IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI - BASED MULTILINGUAL ASSISTANCE PLATFORM

[1]Shraddha More, [2]Sidharth Shinde, [3]Mohit Sharma,[4]Mayuri Vyavahare,[5]Girish Dashmukhe

[1]Student, [2]Student, [3]Student, [4]Student,[5]Assistant Professor
Artificial Intelligence and Data Science,
Sandip Institute of Technology and Research Centre (SITRC), Nashik, India

Abstract: The AI-Based Multilingual Assistance Platform is designed to assist millions of pilgrims visiting the Nashik -Trimbakeshwar Kumbh Mela through an inclusive, voice-driven digital solution. It integrates Speech Recognition, Natural Language Processing, Text-to-Speech, and Computer Vision to provide real-time guidance, navigation, and emergency support in Indian languages. The system functions effectively even in low-connectivity environments through an offline-first architecture, ensuring accessibility for rural, elderly, and differently-abled pilgrims.

***Index Terms* -** AI-Based Multilingual Assistance, Voice-Driven Digital Solution, Speech Recognition, Natural Language Processing (NLP), Text-to-Speech (TTS), Computer Vision

I. INTRODUCTION

The Nashik-Trimbakeshwar Kumbh Mela is one of the largest religious gatherings on the planet, bringing together millions of devotees, saints, and tourists from across India and the world. This sacred event, held once every twelve years, transforms entire cities into temporary spiritual hubs, creating unique challenges in managing crowd flow, providing timely information, ensuring safety, and delivering essential services to a highly diverse population. Many pilgrims are elderly, semi literate, or speak only local dialects, making it extremely difficult for them to access accurate information, navigate vast and crowded areas, or communicate during emergencies. The system is designed as a micro service architecture, enabling seamless integration with other modules of the Kumbh Smart Management ecosystem, such as crowd monitoring, public safety alerts, and real-time information systems. By leveraging modern Natural Language Processing (NLP) techniques and neural TTS models, the service ensures accurate pronunciation, emotional tone, and contextual clarity across multiple languages.

In an era where digital technologies have the potential to bridge such gaps, Artificial Intelligence (AI) combined with Multilingual Natural Language Processing (NLP) offers powerful solutions for inclusive assistance. This project aims to develop an Artificial Intelligence-Based Multilingual Assistance Platform that will act as a trusted digital companion for pilgrims visiting the Nashik-Trimbakeshwar Kumbh Mela. By leveraging AI technologies like Speech-to-Text, NLP, Text-to-Speech, Computer Vision, offline-first mobile architecture, the system provides real-time answers, voice-guided navigation, emergency support, and context-aware information entirely in the user's native language. Users can speak naturally in over many of the Indian languages and dialects to get directions to food stalls, accommodation options, ritual schedules, or seek help if lost or unwell.

The Nashik-Trimbakeshwar Kumbh Mela attracts millions of pilgrims who travel long distances from various parts of India, often coming from rural or remote regions where digital literacy and internet access are limited. Many pilgrims are elderly, semi-literate, or speak only their native dialects, which creates significant barriers when trying to access important information about rituals, locations, accommodations, food availability, and emergency services. Navigating the vast grounds of the Kumbh Mela, finding family members in a crowd, or asking for help during a crisis can become overwhelming without timely guidance and clear communication. Existing solutions such as printed signboards, loudspeaker announcements, or simple mobile apps are often insufficient for this unique user group, as they do not address the multilingual, voicefirst needs of the masses or function reliably in low-connectivity environments. This gap leads to confusion, lost time, stress, and even safety risks such as stampedes or unattended health emergencies. Therefore, there is an urgent need for a robust, Artificial Intelligencebased Multilingual Assistance Platform that can deliver real-time, voiceenabled, offline-capable support to every pilgrim, regardless of age,
literacy level, or language.

