Improving Hindi Audiobook Generation with Inference-Time Interventions

Anonymous Authors[1]

Abstract

Hindi audiobook generation is an accessibility-critical but under-served text-to-speech setting: available systems often sound fluent yet fail to adapt prosody to narrative context, and retraining large multilingual TTS models is expensive for many researchers and developers. We study whether expressive long-form Hindi narration can be improved entirely at inference time using a frozen caption-conditioned backbone. Our pipeline augments IndicParler-TTS with chunk-level sentiment-to-caption mapping, Devanagari punctuation normalization, and a GPT-4o variant that predicts prosodic captions from full-story context while preserving speaker consistency with lightweight gating. Across short-form and long-form evaluations, baseline IndicParler-TTS remains smooth but weakly aligned with textual affect, whereas punctuation normalization improves pause placement and GPT-4o captions raise semantic alignment from 0.042 to 0.146 on a 10-segment Hindi story. Even so, long-form sentiment agreement remains low at 32.1%, showing that caption-only control is not sufficient for robust contextual expressiveness. We present these results as a practical low-resource baseline for Hindi audiobook generation: meaningful gains are possible without retraining, but stronger Hindi-specific expressive modeling is still needed.

1. Introduction

Expressive audiobook generation is a practical accessibility problem for Hindi and other under-served languages. Educational material, public-interest content, and leisure reading are increasingly distributed in audio form, yet high-quality expressive narration remains concentrated in high-resource languages and closed commercial systems. For many researchers and developers in the Global South, retraining a modern TTS model for a new language, domain, or voice is computationally and financially unrealistic.

This paper studies whether a frozen caption-conditioned TTS model can be pushed toward context-aware Hindi narration entirely at inference time. We focus on IndicParler-TTS because it is open-weight, supports natural-language style captions, and already covers Hindi, but its long-form output is often temporally smooth while remaining emotionally flat. The core challenge is that narrative context is not automatically translated into prosody: tragic, neutral, and celebratory passages can all be spoken without contextual paralinguistic attributes like corresponding tone, accent, and stress in the voice.

Our approach compares chunk-level sentiment-to-caption mapping with multilingual BERT, language-specific orthographic normalization for Devanagari punctuation, and a GPT-4o variant that predicts richer prosodic captions from full-story context. Because all interventions act at inference time, the underlying TTS backbone remains frozen, making the pipeline cheaper to reproduce and easier to adapt in constrained settings.

Contributions. We make three contributions. First, we frame Hindi audiobook generation as a low-resource, accessibility-oriented expressive TTS problem rather than only a benchmark optimization task. Second, we restructure expressive control around inference-time captioning, preserving the open-weight Hindi backbone and avoiding model retraining. Third, we provide a multi-part evaluation spanning short-form expressiveness, long-form narration, sentiment alignment, and language-specific punctuation normalization, showing both where lightweight interventions help and where they still fall short.

2. Related Work

Multilingual and Indic Text-to-Speech. Neural text-to-speech has reached near-human naturalness for short English utterances, yet long-form synthesis in low-resource languages remains open. Parler-TTS (Lyth & King, 2024) introduced caption-conditioned generation controlled by natural language descriptions of pitch range, speaking rate, and noise. Its annotation pipeline labels surface acoustics but

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

omits affect: a despairing line and a triumphant one receive identical annotations when their measured properties match. IndicParler-TTS (Lacombe et al., 2024) extends this to 22 Indic languages with named emotion tokens, but provides no mechanism to select tokens from narrative context and is prone to speaker drift in long-form generation. IndicF5 (V et al., 2025; Chen et al., 2024), a non-autoregressive flow-matching model, excels at zero-shot voice cloning but lacks sequential memory to build emotional arcs across segments. svara-TTS (Kenpath Technologies, 2025) appends discrete emotion tags (`<happy>`, `<sad>`) that lack the gradient control needed for continuous emotional transitions in narration. Bulbul V3 (Sarvam AI, 2026) is production-grade but closed-source, preventing architectural intervention. SeamlessExpressive (Seamless Communication et al., 2023) transfers prosody cross-lingually via an expressivity encoder but is optimized for speech-to-speech translation and reverts to flat prosody without a high-quality reference for every line.

Unlike these systems, our approach introduces an *inference-time* sentiment-to-caption pipeline that automatically selects contextually appropriate emotion tokens for a frozen IndicParler-TTS model, requiring no retraining or replacement of the TTS backbone.

