One Head, Three Transcripts: A Controlled Oracle-to-ASR Audit of Bangla Speech–Text Sentiment Analysis

Md Raihan[1] **Nilam Rajak**[2] **Kuntal Ghosh**[3] **Sathyanarayanan Dhorali**[1]
[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]ERP Professionals, Germany [3]WallStreetDocs (WSD), United Kingdom
md.raihan@fau.de, nilam.rajak@erp-professionals.de
kuntal.ghosh@wsd.com, sathyanarayanan8998@gmail.com

Abstract

Speech sentiment systems are routinely benchmarked with gold transcripts, but a deployed cascade must read whatever its recognizer produces. We audit that gap on BanglaMUSE, a Bangla read-speech sentiment corpus, using sentence-grouped cross-validation and a controlled substitution design in which one trained classifier is evaluated three times while only the test-time transcript changes. Replacing gold transcripts with ASR output costs 0.109 macro-F1 with Bengali XLS-R CTC and 0.341 with our Whisper large-v3 configuration, a high-error stress case. Retraining the classifier on matched recognizer output recovers 57–81% of the gap, but re-selecting only the decision threshold on ASR validation data, with weights and probabilities unchanged, already recovers 28–49%. Adding the speech embedding improves over ASR text, yet replacing it with acquisition/nuisance features gives a similar gain for gold-trained heads. These features alone predict sentiment at 0.669 macro-F1 (group-permutation null mean 0.495, $p = 0.001$), and a speech-only classifier that reaches 0.940 falls to 0.621 when both speaker and sentence are unseen. These speech-only and fusion results therefore do not provide clean evidence of prosodic sentiment cues. Low-resource speech–text benchmarks should report oracle and deployable conditions, separate threshold effects from retraining, and test acquisition variables before interpreting audio.

1 Introduction

Many speech sentiment and emotion systems evaluate the text branch on human transcripts. This is reasonable for studying fusion in isolation, but a deployed cascade consumes recognizer output. Prior work has shown that this mismatch can materially affect affective prediction on English benchmarks (Sheikh et al., 2018; Culnan et al., 2021; Li et al., 2024); controlled evidence remains scarce for low-resource languages, where ASR errors are larger and gold text is a less realistic deployment assumption.

Bangla makes a useful case study. It is spoken by more than 230 million people yet remains under-resourced in speech technology (Nayeem et al., 2026). Its writing system also makes normalization consequential for ASR scoring: visually coherent grapheme clusters can span multiple Unicode code points, so we use symmetric normalization and grapheme-cluster CER in addition to WER. We compare Whisper, a general multilingual recognizer (Radford et al., 2023), with a community Bengali CTC model fine-tuned from XLS-R-300M (Babu et al., 2022).

We audit the gap on BanglaMUSE (Nayeem et al., 2026), a corpus of 1,000 sentiment-labelled Bangla sentences each read by four speakers. Our design is deliberately narrow. We fix the encoders, freeze them, and use one lightweight classifier head, because the question is not which architecture wins but how much the evaluation protocol itself inflates reported performance. The central device is a controlled substitution: the *same trained head* is evaluated on gold and on automatic transcripts, so the measured difference is attributable to transcript source rather than to two separately fitted models.

Four questions organize the work. **RQ1:** how accurately do a general multilingual recognizer and a Bangla-specific one transcribe this corpus? **RQ2:** holding the classifier fixed, how much does text-only performance change when gold test transcripts are replaced by automatic ones? **RQ3:** does the same hold for speech–text fusion? **RQ4:** how much of any gap is recovered by training on matched recognizer output, and how much of that recovery needs only a different decision threshold? A fifth question turned out to matter as much: whether speech-input results can be interpreted cleanly as evidence of prosodic sentiment. They cannot, because acquisition conditions are strongly

Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages, pages 40–54
November 9, 2026 ©2026 Association for Computational Linguistics

aligned with the labels.

Our contribution is an evaluation and audit study rather than a new method: (i) a sentence-disjoint evaluation protocol for BanglaMUSE with automated leakage assertions; (ii) a controlled oracle-to-ASR substitution audit across two Bangla-capable recognizers with sentence-clustered uncertainty; (iii) an operational adaptation analysis that separates decision-threshold adaptation from head retraining; and (iv) an acquisition audit—a nuisance-feature fusion control, a permutation null and a speaker-transfer diagnostic—showing that speech-input results on this corpus are confounded with recording conditions.

2 Related Work

Imperfect transcripts are known to matter for affective prediction. Sheikh et al. (2018) study sentiment analysis with imperfect ASR and acoustic views on CMU-MOSI/MOSEI. Culnan et al. (2021) compare gold and automatic transcripts across emotion, sarcasm and personality detection and report partial mitigation from audio together with a weak link between WER and downstream performance. Wu et al. (2022) propose multimodal refinement of sentiment-word representations and evaluate it on CMU-MOSI transcripts produced by three ASR systems. Li et al. (2024) benchmark eleven recognizers on IEMOCAP (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2016) and MSP-Podcast (Lotfian and Busso, 2019) using text-only and six bimodal fusion techniques, and propose an ASR-error-robust framework. We therefore do not claim that oracle-to-ASR degradation, training on recognizer output, or holding a classifier fixed is new. We instead ask how large this gap is on a low-resource Bangla dataset, how much can be recovered, and which alternative explanations survive a controlled same-head analysis.

A related line trains downstream systems to tolerate recognition errors: simulated ASR errors can be injected when training spoken language understanding systems (Simonnet et al., 2018), and interventional speech-noise injection improves generalization across recognizers (Jung et al., 2024). Our matched-training conditions use real recognizer output instead and test only this simplest form of the idea.

Self-supervised speech representations (Baevski et al., 2020; Hsu et al., 2021) and their cross-lingual extensions (Babu et al., 2022) underpin most current low-resource recognizers, either fine-tuned with a CTC objective (Graves et al., 2006) or used as encoders. Large multilingual systems also broaden Bangla coverage: Whisper uses large-scale weak supervision (Radford et al., 2023), whereas MMS combines multilingual self-supervised pre-training with supervised ASR adaptation across more than 1,000 languages (Pratap et al., 2024). Bhogale et al. (2023) find that recognizer rankings on one Indic benchmark can correlate weakly with rankings on another, and published Bangla error rates differ in normalization, punctuation handling and language-model use, so we report our own operating points.

Models routinely exploit incidental dataset structure rather than the intended signal (Torralba and Efros, 2011; Geirhos et al., 2020), as documented for annotation artifacts in NLI (Gururangan et al., 2018) and spurious source tags in image classification (Lapuschkin et al., 2019). Our acquisition audit applies this concern to speech: we ask whether recording conditions, rather than content or delivery, carry the label.

3 The BanglaMUSE Corpus and Its Preparation

BanglaMUSE (Nayeem et al., 2026) contains 1,000 sentiment-labelled Bangla sentences, balanced 500 positive and 500 negative, each read by four speakers (two female, two male). We use it because every recording comes with its exact reference text and a sentence-level sentiment label and every sentence is read four times, which a transcript-substitution audit requires. Prominent public Bangla affective-speech alternatives, such as SUBESCO (Sultana et al., 2021) and BanglaSER (Das et al., 2022), annotate categorical emotions on acted readings of a small fixed set of sentences.