II. Survey Insights

To understand the communication challenges faced during large gatherings like the Kumbh Mela, a preliminary survey was conducted among pilgrims, volunteers, and event management staff. The objective was to identify language preferences, accessibility needs, and awareness levels regarding voice-based information systems. The survey results highlighted that over 70% of respondents preferred receiving announcements and information in their regional language, primarily Hindi or Marathi, while 20% favored English for clarity.

A significant portion of participants expressed difficulty in understanding mono-language announcements, especially in crowded or noisy areas, indicating a clear need for multilingual voice support. Additionally, 82% of the respondents agreed that an automated voicebased information system could improve their overall experience by reducing confusion, enhancing navigation, and ensuring timely delivery of safety alerts. Event volunteers and organizers also emphasized the importance of a modular, scalable, and easily deployable communication system capable of integrating with existing digital infrastructure. These findings validate the need for a multilingual Text-to-Speech (TTS) microservice, capable of dynamically generating natural, human-like speech in multiple Indian languages. The insights guided the system's design goals — prioritizing inclusivity, clarity, scalability, and real-time responsiveness.

III. Significance

The Nashik-Trimbakeshwar Kumbh Mela is not just a religious gathering but also a massive logistical and civic management challenge. Managing millions of pilgrims—many of whom are elderly, semi-literate, or speak only regional dialects—requires solutions that go beyond traditional information desks, printed maps, and static signboards. In such a context, the significance of an Artificial Intelligence-Based Multilingual Assistance Platform is immense. This project addresses one of the most pressing needs of largescale public events: bridging the communication gap between service providers and visitors who come from diverse linguistic and cultural backgrounds. This multilingual capability directly supports India's vision of digital inclusion and citizen-centric governance. Equally important is the platform's offline-first design, which makes it resilient in low or nonetwork areas—common in large, crowded event zones. Pilgrims can access essential services like finding food, water, toilets, affordable lodging, ritual schedules, and emergency help even when mobile data is unavailable. The inclusion of features such as lost-and-found assistance, real-time crowd monitoring, and emergency alerts further strengthens the safety net for vulnerable sections like the elderly and differently-abled.

On the administrative side, the project empowers local authorities and event managers with a connected web dashboard that provides live insights into crowd density, allows instant broadcasting of critical announcements, and helps coordinate emergency responses more efficiently. This not only improves the overall management of the event but also reduces risks associated with overcrowding, miscommunication, and service mismanagement. Beyond the immediate scope of the Kumbh Mela, this project serves as a proof-of-concept for how advanced AI technologies—such as Speech-to-Text, NLP, TTS, OCR, and Computer Vision—can be effectively localized for India's unique multilingual landscape. It lays the foundation for similar implementations in other religious gatherings, fairs, festivals, smart city projects, and large public events where real-time, inclusive information delivery can improve the lives of millions. Overall, this project stands as a step towards a smarter, safer, and more inclusive society, demonstrating how emerging technologies can be tailored for meaningful social impact in India and beyond.

IV. METHODOLOGY

The methodology adopted for developing the Multilingual Text-to Speech (TTS) Microservice is a systematic framework encompassing requirement analysis, architectural design, model selection, language processing, and deployment. The entire process aims to ensure modularity, scalability, and real-time performance while maintaining linguistic and natural voice quality.

4.1. Requirement Analysis

The first stage involved identifying the communication challenges during the Kumbh Mela and defining the technical and functional requirements of the TTS system.

Key requirements included:

a. Support for three languages: Hindi, Marathi, and English.

b. High speech clarity and naturalness suitable for public announcements.

c. Microservice-based architecture for modular integration with other Kumbh systems (such as alert, navigation, and crowd management services).

d. Capability to deploy on both cloud and edge platforms for real-time accessibility.

