A Foundational Study of Multimodal Speech–Text Alignment in Malayalam

Meharuniza Nazeem* **Navaneeth S** **Adithyakiran A S** **Rajeev R R** **Ashbi Biju**
International Centre for Free and Open Source Solutions (ICFOSS)
Kerala, India

Abstract

Speech–text retrieval is a multimodal alignment task in which spoken audio and written transcripts are mapped into a shared representation space. Standard retrieval metrics test whether an audio embedding recovers its transcript from a large candidate pool, but not whether the shared representation preserves small grammatical contrasts inside agglutinated words. We construct an automatic Malayalam benchmark with 12,611 controlled audio–text comparison rows derived from 6,611 held-out audio queries. The benchmark contains three morphosyntactic categories—case, tense, and negation—and a lexical numeral-control category. Alternatives are generated with the `mlmorph` finite-state analyser and generator and retained after a generate–reanalyse consistency check. A CLIP-style model without explicit hard negatives obtains 37.45% full-pool Recall@1 but only 53.66% pairwise accuracy, close to the 50% pairwise baseline. We diagnose a fixed-first-candidate loader defect that under-delivers later-generated categories and replace it with per-transcript category-aware, access-time resampling. At the matched epoch count, pairwise accuracy rises from 73.50% to 79.76%, and tense accuracy rises from 52.83% to 77.70%. The complete corrected procedure reaches 89.69% pairwise accuracy after extended training, while full-pool retrieval remains below the original checkpoint. A stratified human review of 797 rows retains or corrects 731 (91.7%); the final checkpoint obtains 74.01% on this deliberately reweighted and partly rewritten subset. These results show that global retrieval, controlled grammatical discrimination, candidate-delivery logic, and human benchmark validation are distinct components of multimodal evaluation.

*Corresponding author: `meharuniza@icfoss.org`

1 Introduction

Speech–text retrieval is a multimodal learning problem connecting spoken audio with written language. A speech encoder and a text encoder map the two modalities into a shared embedding space, enabling spoken-archive search, recording–caption alignment, and audio–language transfer. Standard evaluation asks whether the correct transcript appears among the highest-scoring candidates. This measures global retrieval, but it can conceal a language-specific failure: a model may identify the topic and most lexical content while ignoring a short segment that changes the utterance's grammatical interpretation.

Malayalam is a morphologically rich Dravidian language in which case, tense, polarity, number, and auxiliary structure can be expressed within an orthographic word. For example, പോയി *poyi* 'went' and പോയില്ല *poyilla* 'did not go' share most of their surface and acoustic material, yet differ in polarity. Similarly, സ്കൂളിൽ *skuulil* 'at school' and സ്കൂളിലേക്ക് *skuulilekkŭ* 'to school' differ in case-related material. A model that scores both alternatives almost equally may still achieve high Recall@K if all other candidates are unrelated.

The distinction is expressed differently in the two modalities. In written text, case, tense, and negation are represented by explicit orthographic material, whereas in speech the corresponding distinction is realized acoustically rather than as a symbolic morphological label. The controlled evaluation therefore asks whether acoustic evidence from the fixed utterance is aligned with the appropriate written morphological form in the shared representation space.

In this multimodal setting, dual encoders and contrastive objectives commonly map audio and text into a shared space. This paradigm builds on contrastive predictive coding and CLIP-style symmetric training (van den Oord et al., 2018; Radford

Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages, pages 18–27
November 9, 2026 ©2026 Association for Computational Linguistics

et al., 2021), and has been adapted to speech–text retrieval, speech translation, and audio–language representation learning (Khurana et al., 2022; Shih et al., 2023; Ye et al., 2022; Elizalde et al., 2023). Pretrained multilingual speech and text models make such training feasible for Indian languages (Babu et al., 2022; Khanuja et al., 2021). However, ordinary in-batch negatives tend to be far from the correct transcript. They teach coarse discrimination without necessarily requiring the model to resolve a suffix-sized contrast.

This paper examines evaluation, data delivery, and benchmark validity. We first ask whether full-pool Recall@K is a reliable proxy for fine-grained sensitivity to Malayalam grammatical contrasts. We then inspect the path through which generated hard negatives reach the loss. Although the source file contained multiple categories, the original loader attached only the first generated candidate to each transcript. Because candidate generation scans from left to right and some triggers occur more frequently, this positional rule systematically altered the effective training distribution. Finally, because finite-state round-trip consistency does not guarantee sentence-level naturalness, we perform a stratified human review and re-evaluate all supplied checkpoints on the accepted or corrected subset.

