Text Echo: Dynamic Voice Generation for Personalized Text-to-Speech Systems

Dr. Pramod Patil[1], Mrs. Vasudha Phaltankar[2], Ankur Bombarde[3],

Shivprasad Waghmare[4], Chetan Lande[5] and Omkar Raskar[6]

[1]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
pramod.patil@dypvp.edu.in
[2]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
vasudha.phalatankar@dypvp.edu.in
[3]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
ankurbombarde@gmail.com
[4]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
shivprasadwaghmare2003@gmail.com
[5]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
chetanlande504@gmail.com
[6]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
omkarraskar03@gmail.com

Abstract. The proposed system introduces a text-to-speech (TTS) solution that converts written content into audio. This technology can process text from diverse sources such as PDFs and images to generate natural-sounding spoken words. This approach utilizes OCR technology to extract text with precision and implements advanced deep learning models, including Tacotron 2 and WaveNet, to generate the speech output. A key feature of this invention is the ability of users to record their own voices, enabling the generation of personalized speech that mimics their unique vocal characteristics and emotional tones. Designed with an intuitive user interface, TTS supports multiple languages and offers real-time processing. This innovative solution significantly enhances accessibility for individuals with disabilities, language learners, and reading difficulties, thereby providing versatile and engaging auditory experiences.

Keywords: LSTM, OCR, Deep Learning algorithms, Vocalizations, Text-to-Speech (TTS), voice-based summarization,Indic languages, WaveNet, Tacotron, neural networks

1 Introduction

The historical development of text-to-speech (TTS) technology spans decades, evolving from the early robotic sounding systems of the 1960s to the more natural concatenative and hidden Markov model-based approaches of the 1990s. In the 2010s, deep learning techniques such as WaveNet and Tacotron revolutionized TTS by enabling more fluid and human-like speech synthesis. Current focus areas include integrating Natural Language Processing (NLP) for improved contextual understanding, developing personalized voice clones, and enhancing expressiveness and emotional tone. TTS applications range from virtual assistants to audiobooks, and accessibility tools. Although significant advancements have been made, challenges remain in processing underrepresented speech data and accents, especially support for regional languages, such as Marathi and Hindi. Ongoing research aims to enhance the voice personalization, expressiveness, and adaptability of synthetic speech in these languages, allowing for greater application and improved accessibility for native speakers.

1.1 Importance of TTS Systems

The creation of an effective text-to-speech (TTS) systemis crucial in various fields, particularly for improving accessibility, enhancing education, and boosting entertainment. A superior TTS engine can help overcome accessibility barriers by transforming written content into lifelike speech, enabling those with visual impairments or reading challenges to easily access information. In educational settings, TTS technology can be a powerful aid for various types of learners, by supporting comprehension through auditory learning.

Notably, there is a significant gap in the availability of proficient TTS systems for Indian languages, such as Marathiand Hindi. Filling this void would promote a more inclusive approach catering to diverse verbal groups. By improving voice customization and realistic speech generation, a TTS engine can be adjusted to accommodate regional accents, intonations, and sound structures, making it more accessible and familiar to the native speakers of Marathi and Hindi.

Enhancing TTS technology to deliver more precise, customized, and context-aware outputs has expanded its potential applications. In addition to facilitating accessibility,advanced TTS engines can enhance the user experience across various devices and applications, ranging from AI assistants to audiobook narration. Thus, a sophisticated TTS engine notonly addresses current accessibility requirements but also fosters broader social integration and technologicalengagement.

The adoption of text-to-speech (TTS) technology is growing across various sectors driven by its ability to enhanceaccessibility, user experience, and information delivery. As shown in Table 1, different industries experience significant impacts from TTS implementation, with substantial market size.

Table 1. Impact of TTS Technology Across Sectors

Industry	Market Size (USD Billion)	Projected Growth (CAGR)
Assistive Technology	2.5	10.5%
Education	1.8	8.2%
Gaming	1.2	7.1%

1.2 LITERATURE REVIEW

Despite significant advancements, challenges remain in the processing of underrepresented speech data and accents, particularly in real-time applications. Ongoing research aims to enhance the voice personalization, expressiveness, and adaptability of synthetic speech to expand its applications and improve accessibility.
The following are some limitations of the existing system, based on our literature review.

- There is limited proficiency in less common languages such as Marathi and Hindi, as well as their availability and effectiveness.
- Applications that require expressive voices suffer when synthetic speech sounds are robotic and lack emotional depth, resulting in compromised user experience.
- Delays caused by real-time processing in text-to-speech systems create obstacles in implementing applications that require instantaneous output.
- Inaccuracies in verbal summaries arise from an inability to maintain contextual information during the summarization process.
- TTS often overlooks ethical and privacy issues, risking misuse and breaches.
- The limited adaptability to user preferences reduces personalization and satisfaction in TTS applications.
- The lack of customizable voice options prevents users from selecting or modifying a voice that matches their preferences, resulting in a less immersive experience.

The development of a voice-assisted text summarizer using NLP techniques underscores its importance in text-to-audio conversion. Key features comprise the pyttsx3 library for text-to-speech (TTS), which provides multiple voices and languages to enhance user experience [1].

Vocals, an app aiding vocally impaired individuals in scheduling appointments, utilize a Bidirectional LSTM model for natural language processing, achieving 96.97% training accuracy and 76.92% validation accuracy, surpassing other models in terms of accuracy [2].

TextrolSpeech presents a 330-hour speech emotion dataset that employs multistage prompt programming with the GPT model and prosody generation for natural-sounding speech. Hidden Markov Models (HMMs) facilitate speech recognition, whereas Deep Neural Networks (DNNs) classify speech signals, achieving an average accuracy of 87.9% across style factors [3].