Prosody Control and Expressive Speech Synthesis. StyleTTS 2 (Li et al., 2023) models style as a latent variable via diffusion, using a WavLM adversarial discriminator that learns paralinguistic features, pauses, intonation, stress, on audiobook data but not on neutral read-speech corpora, supporting the hypothesis that expressive training data drives prosodic quality more than architectural complexity. Daft-Exprt (Zaïdi et al., 2022) places FiLM conditioning layers throughout FastSpeech 2 to transfer prosody from a reference utterance across speakers via adversarial disentanglement, but requires a reference with the desired stress pattern at inference. For Hindi, E-TTS (Gupta & Murthy, 2023) augments 8.5 hours of neutral read speech with synthetic interrogative data and 11 minutes of story narration, forcing the model to learn rising intonation patterns and achieving a MOS of 4.42 - though evaluation is limited to conversational style, not long-form narration. Sigurgeirsson and King (Sigurgeirsson & King, 2023) use an LLM to predict per-word F0 and energy overrides for FastSpeech 2 without retraining, rated more contextually appropriate than the baseline in 50% of preference tests versus 31%, but bounded by the acoustic model's expressive range.

Our BERT-based pipeline operates at the *caption level* of a frozen model rather than at the acoustic-feature level. Our GPT-4o variant is closest to (Sigurgeirsson & King, 2023) but targets caption-level style selection with full-transcript context rather than phone-level overrides. Neither variant requires retraining, reference audio, or expressive training data.

Retrieval-Augmented Expressiveness for Long-Form Synthesis. A recent line of work uses CLAP-based retrieval to find emotionally matched reference audio before synthesis. CA-CLAP (Xue et al., 2024) trains a context-aware joint text-audio embedding space and retrieves, for each sentence, the single clip from the same speaker's earlier recordings whose emotional tone best matches the current passage. Evaluation on 48 English audiobooks shows gains over random retrieval on naturalness MOS (3.92 vs. 3.62) and similarity MOS (3.78 vs. 3.31). REA-TTS (Lu et al.) extends this by concatenating *multiple* matched clips per segment, achieving higher expressiveness MOS (4.16 vs. 3.85) but introducing naturalness drops at concatenation boundaries, a problem not reported with CA-CLAP's single-reference approach.

Our pipeline is complementary: where CA-CLAP and REA-TTS select reference *audio*, our pipeline selects reference *captions*, a text-to-text bridge directly applicable to caption-conditioned models without architectural changes or retrieval database construction. We return to retrieval-based extensions as future work in the conclusion.

3. Method

Current open-weight Hindi TTS models produce temporally smooth but expressively flat narration. We trace this to the DataSpeech annotation pipeline used to train IndicParler-TTS, which labels surface acoustic properties such as pitch range and speaking rate but omits emotional affect and intonation type. The model therefore has little signal connecting narrative context to prosodic delivery. Our method addresses this through inference-time expressiveness control, requiring no model retraining.

3.1. Models

Text-to-speech systems. **IndicParler-TTS** (AI4Bharat, 2024) is an instruction-conditioned multilingual TTS model that can officially speak 20 Indic languages, including Hindi and English. It is the primary open-weight system studied in this paper. Alongside the transcript, IndicParler-TTS accepts a natural-language description prompt that controls speaker characteristics, emotional tone, speaking style, and delivery. To improve speaker consistency during long-form generation, we use the pre-trained speaker name "Rohit" for all Hindi generations. Although IndicParler-TTS exposes emotion-specific prompts for some languages, these controls do not reliably transfer to Hindi.

ElevenLabs Eleven v3 is a commercial TTS system supporting long-form generation with strong built-in speaker consistency. We use the voice profile "George-Warm, Capti-

vating Storyteller" across all tasks. Unlike IndicParler-TTS, Eleven v3 exposes caption-level emotion control for Hindi via explicit transcript prefixes such as "[happy]." Because it is closed and commercial, we treat it as an upper bound rather than as the target deployment setting for low-resource researchers.

Voxtral 4B TTS (AI, 2026) is a recently released open-weight TTS model that we evaluate as an additional baseline. According to its authors, it performs strongly on multilingual speech generation and outperforms ElevenLabs Flash 2.5 across languages with a 68.4% win rate.

Analysis and consistency models. **Praat** is used for pitch-based validation by extracting fundamental-frequency (F0) statistics from synthesized speech, including mean pitch, pitch variance, and voiced-frame ratio. These measurements provide a lightweight consistency check to detect prosodic drift across generated chunks.