We position this work as a foundational study in the sense that it establishes an initial evaluation framework and empirical reference point for Malayalam multimodal speech–text alignment, rather than claiming a comprehensive treatment of architectures or linguistic phenomena.

Our contributions are:

1. We construct a 12,611-row automatic multimodal diagnostic over 6,611 held-out audio queries, covering case, tense, negation, and a lexical numeral control.
2. We demonstrate a divergence between full-pool retrieval and controlled discrimination: the no-hard-negative model reaches 37.45% full-pool R@1 but only 53.66% pairwise accuracy.
3. We diagnose a fixed-first-candidate sampling defect and replace it with per-transcript category-aware retention and access-time resampling, without changing the loss architecture.
4. We separate matched-epoch evidence for the delivery correction from the extended-training result and report the unresolved gap between local discrimination and global retrieval.
5. We human-review a stratified 797-row sample, retain or correct 731 rows, and report checkpoint performance on this independently checked subset.

2 Related Work

2.1 Multimodal speech–text representation learning

Contrastive predictive coding learns representations by distinguishing observed future structure from sampled alternatives (van den Oord et al., 2018). CLIP scales a symmetric dual-encoder objective to image-text pairs and establishes the shared-embedding template used by many later multimodal systems (Radford et al., 2021). Wav2CLIP and SpeechCLIP transfer this idea to audio or speech by connecting pretrained audio features with language-aligned spaces (Wu et al., 2022; Shih et al., 2023). CLAP learns a general audio-text space from natural-language supervision (Elizalde et al., 2023). For multilingual speech, SAMU-XLSR aligns utterance-level speech representations with sentence embeddings (Khurana et al., 2022), while ConST adds cross-modal contrastive learning to speech translation (Ye et al., 2022). ATRI focuses on inconsistencies caused by multilingual audio-text data distributions (Yin et al., 2025). Our study differs by targeting controlled, within-language grammatical contrasts rather than broad semantic compatibility.

The backbones used here reflect two relevant lines of pretrained modelling. Wav2vec 2.0 and XLS-R learn contextual speech representations from unlabeled audio, with XLS-R extending the approach across languages at scale (Baevski et al., 2020; Babu et al., 2022). MuRIL provides text representations tailored to Indian languages and transliterated text (Khanuja et al., 2021). Benchmarks and datasets such as Common Voice, FLEURS, and IndicSUPERB illustrate the growing infrastructure for multilingual and Indian-language speech evaluation (Ardila et al., 2020; Conneau et al., 2023; Javed et al., 2023); our contribution is a targeted diagnostic rather than a new general speech corpus.

2.2 Negative sampling and dense retrieval

Dual encoders depend on the negative distribution. Dense Passage Retrieval shows the effectiveness of in-batch and retrieved negatives for scalable ranking (Karpukhin et al., 2020); RocketQA

further develops denoised and cross-batch negative strategies (Qu et al., 2021). Theoretical and empirical work shows that hard-negative choice changes the bias and optimization behaviour of contrastive objectives (Zhang and Stratos, 2021; Robinson et al., 2021). In our setting, the challenge precedes the scoring function: generated categories existed in storage, but the loader's positional choice prevented many from participating. This makes the sampler an experimentally consequential part of the model specification.

2.3 Targeted and minimal-pair evaluation

Aggregate test performance can hide narrow decision-boundary failures. Contrast sets perturb examples in controlled ways to expose such failures (Gardner et al., 2020); CheckList organizes behavioural tests around capabilities and perturbation templates (Ribeiro et al., 2020); Dynabench emphasizes iterative, failure-driven benchmarking (Kiela et al., 2021). BLiMP demonstrates how minimal pairs can isolate particular linguistic phenomena (Warstadt et al., 2020). Our benchmark imports this diagnostic logic into speech-text retrieval: the audio is held fixed, and the model must prefer the attested transcript over one controlled text alternative.

2.4 Morphological resources and finite-state control

Finite-state morphology provides reversible analysis and generation for morphologically complex languages (Beesley and Karttunen, 2003). For Malayalam, `mlmorph` implements a reversible finite-state analyser and generator covering inflectional, derivational, and phonological processes (Thottingal, 2019). Its reversibility supports the mutation protocol used here: analyse a word, alter a designated feature, generate a new surface form, and reanalyse it. This round trip checks consistency with the finite-state system; it is not equivalent to human validation of sentence naturalness.