The delivered corpus required repair, which we performed by content-derived rules rather than by hand (Appendix A.1). The prepared corpus contains 3,996 usable recordings (1,997 positive, 1,999 negative; 4.08 hours) over 998 sentence groups; two sentence pairs with identical token multisets are merged into shared groups. Looser paraphrases are deliberately not merged: 60 sentence pairs have token Jaccard similarity of at least 0.7, and 13 of those carry opposite labels, so such overlap may inflate *or* deflate absolute scores; Section 6.2 shows the reported gaps are insensitive to these items. The recordings mix 16, 44.1 and 48 kHz, mono and

stereo, and two codecs. All audio is resampled to 16 kHz mono, but as Section 6.5 shows, acquisition traces survive.

4 Evaluation Protocol

All recordings of a sentence group share one outer fold. We use 5-fold stratified group cross-validation, grouped on sentence group and stratified on sentiment (about 200 groups and 800 recordings per test fold), with a deterministic group-disjoint 10% inner validation split for early stopping and model selection (Appendix A.2). An utterance-random split would test the text branch on sentences whose exact wording it had seen through another speaker. Because this failure is silent, 44 automated assertions confirm zero sentence-group and normalized-transcript overlap between every training, validation and test partition, and exactly one out-of-fold prediction per recording and condition.

A single frozen normalizer is applied identically to gold text and to both recognizers: Unicode NFC, removal of model-specific and zero-width characters, Bangla digit mapping, whitespace collapsing, and punctuation removal. Punctuation is removed because Whisper emits it and CTC models do not; retaining it would charge the CTC system for a stylistic difference rather than a recognition error. WER is computed over whitespace tokens and CER over extended grapheme clusters, since one visual unit of Bangla routinely spans several code points. A narrowly scoped post-hoc canonicalization check is reported in Appendix A.18.

Unless stated otherwise, scores are pooled recording-level out-of-fold macro-F1. Four speakers reading the same sentence are not independent observations, so the bootstrap resamples sentence *groups* with all their recordings (10,000 iterations), shares each resample across conditions, and reports paired 95% intervals on differences between conditions (Efron and Tibshirani, 1994; Dror et al., 2018).

The four substitution gaps, the two fusion-versus-ASR-text comparisons (C6−C3, C7−C4) and the four matched-training recovery comparisons were declared before the primary analysis; the high-similarity exclusion, X1–X2 and the C1-versus-fusion comparisons were post hoc. The nuisance-fusion control, threshold-only adaptation, lexical baseline, permutation null, seed variability, recovery intervals and speaker-transfer diagnostic were added later. Their settings were fixed before they were run, but after the primary results were known, so we label them *secondary* throughout.

5 Experimental Setup

ASR-A is the multilingual Whisper large-v3 model of Radford et al. (2023),[1] run through Hugging Face `transformers` 5.14.1 with beam 2, early stopping, at most 200 new tokens, Bengali and transcription forced, and sampling disabled. Every clip is shorter than 8.4 s, so Whisper's long-form options (conditioning on previous text and the compression-ratio, log-probability and no-speech thresholds) are inactive (Appendix A.16). Beam search was preferred over greedy decoding on a 300-utterance diagnostic sample drawn from this corpus, so this choice used evaluation references: greedy decoding emits romanized Latin-script output for 31% of that sample and reaches 0.744 WER, against 0.692 for beam 2 and 0.679 for beam 5. We retained beam 2 as the primary, compute-motivated operating point; beam 5 is reported only to show decoding sensitivity. Whisper's tokenizer needs a median of 72 new tokens for these sentences; a 48-token budget silently truncates 99.5% of them.

ASR-B is Bengali XLS-R CTC[2], a community Bengali fine-tune of XLS-R-300M (Babu et al., 2022) with a CTC head (Graves et al., 2006), decoded greedily and with no external language model; its WER is therefore not comparable to published LM-rescored numbers. Neither recognizer was fine-tuned on BanglaMUSE: each recording was transcribed once by the unadapted recognizers, and the transcripts are indexed by the downstream folds.

The text branch uses the frozen `csebuetnlp/banglabert` checkpoint (Bhattacharjee et al., 2022) with attention-mask-aware mean pooling over the final layer. The speech branch uses the frozen `facebook/wav2vec2-xls-r-300m` checkpoint (Babu et al., 2022) with mask-aware temporal mean and standard-deviation pooling. Embeddings are computed once and reused across every fold, seed and condition. ASR-B is a fine-tune of the same self-supervised backbone as the speech encoder; thus, in the Bengali XLS-R condition, the recognizer and speech encoder do not provide acoustic-backbone diversity.

[1] `openai/whisper-large-v3`, revision `06f233fe`.

[2] `arijitx/wav2vec2-xls-r-300m-bengali`, revision `45ed7c70`.

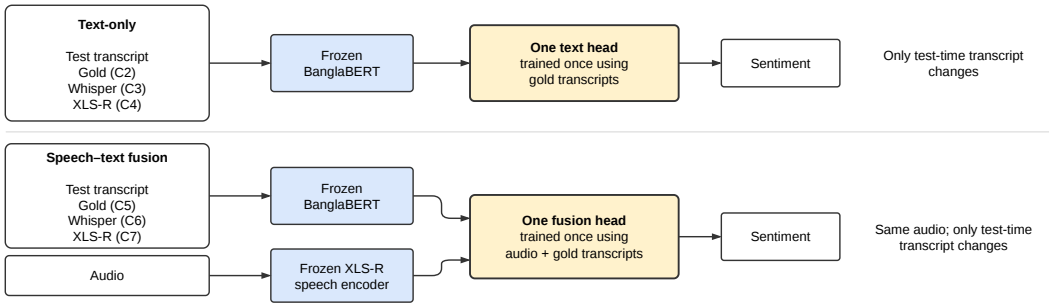


Figure 1: Controlled substitution design. C2–C4 reuse one text head trained on gold transcripts, and C5–C7 reuse one fusion head trained on audio and gold transcripts; only test-time transcripts change, with audio fixed for fusion. The secondary control N5–N7 follows the same design with acquisition/nuisance features in place of the speech encoder.

The classifier is a nonlinear head, LayerNorm $\rightarrow$ Linear(256) $\rightarrow$ GELU $\rightarrow$ Dropout(0.2) $\rightarrow$ Linear(2), trained with AdamW (Loshchilov and Hutter, 2019) (learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, at most 60 epochs, early-stopping patience 8) on unweighted cross-entropy; the epoch with the highest validation macro-F1 is retained. Embeddings are L2-normalized per sample, and fusion concatenates the normalized speech and text embeddings. The primary heads fit no scaler or other data statistic; the secondary nuisance and lexical baselines fit their preprocessing on training partitions only. Results average the probabilities of three seeds (13, 42, 87) before applying the rule $p(\text{positive}) \geq 0.5$.

Condition definitions: C1 is an independently trained speech-only head. C2, C3 and C4 reuse one gold-trained text head evaluated on gold, Whisper and Bengali XLS-R transcripts; C5–C7 reuse one gold-trained fusion head in the same way (Figure 1). Training and evaluation occur within a single process, so head reuse is a property of the code rather than of a checkpoint round-trip. Gold-trained heads are early-stopped on gold validation transcripts. A1–A4 are text and fusion heads trained and early-stopped on the matching recognizer's transcripts; X1 and X2 reuse the A1 and A2 heads on the *other* recognizer's transcripts (post hoc; Appendix A.5). The secondary conditions are defined where they are used. We keep three comparison types apart: *controlled substitutions* (same head; the test transcript changes), *threshold-only adaptation* (same weights and probabilities; a validation-selected cutoff changes) and *system-level comparisons* (different fitted heads or inputs).