ECAPA-TDNN (Desplanques et al., 2020) provides chunk-level speaker consistency gating. A cosine similarity score is computed between each synthesized chunk and an anchor embedding; chunks below a threshold of 0.82 trigger a patch-and-retry loop. This gate operates jointly with an F0-based check using Praat features (mean drift $\leq 15\%$, standard-deviation ratio $\leq 35\%$).

Multilingual BERT (tabularisai, 2024) assigns per-chunk sentiment scores on a continuous $[-1, +1]$ scale. Scores are mapped to one of twelve IndicParler-TTS emotion tokens via polarity-bin partitioning, with lexical keyword overrides for high-confidence edge cases. Language-specific keyword sets and Devanagari danda-aware chunking are applied for Hindi inputs.

Wav2Vec2 is used for audio-side sentiment evaluation by predicting dimensional emotion scores from synthesized speech, specifically valence, arousal, and dominance (VAD). We use the audeering wav2vec2 emotion checkpoint to measure how closely the emotional delivery of generated audio aligns with the intended sentiment of the source text.

3.2. Dataset

We evaluate on two complementary settings. The first is a short-form expressive benchmark built from the Hindi subset of the Rasa multilingual expressive TTS dataset. Individual utterances span roughly 3–20 seconds. From the 13.5k rows of Hindi male speech, we select a balanced subset of 48 samples: 8 rows for each of 6 emotions (happy, anger, surprise, sad, fear, and disgust). This smaller subset keeps inference tractable while preserving broad emotional coverage.

The second setting is long-form story narration, which better reflects audiobook generation because it contains emotional progression across consecutive segments. We evaluate both Hindi and English stories. English serves as a high-resource sanity check and uses the British Council Story Zone story "Bad Blood." The Hindi evaluation uses internally curated story-style transcripts, including `story120_multi` for sentiment analysis and a 10-segment slice of Hindi Story 23 for feature-impact analysis. Long transcripts are chunked only at sentence boundaries. If appending the next sentence would exceed a language-specific character budget, a new chunk is started; we use 200 characters for English and 180 for Hindi to stay under the roughly 30-second generation ceiling we observed for IndicParler-TTS. Individual chunks are then stitched sequentially to form the full narration.

3.3. Evaluation

Compared conditions. We compare flat-caption IndicParler-TTS, emotion-conditioned IndicParler-TTS, ElevenLabs Eleven v3, and Voxtral 4B TTS, then study three inference-time interventions on IndicParler-TTS: BERT-based sentiment-to-caption mapping, orthographic symbol replacement for Devanagari punctuation, and GPT-4o caption generation from full-story context. We also evaluate the combined GPT-4o-plus-punctuation setting and an ablated prompt-only variant.

Baseline alignment metrics. We compute L_{sem} and L_{dyn} to characterize baseline long-form behavior. L_{sem} measures semantic alignment between the text and generated audio. For each segment, both the transcript and synthesized speech are embedded with CLAP, and we average the cosine similarity between the paired embeddings:

$$L_{\text{sem}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi_{\text{text}}(t_i) \cdot \phi_{\text{audio}}(a_i)}{\|\phi_{\text{text}}(t_i)\|_2 \, \|\phi_{\text{audio}}(a_i)\|_2}.$$

Higher L_{sem} indicates stronger alignment between textual meaning and spoken delivery.

L_{dyn} measures inter-segment expressive smoothness across long-form narration. It is the average squared ℓ_2 distance between CLAP audio embeddings of consecutive generated segments:

$$L_{\text{dyn}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|\phi_{\text{audio}}(a_{i+1}) - \phi_{\text{audio}}(a_i)\|_2^2 .$$

Lower L_{dyn} indicates smoother transitions and more stable long-form narration.

Short-form and long-form protocols. For short-form evaluation, we generate audio for the 48 Rasa utterances and perform a subjective comparison of expressiveness, speaker consistency, and overall acoustic quality. When evaluating

IndicParler-TTS, the target emotion is inserted into the description prompt. For ElevenLabs, the intended emotion is prepended directly to the transcript.

For long-form evaluation, we assess whether a model maintains contextual emotional alignment and coherent expressive delivery across multi-paragraph narration. Speaker consistency is controlled through fixed speaker identities and the consistency gates described above. Hindi and English are both included so that Hindi performance can be interpreted against a higher-resource reference language.

