-speaker diarization ribbon beneath it, the speaker roster for the recording, and the segment list where each row carries a speaker assignment, timestamps, and an editable Bengali transcript. Annotators can split, merge, or delete segments and reassign speakers from the dropdown. Identifying details of the source video, including the video identifier and all channel branding, are masked for double-blind review; annotators saw the unmasked frame.

gold split and the weakly supervised corpus share one format and the gold clips can be substituted for their noisy counterparts without any change to a training pipeline.

C Diarization Output Licensing

Speaker-diarization annotations were generated using the commercial pyannoteAI Precision-2 service. Under pyannoteAI’s Terms of Use, intellectual property rights in results produced by the service are assigned to the customer and may be reproduced, modified, reused, and transferred to third parties. Accordingly, we release the generated Lipi-Ghor diarization outputs under CC-BY-4.0. This license applies only to the generated annotations and does not apply to the proprietary Precision-2 model or the pyannoteAI service itself.