System	WER	CER	Sub.	Del.	Ins.
Whisper	0.687	0.360	0.590	0.012	0.086
Bengali XLS-R	0.276	0.111	0.230	0.028	0.017

Table 1: ASR quality on BanglaMUSE: Whisper large-v3 and Bengali XLS-R CTC. Neither recognizer was fine-tuned on the corpus. All rates are corpus-level (edits pooled over recordings).

6 Results

6.1 Recognition quality

Table 1 reports RQ1. Whisper large-v3 has a WER of 0.687, 2.5$\times$ the 0.276 WER of Bengali XLS-R CTC; CER gives the same ranking (0.360 versus 0.111). Whisper's insertion rate is about 5.1$\times$ that of the CTC system (0.086 versus 0.017), but repetition loops do not explain it: only 2 of 3,996 Whisper hypotheses show an immediate n-gram loop or an excessive compression ratio, and they contain none of the insertions. Romanized output matters more: 9.8% of Whisper hypotheses contain Latin script, including 11 of the 20 with the highest WER (Appendix A.16). We treat the two systems as tested operating points, not as representative of Bangla ASR.

6.2 Controlled oracle-to-ASR substitution

Table 2 and Figure 2 give the principal result, addressing RQ2 and RQ3. Within each column, the same trained head is reused, so only the supplied test-time transcript source changes. For text alone, gold-transcript evaluation overstates performance under ASR input by 0.109 macro-F1 with Bengali XLS-R and 0.341 with Whisper; the corresponding fusion gaps are 0.086 and 0.304. Every interval excludes zero, and every gap keeps its sign and approximate size in each of the three training

	Text only		Text + nuisance (secondary)		Text + speech	
Test transcript	Macro-F1	Gap [95% CI]	Macro-F1	Gap [95% CI]	Macro-F1	Gap [95% CI]
Gold	C2 0.976	–	N5 0.980	–	C5 0.981	–
Whisper	C3 0.635	0.341 [0.316, 0.365]	N6 0.692	0.289 [0.266, 0.311]	C6 0.677	0.304 [0.282, 0.328]
Bengali XLS-R	C4 0.867	0.109 [0.095, 0.124]	N7 0.890	0.090 [0.078, 0.103]	C7 0.895	0.086 [0.074, 0.099]

Table 2: Principal conditions and acquisition/nuisance control. Within each column, the same gold-trained head is reused and only the test transcript changes. Gaps are paired sentence-clustered bootstrap differences from the gold condition. Full scores are in Appendix Table 6.

seeds (C2−C3 ranges from 0.311 to 0.353; Appendix A.10). Excluding from evaluation every sentence with token Jaccard similarity of at least 0.7 to another sentence (84 sentences, 333 recordings) moves the four gaps by at most 0.004, with every interval still excluding zero (Appendix Table 7); this check changes the scored items, not the training data.

Adding the speech embedding improves macro-F1 over ASR text by 0.042 [0.030, 0.053] under Whisper (C6−C3) and by 0.028 [0.020, 0.035] under Bengali XLS-R (C7−C4). Because text-only and fusion use different trained heads, these are system-level comparisons. As a secondary control, we replaced the speech embedding with the 17-dimensional acquisition/nuisance feature vector of Section 6.5, whose imputation, encoding and standardization are fitted on training partitions only (N5–N7; Appendix A.6). This text+nuisance head improves over ASR text by 0.057 [0.045, 0.068] and 0.023 [0.016, 0.030], and its direct differences from speech fusion are small (C6−N6 = −0.015 [−0.025, −0.004]; C7−N7 = 0.005 [−0.001, 0.011]) and change sign across training seeds. The measured acquisition/nuisance features produce a fusion gain of similar size. We therefore cannot attribute the speech-fusion gain specifically to prosody or other sentiment-bearing acoustics. This control also does not tell us which cues the speech encoder actually uses, and the acquisition/nuisance feature vector is not guaranteed to be content-free: its trailing 0.15 s window lies within 20 dB of the utterance's level in 47% of recordings.

The oracle score itself is near ceiling. A secondary TF-IDF (word uni- and bigrams over whitespace tokens) and logistic-regression baseline on gold text (Appendix A.8), fitted on training partitions only under the same folds, reaches 0.854, which is 0.122 [0.101, 0.145] below the frozen-encoder head (C2): simple lexical features separate the classes well under this protocol, though less well than the pretrained encoder. This does not rule out lexical shortcuts, and the high-similarity check above is not a family-disjoint retraining.

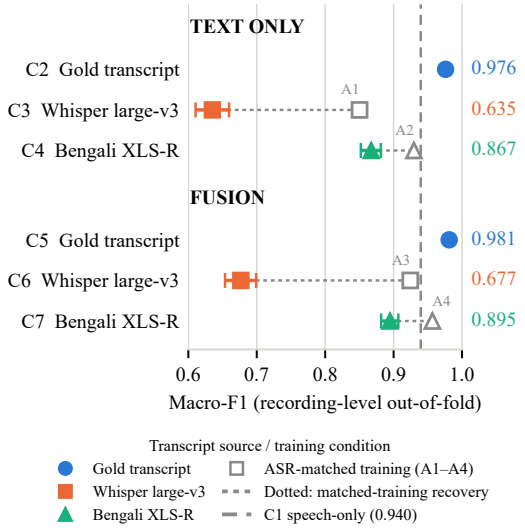


Figure 2: Macro-F1 for C2–C7. Error bars show 95% paired sentence-clustered bootstrap intervals for oracle-to-ASR gaps, mapped to scores relative to C2 or C5. Hollow markers show ASR-matched heads (A1–A4); dotted segments show recovery. The dashed line is the independently trained speech-only head C1.

6.3 Degradation tracks transcription quality

Is the degradation simply item difficulty? Across bins fixed before downstream performance was inspected, substituted-text accuracy for Bengali XLS-R falls from 0.971 at zero WER to 0.738 in the highest-WER bin, while oracle-text accuracy on the *same items* stays between 0.971 and 0.982 and speech-only accuracy between 0.930 and 0.946 (Table 3); Whisper shows the same pattern across its positive-WER bins. Normalized transcripts are exact for 801 of 3,996 XLS-R but only 10 Whisper recordings, so the table omits Whisper's zero-WER bin; its 10 items remain in every other analysis. This is an observed association between transcription quality and downstream accuracy, not a causal dose-response: the near-ceiling oracle and the con-

ASR	Bin	n	WER	Oracle	ASR text	ASR fusion	Speech
Whisper	low	1392	0.424	0.971	0.747	0.787	0.940
Whisper	mid	1642	0.732	0.975	0.645	0.678	0.938
Whisper	high	952	1.044	0.985	0.601	0.623	0.942
XLS-R	WER = 0	801	0.000	0.971	0.971	0.981	0.945
XLS-R	low	1152	0.167	0.982	0.909	0.927	0.930
XLS-R	mid	1092	0.333	0.978	0.865	0.891	0.940
XLS-R	high	951	0.579	0.971	0.738	0.791	0.946

Table 3: Accuracy by transcription-quality bin. Bins are zero WER followed by tertiles of positive WER; Whisper's zero-WER bin ($n = 10$) is omitted. WER is the unweighted mean of per-recording WERs; Table 1 reports corpus-level WER. WER can exceed 1 when insertions make the total edit count larger than the number of reference words.

founded speech-only condition are imperfect controls for item difficulty. Fusion beats text alone in every bin shown (descriptive error patterns: Appendix A.15).

