# Expressive Hindi Audiobook Generation with CLAP-Based Retrieval

**William Xing** [1] **Kiran Raja** [1] **Pranav Anuraag** [1] **Arjun Bahuguna** [1] **Vasu Sharma** [1]

## Abstract

Open-weight text-to-speech models for Hindi produce fluent narration but fail to adapt prosody to narrative context, generating every sentence in a near-identical tone regardless of emotional or dramatic content. We address this with a retrieval-augmented pipeline that requires no fine-tuning of the synthesis model. We train a Hindi CLAP model (MuRIL text encoder, HTS-AT audio encoder) on IndicVoices and Rasa to learn a joint text-audio embedding space. At inference, each sentence is embedded and matched against a precomputed library of Hindi audio clips via cosine similarity. The retrieved audio conditions IndicF5, a zero-shot TTS model that accepts reference waveforms natively, for expressive speech generation. Sentence-level outputs are then concatenated into full audiobook narrations. Our pipeline achieves a MOS of 4.00, NMOS of 3.89, and state-of-the-art intelligibility with 17.07% WER and 5.19% CER on held-out Hindi test data. The full pipeline is open-weight and runs on a single GPU.

## 1. Introduction

Audio-first content platforms serving Hindi audiences have grown rapidly, yet open-weight TTS systems narrate every sentence in a near-identical tone regardless of narrative context. A character's whispered confession and a battle cry receive the same flat delivery. Retraining with emotion-labeled data requires prohibitive compute for under-resourced languages, and commercial APIs such as ElevenLabs and Bulbul V3 (Sarvam AI, 2026) are closed-source. Improving expressiveness at inference time without retraining remains both a practical and a research challenge.

Existing approaches fall into three categories. First, caption-conditioned models such as IndicParler-TTS (Lacombe et al., 2024) label surface acoustics (pitch, speed) without narrative affect and drift over long-form generation. Second, prosody-control methods such as StyleTTS 2 (Li et al., 2023) and E-TTS (Gupta & Murthy, 2023) require retraining on expressive corpora. Third, retrieval-augmented systems such as CA-CLAP (Xue et al., 2024) and REA-TTS (Lu et al., 2026) retrieve emotionally matched reference audio per sentence, but operate exclusively in English or Mandarin and assume synthesis backends fine-tuned jointly with retrieval. No existing system provides retrieval-augmented expressive generation for Hindi with a frozen synthesis model.

We observe that recent zero-shot TTS models such as XTTS-v2 (Casanova et al., 2024) and IndicF5 (V et al., 2025) condition generation directly on reference waveforms, fully decoupling retrieval from synthesis. This suggests that a strong retrieval module can improve expressiveness without modifying the synthesis model at all. This raises the question of whether a CLAP model trained on Hindi data can retrieve sufficiently expressive references to produce natural audiobook narration through this decoupled approach.

We propose a simple two-stage pipeline. First, we train a Hindi CLAP model using MuRIL (Khanuja et al., 2021) as the text encoder and HTS-AT (Chen et al., 2022) as the audio encoder, learning a shared 512-dimensional embedding space on the Hindi subsets of IndicVoices-R (Sankar et al., 2024) and Rasa (Varadhan et al., 2024). Second, at inference time, we segment input text into sentences, embed each sentence, retrieve the most semantically similar audio clip via cosine similarity, and pass it as the reference to IndicF5 for generation. The resulting sentence-level outputs are concatenated into a continuous narration.

**Contributions.**

1. A Hindi CLAP model (MuRIL + HTS-AT) trained contrastively on IndicVoices-R and Rasa, achieving 50.8% R@1 and 89.2% R@10 on held-out Hindi test data with a T2A mean rank of 6.09
2. A retrieval-augmented inference pipeline connecting CLAP retrieval to frozen IndicF5, showing semantic retrieval alone improves expressive Hindi narration.
3. Evaluation of retrieval accuracy, speaker consistency, and mean opinion score against IndicF5, IndicParler-TTS, and an ElevenLabs commercial baseline.

[1]PocketFM. Correspondence to: William Xing <williamlelexing@gmail.com>, Kiran Raja <kiran.rajatx1@gmail.com>.

*Proceedings of the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea. PMLR 306, 2026. 