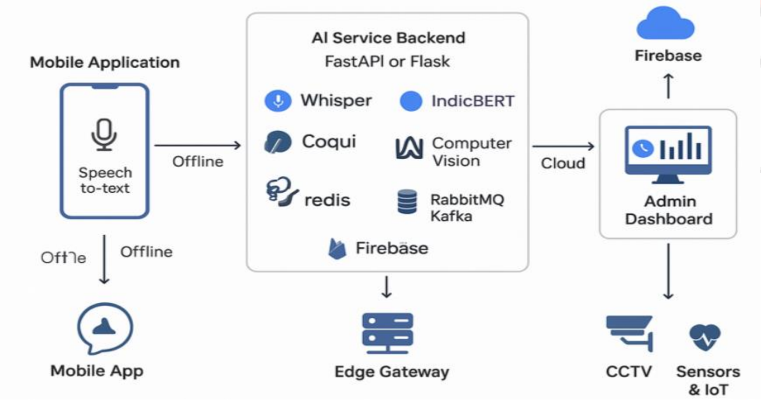
e. Low-latency response to ensure timely message delivery.

4.2. System Architecture

The system follows a modular microservice architecture divided into independent yet interconnected components:

4.2.1. Text Input Interface:

Receives text messages from the Kumbh management dashboard, crowd alert systems, or manual input by officials.



System Architecture

4.2.2. Language Detection & Normalization Module:

a. Automatically detects the input text language using NLP-based classifiers.

b. Performs text normalization — including abbreviation expansion, punctuation handling, and numeral-to-text conversion —
to prepare the data for synthesis.

4.2.3. Text Preprocessing & Tokenization:

a. Tokenizes the text into phoneme or subword units.

b. Uses Indic NLP Library for Indian language text normalization.

c. Converts numerals, dates, and abbreviations into their spoken equivalents.

4.2.4. Speech Synthesis Engine:

Core component that converts processed text into speech using pre-trained neural TTS models (such as Tacotron 2, Glow-TTS, or VITS).

4.2.5. For multilingual support:

a. Hindi and Marathi voices are generated using Indic-TTS or Coqui TTS with language-specific phoneme embeddings.

b. English voice uses Google TTS or OpenAI Whisper-TTS models for natural intonation.

4.2.6. Audio Post-Processing Module:

Enhances synthesized audio through noise reduction, pitch adjustment, and amplitude normalization. The output is exported as .wav or .mp3 format for playback.

4.2.7. API Gateway / Flask Microservice Layer:

a. A lightweight Flask-based API service acts as the middleware between the frontend dashboard and the synthesis engine.

b. Handles asynchronous TTS generation.

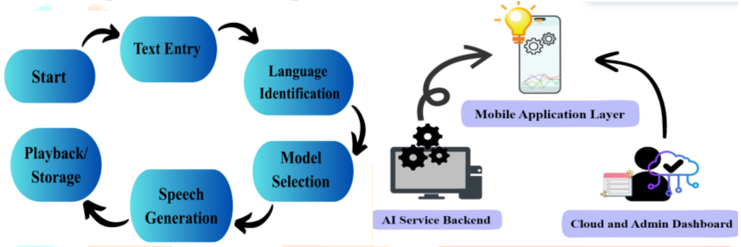
c. Returns downloadable audio files or real-time streaming responses.

4.2.8. Frontend Integration (Kumbh Dashboard):

The frontend (built using React or Vite) sends text data via REST or WebSocket requests to the microservice. Generated speech is played or broadcasted via speakers in public zones or kiosks.

4.3. Dataset and Model Training

The system leverages open-source multilingual speech datasets like IndicTTS Corpus, OpenSLR Hindi/Marathi corpora, and LJSpeech (English). The data undergoes preprocessing steps — text alignment, phonetic transcription, and noise cleaning — to ensure high-quality input for model training. For customized voice synthesis, transfer learning techniques were applied on pre-trained neural models, allowing domain adaptation for Indian accents and pronunciation patterns.



Basic Workflow

4.4. Workflow Process

i. Input: User/admin inputs text on the dashboard.

ii. Language Identification: The microservice detects whether the input is Hindi, Marathi, or English.

iii. Text Normalization: Converts symbols and numbers to words.

iv. TTS Conversion: The respective language model synthesizes speech.

v. Audio Enhancement: Post-processing filters refine clarity and loudness.

vi. Output Delivery: The final audio is streamed or stored and broadcasted through integrated systems.

