



# End-to-End Indian Language Dubbing with Zero-Shot Speaker Preservation

Giri Raju<sup>1</sup>, Sandeep Konam<sup>1</sup>

<sup>1</sup>R&D, Hitloop, USA

giriraju@hitloop.com, san@hitloop.com

## Abstract

This paper presents an end-to-end AI-driven dubbing platform designed specifically for Indian languages. The system leverages state-of-the-art speech models (ASR, MT, TTS) to streamline the dubbing process, minimizing manual effort while enabling precise output control. It generates high-quality, natural-sounding speech, preserving speaker characteristics through zero-shot synthesis irrespective of accent, age, or gender. Already successfully deployed with creators and educators, the platform enhances content accessibility and cross-lingual adaptation in education and entertainment for India's diverse linguistic communities.

**Index Terms:** AI-based Dubbing, Speech Synthesis, Speech Recognition, Translation

## 1. Introduction

The proliferation of digital content, particularly in education and entertainment, faces significant accessibility hurdles due to language barriers. In linguistically diverse nations like India, much valuable content remains confined to English or dominant regional languages, limiting reach. Traditional dubbing methods are resource-intensive, time-consuming, and dependent on specialized voice talent. Furthermore, delivering educational material in native languages is crucial for inclusive learning, while preserving the original speaker's vocal identity enhances engagement in entertainment.

To address these challenges, we developed an automated end-to-end dubbing platform. Its purpose is to enable seamless, high-quality content adaptation across multiple Indian languages, crucially preserving the original speaker's timbre and intonation for natural and authentic results using zero-shot synthesis. The platform has already demonstrated its utility and impact, being successfully used by influencers, educators, and creators, with plans for broader deployment underway. This work details its architecture, underlying technology, and features.

## 2. Dubbing Platform

The platform integrates Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) synthesis into an automated pipeline. It supports multilingual adaptation involving English, and nine Indic languages: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu. Key features include explicit duration control and utterance-level synthesis, managed through a user-friendly interface.

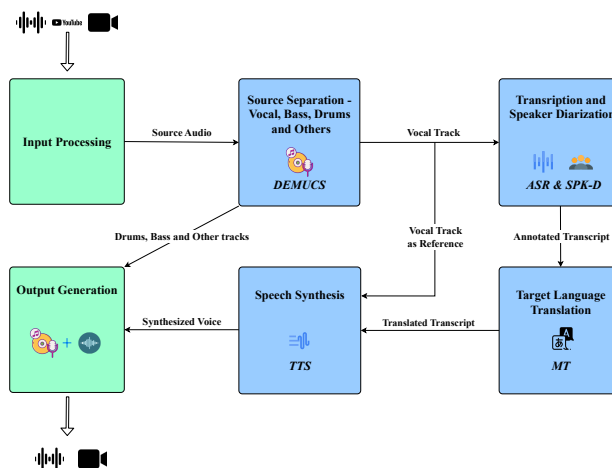


Figure 1: Workflow of the Dubbing Platform

### 2.1. Platform Workflow and Architecture

The platform [Fig 2] employs a sequential pipeline [Fig 1] designed for efficiency and ease of use, transforming input audio/video into dubbed content in the target language.

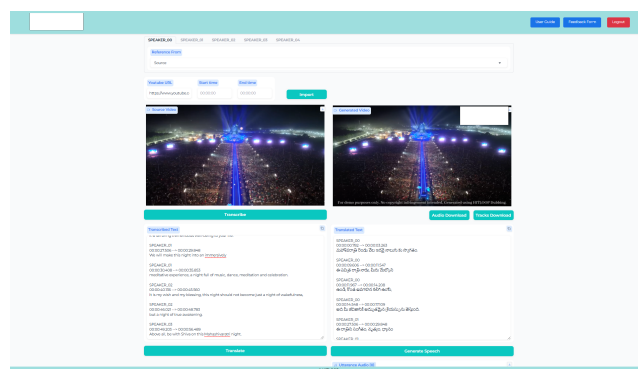


Figure 2: Interface of the Dubbing platform

**Step 1: Input Processing** The platform supports multiple input formats, enabling users to upload audio, video, or YouTube URLs. Regardless of the original format, all inputs are converted into a standardized audio format. A source separation [1] technique is then applied to isolate the vocal components from any noise or background audio. The same process is applied when an external speaker reference is provided.

**Step 2: Transcription** Once the vocal track is extracted, it is transcribed into text based on the detected language. State-of-the-art transcription models [2] [3] are employed to ensure accurate speech-to-text conversion across diverse language groups.