## 2. Related Work

**Multilingual and Indic TTS.** IndicParler-TTS (Lacombe et al., 2024) extends Parler-TTS (Lyth & King, 2024) to 22 Indic languages with caption-conditioned generation, but labels surface acoustics without narrative affect and drifts in long-form settings. XTTS-v2 (Casanova et al., 2024) performs zero-shot voice cloning from reference audio and supports Hindi natively. svara-TTS (Kenpath Technologies, 2025) appends discrete emotion tags but lacks gradient control for continuous transitions. Bulbul V3 (Sarvam AI, 2026) is production-grade but closed-source. Our pipeline solves expressiveness through learned retrieval rather than caption engineering or architectural modification.

**CLAP and retrieval-augmented expressiveness.** CLAP (Elizalde et al., 2022) learns joint text-audio embeddings via contrastive learning. CA-CLAP (Xue et al., 2024) retrieves emotionally matched clips per sentence, improving naturalness MOS from 3.62 to 3.92 on English audiobooks. REA-TTS (Lu et al., 2026) concatenates multiple matched clips per segment (MOS 4.16 vs. 3.85) but introduces boundary artifacts. Both are English-only and require jointly trained synthesis backends. We adapt CLAP retrieval to Hindi with a MuRIL text encoder and pair it with a synthesis model that accepts reference audio natively, removing the need for joint training.

**Prosody control without retraining.** StyleTTS 2 (Li et al., 2023) requires retraining on expressive data. Daft-Exprt (Zaïdi et al., 2022) needs a reference with the desired stress pattern. Sigurgeirsson and King (Sigurgeirsson & King, 2023) predict per-word F0 overrides via an LLM without retraining (50% preference vs. 31% baseline) but remain bounded by the acoustic model's range. E-TTS (Gupta & Murthy, 2023) learns Hindi intonation patterns but evaluates only conversational style. These methods operate at the acoustic-feature level. Our pipeline instead retrieves a real utterance and provides it as waveform-level conditioning, giving the synthesis model a direct acoustic example.

## 3. Method

### 3.1. Overall System

Our proposed audio generation pipeline primarily consists of two parts: a CLAP-based audio retrieval module, and a speech synthesis module.

### 3.2. CLAP Training Datasets

**Rasa** (Varadhan et al., 2024) is a large-scale expressive speech dataset developed for text-to-speech research in Indian languages. The Hindi subset of Rasa consists of around 24 hours of high-quality, single-speaker speech, with recordings covering multiple expressive speaking styles including neutral narration, conversational speech, and Ekman emotions such as happiness, sadness, and anger. Rasa was mainly designed to support expressive TTS systems, making it ideal for our use case of learning semantic and expressive relationships between Hindi speech and text for audiobook narration.

**IndicVoices-R** (Sankar et al., 2024) is a large-scale multi-speaker speech corpus developed for Indian text-to-speech research. The Hindi subset of IndicVoices-R contains approximately 75 hours of high-quality speech from almost 400 speakers, with recordings consisting primarily of natural extempore and conversational speech. The diversity of speakers and expressive conversational speech makes the dataset particularly useful for learning robust semantic and acoustic representations across a wide variety of speech.

### 3.3. CLAP-Based Retrieval Module

Contrastive Language-Audio Pretraining (CLAP) is a multimodal representation learning framework that learns a shared embedding space between audio and text. The central idea is to map semantically related audio clips and textual descriptions to nearby points in the latent space while pushing unrelated pairs farther apart. As a result, the model learns high-level semantic and expressive relationships between speech and language without requiring explicit class labels.

CLAP consists of two separate encoders: an audio encoder and a text encoder. The audio encoder processes spectrogram representations of audio clips and produces fixed-dimensional audio embeddings. In our implementation, the audio encoder is based on HTS-AT (Chen et al., 2022), which uses transformer-based hierarchical attention mechanisms to capture both local acoustic structure and higher-level semantic information. The text encoder, based on (Khanuja et al., 2021), converts corresponding transcripts or captions into text embeddings of the same dimensionality. Both the text and audio embeddings are then projected into a shared latent space of 512 dimensions through learned projection layers. During training, the entirety of each encoder is finetuned for this task of Hindi audio retrieval.