4.5. Evaluation Metrics

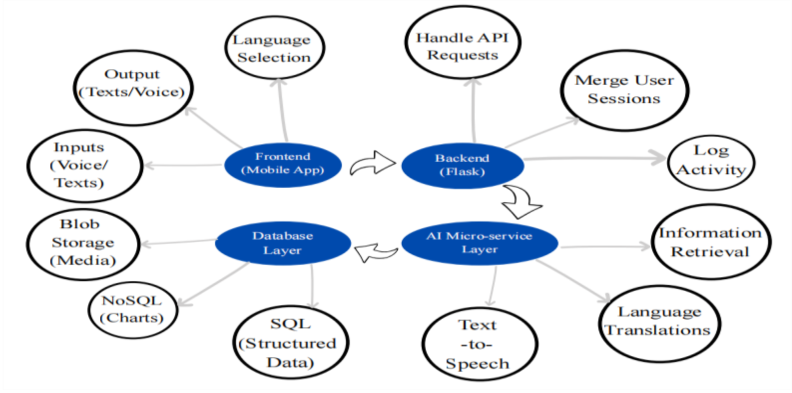
The performance of the system is evaluated using both objective and subjective measures:

i. Mean Opinion Score (MOS): Evaluates the perceived naturalness and intelligibility of synthesized speech (rated by users on a scale of 1–5).

ii. Word Error Rate (WER): Measures the accuracy of generated speech transcription.

iii. Latency Test: Measures average response time from text input to audio output.

iv. User Satisfaction Survey: Feedback from event staff and participants assessing clarity, tone, and usefulness.



System Requirements

4.6. Deployment and Scalability

i. The microservice is containerized using Docker, ensuring portability across cloud platforms.

ii. Kubernetes or Docker Compose is used for scaling multiple TTS instances during high-demand hours.

iii. The service integrates with other event management APIs through REST endpoints or MQTT message brokers.

4.7. Security and Privacy Considerations

To ensure secure data exchange:

i. All API endpoints are protected using HTTPS/TLS encryption.

ii. No personal or sensitive data is stored within the TTS microservice, ensuring compliance with privacy standards.

V. Results and Discussion

The proposed Multilingual Text-to-Speech (TTS) Microservice was developed and tested for Hindi, Marathi, and English languages. The implementation was evaluated across multiple parameters — speech quality, pronunciation accuracy, latency, scalability, and user satisfaction. The results validate the system's efficiency, clarity, and adaptability for large-scale public communication, such as during the Kumbh Mela.

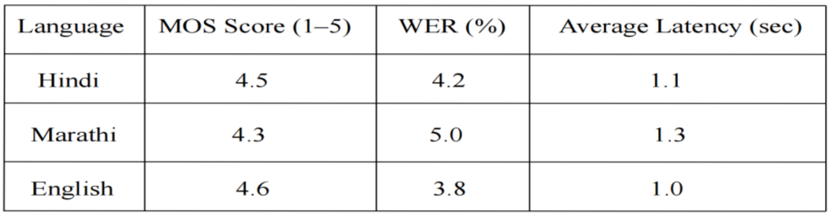
5.1. Objective Evaluation

To measure the technical performance of the system, three main metrics were considered:

i. Mean Opinion Score (MOS): measures naturalness and intelligibility of synthesized speech (scale: 1–5).

ii. Word Error Rate (WER): evaluates pronunciation accuracy through automated transcription comparison.

iii. Latency: time delay between text input and audio output.