**Step 3: Speaker Diarization** Speaker diarization [4] is applied to the extracted vocal track to distinguish and annotate individual speakers. The transcribed text is then aligned with speaker labels and timestamps, providing an editable interface where users can review and modify the content as necessary.

**Step 4: Translation** The transcribed content is translated into the target language using advanced MT models [5] tailored to specific language groups. The interface permits users to refine the translation for linguistic accuracy and contextual fidelity while maintaining speaker alignment.

**Step 5: Speech Synthesis** The final translated text is synthesized into speech using a fine-tuned F5TTS [6] model, conditioned on either the separated source vocal track, an externally uploaded reference audio, or a selected preset voice. This reference-guided approach enables zero-shot speech synthesis, effectively preserving the speaker’s characteristics and prosodic style.

**Step 6: Output Generation** The synthesized speech is merged with the non-vocal tracks (music, effects) obtained from source separation and re-encoded into the desired output format (audio/video).

## 2.2. Data and Training Details

The pipeline’s core synthesis component is based on F5TTS [6], a model capable of zero-shot, expressive speech synthesis that preserves speaker identity.

### 2.2.1. Data

To support Indic languages, the base F5TTS model was fine-tuned using a combination of open-source datasets (IndicTTS, LIMMITS, SnowMountain, Rasa, IndicVoices-R) and a curated in-house dataset. The in-house data comprises 79.2 hours (70,767 utterances) across English, Hindi, Kannada, Malayalam, Tamil, and Telugu languages from native speaker personalities. The combined dataset totals 1,626 hours (737,632 utterances) across English, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu languages from 4,525 diverse speakers (age, gender, accent), enhancing model robustness and generalization.

### 2.2.2. Training

The F5TTS [6] model was fine-tuned for Indic languages using the aforementioned dataset. To adapt the model effectively, the vocabulary was expanded to include Indian languages, and the embedding layer was resized accordingly. Additional layers were initialized with random weights to accommodate newly introduced tokens. The base architecture was retained, while character-based tokenization was replaced with phoneme-based tokenization to enhance linguistic representation, improving pronunciation accuracy and speech synthesis quality.

Training was conducted for 1.84 million steps with a batch size of 1,600 using two NVIDIA L40 GPUs, maintaining default hyperparameter settings. The Vocos model was employed as the vocoder, transforming log mel spectrograms from the F5TTS model into audio waveforms. Trained from scratch on Indic languages for 2 million steps, it ensures improved linguistic adaptation and pronunciation accuracy.

## 2.3. Platform Features

The platform offers several user-centric controls:

**Duration Alignment:** A toggleable option aligns synthesized speech duration with source timestamps for precise lip-sync in videos. If disabled, utterances are synthesized sequentially with a 1s pause.

**Utterance-Level Control:** Users can synthesize individual utterances and modify the speaker reference per utterance (upload external audio, select preset voice, use source speaker strict match).

**Speech Rate & Fine-tuning:** Speaking rate is adjustable (0.5x to 1.5x). Controls for loudness, pitch, and speed allow further refinement of the synthesized output.

**Output Variants:** Provides "Base", "Crisp", and "Balanced" output formats, each available "with tracks" (synthesized vocals merged with original non-vocal audio) or "without tracks" (synthesized vocals only).

**Session Management:** Users can save, load, and switch between dubbing projects.

## 3. Conclusion

This end-to-end AI-driven dubbing platform provides an effective solution for cross-lingual content localization, specifically addressing the needs of Indian languages with zero-shot speaker preservation. By integrating robust ASR, MT, and expressive, identity-preserving TTS, it automates high-quality dubbing while offering user control over critical aspects like duration alignment and speaker identity. Its successful application by creators and educators validates its utility in enhancing accessibility for education and entertainment content. Aligned with the goals of Fair and Inclusive Speech Technology, the platform helps democratize content consumption and knowledge sharing across linguistic divides in India. Future directions include improving nuanced cross-lingual expressivity, and expanding linguistic coverage. We acknowledge the vital role of open-source AI advancements and models in enabling the development of this platform.

## 4. References

- [1] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," 2022. [Online]. Available: <https://arxiv.org/abs/2211.08553>
- [2] V. S. Lodagala, S. Ghosh, and S. Umesh, "Ccc-wav2vec 2.0: Clustering aided cross contrastive self-supervised learning of speech representations," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1–8.
- [3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [4] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," 2019. [Online]. Available: <https://arxiv.org/abs/1911.01255>
- [5] J. Gala and P. A. Chitale, "Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages," *Transactions on Machine Learning Research*, 2023. [Online]. Available: <https://openreview.net/forum?id=vfT4YuzAYA>
- [6] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," 2024. [Online]. Available: <https://arxiv.org/abs/2410.06885>