3 Multimodal Model and Implementation

3.1 Cross-modal retrieval and controlled discrimination

Let $a_i = f_a(x_i)$ be the normalized embedding of Malayalam waveform x_i, and let $t_j = f_t(y_j)$ be the normalized embedding of transcript y_j. Global retrieval ranks all candidate transcripts by cosine similarity $a_i^\top t_j$. We report Recall@1, Recall@5, and Recall@10 over the 7,953-transcript held-out pool. The targeted benchmark instead supplies a controlled alternative y_i^- for each true transcript y_i^+ and records whether

$$a_i^\top f_t(y_i^+) > a_i^\top f_t(y_i^-). \tag{1}$$

Pairwise accuracy is the proportion of rows satisfying this inequality; its random baseline is 50%. The full-pool and pairwise tasks use the same audio embeddings but test different properties of the shared multimodal space. The pairwise probe is inherently cross-modal: the audio is held fixed while the two text candidates differ only in a controlled linguistic feature. Correct discrimination therefore requires the audio embedding to preserve acoustic evidence sufficient to prefer the written form that corresponds to what was actually spoken.

3.2 Architecture and optimization

The model is a CLIP-style dual encoder. A `wav2vec2-xls-r-300m` branch consumes 16 kHz raw waveforms, and a shared `muril-base-cased` text branch encodes both true and controlled-alternative transcripts. Attention-masked mean pooling is used for text and temporal mean pooling for audio. Linear projection heads map both branches into a shared 256-dimensional space, followed by ℓ_2 normalization. No explicit spectrogram branch is used: the audio modality enters as waveform samples.

Training minimizes symmetric audio-to-text and text-to-audio InfoNCE, with a learnable temperature. When a row has an attached hard negative t_i^-, a hinge term is added:

$$\mathcal{L}_{\text{HN}} = \max\Big(0, 0.2 - (a_i^\top t_i^+ - a_i^\top t_i^-)\Big). \tag{2}$$

Table 1 records the supplied implementation settings. The temperature is clamped after each update to prevent bf16 overflow. Gradient clipping was added when the corrected run was extended beyond epoch 2; it is therefore part of the final procedure, not of the matched-epoch sampler comparison.

4 Data and Agglutination-Aware Benchmark

4.1 Corpus and held-out pool

The processing pipeline begins with 89,200 Malayalam audio–transcript pairs aggregated from multiple publicly available speech datasets identified