6.4 How much of the gap is recovered, and how

Retraining the head on matched ASR transcripts (A1–A4) recovers 57–81% of the gap (Table 4). The gold-trained heads, however, also shift their decisions under ASR input: the predicted-positive rate falls from 50.0% on gold text to 18.7% on Whisper and 38.9% on XLS-R text against a true rate of 50%, and C3's recall on positive recordings is only 0.358. We therefore asked, in a secondary analysis, how much of the recovery needs only a new threshold. For each fold we re-selected the decision threshold of the unchanged gold-trained head on its inner-validation ASR transcripts (seed-averaged probabilities; grid 0.000–1.000; ties broken towards 0.5) and applied it to the test fold; weights, probabilities and test labels are untouched (Appendix A.7).

Threshold-only adaptation restores balanced predictions and recovers 28–49% of the gap, for example 0.149 [0.124, 0.174] macro-F1 for Whisper text (T3−C3) with thresholds between 0.010 and 0.099. A threshold selected on *gold* validation data gains at most 0.008, so the effect comes from adapting the cutoff to ASR input rather than from optimizing a cutoff as such. Head retraining still adds 0.032 to 0.097 over threshold-only adaptation in every setting, and 0.031 to 0.097 when the matched heads also receive their own tuned thresholds, which barely change them; the retraining gain therefore does not come from the cutoff alone. All conditions use the same test labels, so the change is not a change in the label prior. Threshold adaptation also leaves the predicted probabilities unchanged, so we do not interpret it as calibration. With the encoders frozen, the remaining gain comes from retraining the classifier head. For each setting, let O be the oracle score, B the gold-trained head's score on ASR input with the default threshold, T its score after ASR-validation threshold adaptation, and A the ASR-matched head's score with the default threshold. We use the decomposition $O-B = (T-B)+(A-T)+(O-A)$ only as an accounting device; it should not be interpreted as a causal separation of mismatch and information loss.

Recovery is larger for fusion (72–81%) than for text (57–63%). A secondary matched text+nuisance head (NA3, NA4) also improves over matched text (by 0.042 and 0.012), but less than matched speech fusion (A3−NA3 = 0.032 [0.023, 0.040]; A4−NA4 = 0.016 [0.009, 0.022]; positive in every seed). This control therefore does not reproduce the whole matched-fusion gain; unmeasured acquisition effects and lexical information in the speech signal remain possible explanations, and the larger fusion recovery is not evidence of complementary sentiment acoustics. Finally, recovery partly transfers across recognizers: a head trained on XLS-R transcripts raises Whisper-test macro-F1 from 0.635 to 0.815, against 0.850 under matched training (post hoc; Appendix A.5).

6.5 Acquisition audit

Speech-only reaches 0.940 macro-F1, which invites a prosodic reading—that expressive delivery carries the sentiment. That reading is not available on this corpus. The corpus is label-blocked: sentences 1–500 are positive and 501–1000 negative. Using the filesystem creation timestamps retained in the delivered files as a proxy for acquisition order, speakers m1, f2, m2 and f1 show only 1, 2, 9 and 11 label switches, respectively, versus roughly 500 per speaker under random interleaving (Appendix Table 15). These timestamps are not authoritative acquisition logs, but their pattern is consistent with class-blocked recording.

A logistic regression given **only** the acquisition and nuisance features listed in Table 5 reaches 0.669 macro-F1. In a secondary whole-pipeline permutation test that shuffles labels between sentence groups and regenerates the folds (1,000 permutations; Appendix A.9), the null scores average 0.495 (95% of them between 0.476 and 0.512) and none reaches the observed score

		Gold-trained head			Matched head		Share of gap recovered (%)	
Setting	Oracle	Default	ASR thr.	Gold thr.	Default	ASR thr.	Matched head	Threshold only
Whisper text	0.976	0.635 (18.7)	0.784 (51.9)	0.635	0.850	0.845	63.1 [58.6, 67.6]	43.6 [37.9, 48.9]
XLS-R text	0.976	0.867 (38.9)	0.898 (47.7)	0.870	0.930	0.929	57.4 [47.2, 67.4]	28.2 [18.1, 37.8]
Whisper fusion	0.981	0.677 (21.1)	0.827 (48.3)	0.685	0.924	0.924	81.3 [77.8, 85.0]	49.4 [44.4, 54.3]
XLS-R fusion	0.981	0.895 (41.0)	0.921 (47.8)	0.897	0.957	0.952	71.6 [62.7, 80.5]	30.5 [20.7, 39.7]

Table 4: Decision rule versus retraining (RQ4). Macro-F1 for gold-trained heads and ASR-matched heads (A1–A4), with default or per-fold validation-selected thresholds. Parentheses: predicted-positive rate (%). Recovery shares $(A-B)/(O-B)$ (retraining) and $(T-B)/(O-B)$ (threshold only) are computed within each bootstrap resample; brackets show 95% intervals.

Model	Shared speakers	Held-out speaker and sentence
Speech-only head (C1)	0.940	0.621
Acquisition + nuisance LR	**0.669**	0.333
metadata only	0.521	–
extended exploratory	0.699	–

Table 5: Acquisition audit (pooled out-of-fold macro-F1). The primary logistic regression (LR) combines metadata with encoding- and signal-level features (Appendix A.6); the exploratory variant adds duration, RMS, and spectral centroid. Group-permutation null for the primary LR: mean 0.495, $p = 0.001$.

($p = 0.001$, the smallest attainable). This supports predictability relative to that null only; it neither removes session effects nor licenses inference to independently recorded sessions. Metadata alone reaches only 0.521, whereas adding the encoding- and signal-level features raises performance to 0.669. This is the speech analogue of the artifact literature discussed in Section 2.

Our folds share speakers between training and test. In a secondary joint speaker-and-sentence holdout, each recording is predicted by models trained on the other three speakers and on other sentence groups (20 partitions; Appendix A.11). The speech-only head then falls from 0.940 to 0.621 (0.555–0.686 per speaker), and the acquisition/nuisance classifier from 0.669 to 0.333: the feature–label mapping learned from three speakers inverts or collapses on the fourth. The high speech-only score therefore does not carry over to an unseen speaker reading unseen sentences under this protocol, which also trains on fewer speakers. With four speakers this is descriptive, as are the per-speaker results in Appendix A.13, and neither drop by itself identifies an acquisition mechanism.

We report this as an evaluation lesson rather than a criticism of the resource: sequential collection is practical, but class-blocked collection can make nuisance conditions label-predictive. The caveat applies to every speech-input condition, including fusion (Section 6.2), and reaches the ASR-text conditions, because transcripts are produced from the audio; Appendix Table 12 shows no aggregate class-level WER pattern that explains the class asymmetry, but finer-grained effects remain possible. The controlled substitutions avoid a refitting confound; the adaptation and cross-recognizer comparisons are system-level evidence.