Sentiment alignment. We measure how well synthesized speech preserves textual sentiment using a matched-pair design across all generated segments. Text sentiment is predicted with multilingual BERT, which maps each segment into a score on $[-1, 1]$. The corresponding audio segment is evaluated with the Wav2Vec2 VAD model. Both modalities are then reduced to three polarity classes, negative, neutral, and positive, using valence thresholds of 0.4 and 0.6 for audio classification. Final metrics include three-class agreement rate, Cohen's κ, Pearson correlation, and Spearman correlation.

Inference-time control pipeline. The BERT-guided narrator pipeline divides the long-form transcript into sentence-level chunks, predicts a sentiment score and confidence value for each chunk, maps that signal to an IndicParler emotion token, and appends the resulting prompt to a fixed Character Voice Caption that preserves speaker identity. Each synthesized chunk is then passed through a two-stage consistency gate: **Stage 1** compares Praat F0 statistics against an anchor clip, and **Stage 2** requires ECAPA-TDNN cosine similarity above 0.82. If either gate fails, the system patches the pitch descriptor and retries up to twice.

To capture story-level patterns that the chunk-local BERT pipeline may miss, GPT-4o receives the full transcript and predicts per-segment pitch level, pace, focus words, intonation contour, and a style caption. These replace the emotion token in the generation caption. We also apply a deterministic orthographic replacement step before generation: Devanagari dandas, commas, and question marks are converted to ASCII equivalents. This preserves text content while exposing sentence boundaries more clearly to the TTS model.

4. Experiments

We found that Voxtral is inadequate for Hindi audiobook generation in its current form. The primary issue is frequent pauses every few words, which create highly unnatural pacing and substantially reduce narration quality. We therefore exclude Voxtral from the detailed result discussion below.

4.1. Baseline Quality of IndicParler-TTS

Computing L_{dyn} on IndicParler-TTS yields approximately 0, while L_{sem} is 0.35. Together, these metrics show that the model produces very smooth segment-to-segment narration but only limited semantic alignment between text and delivered prosody. This smooth-but-flat behavior motivates the inference-time control strategies studied in the rest of the paper.

4.2. Short-Form Expressiveness on Rasa

Overall, none of the evaluated systems closely match the Rasa ground truth in expressiveness. For audiobook generation, however, extremely strong expressiveness is not always desirable because long-form narration generally benefits from a controlled and only gradually varying tone.

Eleven v3 performs well as a short-form reference system. Based on subjective evaluation, its tone is very flat and shows limited expressive variation, but this may be intentional: a conservative delivery avoids obviously incorrect emotions and leaves more control to prompting.

IndicParler-TTS is more expressive than ElevenLabs in short clips. It often maintains a slightly elevated volume that gives the speech a cheerful quality while otherwise retaining a narration style similar to ElevenLabs' flatter delivery.

4.3. Long-Form Audiobook Generation

Eleven v3 is the strongest overall system for audiobook generation. Its comparatively flat tone is actually useful for long-form narration because it reduces listener fatigue, introduces almost no phoneme- or word-level errors, and still adopts subtle tonal shifts when important events occur. For this paper, however, ElevenLabs is best understood as a ceiling: it is closed and commercial, making it difficult to treat as a realistic solution for many Global South researchers and developers.

IndicParler-TTS is more expressive than ElevenLabs in the short-form Rasa setting, but its expressiveness is not contextual in long-form narration. In practice, it can describe tragic and happy events in nearly the same tone. This weakens the overall storytelling effect and makes ElevenLabs' flatter narration sound more appropriate to the text even when it is less vivid.

4.4. BERT-Guided Captioning

The BERT-guided experiment yields a mixed result. Pronunciation and overall clarity improve substantially, making the generated audio sound more like natural Hindi speech. However, the custom emotion-based description prompt for each chunk does not produce a clear improvement in contextual expressiveness.

We also observe higher-intensity chunks (that is, $|\text{BERT score}| > 0.6$), implying that the model is attempting stronger variation but may be exaggerating the wrong cues or producing prosody that is not well aligned with the text. This suggests that prompt-conditioned IndicParler-TTS can absorb coarse emotional hints without retraining, yet still lacks the fine-grained Hindi expressiveness needed for reliable audiobook narration.

4.5. Sentiment Alignment

For the Hindi IndicParler-TTS evaluation (`story120_multi`), sentiment alignment remains weak, with a three-class agreement rate of 32.1% and Cohen's $\kappa = -0.050$, indicating little meaningful agreement beyond chance. The confusion matrix shows strong collapse toward the neutral class, with most negative and positive text segments being classified as neutral in speech. Positive samples are relatively rare overall, and only 3 positive audio predictions are correctly matched. This suggests that generation remains smooth while contextual emotional alignment is still limited; see Figure 1.