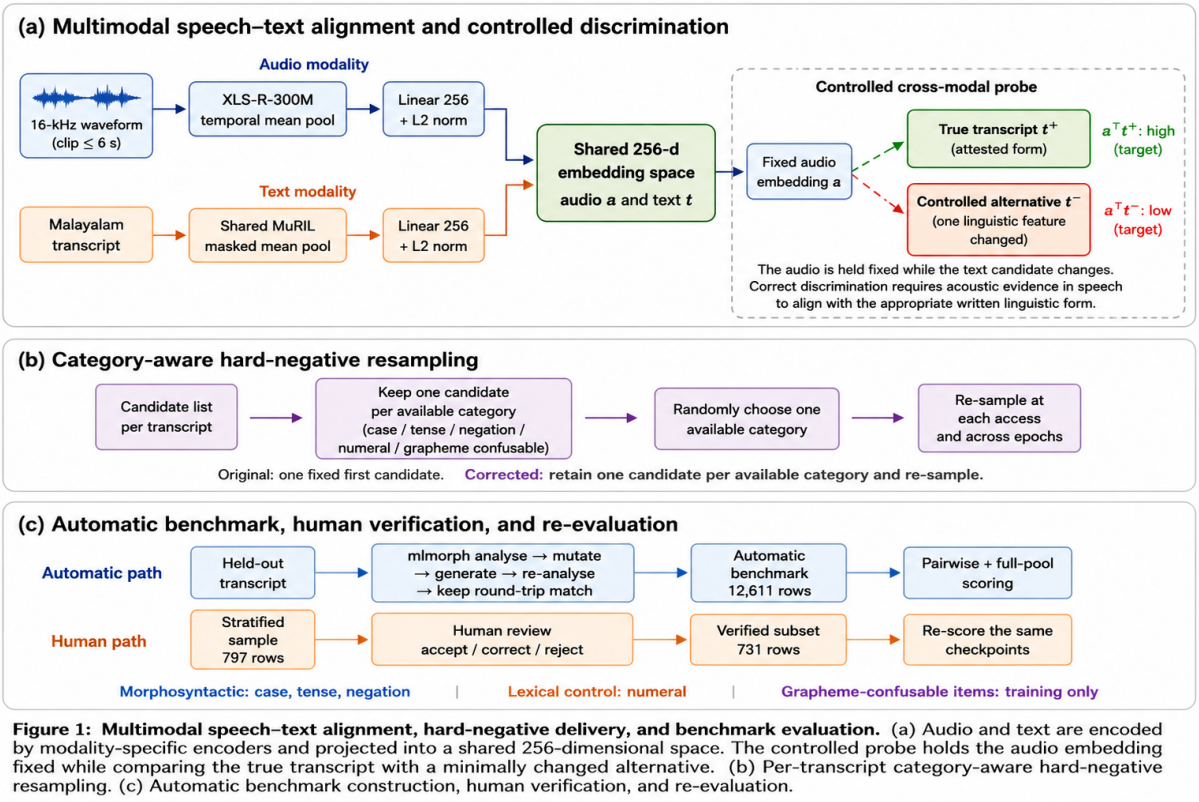


Figure 1: Multimodal speech–text alignment, hard-negative delivery, and benchmark evaluation. (a) Audio and text are encoded by modality-specific encoders and projected into a shared 256-dimensional space. The controlled probe holds the audio embedding fixed while comparing the true transcript with a minimally changed alternative. (b) Per-transcript category-aware hard-negative resampling. (c) Automatic benchmark construction, human verification, and re-evaluation.

Table 1: Implementation settings used in the reported experiments.

Component	Setting
Audio encoder	`wav2vec2-xls-r-300m`; temporal mean pool
Text encoder	`muril-base-cased`; masked mean pool; shared for positive and negative text
Shared space	Linear projection to 256 dimensions; ℓ_2 normalization
Loss	Symmetric InfoNCE + hard-negative hinge, margin 0.2
Optimization	AdamW, learning rate 10^{-4}, batch 32, bf16
Audio input	16 kHz waveform, maximum 6 seconds
Stabilization	Temperature clamped to $[0, \ln 100]$; gradient norm clipped at 1.0 in the extended corrected run
Memory	Gradient checkpointing on both encoders

during corpus construction. Because the corpus combines data from several sources, the original provenance and source-specific licensing conditions are retained for each constituent dataset. Deduplication and quality control yield 55,059 unique transcript records used for linguistic annotation and hard-negative generation. The supplied split contains a 5,225-record anchor subset and uses speaker-disjoint partitioning; all reported benchmark evaluations use a held-out pool of 7,953 audio–transcript pairs.

The linguistic-processing record contains 760,630 generated alternatives before benchmark filtering. These inventory counts describe stored candidates; they do not imply that every category reached the loss under the original loader, which is the defect studied in Section 5.

4.2 Controlled candidate generation

For each content word, `mlmorph` analyses the surface form, applies an eligible category-specific mutation, regenerates a replacement, and reanalyses it. A candidate is retained when the generated analysis matches the intended mutation. Case candidates cycle among configured case analyses; tense candidates flip configured past/present analyses; negation is added or removed where licensed by the analyser; and numeral candidates use a fixed replacement lexicon.

A fifth training category, called `phon_conf` in the implementation, substitutes configured grapheme-confusable sets such as ല/ള/ഴ and ന/ണ. Because this operation is orthographic and is not independently pronunciation-validated, we refer to it as a *grapheme-confusable* category. It con-

Table 2: Automatic benchmark composition and stratified human verification. "Retained" includes rows accepted unchanged or after reviewer correction.

Category	Auto. rows	Reviewed	Retained	Rate
Case	4,890	150	148	98.7
Tense	1,870	150	142	94.7
Negation	5,504	150	138	92.0
Numeral control	347	347	303	87.3
Total	12,611	797	731	91.7
Covered audio queries: 6,611 / 7,953 (83%)				

tributes to training but is excluded from the reported morphosyntactic benchmark. Round-trip validation establishes consistency with the finite-state analyser and generator; it does not by itself establish sentence naturalness or contextual appropriateness.