7 Discussion

The recognizer ranking by WER is preserved downstream, but not proportionally: WER differs by 2.5× whereas the text-only deployment gap differs by 3.1×, and within each recognizer accuracy declines as WER rises, which supports reporting downstream impact alongside WER (Culnan et al., 2021; Bhogale et al., 2023). Threshold adjustment alone recovers a substantial part of the improvement obtained with matched-head retraining. For this reason, we recommend reporting a threshold-only baseline alongside retraining results. For low-resource benchmark design, we recommend reporting oracle and deployable transcript conditions with controlled substitution rather than separately fitted per-condition models, and testing acquisition predictability, nuisance-feature fusion and speaker transfer before interpreting audio. When collection is sequential, interleaving classes or releasing session structure lets such effects be audited.

8 Conclusion

On BanglaMUSE, the oracle-to-ASR macro-F1 gap is 0.109 for Bengali XLS-R and 0.341 for the high-error Whisper configuration. Matched retraining recovers 57–81% of the gap, re-selecting only the decision threshold already recovers 28–49%, and a residual oracle advantage remains. Speech fusion performs better than ASR text alone, but

a control using only acquisition/nuisance features also improves performance. This matters because the corpus is label-blocked, and its filesystem timestamps are consistent with class-blocked recording. The nuisance features predict sentiment well above the permutation null, while the speech-only score falls from 0.940 to 0.621 when both speaker and sentence are unseen. Together, these results make it difficult to interpret the speech-input performance as evidence of prosodic sentiment cues. Low-resource speech–text evaluations should report oracle and deployable conditions, separate threshold effects from retraining, and audit acquisition variables before interpreting audio.

Limitations

The corpus has four speakers, uses read rather than spontaneous speech, and carries binary sentence-level labels. All conclusions apply to BanglaMUSE, these two recognizer configurations, these models and this protocol, and should not be read as claims about Bangla speech sentiment in general. The main protocol shares speakers between training and test folds, and the speaker-and-sentence holdout has only four speakers.

Every speech-input condition is confounded with acquisition conditions, and ASR output inherits recording conditions from the audio. The measured acquisition/nuisance feature vector is incomplete and not guaranteed to be content-free, so the nuisance-fusion control tests one acquisition-based explanation rather than identifying what the speech encoder uses. The permutation null does not provide inference to independently recorded sessions, and the timestamp reconstruction is a proxy rather than a verified acquisition log.

A changed threshold identifies neither label shift nor calibration, and it is selected on the validation split already used for early stopping; the split into threshold, retraining and residual parts is operational, not causal. Bootstrap intervals are conditional on the fitted heads, splits and thresholds and exclude training-seed variation, which we report separately. Neither the lexical baseline nor the high-similarity check establishes independence from paraphrase families. Neither recognizer is at its ceiling (XLS-R runs without a language model, Whisper uses beam 2, and decoding was chosen on a sample of this corpus). The speech encoder shares a backbone with ASR-B, punctuation removal discards a potentially informative cue from all conditions, and the secondary analyses were specified after the primary results were known.

Ethical Considerations

We use the public BanglaMUSE data release, licensed CC BY 4.0 (Nayeem et al., 2025), and collect no new personal data. Speakers are identified only by pseudonymous codes and we make no attempt at identification or at inferring speaker attributes. Results describe only the four speakers in this corpus; they must not be extrapolated to dialectal, spontaneous, conversational or code-switched Bangla, nor used to support claims about groups of speakers, and the per-speaker results are descriptive only, since speaker is confounded with recording condition. The acquisition audit reports a property of the released resource, not a judgement of the people who built it: class-blocked sequential collection is a practical choice for low-resource corpus building, and our recommendation is a methodological check, not a standard that existing resources have failed.

All experiments ran on a single consumer GPU; the dominant costs were approximately 1.3 hours of Whisper inference and 13 minutes for Bengali XLS-R. Code, the split definitions with their leakage assertions, per-recording out-of-fold predictions and the secondary-analysis scripts are released under open licenses (MIT for code, CC BY 4.0 for data) at `https://github.com/raihansaa/one-head-three-transcripts`.

Acknowledgments

Generative AI tools were used to assist with drafting and language refinement throughout the manuscript, as well as structural feedback, literature search, suggestions for candidate analyses, and implementation of analysis code. The authors reviewed and revised all AI-assisted material, independently verified cited sources and reported results, and take full responsibility for the work.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Kaushal 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 *Interspeech 2023*, pages 4384–4388.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

John Culnan, Seongjin Park, Meghavarshini Krishnaswamy, and Rebecca Sharp. 2021. Me, myself, and ire: Effects of automatic transcription quality on emotion, sarcasm, and personality detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 250–256, Online. Association for Computational Linguistics.

Rakesh Kumar Das, Nahidul Islam, Md. Rayhan Ahmed, Salekul Islam, Swakkhar Shatabda, and A. K. M. Muzahidul Islam. 2022. BanglaSER: A speech emotion recognition dataset for the Bangla language. *Data in Brief*, 42:108091.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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

YeonJoon Jung, Jaeseong Lee, Seungtaek Choi, Dohyeon Lee, Minsoo Kim, and Seung-won Hwang. 2024. Interventional speech noise injection for ASR generalizable spoken language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20642–20655, Miami, Florida, USA. Association for Computational Linguistics.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.

Yuanchao Li, Peter Bell, and Catherine Lai. 2024. Speech emotion recognition with ASR transcripts: A comprehensive study on word error rate and fusion techniques. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 518–525. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Reza Lotfian and Carlos Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Md Darun Nayeem, Zarin Rafa, Tasnuva Tasnim Nova, Yasin Rahman, Abdul Mumeet Pathan, and Md Masudul Islam. 2026. BanglaMUSE: A multimodal Bangla sentiment dataset of text–audio pairs for speech and sentiment analysis. *Data in Brief*, 65:112458.

Md Darun Nayeem, Zarin Rafa, Tasnuva Tasnim Nova, Yasin Rahman, Abdul Mumeet Pathan, and Nusrat Sultana. 2025. A multimodal Bangla text–audio dataset for sentiment analysis. Mendeley Data. Version 1; CC BY 4.0.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng 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.

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Imran Sheikh, Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu. 2018. Sentiment analysis using imperfect views from spoken language and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 35–39, Melbourne, Australia. Association for Computational Linguistics.

Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, and Yannick Estève. 2018. Simulating ASR errors for training SLU systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sadia Sultana, M. Shahidur Rahman, M. Reza Selim, and M. Zafar Iqbal. 2021. SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLOS ONE*, 16(4):e0250173.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528.

Yang Wu, Yanyan Zhao, Hao Yang, Song Chen, Bing Qin, Xiaohuan Cao, and Wenting Zhao. 2022. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1397–1406, Dublin, Ireland. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

A Supporting Analyses

A.1 Corpus preparation

The delivered archive contained 4,735 audio files. Of these, 737 were byte-identical extra copies of base recordings and were dropped after verifying that their hashes matched. Identifiers appeared in three inconsistent forms—zero-padded, unpadded, and one file with a five-digit identifier—and were parsed as integers and re-padded, which collapsed the variants automatically. 244 recordings were delivered as 16 kHz Ogg/Opus audio, 242 of them carrying `.mp3` filenames; we decode by content rather than by extension. The same audio payload appears under two different sentence identifiers whose texts differ by one word, making the reference assignment ambiguous; both recordings were excluded rather than forcing an arbitrary choice. Two further nominal recordings are absent from the delivery. Thus the 4,000 nominal recording slots comprise 3,996 usable recordings, the two ambiguous recordings and two missing recordings. Two sentence pairs have identical token multisets and differ only in word order; placing such near-identical lexical content in different folds would create an unusually strong overlap shortcut, so each pair is merged into a shared group by this objective rule, declared in advance. The 998 groups result from these merges; recording exclusions do not determine the group count. No label conflicts were found across the four recordings of any sentence.