For the English ElevenLabs baseline (`elevenlabs_vampire_story`), results are also weak, with 29.7% agreement and Cohen's $\kappa = 0.007$. Although the text contains a wider emotional range, the generated audio is heavily biased toward negative and neutral predictions, with almost no positive audio classifications despite many positive text segments. In particular, 59 positive text segments are present, yet none are correctly classified as positive in audio output. Even a strong commercial system therefore struggles to preserve sentiment polarity across long-form narration, especially for positive affect; see Figure 2.

4.6. Orthographic Normalization and GPT-4o Captions

Orthographic symbol replacement is more than preprocessing: it injects script-specific knowledge that English-centric TTS pipelines often ignore. Replacing Devanagari punctuation with ASCII equivalents increases average trailing pause duration from 38.3 ms to 97.5 ms and nearly doubles pitch variance from 19.63 to 38.77, confirming that the model inserts contextually appropriate pauses once it can parse sentence boundaries. However, L_{sem} drops from 0.0422 to 0.0194, indicating that improved pacing alone does not yield better semantic alignment. Punctuation normalization is therefore necessary but not sufficient.

GPT-4o style captions produce the largest semantic-alignment gains. With Devanagari punctuation intact (*prosody_only*), L_{sem} rises 3.5× over baseline, from 0.0422 to 0.1463, and pause duration increases to 219.4 ms, but pitch variance *decreases* to 16.49. Without normalized punctuation, the model interprets rich captions primarily as pacing cues rather than intonation cues.

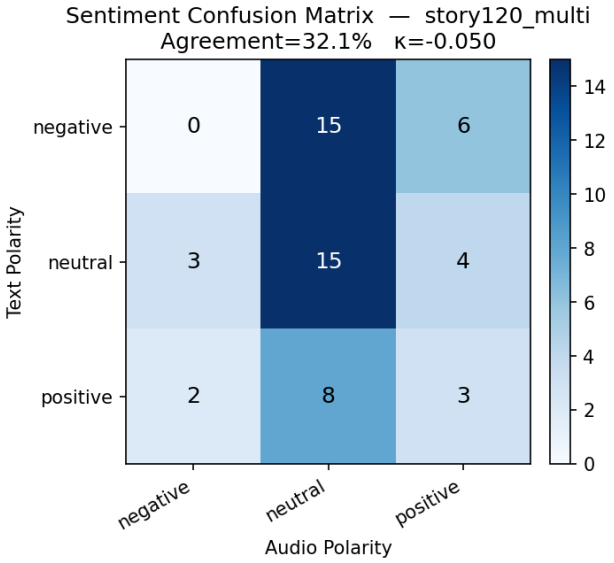

Figure 1. Sentiment confusion matrix for Hindi IndicParler-TTS. The model shows strong collapse toward neutral predictions, with limited positive sentiment preservation and only 32.1% overall agreement.

Table 1. Feature impact on 10 segments of Hindi Story 23. Punctuation normalization improves pauses; GPT-4o captions improve alignment; combining both yields 5.1× pitch variance and 3.3× L_{sem} over baseline.

Condition	Pause (ms)	Pitch Var	L_{sem}
Baseline	38.3	19.63	0.042
+ Punct. norm.	97.5	38.77	0.019
+ GPT-4o caption	219.4	16.49	0.146
+ Both	**206.7**	**100.07**	**0.138**

Combining both features resolves this dissociation. Pitch variance jumps to 100.07, or 5.1× baseline, L_{sem} remains high at 0.1384, and pause duration stays elevated at 206.7 ms. The interaction is clear: punctuation normalization enables boundary parsing, and GPT-4o captions then steer prosodic contour within each segment. Neither feature alone achieves both effects. Results are summarized in Table 1.

Absolute L_{sem} values remain modest given the $[-1, 1]$ cosine scale, reflecting IndicParler-TTS's limited responsiveness to fine-grained style instructions. The evaluation is also restricted to 10 segments of a single Hindi story, so generalization remains untested.