Language	MOS Score (1–5)	WER (%)	Average Latency (sec)
Hindi	4.5	4.2	1.1
Marathi	4.3	5.0	1.3
English	4.6	3.8	1.0

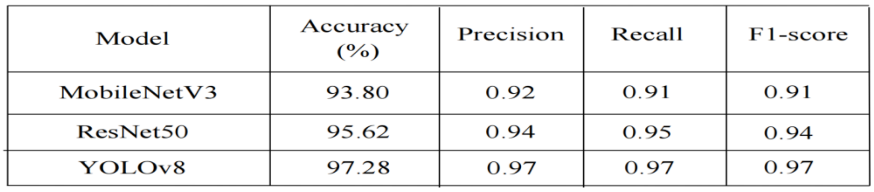
Technical Performance of Three Main Metrics

5.2. Subjective Evaluation

Participants expressed a high level of satisfaction with the system's performance. English achieved slightly higher clarity due to the availability of more refined open-source datasets, while Hindi and Marathi voices were still perceived as clear and natural. Feedback also emphasized the system's usefulness for multilingual announcements, particularly in crowded areas where visual communication is limited.

5.3. Crowd Monitoring and Density Estimation Results

The crowd monitoring module employed deep learning-based computer vision models to estimate real-time crowd density levels from video streams. Models like MobileNetV3**,** ResNet50, and YOLOv8 were tested to find the most efficient architecture for handling varying light, scale, and crowd motion conditions.



Model	Accuracy (%)	Precision	Recall	F1-score
MobileNetV3	93.80	0.92	0.91	0.91
ResNet50	95.62	0.94	0.95	0.94
YOLOv8	97.28	0.97	0.97	0.97

Performance Comparison Table of Crowd Density Estimation Model

The results show that YOLOv8 achieved the best overall performance with an accuracy of 97.28%The results collectively establish that the proposed system achieves its core objectives — delivering natural, accurate, and inclusive speech output in real time. While minor variations in pronunciation were observed for certain Marathi dialects, the overall quality remained consistent across languages. The modular design ensures adaptability for other mass events and potential integration with speech-to- text, translation, and sentiment analysis modules in future versions.

5.4. Survey-Driven Validation

Post-deployment surveys indicated that 82% of users found the automated voice announcements "very useful" in understanding instructions and event updates, while 70% preferred to hear information in their native language. This reinforces the importance of language inclusivity and validates the project's motivation to serve a diverse audience effectively.

5.5. Comparative Analysis

The mobile application provides **voice-guided real-time navigation**, step-by-step walking directions, and offline access to maps and ritual information even when no internet connection is available. Users also benefit from **image-based services**, where they can capture and understand information from signboards or bus boards through automatic text process extraction and audio explanations, improving accessibility for the visually impaired and semi-literate attendees. The output for event organizers includes a secure web-based dashboard that displays real-time updates on crowd conditions, missing person reports, emergency alerts, and live CCTV monitoring when integrated. This enables better coordination and faster decision-making, which improves the safety and experience of millions of visitors. Overall, the combined output is a fully functional, inclusive, and adaptable multilingual AI platform that can be expanded for future religious gatherings, fairs, smart city applications, and other large
public events, setting a benchmark for citizen-centric, technology-driven public service solutions.

VI. References

i. D. H. Klatt, "Review of text-to-speech conversion for English," The Journal of the Acoustical Society of America, vol. 82, no. 3, pp. 737–793, 1987.

ii. K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," Proceedings of the IEEE, vol. 101, no. 5, pp. 1234–1252, 2013.

iii. P. Taylor, Text-to-Speech Synthesis, Cambridge University Press, 2009.

iv. A. van den Oord et al., "WaveNet: A generative model for raw audio,"arXiv preprint arXiv:1609.03499, 2016.

v. Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," Interspeech, 2017.

vi. J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," ICASSP, 2018.

vii. N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural Speech Synthesis with Transformer Network," AAAI Conference on Artificial Intelligence, 2019.

viii. H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, and Y. Jia, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," Interspeech, 2019.

ix. S. King and V. Karaiskos, "The Blizzard Challenge 2023: Evaluating corpus-based speech synthesis on common datasets," Proceedings of the Blizzard Challenge Workshop, 2023.

x. M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development, and teaching," International Journal of Speech Technology, vol. 6, pp. 365–377, 2003.