A.2 Reproducibility

Outer folds come from `StratifiedKFold` (5 folds, shuffled, seed 42) over the 998 sentence groups, stratified on the label; each outer-training partition holds out a label-stratified 10% of its groups for inner validation (seed 1000 + fold). The heads use seeds 13, 42 and 87 and the bootstrap seed 20260807. The neural-head architecture and optimizer settings in Section 5 were fixed before the primary experiments and were not tuned; validation macro-F1 was used only to select the checkpoint epoch. Gold-trained heads are early-stopped on gold validation transcripts, matched heads on the matching recognizer's validation transcripts, and the threshold-only conditions select their cutoff on the gold-trained head's seed-averaged probabilities for the ASR validation transcripts. Software: Python 3.12, PyTorch 2.11 (CUDA 12.8), `transformers` 5.14.1 and scikit-learn 1.9.0 on one consumer GPU (RTX 5060 Laptop). Checkpoint revisions: whisper-large-v3 `06f233fe`, wav2vec2-xls-r-300m-bengali `45ed7c70`, banglabert `9ce791f3`, wav2vec2-xls-r-300m `1a640f32`. On this machine, re-training every primary head reproduces all 155,844 stored test probabilities bit for bit. The code release includes the secondary-analysis specification, fixed before those analyses were run, and a manifest of all settings and result hashes.

A.3 All conditions

Table 6 lists every condition with its accuracy, class-wise F1 and predicted-positive rate.

A.4 High-similarity exclusion

Table 7 reports the high-similarity exclusion check discussed in Section 6.2: the four controlled gaps are essentially unchanged and every paired interval still excludes zero.

A.5 ASR-trained and cross-recognizer conditions

Table 8 reports the ASR-matched conditions A1–A4 and the post-hoc cross-recognizer conditions X1–X2. Training on Bengali XLS-R transcripts raises Whisper-test macro-F1 from 0.635 to 0.815 (X2−C3 = 0.180 [0.152, 0.207]) against 0.850 under matched training; training on Whisper transcripts raises Bengali XLS-R-test performance from 0.867 to 0.909 (X1−C4 = 0.042 [0.026, 0.058]) against 0.930. Matching still adds a significant increment (A1−X2 = 0.035 [0.020, 0.051]; A2−X1 = 0.021 [0.011, 0.031]), but a smaller one. Substantial transfer shows that the adaptation benefit is not wholly recognizer-specific for these two systems. These are comparisons between separately trained heads, not controlled substitutions, and two recognizers are not enough to infer transfer to arbitrary ASR systems.

A.6 Nuisance-fusion control

The acquisition/nuisance feature vector is the primary Table 5 feature set: eight numeric features (bytes per second, bitrate, clipping fraction, DC offset, and the RMS and 10th-percentile amplitude of the leading and trailing 0.15 s) and one-hot codec, sample rate, channel count and container format, 17 dimensions in total. Median imputation, one-hot categories and standardization are fitted on each fold's training partition and frozen for validation, test and every transcript substitution; the standardized vector is then L2-normalized per row and concatenated with the text embedding in place of the speech embedding. No fold had missing values, zero-variance training columns, unseen categories or zero-norm rows. The resulting head has 785 inputs and 203,300 parameters, against 2,816 inputs and 727,298 parameters for speech fusion, so the two are not architecture-identical. As a pipeline check, the Table 5 logistic regression on this normalized vector reaches 0.653, close to its original 0.669. The edge windows are not guaranteed to be silent: the leading window lies a median 69 dB below the utterance's RMS, but the trailing window lies within 20 dB of it for 47% of recordings and within 10 dB for 22%. Per seed, C6−N6 is −0.009, 0.009 and −0.036 and C7−N7 is −0.002, 0.019 and −0.004, whereas A3−NA3 (0.033, 0.034, 0.040) and A4−NA4 (0.014, 0.014, 0.012) are positive in every seed.

A.7 Threshold-only adaptation

For each outer fold, the three seeds' inner-validation probabilities are averaged and the threshold t on the grid $\{0.000, 0.001, \ldots, 1.000\}$ that maximizes validation macro-F1 under the rule $p \geq t$ is chosen, breaking ties by the smallest $|t - 0.5|$ and then the smaller t. The threshold is applied to that fold's seed-averaged test probabilities, which are exactly those of the corresponding default condition. The selected thresholds range from 0.010 to 0.099 for T3, 0.119 to 0.268 for T4, 0.002 to 0.042 for T6 and 0.039 to 0.177 for T7; the gold-validation thresholds range from 0.343 to 0.623 and the matched-head thresholds from 0.177 to 0.636. Table 9 gives every term of the operational decomposition with paired intervals.

A.8 Lexical baseline

The secondary lexical baseline uses a TF-IDF vectorizer with whitespace tokens, no lowercasing, word uni- and bigrams, min_df = 1, smoothed IDF and L2 normalization, fitted on the training partition only (about 4,200 features per fold). As for C2, training and validation rows are one per sentence group and test rows are recordings. Logistic regression (lbfgs) selects $C \in \{0.1, 1, 10\}$ on inner-validation macro-F1 (ties to the smaller C); the selected values were 1, 10, 0.1, 10 and 10, and every fit converged. The baseline reaches 0.854 macro-F1 ($F1_-$ 0.866, $F1_+$ 0.841).

A.9 Permutation null

Each of 1,000 permutations shuffles the 998 sentence-group labels between groups, broadcasts them to the groups' recordings, regenerates the outer folds with the original stratified splitter and seed, and refits the unchanged Table 5 pipeline; the statistic is pooled out-of-fold macro-F1. Scikit-learn's `permutation_test_score` with sentence groups would shuffle labels only within groups, which randomizes nothing here. The same procedure on the true labels reproduces the original folds and gives 0.669. The null has mean 0.495,

ID	Condition	Macro-F1	Acc.	$F1_{-}$	$F1_{+}$	Pred. pos. (%)
C1	Speech only	0.940	0.940	0.940	0.940	50.2
C2	Gold text	0.976	0.976	0.976	0.976	50.0
C3	Gold-trained text, Whisper test	0.635	0.671	0.749	0.521	18.7
C4	Gold-trained text, XLS-R test	0.867	0.869	0.882	0.852	38.9
C5	Gold fusion	0.981	0.981	0.981	0.981	49.7
C6	Gold-trained fusion, Whisper test	0.677	0.704	0.770	0.583	21.1
C7	Gold-trained fusion, XLS-R test	0.895	0.896	0.904	0.885	41.0
A1	Whisper-trained text	0.850	0.850	0.853	0.847	48.0
A2	XLS-R-trained text	0.930	0.930	0.929	0.930	50.3
A3	Whisper-trained fusion	0.924	0.924	0.925	0.924	49.7
A4	XLS-R-trained fusion	0.957	0.957	0.957	0.957	49.9
X1	Whisper-trained text, XLS-R test (post hoc)	0.909	0.909	0.906	0.911	52.9
X2	XLS-R-trained text, Whisper test (post hoc)	0.815	0.816	0.801	0.828	57.3
N5	Gold-trained text+nuisance, gold test	0.980	0.980	0.980	0.980	50.1
N6	Gold-trained text+nuisance, Whisper test	0.692	0.714	0.774	0.609	23.3
N7	Gold-trained text+nuisance, XLS-R test	0.890	0.891	0.900	0.880	40.8
NA3	Whisper-trained text+nuisance	0.893	0.893	0.892	0.893	50.3
NA4	XLS-R-trained text+nuisance	0.941	0.941	0.941	0.942	50.7
T3	C3 head, ASR-validation threshold	0.784	0.784	0.780	0.788	51.9
T4	C4 head, ASR-validation threshold	0.898	0.898	0.900	0.895	47.7
T6	C6 head, ASR-validation threshold	0.827	0.827	0.830	0.824	48.3
T7	C7 head, ASR-validation threshold	0.921	0.921	0.923	0.919	47.8
	C3/C4/C6/C7 heads, gold-validation threshold	0.635 / 0.870 / 0.685 / 0.897				
	A1–A4 heads, ASR-validation threshold	0.845 / 0.929 / 0.924 / 0.952				
	TF-IDF + logistic regression, gold text	0.854	0.855	0.866	0.841	41.5

