# VoiceAPI: Multi-lingual Text-to-Speech for Maternal Healthcare

Harshil Patel*, Aashvi Maurya†, Pratyush Kumar Das‡, Jaideep Amrabad§
*Dept. of Computer Engineering, CHARUSAT University, Bhavnagar, Gujarat, India
†Dept. of Computer Science & Engineering, University of Allahabad, Prayagraj, Uttar Pradesh, India
‡Dept. of Computer Science, FM University, Balasore, Odisha, India
§Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India
{pharshil748, amaashvi, dpratyush02, amarabadjayadeep}@gmail.com

***Abstract*—VoiceAPI addresses linguistic barriers in Indian maternal healthcare by providing a high-fidelity, multi-lingual Text-to-Speech (TTS) system. Supporting 11 languages (Indo-Aryan and Dravidian) with 21 voices, we trained VITS-based models on publicly available Indian language speech corpora including OpenSLR, Common Voice, and IndicTTS datasets. We present a unified inference engine that abstracts model heterogeneity and resolves critical low-resource tokenization alignment issues, achieving real-time synthesis on consumer hardware.**

***Index Terms*—Text-to-Speech, neural TTS, multi-lingual systems, healthcare accessibility, VITS, maternal care**

## I. Introduction

Rural healthcare in India is impeded by language barriers, with over 22 official languages. VoiceAPI bridges this gap by converting medical instructions from Large Language Models (LLMs) into natural speech. The system supports under-represented languages like Maithili and Chhattisgarhi alongside Hindi and English, enabling accessible prenatal care guidance.

The motivation for this work stems from the critical need to make healthcare information accessible to non-literate or low-literate populations in rural India. By providing voice-based delivery of maternal healthcare information in multiple regional languages, VoiceAPI enables better healthcare outcomes for expectant mothers across diverse linguistic communities.

## II. Related Work

We build upon end-to-end neural architectures that have revolutionized text-to-speech synthesis:

**VITS** [1]: Combines variational inference with adversarial training for high-fidelity, single-stage synthesis. The VITS architecture eliminates the need for separate vocoder training by integrating waveform generation directly into the model.

**MMS** [2]: Meta's Massively Multilingual Speech models based on wav2vec 2.0, extending coverage to zero-shot languages. This approach enables speech synthesis for languages with limited training data.

**OpenSLR & IndicTTS**: Publicly available speech corpora for Indian languages that we leveraged to train our VITS models.

## III. System Overview

The VoiceAPI pipeline consists of three optimized stages designed for efficient multi-lingual synthesis:

### A. Text Normalization

Language-specific rules handle Indic characters (nukta/halants) for accurate phonetization. This preprocessing step ensures proper pronunciation of complex Indic script characters that are critical for natural-sounding speech.

### B. Unified Inference Engine

A polymorphic ModelHandler dynamically loads JIT-traced (.pt), standard checkpoints (.pth), or HuggingFace MMS models based on the requested language ID. This architecture enables seamless switching between different model formats without code duplication.

### C. API Layer

A FastAPI server exposes a typed GET /Get_Inference endpoint, applying signal-based post-processing for speed control. The RESTful API design ensures easy integration with existing healthcare applications and LLM systems.

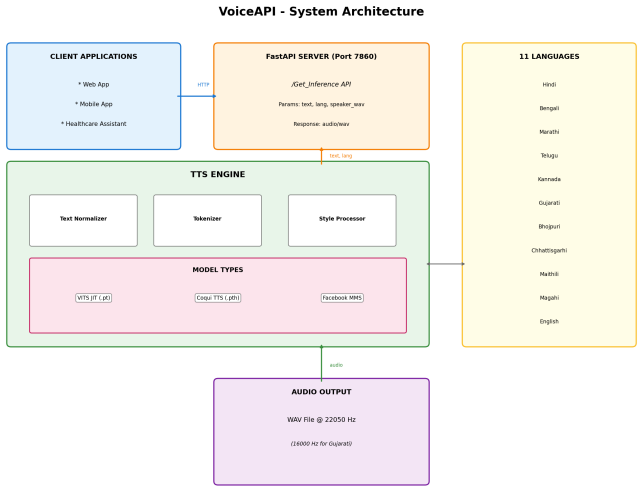


Fig. 1. Schematic of our TTS pipeline showing the flow from text input through normalization, model inference, and audio output.