4.3 Automatic benchmark composition and metrics

Emitting one row per available query–category pair yields 12,611 automatic rows. A query can appear in more than one row, but each row contains one true transcript and one controlled alternative. Table 2 combines the automatic benchmark composition with the stratified human-verification outcome. Case, tense, and negation are morphosyntactic contrasts; numeral replacement is a lexical control. The automatic imbalance reflects trigger availability in held-out text rather than a balanced test design.

Pairwise accuracy is the primary diagnostic: for each row, the model must score the attested transcript above its controlled alternative. The random baseline is 50%. Full-pool Recall@K is computed for the same 6,611 covered audio queries against all 7,953 held-out transcripts. These metrics answer different questions and have different chance levels; their raw percentages should not be compared as if they shared a common scale.

4.4 Human-verification protocol

Finite-state consistency does not guarantee that a generated sentence is natural or that only the intended category changed. We therefore drew a stratified sample of 797 automatic rows: 150 each for case, tense, and negation, together with all 347 numeral rows. Four category-dedicated Malayalam reviewers assessed their assigned rows, and a lead reviewer independently checked the complete sample. Each pair was judged for sentence naturalness, agreement with the labelled change, and final disposition: accept unchanged, accept with a reviewer-written correction, or reject. The lead review supplied a second judgement for every item and the final accept/reject decision; reviewer correction text was retained where needed. As summarized in Table 2, of the 797 reviewed rows, 353 were accepted unchanged, 378 were retained after reviewer correction, and 66 were rejected; thus 731/797 rows (91.7%) were retained overall.

5 Per-Transcript Category-Aware Hard-Negative Resampling

5.1 Fixed-first-candidate defect

The hard-negative file stores multiple candidates in generation order, but the original loader attached only `negatives[0]` and reused that text for every epoch. Since generation scans words left to right and `phon_conf`/case triggers occur frequently, they commonly occupied position zero. Tense requires an eligible explicitly tensed verb and was under-delivered by the fixed-first selection rule. The held-out benchmark nevertheless contains 1,870 tense contrasts, allowing this failure mode to be evaluated. The defect changes the effective supervision distribution without changing the model or loss.

5.2 Corrected candidate-delivery procedure

We use a simple category-aware resampling correction to ensure that available contrast categories are not deterministically excluded by candidate order. We do not claim this procedure as a novel sampling algorithm; its role is to isolate the consequences of an experimentally consequential data-delivery defect. To correct the fixed-first-candidate defect, we reorganize the generated alternatives by transcript and category. For each transcript, at most one candidate is retained from every available category. Rather than attaching one fixed alternative when the dataset is loaded, the training dataset randomly selects one of the available category-specific candidates whenever the item is accessed. Consequently, a transcript with multiple available contrast types can expose different hard negatives across training epochs. The procedure does not globally equalize category frequencies because each transcript exposes only the categories licensed by its linguistic content; we therefore refer to it as per-transcript category-aware resampling rather than global category balancing.

The original and corrected epoch-2 checkpoints

provide the closest available comparison of this delivery change. The corrected epoch-6 checkpoint additionally includes four further epochs and gradient clipping, and is therefore treated as the complete corrected training procedure rather than a sampler-only ablation.

6 Experimental Setup

6.1 Systems and comparison logic

We evaluate four supplied checkpoints: (i) **Stage 1**, symmetric contrastive training without attached hard negatives; (ii) **Stage 2 original**, fixed first-candidate training stopped at epoch 2; (iii) **resampled epoch 2**, per-transcript category-aware resampling at the matched epoch count; and (iv) **resampled epoch 6**, the corrected run after four additional epochs, with gradient clipping during the extension.

Stage 1 reached 65.45% development R@1 after five epochs. The original Stage 2 development R@1 trajectory was 28.05%, 75.65%, and 82.20% at epochs 0–2. The corrected run developed more slowly, from 2.85% at epoch 0 to 41.35% at epoch 2 and 71.80% at epoch 6.

Three comparisons answer distinct questions. Stage 1 versus Stage 2 original describes the association between the supplied hard-negative stage and controlled discrimination. Original versus resampled epoch 2 is the closest available comparison of candidate delivery at a matched epoch count. Resampled epoch 2 versus epoch 6 measures progress under the corrected procedure. Original versus resampled epoch 6 is descriptive but cannot isolate sampling because duration and gradient clipping also differ. All reported checkpoints are from single training runs. Repeated-seed estimates were not available; therefore, the reported differences are descriptive and we do not report standard deviations, confidence intervals, or significance tests.