4.7. Ablation

We conduct an ablation study to evaluate whether the sentiment-classification and F0-based regeneration stages

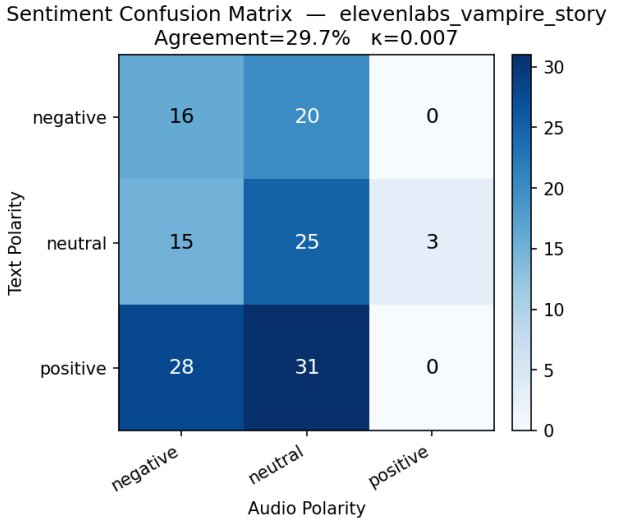

Figure 2. Sentiment confusion matrix for English ElevenLabs. Despite stronger baseline quality, the model remains heavily biased toward negative and neutral predictions, with almost no correctly preserved positive sentiment.

contribute meaningfully beyond prompt engineering alone. In this simplified setting, all BERT-based sentiment prediction, confidence filtering, and Praat-guided F0 correction are removed, and generation is performed using only manually designed description prompts appended to the transcript. The resulting audio shows the same improvements in pronunciation quality and overall expressiveness as the full pipeline, indicating that the primary gains come from the prompt itself rather than the additional processing stages. This suggests that IndicParler-TTS is largely driven by prompt conditioning, while post-generation correction provides limited benefit. More broadly, it highlights a core limitation of IndicParler-TTS: since the model does not support true audio-conditioned regeneration or refinement, corrective stages can only modify future prompts rather than directly repair generated audio.

5. Conclusion

We restructure Hindi audiobook generation around a low-resource question: how far can a frozen caption-conditioned TTS model be pushed without retraining? Across short-form and long-form evaluations, the answer is mixed. Inference-time interventions, especially Devanagari punctuation normalization and full-context GPT-4o captions, improve pause placement and semantic alignment, while the BERT-guided pipeline improves pronunciation and clarity. However, sentiment agreement remains weak, and even the best configuration still falls short of robust contextual expressiveness.

For the GlobalSouthML setting, the key takeaway is practical. Meaningful gains are possible with lightweight, language-aware methods that do not require expensive fine-tuning, and language-specific preprocessing can matter as much as model scale. At the same time, the strongest variant currently relies on GPT-4o and is evaluated on limited Hindi material. Future work should replace proprietary components with open alternatives, scale evaluation across more Indic languages supported by IndicParler-TTS, and study downstream accessibility benefits for education, oral storytelling, and low-vision readers.

References

AI, M. Voxtral: A fully open-weight and end-to-end speech model family. *arXiv preprint arXiv:2603.25551*, 2026. URL https://arxiv.org/abs/2603.25551.

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

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., and Chen, X. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching, 2024.

Desplanques, B., Thienpondt, J., and Demuynck, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*, pp. 3830–3834, 2020.

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.

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. 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.

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

Seamless Communication, Barrault, L., Chung, Y.-A., Coria Meglioli, M., Dale, D., Dong, N., Duppenthaler, M., Duquenne, P.-A., Ellis, B., Elsahar, H., Haaheim, J., Hoffman, J., Hwang, M.-J., Inaguma, H., Klaiber, C., Kulikov, I., Li, P., Licht, D., Maillard, J., Mavlyutov, R., Rakotoarison, A., Ram Sadagopan, K., Ramakrishnan, A., Tran, T., Wenzek, G., Yang, Y., Ye, E., Evtimov, I., Fernandez, P., Gao, C., Hansanti, P., Kalbassi, E., Kallet, A., Kozhevnikov, A., Mejia Gonzalez, G., San Roman, R., Touret, C., Wong, C., Wood, C., Yu, B., Andrews, P., Balioglu, C., Chen, P.-J., Costa-jussà, M. R., Elbayad, M., Gong, H., Guzmán, F., Heffernan, K., Jain, S., Kao, J., Lee, A., Ma, X., Mourachko, A., Peloquin, B., Pino, J., Popuri, S., Ropers, C., Saleem, S., Schwenk, H., Sun, A., Tomasello, P., Wang, C., Wang, J., Wang, S., and Williamson, M. Seamless: Multilingual expressive and streaming speech translation. In *arXiv*, 2023.

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

tabularisai. Multilingual sentiment analysis. https://huggingface.co/tabularisai/multilingual-sentiment-analysis, 2024.

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

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.