Training is performed on the Hindi train subsets of the Rasa and IndicVoices-R datasets using a contrastive InfoNCE objective. Given a batch of $N$ paired audio and text samples, cosine similarities are computed between every audio embedding $a_i$ and text embedding $t_j$. These similarity values are scaled by a learnable temperature parameter $\gamma$ used for controlling the sharpness of the similarity distribution. The model is optimized such that matching pairs receive high similarity scores while non-matching pairs receive low similarity scores. The symmetric contrastive loss is defined as:

$$L = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\exp(\gamma)\,\mathrm{sim}(a_i,t_i))}{\sum_{j=1}^{N}\exp(\exp(\gamma)\,\mathrm{sim}(a_i,t_j))} + \log\frac{\exp(\exp(\gamma)\,\mathrm{sim}(a_i,t_i))}{\sum_{j=1}^{N}\exp(\exp(\gamma)\,\mathrm{sim}(t_i,a_j))}\right]$$

Key training hyperparameters include a batch size of 128, a learning rate of $1\times10^{-5}$ for the encoders, and a learning rate of $1\times10^{-4}$ for the rest of the CLAP model. We use a smaller learning rate for the encoders as they are already pretrained on vast amounts of data, while the other parameters are initialized from scratch. The model was trained for 120 epochs.

### 3.4. Text-to-Speech Models

**IndicParler-TTS** (AI4Bharat, 2024) is a multilingual TTS model that can officially speak 20 Indic languages, including Hindi. Alongside the transcript, IndicParler-TTS accepts natural-language prompts for controlling speaking style and delivery. To improve speaker consistency during long-form generation, we use the pre-trained speaker name "Rohit" for all Hindi generations. Because IndicParler-TTS doesn't take in reference audio, we treat it as an open source baseline.

**IndicF5** (V et al., 2025) is a multilingual zero-shot TTS model trained on various Indic datasets, including Rasa and IndicVoices-R. It performs speech generation using a reference audio clip and the reference audio's corresponding transcript for speaker and prosody conditioning. Due to its strong expressive generation capabilities when conditioned on reference audio, IndicF5 is the primary open-weight TTS model studied in this paper.

**XTTS-v2** (Casanova et al., 2024) is a multilingual zero-shot text-to-speech model that also performs speech generation with a reference audio sample. We also test our CLAP retrieval module on XTTS-v2 for comparison against IndicF5.

**ElevenLabs Eleven v3** is a commercial, natively multilingual TTS system supporting long-form generation with strong built-in speaker consistency. Unlike zero-shot models, it processes Hindi text directly without requiring user-provided reference audio waveforms. We use the voice profile "George-Warm, Captivating Storyteller" across all tasks. While other highly expressive commercial platforms exist (such as Cartesia Sonic or Resemble AI Sesame) we utilize ElevenLabs as a representative closed-source, high-fidelity upper bound for multi-sentence narration.

### 3.5. Speech Synthesis Module

After training the CLAP model, we construct a reference audio library by embedding all audio samples from the Hindi subsets of Rasa and IndicVoices-R into the shared audio-text latent space. These projected embeddings enable direct similarity comparisons between textual queries and candidate audio samples for semantic and expressive retrieval. However, a clear issue with generating multiple sentences with different reference audio is speaker consistency. Therefore, we also test constructing the reference library solely with Rasa, a single-speaker dataset.

During inference, input text is first segmented into sentence-level chunks to enable fine-grained expressive retrieval and generation. Each sentence is then tokenized and passed through the trained CLAP text encoder to obtain a text embedding in the shared latent space. This embedding is compared against the precomputed reference library of audio embeddings using cosine similarity. The most semantically similar audio sample is selected as the reference audio. Then, the retrieved reference audio and its corresponding transcript are provided to a TTS model to guide zero-shot expressive speech generation for the target sentence. We test both IndicF5 and XTTS-v2 as the TTS model for reference-conditioned generation. Finally, all generated sentence-level audio segments are concatenated into a single continuous audiobook narration.

### 3.6. Evaluation Metrics

On the retrieval side, we mainly evaluate objective metrics. T2A and A2T mean rank measure the average rank position of the correct match during retrieval. T2A retrieves the most semantically similar audio clips from text, while A2T retrieves the most similar text from audio. All subsequent retrieval metrics were computed using the T2A pathway, which is the retrieval setting used during inference. Recall@K (R@K) measures the fraction of queries for which the correct sample appears within the top K retrieved results, while mAP summarizes overall ranking quality by averaging the inverse of each rank.

Each TTS model generated audiobook narrations for 12 Hindi stories collected from the Pratham Books StoryWeaver platform.[1] To evaluate each model's speech generation capabilities, we use both objective and subjective metrics. The objective metrics tested include word error rate (WER) and character error rate (CER). The subjective metrics tested include mean opinion score (MOS), naturalness mean opinion score (NMOS), and expressiveness mean opinion score (EMOS). Subjective scores were evaluated via a listening survey completed by 15 fluent Hindi speakers who graded randomized, blinded audio samples from each system.