The 12,611-row automatic benchmark is the primary evaluation. The same checkpoints are then re-scored on the 731-row human-verified subset. Because the verification sample intentionally oversamples the 347-row numeral category and includes reviewer-corrected alternatives, its aggregate score is not directly comparable with the automatic benchmark's micro-average.

7 Results

7.1 Global retrieval and local discrimination diverge

The central result is visible without a separate table. Stage 1 retrieves the true transcript at rank 1 for 37.45% of benchmark queries in the 7,953-candidate pool, but scores the true transcript above its controlled alternative in only 53.66% of rows, just 3.66 points above the 50% pairwise baseline. Stage 2 original improves these values to 55.01% full-pool R@1 and 73.50% pairwise accuracy. Thus, non-trivial global retrieval can coexist with weak sensitivity to a controlled grammatical change; the fuller category-wise comparison appears in Table 3.

7.2 Matched-epoch effect of corrected delivery

At epoch 2, aggregate pairwise accuracy is 79.76% with the resampled procedure and 73.50% with the original. The most diagnostic change is tense, which rises from 52.83% to 77.70%. Negation rises from 77.25% to 90.44%. These are consistent with the purpose of exposing alternatives that were available in storage but systematically underdelivered by fixed-first selection.

The matched checkpoint is not uniformly superior. Case is 70.12% versus 77.16%, numeral is 57.06% versus 73.78%, and full-pool R@1 is 18.35% versus 55.01%. The convergence trajectories in Figure 2 indicate that the corrected negative distribution is associated with slower global-retrieval learning over the observed epochs. The epoch-2 result should therefore be interpreted as a redistribution toward controlled discrimination, not as a globally superior checkpoint.

7.3 Extended corrected training

At epoch 6, aggregate pairwise accuracy reaches 89.69%. Category scores are 83.78% for case, 86.52% for tense, 97.33% for negation, and 68.88% for the numeral control. Relative to the original checkpoint, the supplied differences are +6.62, +33.69, +20.08, and -4.90 points, respectively, with +16.19 points overall. Full-pool R@1/R@5/R@10 reaches 42.47/82.08/89.05%, below the original 55.01/90.06/94.01% checkpoint. Because this checkpoint additionally includes four epochs and gradient clipping, these values characterize the complete corrected training procedure rather than a sampler-only ablation.

Table 3: Automatic-benchmark pairwise accuracy and full-pool recall. All values are percentages. Δ is the supplied original-to-resampled-epoch-6 difference.

Metric	Orig.	Resp.-2	Resp.-6	Δ
Case	77.16	70.12	**83.78**	+6.62
Tense	52.83	77.70	**86.52**	+33.69
Negation	77.25	90.44	**97.33**	+20.08
Numeral control	**73.78**	57.06	68.88	−4.90
All pairwise	73.50	79.76	**89.69**	+16.19
Full-pool R@1	**55.01**	18.35	42.47	−12.54
Full-pool R@5	**90.06**	50.81	82.08	−7.98
Full-pool R@10	**94.01**	64.04	89.05	−4.96

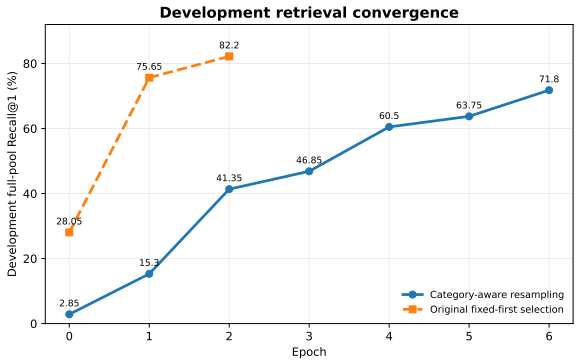


Figure 2: Development full-pool Recall@1 by epoch. These are observed trajectories from separate runs, not repeated-run uncertainty estimates.

Progress within the corrected run. From resampled epoch 2 to epoch 6, every supplied metric increases. Pairwise accuracy moves from 79.76% to 89.69%. Full-pool R@1/R@5/R@10 moves from 18.35/50.81/64.04% to 42.47/82.08/89.05%, corresponding to supplied gains of +24.12, +31.27, and +25.01 points. Pairwise accuracy gains 9.93 points. This shows that the severe epoch-2 global-retrieval deficit is at least partly associated with training progress. It does not establish that the corrected model would match or exceed the original after convergence, because no later checkpoint was supplied.