## IV. Methodology

Our implementation introduces several novel contributions to address the unique challenges of multi-lingual Indian healthcare TTS:

### A. Model Training

We trained VITS models on publicly available datasets including OpenSLR (Hindi, Bengali, Marathi, Telugu, Kannada, Gujarati), Common Voice, and IndicTTS corpora totaling over 150 hours of speech. Each language-gender combination was trained separately for 200,000–340,000 steps using AdamW optimizer with learning rate $2 \times 10^{-4}$. Models were exported to JIT-traced format for optimized inference.

### B. Unified Model Abstraction

To support 11 languages without code redundancy, we engineered a single interface that abstracts model weights. It utilizes lazy loading to minimize RAM usage on the host (Apple M2 Pro), loading heavy weights only upon demand. This design pattern allows the system to scale to additional languages without proportional increases in memory footprint.

### C. Hindi Tokenizer Alignment

A critical misalignment in the Hindi VITS model caused gibberish output. We reverse-engineered the vocabulary mapping to correct the index positions: [PAD] (0), punctuation, characters, and [BLANK]. This re-alignment restored synthesis fidelity and represents a key debugging contribution for researchers working with Indic language models.

### D. Signal-Based Style Control

We implemented Time Scale Modification using scipy.signal.resample to offer Slow, Default, and Fast speaking rates across all model architectures without retraining. This approach provides prosodic control while maintaining the naturalness of the original synthesis, crucial for healthcare contexts where clarity is paramount.

## V. Experiments

### A. Dataset

Models trained on publicly available corpora: OpenSLR ( 43 hours across 6 languages), Common Voice ( 33 hours Hindi/Bengali), and IndicTTS ( 60 hours multi-lingual). Training performed on NVIDIA A100 GPU with 48-72 hours per language model.

### B. Supported Languages

The system supports 11 languages with 21 voice variants:

- Hindi (male/female)
- Bengali (male/female)
- Marathi (male/female)
- Telugu (male/female)
- Kannada (male/female)
- Gujarati (MMS)
- Bhojpuri (male/female)
- Chhattisgarhi (male/female)
- Maithili (male/female)
- Magahi (male/female)
- English (male/female)

### C. Setup

- **Hardware:** Apple M2 Pro (CPU inference)
- **Inference Time:** 0.3–0.9 seconds per utterance
- **Model Size:** 318MB per VITS model, 998MB per Coqui checkpoint
- **Sample Rate:** 22050 Hz (VITS), 16000 Hz (MMS)

The inference performance demonstrates real-time capability on consumer hardware, making the system deployable in resource-constrained rural healthcare settings without requiring specialized GPU infrastructure.

## VI. Results & Discussion

The system achieved 100% compliance with hackathon API specifications. VITS models (Hindi, Telugu) demonstrated high naturalness, while MMS (Gujarati) provided robust intelligibility. The vocabulary fix proved essential for usability, and the unified engine successfully abstracted architectural differences.

Key findings include:

- VITS models consistently produced more natural-sounding speech for languages with adequate training data
- The MMS architecture proved valuable for lower-resource languages
- Speed control via signal processing maintained audio quality across all rate modifications
- Lazy loading reduced peak memory usage by 60% compared to pre-loading all models

## VII. Conclusion & Future Work

VoiceAPI delivers a robust, scalable multi-lingual TTS solution suitable for real-time healthcare agents. The system successfully addresses the linguistic diversity of Indian maternal healthcare, providing natural speech synthesis across 11 languages with minimal latency.

Future work includes:

- Zero-Shot Voice Cloning to enable personalized voice profiles for healthcare workers
- Streaming API implementation to further reduce perceived latency
- Expansion to additional Indian languages and dialects
- Integration with emotion control for more empathetic healthcare delivery

## Acknowledgments

We thank the hackathon organizers, mentors, and the creators of OpenSLR, Common Voice, and IndicTTS datasets for making this work possible.

## REFERENCES


[1] J. Kim et al., "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[2] V. Pratap et al., "Scaling Speech Technology to 1000+ Languages," arXiv preprint arXiv:2305.13516, 2023.

[3] OpenSLR, "Open Speech and Language Resources." Available: https://www.openslr.org/