*Table 1.* Retrieval performance comparison between our CLAP model and REA-TTS

| Model | T2A mean rank ↓ | A2T mean rank ↓ | R@1 ↑ | R@5 ↑ | R@10 ↑ | mAP ↑ |
|---|---|---|---|---|---|---|
| REA-TTS | 12.58 | 14.02 | **0.53** | 0.76 | 0.83 | 0.63 |
| Ours | **6.09** | **5.47** | 0.51 | **0.82** | **0.89** | **0.64** |

*Table 2.* Evaluation results for MOS, CER, and WER of different models

| Model | MOS ↑ | NMOS ↑ | EMOS ↑ | WER/% ↓ | CER/% ↓ |
|---|---|---|---|---|---|
| IndicParler-TTS | $3.70 \pm 0.67$ | $3.61 \pm 0.78$ | $2.42 \pm 0.90$ | 26.47 | 9.13 |
| ElevenLabs | $\mathbf{4.29} \pm 0.95$ | $\mathbf{4.38} \pm 0.88$ | $\mathbf{3.44} \pm 1.13$ | 18.98 | 5.66 |
| XTTS-v2 + Rasa, IndicVoices-R | $2.43 \pm 1.13$ | $2.71 \pm 1.35$ | $3.00 \pm 0.87$ | 31.31 | 17.48 |
| XTTS-v2 + Rasa | $2.20 \pm 1.14$ | $2.70 \pm 1.10$ | $2.50 \pm 1.17$ | 30.42 | 18.27 |
| IndicF5 + Rasa, IndicVoices-R | $3.25 \pm 1.04$ | $3.70 \pm 1.08$ | $2.67 \pm 0.87$ | 18.06 | 5.56 |
| IndicF5 + Rasa only | $4.00 \pm 1.07$ | $3.89 \pm 1.10$ | $3.11 \pm 1.27$ | **17.07** | **5.19** |

# 4. Experiments

## 4.1. CLAP Retrieval Results

Table 1 reports text-to-audio retrieval performance for the proposed CLAP-based framework. Retrieval metrics were computed on 1700 held-out Hindi samples.

Our approach achieves lower mean rank values and improved R@5, R@10, and mAP scores compared to REA-TTS, indicating stronger semantic alignment within the shared embedding space. While REA-TTS was evaluated on Mandarin speech, this cross-lingual comparison serves strictly as an architectural benchmark to validate alignment performance in the joint embedding space; the downstream contribution of CLAP retrieval to Hindi synthesis quality is isolated and demonstrated across the metrics in Table 2.

The T2A and A2T mean ranks are relatively similar (small difference of 0.62), demonstrating that the embedding space isn't biased in any specific direction. Furthermore, nearly 90% of the correct audio matches were retrieved within the top 10 results, demonstrating the CLAP model's strong semantic retrieval capability. The fact that the mAP is 0.64 while the T2A mean rank is 6.09 (expected mAP of 1/6.09=0.164) implies that many queries retrieve the correct audio very early, but a small number of outliers increase the average rank.

## 4.2. Speech Synthesis Results

We evaluate six speech synthesis systems: IndicParler-TTS, ElevenLabs Eleven v3, XTTS-v2 with CLAP-based retrieval using the full reference library, XTTS-v2 with CLAP-based retrieval using only Rasa references, IndicF5 with CLAP-based retrieval using the full reference library, and IndicF5 with CLAP-based retrieval using only Rasa references.

Table 2 reports all the evaluation metrics for audio generation with these models. ElevenLabs performs the best in terms of MOS, NMOS, and EMOS. First, we noticed that it actually uses a flatter tone compared to our CLAP-augmented IndicF5 model. This may reduce listener fatigue while still adopting subtle tonal shifts that are key for audiobook narration. Furthermore, its WER and CER metrics are only marginally behind IndicF5. However, IndicF5 using only the Rasa reference library is still second overall across all metrics, beating out the other open-source models. While the difference in WER and CER is marginal when including and not including IndicVoices-R in the reference library, excluding IndicVoices-R greatly improved speaker consistency, which is very important in long-form audio generation. This occurred because the remaining dataset, Rasa, is a single-speaker dataset. Because IndicF5 was trained on datasets overlapping with our retrieval reference library (Rasa and IndicVoices-R), its generations may also benefit from increased familiarity with the acoustic and stylistic characteristics of the evaluation domain. XTTS-v2 received lower MOS scores due to noticeable degradation in audio quality. While MOS and NMOS scores are relatively high, EMOS scores remain below 3.50 across all configurations. This modest performance underscores a structural limitation of an inference-time pipeline: while retrieval provides excellent semantic guidance, the absolute expressive ceiling is tightly bounded by the acoustic range of the frozen synthesis model.