7.4 Human-verified evaluation

The verification outcome is summarized together with benchmark composition in Table 2. Across category reviewer–lead pairs, supplied raw agreement ranges from 82% to 95%, while Cohen's κ ranges from 0.11 to 0.50; exact per-category values were not supplied. The lower end of the κ range indicates weak chance-corrected agreement for at least one category; therefore, the 91.7% figure should be interpreted as a lead-adjudicated retention rate rather than evidence of uniformly strong annotator consensus. Most disagreements concern "accept unchanged" versus "accept with correction," rather than accept versus reject.

Table 4: Final checkpoint on the full automatic benchmark and the stratified human-verified subset. The columns are not directly comparable micro-averages because category mixture and text provenance differ.

Category	Automatic	Verified	Δ
Case	83.78	63.51	−20.27
Tense	86.52	81.69	−4.83
Negation	97.33	84.78	−12.55
Numeral control	68.88	70.63	+1.75
All pairwise	89.69	74.01	−15.68

Table 4 re-scores the final checkpoint on the 731 retained rows. The 74.01% aggregate should not be read as a paired 15.68-point regression from the 89.69% automatic micro-average: the verified sample includes every numeral row but only 150 rows from each larger category, and 378 alternatives were rewritten by reviewers. Reweighting the automatic category scores to the verified sample's category proportions gives 80.69%. On the 353 rows accepted unchanged, observed accuracy is 75.35%, compared with a supplied category-mix-adjusted expectation of 74.98% (a +0.37-point residual). This supports the narrower conclusion that automatically generated rows confirmed by humans behave consistently with the full automatic benchmark, while reviewer corrections produce a different and generally harder diagnostic subset.

8 Analysis and Discussion

8.1 Global retrieval is not a morphology test

Full-pool retrieval rewards separation from thousands of candidates that often differ in topic, length, and lexical content. A model can exploit those broad cues without assigning a robust margin to a suffix-level contrast. Stage 1 combines 37.45% full-pool R@1 with 53.66% pairwise accuracy, demonstrating that the capabilities separate in this setting. Recall@K remains operationally necessary, but it should be accompanied by targeted diagnostics when a system is expected to preserve fine-grained grammar, as in contrast sets and minimal-pair suites for text NLP (Gardner et al., 2020; Ribeiro et al., 2020; Warstadt et al., 2020).

8.2 Candidate delivery is part of the method

The source file contained category diversity, but fixed position-zero selection converted it into a narrower effective distribution. At the matched epoch, tense rises from 52.83% to 77.70% under corrected

delivery, consistent with the diagnosis that later-triggering candidates were under-delivered. The corrected procedure does not equalize global category counts; it merely ensures that every category available for a transcript can reach the hinge loss. Reproducible reporting must therefore include when negative selection occurs, how many candidates are retained, and, where available, the category distribution actually observed in mini-batches rather than only the stored inventory.

8.3 What human verification changes

The 91.7% retained rate shows that the finite-state procedure generates a large proportion of usable controlled alternatives in the sampled data, but the 378 reviewer corrections also show that round-trip consistency is not sufficient for a high-precision linguistic benchmark. The verified subset is intentionally not a miniature random copy of the 12,611-row benchmark: it over-represents numeral controls and replaces many generated texts. Consequently, the verified aggregate should be interpreted as a secondary diagnostic of robustness to human-curated contrasts. The close match between observed and mix-adjusted expected accuracy on the 353 unchanged rows argues against the specific concern that obviously invalid automatic rows alone inflated the reported automatic score. It does not validate every unsampled row in the full benchmark.

8.4 Observed local–global behavior

The resampled model becomes substantially stronger on controlled pairs while global retrieval develops more slowly. At epoch 6, automatic pairwise accuracy is 89.69%, but full-pool R@1 remains below the original checkpoint. The supplied trajectory is still increasing, so incomplete convergence is plausible; an optimization trade-off between close negative discrimination and broad ranking is also possible. The available measurements do not distinguish these explanations. Numeral is the only category below the original at epoch 6. Its 347 automatic rows make scarcity a plausible concern, but no repeated runs or category-size ablation were supplied, so the cause remains unresolved.