Table 6: All conditions (pooled recording-level out-of-fold scores). N, NA, T and the lexical baseline are secondary. Predicted-positive rates are in %; the true rate is 50.0%.

standard deviation 0.009 and maximum 0.524, so $p = (1 + 0)/(1{,}000 + 1) = 0.001$.

A.10 Seed variability

Table 10 scores each training seed separately, pooled over the five outer folds, next to the probability-averaged ensemble used elsewhere. The three seeds are not independent replications of the test set, and the bootstrap intervals in the paper do not include this variation.

A.11 Speaker-transfer diagnostic

For each speaker s and outer fold f, the test set is speaker s's recordings of fold f's test groups; training and inner validation use the other three speakers' recordings of fold f's training and inner-validation groups. The speech head uses the C1 recipe and seeds; the logistic regression, which has no validation step, is fitted on the other speakers' recordings of all of fold f's training groups. Every recording receives exactly one prediction (Table 11). Under this holdout the logistic regression predicts positive for 1.0% of f1's positive and 50.0% of its negative recordings, and for 23.7% and 79.0% of m2's, while it labels almost all of f2's and most of m1's recordings positive.

A.12 ASR quality by sentiment class

Table 12 reports ASR quality by sentiment class; Appendix A.14 uses it to assess the per-class asymmetry of the Whisper conditions.

A.13 Per-speaker descriptive results

Table 13 lists per-speaker recognition and downstream results. Speaker is confounded with recording condition in this corpus, so these numbers are descriptive only.

A.14 Post-hoc system-level observations

Both substituted fusion systems score below the speech-only system (C1−C6 = 0.263 [0.240, 0.286]; C1−C7 = 0.045 [0.031, 0.059]). These comparisons were not predeclared and, unlike C5→C6, contrast two *separately trained* systems, so they answer a deployment question—is this fusion system better than simply using speech?—rather than a controlled substitution question. Given Section 6.5, C1 is additionally affected by acquisition confounding, so we report the observation as directional and build no claim on it. There is no numerical contradiction with C6 > C3: the fusion system scores higher than the text-only system while still scoring below C1. As Section 6.2

Comparison	Full-set gap [95% CI]	Excluded-set gap [95% CI]	Change
C2−C3	0.341 [0.316, 0.365]	0.344 [0.318, 0.370]	+0.004
C2−C4	0.109 [0.095, 0.124]	0.109 [0.094, 0.125]	+0.001
C5−C6	0.304 [0.282, 0.328]	0.309 [0.285, 0.332]	+0.004
C5−C7	0.086 [0.074, 0.099]	0.087 [0.075, 0.101]	+0.001

Table 7: Post-hoc robustness check after excluding from evaluation all items with a token-Jaccard-similar neighbour (≥ 0.7). The same 10,000 sentence-grouped paired bootstrap is used; changes are computed from unrounded gaps.

ID	Train transcripts	Test transcripts	Branch	Macro-F1	Gain over gold-trained [95% CI]
A1	Whisper	Whisper	text	0.850	0.215 [0.192, 0.239]
A2	Bengali XLS-R	Bengali XLS-R	text	0.930	0.062 [0.048, 0.077]
A3	Whisper	Whisper	fusion	0.924	0.248 [0.225, 0.270]
A4	Bengali XLS-R	Bengali XLS-R	fusion	0.957	0.062 [0.050, 0.074]
X2	Bengali XLS-R	Whisper	text	0.815	0.180 [0.152, 0.207]
X1	Whisper	Bengali XLS-R	text	0.909	0.042 [0.026, 0.058]

Table 8: ASR-matched adaptation (A1–A4) and post-hoc cross-recognizer transfer (X1–X2). X1 and X2 reuse the A1 and A2 text heads on the other recognizer's transcripts. Gains use unrounded scores and may differ by 0.001 from subtraction of the displayed values.

	Whisper text	XLS-R text	Whisper fusion	XLS-R fusion
$T - B$ (threshold only)	0.149 [0.124, 0.174]	0.031 [0.019, 0.043]	0.151 [0.129, 0.172]	0.026 [0.017, 0.036]
$A - T$ (added by retraining)	0.067 [0.050, 0.083]	0.032 [0.020, 0.044]	0.097 [0.084, 0.111]	0.036 [0.026, 0.045]
$O - A$ (residual)	0.126 [0.109, 0.142]	0.046 [0.035, 0.058]	0.057 [0.045, 0.068]	0.025 [0.016, 0.033]
$A_{\text{thr}} - T$	0.061 [0.045, 0.078]	0.031 [0.019, 0.043]	0.097 [0.084, 0.111]	0.031 [0.022, 0.041]
Gold-val. threshold $- B$	0.000 [−0.005, 0.005]	0.003 [0.000, 0.007]	0.008 [0.005, 0.012]	0.002 [0.000, 0.004]
$(A - T)/(O - B)$ (%)	19.5 [14.7, 24.7]	29.3 [19.3, 39.5]	31.9 [27.5, 36.7]	41.2 [30.9, 52.1]

Table 9: Operational decomposition of each oracle gap $O - B$ (secondary). O: oracle; B: gold-trained head, default rule; T: the same head with an ASR-validation threshold; A: matched head, default rule; A_{thr}: matched head with its own threshold. Paired 95% intervals from the shared bootstrap resamples; terms are rounded separately.

shows, this system-level gain cannot be attributed to sentiment-bearing acoustics, because measured acquisition/nuisance features reproduce a similar improvement.

The Whisper conditions show a large per-class asymmetry (C3: F1 0.749 negative against 0.521 positive), which is substantially reduced by threshold-only adaptation (Section 6.4 and Table 6). Aggregate class-conditioned errors are mixed: positive utterances have lower WER than negative ones (0.676 versus 0.699) but higher CER (0.367 versus 0.352), with the WER advantage driven by fewer insertions (0.069 versus 0.103; Table 12).

A.15 Error-pattern diagnostic

Among the 1,725 recording–recognizer cases that are correct under oracle text and wrong under an ASR-substituted text condition, 1,267 involve Whisper and 458 involve Bengali XLS-R. A heuristic edit-pattern categorization flags 887 substitution-only cases, 561 with inserted content, 227 with deleted content and 130 with residual Latin-script material, whereas detected negation changes are rare (6 added and 3 lost). These counts are descriptive: the flags are neither mutually exclusive nor exhaustive, so they need not sum to 1,725, and they are not causal explanations. They nevertheless indicate that the observed failures are associated much more often with broad lexical corruption than with a negation-specific polarity reversal.