In contrast, for WER and CER, IndicF5 using only Rasa references performs the best out of all the models, outperforming even the commercial ElevenLabs baseline. While massive, generalized multilingual foundation models can occasionally introduce mild phonetic drift or accent artifacts on regional languages like Hindi, our pipeline pairs a

[1] https://storyweaver.org.in/en/stories?language=Hindi&level=4

language-specific retriever with a localized backend. The semantically matched reference audio provides targeted, native phonetic and prosodic guidance, which directly explains the superior pronunciation accuracy and clarity in the generated speech.

## 5. Conclusion

In this paper, we present a retrieval-augmented framework for expressive Hindi audiobook generation using CLAP-based audio retrieval and frozen zero-shot TTS models. By conditioning generation on retrieved reference audio, our approach improves expressive delivery without fine-tuning the synthesis model. Experimental results show improved retrieval accuracy, pronunciation, clarity, and expressive speech generation over several open-weight baselines, highlighting inference-time retrieval as a promising direction for expressive TTS in low-resource Indic languages.

## References


AI4Bharat. Indic parler-tts. https://huggingface.co/ai4bharat/indic-parler-tts, 2024.

Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., and Weber, J. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. *arXiv preprint arXiv:2202.00874*, 2022.

Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H. Clap: Learning audio concepts from natural language supervision, 2022. URL https://arxiv.org/abs/2206.04769.

Gupta, I. and Murthy, H. A. E-TTS: Expressive text-to-speech synthesis for Hindi using data augmentation. In *Speech and Computer: 25th International Conference, SPECOM 2023, Dharwad, India, November 29–December 2, 2023, Proceedings, Part II*, volume 14339 of *Lecture Notes in Computer Science*, pp. 243–257. Springer, Cham, 2023. doi: 10.1007/978-3-031-48312-7_20.

Kenpath Technologies. svara-TTS v1: Open multilingual TTS for india's voices. https://huggingface.co/kenpath/svara-tts-v1, 2025.

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., Gupta, S., Gali, S. C. B., Subramanian, V., and Talukdar, P. Muril: Multilingual representations for indian languages, 2021.

Lacombe, Y., Sankar, A., Thomas, S., Varadhan, P. S., Gandhi, S., and Khapra, M. Indic parler-TTS. https://huggingface.co/ai4bharat/indic-parler-tts, 2024.

Li, Y. A., Han, C., Raghavan, V., Mischler, G., and Mesgarani, N. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19594–19621. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf.

Lu, Q., Bai, B., Xue, J., Li, Y., and Gao, Y. REA-TTS: Retrieval-augmented expressive audiobook text-to-speech generation with contrastive language-audio learning. 2026. ISSN 1995-8188. doi: 10.1007/s12204-026-2904-2. URL https://doi.org/10.1007/s12204-026-2904-2.

Lyth, D. and King, S. Natural language guidance of high-fidelity text-to-speech with synthetic annotations, 2024.

Sankar, A., Anand, S., Varadhan, P. S., Thomas, S., Singal, M., Kumar, S., Mehendale, D., Krishana, A., Raju, G., and Khapra, M. M. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian TTS. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7dfcaf4512bbf2a807a783b90afb6c09-Abstract-Datasets_and_Benchmarks_Track.html.

Sarvam AI. Bulbul V3: Production-grade text-to-speech for indian languages. https://www.sarvam.ai/blogs/bulbul-v3, 2026.

Sigurgeirsson, A. T. and King, S. Controllable speaking styles using a large language model, 2023.

V, P. S., Anand, S., Siddhartha, S., and Khapra, M. M. Indicf5: High-quality text-to-speech for indian languages, 2025. URL https://github.com/AI4Bharat/IndicF5.

Varadhan, P. S., Sankar, A., Raju, G., and Khapra, M. M. Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Low-resource Settings. In *Proc. INTERSPEECH 2024*, 2024.

Xue, J., Deng, Y., Gao, Y., and Li, Y. Retrieval augmented generation in prompt-based text-to-speech synthesis with context-aware contrastive language-audio pretraining, 2024.

Zaïdi, J., Seuté, H., van Niekerk, B., and Carbonneau, M.-A. Daft-Exprt: Cross-speaker prosody transfer on any text for expressive speech synthesis, 2022. v2, revised April 2022.