8.5 Practical implications

The targeted failure mode is operationally relevant: an audio–text system may retrieve a broadly related recording while remaining insensitive to a case marker, tense marker, negation, or numeral that changes the intended interpretation. This matters in spoken-archive search, transcript quality assurance, call-centre compliance, dictation review, and confirmation workflows involving amounts or dates. The benchmark does not certify deployment readiness, but it demonstrates why full-pool retrieval should be paired with controlled contrast evaluation in morphologically rich languages.

9 Conclusion

We introduce an automatic multimodal Malayalam speech–text diagnostic that separates global retrieval from fine-grained controlled discrimination. A no-hard-negative model reaches 37.45% full-pool Recall@1 while remaining close to chance on one-change comparisons, showing that aggregate retrieval is not a proxy for sensitivity to grammatical contrasts. We identify a fixed-first-candidate loader defect that restricted which generated negatives reached training. Per-transcript category-aware retention with access-time resampling improves matched-epoch aggregate and tense pairwise accuracy, and the complete extended corrected procedure reaches 89.69% automatic pairwise accuracy. A stratified human review retains or corrects 731 of 797 sampled rows and shows that reviewer-edited contrasts form a harder secondary evaluation set. The remaining full-pool retrieval gap and the difference between automatic and human-curated evaluation remain open. More broadly, targeted linguistic diagnostics, exact negative-delivery logic, and human validation should be considered jointly when evaluating multimodal speech–text models for morphologically rich languages.

10 Limitations

The full 12,611-row benchmark is automatically generated; human review covers a stratified 797-row sample rather than every row. The lead reviewer adjudicated all sampled items, but the supplied record does not provide per-category agreement values beyond ranges, blinded reviewer details, or a fully independent third adjudication. The verified subset deliberately over-represents numeral controls and contains reviewer-corrected alternatives, so its micro-average is not directly comparable with the full automatic benchmark. Round-trip validation establishes finite-state consistency, not human naturalness, dialectal acceptability, or

semantic felicity. The benchmark contains three morphosyntactic categories and one lexical control and is not a complete analysis of Malayalam morphology. The grapheme-confusable category remains outside the reported benchmark. All results use one audio backbone and one text backbone. The original and corrected trajectories are separate runs, and the extended corrected run additionally uses gradient clipping. No repeated-run estimates were supplied, and the corrected model was not trained to a demonstrated plateau. Because each training condition was observed under a single seed, the magnitude and stability of the reported improvements across random initializations remain unquantified.

References

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*, pages 4218–4222, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.520/.

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 *Proceedings of Interspeech 2022*, pages 2278–2282. https://doi.org/10.21437/Interspeech.2022-143.

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 33*, pages 12449–12460.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, California.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop*, pages 798–805. https://doi.org/10.1109/SLT54892.2023.10023141.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP: Learning audio concepts from natural language supervision. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095889.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323. https://doi.org/10.18653/v1/2020.findings-emnlp.117.

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*, 37(11):12942–12950. https://doi.org/10.1609/aaai.v37i11.26521.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *Computing Research Repository*, arXiv:2103.10730.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. SAMU-XLSR: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1493–1504. https://doi.org/10.1109/JSTSP.2022.3192714.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124. https://doi.org/10.18653/v1/2021.naacl-main.324.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5835–5847. https://doi.org/10.18653/v1/2021.naacl-main.466.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*. https://arxiv.org/abs/2010.04592.

Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath 2022. SpeechCLIP: Integrating speech with pretrained vision and language model. In *2022 IEEE Spoken Language Technology Workshop.*

Santhosh Thottingal. 2019. Finite state transducer based morphology analysis for Malayalam language. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 1–5.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *Computing Research Repository*, arXiv:1807.03748.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392. https://doi.org/10.1162/tacl_a_00321.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2CLIP: Learning robust audio representations from CLIP. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4563–4567. https://doi.org/10.1109/ICASSP43922.2022.9747669.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113. https://doi.org/10.18653/v1/2022.naacl-main.376.

Yuguo Yin, Yuxin Xie, Wenyuan Yang, Dongchao Yang, Jinghan Ru, Xianwei Zhuang, Liming Liang, and Yuexian Zou. 2025. ATRI: Mitigating multilingual audio text retrieval inconsistencies by reducing data distribution errors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5491–5504, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.273.

Wenzheng Zhang and Karl Stratos. 2021. Understanding hard negatives in noise contrastive estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1090–1101. https://doi.org/10.18653/v1/2021.naacl-main.86.