A.16 Whisper decoding diagnostic and settings

Table 14 reports the decoding diagnostic behind the Whisper configuration choices described in Section 5, with the secondary beam-2 row. The sample is re-drawn with the diagnostic's own code and seed (75 utterances per speaker); its stored example outputs match. The other three rows used a 128-token budget: in cached greedy transcripts of the full corpus, 4 of the 300 sampled outputs come within one token of that budget, so the Bengali-forced greedy row may include a few truncated outputs, and 5

	Seeds 13 / 42 / 87	SD	Ens.
C1	0.936 / 0.934 / 0.937	0.002	0.940
C2	0.966 / 0.980 / 0.972	0.007	0.976
C3	0.654 / 0.627 / 0.628	0.016	0.635
C4	0.872 / 0.869 / 0.861	0.006	0.867
C5	0.980 / 0.983 / 0.979	0.002	0.981
C6	0.682 / 0.680 / 0.673	0.005	0.677
C7	0.892 / 0.898 / 0.894	0.003	0.895
A1	0.850 / 0.848 / 0.847	0.001	0.850
A2	0.926 / 0.932 / 0.919	0.006	0.930
A3	0.927 / 0.922 / 0.919	0.004	0.924
A4	0.953 / 0.953 / 0.950	0.002	0.957
N5	0.976 / 0.977 / 0.978	0.001	0.980
N6	0.691 / 0.671 / 0.709	0.019	0.692
N7	0.894 / 0.879 / 0.899	0.010	0.890
NA3	0.894 / 0.887 / 0.879	0.008	0.893
NA4	0.939 / 0.940 / 0.938	0.001	0.941
C2−C3	0.311 / 0.353 / 0.344	0.022	0.341
C2−C4	0.094 / 0.111 / 0.111	0.010	0.109
C5−C6	0.298 / 0.302 / 0.307	0.004	0.304
C5−C7	0.088 / 0.085 / 0.085	0.002	0.086
A1−C3	0.195 / 0.221 / 0.220	0.015	0.215
A2−C4	0.053 / 0.063 / 0.058	0.005	0.062
A3−C6	0.244 / 0.241 / 0.246	0.002	0.248
A4−C7	0.060 / 0.055 / 0.056	0.003	0.062

Table 10: Training-seed variability (secondary): pooled out-of-fold macro-F1 per seed, its sample SD, and the ensemble score. Differences are computed within seed.

Model	Protocol	f1	f2	m1	m2	Pooled
Speech head	joint holdout	0.609	0.686	0.611	0.555	0.621
Speech head	shared CV	0.881	0.918	0.972	0.988	0.940
Acq./nuisance LR	joint holdout	0.207	0.338	0.323	0.223	0.333
Acq./nuisance LR	shared CV	0.711	0.583	0.352	0.922	0.669

Table 11: Macro-F1 per held-out speaker under the joint speaker-and-sentence holdout (secondary), next to the same models under the original shared-speaker cross-validation, which is not a leave-one-speaker-out result. For the shared-CV speech head, macro-F1 agrees with the accuracy in Table 13 to three decimals: each speaker's classes are balanced and its errors are few or nearly symmetric.

of the 300 beam-2 outputs are at least 127 tokens long.

The main run used `transformers` 5.14.1 and the model class `WhisperForConditionalGeneration` for whisper-large-v3 (revision `06f233fe`), calling `generate` with task `transcribe` and language `bn` (forced decoder ids cleared), beam 2, early stopping, 200 new tokens, no sampling, no prompt, no timestamps, repetition and length penalty 1.0 and no n-gram blocking (library defaults), float16, batch size 8, and 16 kHz mono audio padded to 30 s log-Mel features with an attention mask. Because every clip is at most 8.31 s, generation takes the short-form path; Whisper's long-form options `condition_on_prev_tokens`, `compression_ratio_threshold`, `logprob_threshold` and `no_speech_threshold` were not set and would be inactive. No hypothesis reaches the 200-token cap (the longest is about 172 tokens), no audio file failed to load, and no output is empty. Across the full corpus, 9.8% of Whisper hypotheses contain Latin-script characters, including 11 of the 20 with the highest WER. Only 2 of 3,996 show an immediate n-gram loop (an n-gram with $n \leq 4$ repeated at least three times in a row) or a zlib compression ratio above 2.4, and neither contains any of the insertions.

A.17 Acquisition-order diagnostic

Table 15 reports the filesystem-timestamp reconstruction of acquisition order that supports the audit in Section 6.5.

A.18 Orthographic canonicalization check

Table 16 reports the post-hoc orthographic canonicalization check. A canonicalization scheme differing more from our base normalizer might yield a different answer.

ASR system	Class	n	WER	CER	Sub.	Del.	Ins.
Bengali XLS-R CTC	negative	1999	0.262	0.109	0.219	0.026	0.017
Bengali XLS-R CTC	positive	1997	0.290	0.114	0.241	0.031	0.018
Whisper large-v3	negative	1999	0.699	0.352	0.586	0.011	0.103
Whisper large-v3	positive	1997	0.676	0.367	0.595	0.012	0.069

Table 12: ASR quality by sentiment class. For Whisper, aggregate class-level WER runs opposite to the downstream F1 asymmetry, arguing against a simple class-dependent error-rate explanation (Appendix A.14).

Spk	n	WER-W	WER-B	C1	C2	C3	C4	C5	C6	C7
f1	1000	0.713	0.312	0.881	0.976	0.662	0.856	0.975	0.672	0.863
f2	1000	0.668	0.226	0.918	0.976	0.681	0.904	0.978	0.698	0.911
m1	998	0.691	0.267	0.972	0.976	0.675	0.872	0.979	0.687	0.893
m2	998	0.679	0.299	0.988	0.976	0.665	0.844	0.993	0.758	0.916

Table 13: Per-speaker results. The C1–C7 columns report accuracy. Speaker is confounded with recording condition in this corpus, so these results are descriptive only.

Configuration	Budget	WER	CER	Latin-script output	Empty
Greedy, Bengali forced	128	0.744	0.467	31.0%	0.0%
Beam 5, Bengali forced	128	0.679	0.352	2.3%	0.0%
Greedy, autodetect	128	0.750	0.481	31.0%	0.0%
Beam 2, Bengali forced†	200	0.692	0.368	9.0%	0.0%

Table 14: Whisper decoding diagnostic on a 300-utterance speaker-stratified sample of this corpus. †Secondary row: it rescores the main run's transcripts, which every downstream condition uses, on the same utterances with the same normalizer and scoring. Budget is the maximum number of new tokens. WERs are measured on the sample and are not directly comparable to the full-corpus 0.687 in Table 1.

Speaker	Switches	r(order, label)
m1	1	-0.866
f2	2	-0.863
m2	9	-0.784
f1	11	$+0.476$

Table 15: Filesystem creation-timestamp acquisition-order diagnostic. Creation timestamps are used only as a proxy for acquisition order; switch counts and correlations are computed separately for each speaker.

System	Orig. gap	Canon. gap	Changed
Whisper	0.341	0.341	4/3996
XLS-R	0.109	0.109	0/3996

Table 16: Post-hoc orthographic canonicalization check. The externally specified BanglaBERT-oriented canonicalizer changes almost no hypotheses after the frozen base normalization, so this null result has narrow scope.