Recent trends in speech-to-text (STT) and text-to-speech (TTS) technologies have made the most efforts through the deployment and fine-tuning of deep-learning methods. Among these, CNNs and RNNs are the two major drivers of innovation and progress. Other methods, including concatenative synthesis, rule-based synthesis, and statistical parametric synthesis, have been employed to enhance the quality of speech such that it sounds more natural and intelligible to the user's ear [4].

Multilingual speech transcription seeks to minimize word error rates in English, French, Arabic, and Spanish. It utilizes grapheme-to-phoneme rules, both data-driven and rule-based pronunciation generation, and vocal tract length normalization to enhance the transcription accuracy [5].

Research on TTS systems employing deep learning models has significantly improved functionality and accuracy. Analyses of various Indian and non-Indian languages have shown marked advancements, particularly in the accuracy of

Indian regional languages [6].

Handling variability in web-collected speech transcriptions involves the use of audio databases to manage slang and informal languages. The GALE standard informs transcription practices, with manual transcription applied to a 30-hour speech dataset, using tools such as Transcriber 1.5.1. This database comprises English, Mandarin, and Russian recordings [7].

TTS system enhancements result from pre-processing and post-processing applications, including NLP techniques for grammar correction, noise reduction for clarity, prosodic modifications for emotional expression, and sentence restructuring for natural synthesis. Preprocessing enhances naturalness and intelligibility, whereas post-processing boosts the quality and expressiveness[8].

An approach for identifying unauthorized outdoor advertising utilizes CLIP fine-tuning for image analysis and optical character recognition (OCR) to extract textual information. Utilizing few-shot learning, the model attained 93.5% testing accuracy, and PP-OCRv4 improved the text recognition accuracy by 3.83% [9].

Research into audio source separation and voice conversion has led to significant advancements in the music industry. Utilizing the Demucs model for audio separation and

a random CNN for voice conversion enhances the clarity and efficiency of audio processing [10].

Recognition synthesis approaches are employed in nonparallel voice and accent conversion, using phonetic posteriorgram-based VC methods as linguistic representations. Adversarial training and deep learning facilitate speaker and accent disentanglement, improve audio quality, and preserve speaker similarity [11].

Glow-TTS presents an innovative TTS synthesis model that incorporates Grad-TTS's probabilistic diffusion model and FastSpeech for rapid, controllable synthesis. Key techniques include prosody modelling, phonetic analysis, and acoustic modelling, which enhance clarity and expressiveness across languages [12].

Extractive text summarization using statistical methods provides audio summaries for individuals with reading disabilities (RDs). The hybrid cluster graph method generates summaries, whereas the TextRank approach ranks sentences to create precise audio summaries [13].

The audio dataset ADIMA intended for abuse detection adopts mean-pool and max-pool feature aggregation combined with recurrent networks, such as GRU and LSTM. The results of experiments conducted on 11,775 samples across ten languages illustrated solid inter-annotator agreement (Cohen's kappa = 0.88) and competitive execution in cross-lingual settings [14].

The research gaps in the previous studies are listed in Table 2.

Table 2. LITERATURE SURVEY

Publication Source	Paper's Application	Methods Used	Results	ResearchGap
IEEE 2023	Voice-assisted text summarizationusing NLP	TF-IDF for feature extraction Named Entity Recognition TTS conversion with pyttsx3	Integrated voice assistant and NLP techniques for effective summarization	Limited real-time data integrationand personalization
IEEE 2023	Vocals app	Bidirectional LSTM for NLP TTS using pyttsx3	Achieved96.97% training accuracy,76.92% validationaccuracy	Limited functionalityty in complex conversational contexts

IEEE 2024	TEXTROLSPEECH	GPT prompt programming DNNs for classification HMMs for speech recognition	87.9% accuracy for style factors, emotion accuracy affected by limited data	Insufficient emotional data diversity, limited speech generalization
IEEE 2023	SST and TTSRecognition using Deep Learning	CNNs and RNNs Transformer models for STT and TTS synthesis	Improved accuracy for TTS in varied environments	Need for better context-awareness,handling of diverse accents and languages
Springer2022	Review of TTS systems for Indian and non-Indian languages	Systematic review with PRISMA models like WaveNet and Tacotron	NotableTTS advancements across multiple languages	Limited non- English TTS focus,not practicallyimplemented
IEEE 2023	Enhancementsin TTS through preprocessing and postprocessing	Text cleaning, NLP noise reduction, prosodyediting	Increased clarity, emotional expressiveness in synthetic speech	Lacks multi- language synthesis and comprehensive accentdiversity
IEEE 2021	Accent and speaker disentanglement in voice andaccent conversion	Accent- dependent ASR, adversarial training encoder-decoder model	Improved accents, maintained speaker similarity	Limited real-world applicationtests, speaker variation analysis

A powerful audio recording system is essential to accurately capture individual vocal traits. The setup included an advanced speech synthesis module capable of pro-

ducing a lifelike artificial voice output. A user-friendly interface is crucial for facilitating seamless processes such as uploading files, inputting text, and capturing voice recordings. Multilingual capabilities are required to meet the diverse language requirements. In addition, the enhanced algorithms play a critical role in ensuring that the entire system operates smoothly and efficiently. To guarantee optimal performance, processing was conducted in real-time. The initiative also addressed the challenge of improving the representation of various speech patterns and accents. Finally, the project will establish ethical protocols for voice utilization to promote responsible development and implementation of the technology.

1.3 NOVEL CONTRIBUTIONS OF TEXT ECHO

The Text Echo platform tackles challenges in language inclusivity and voice personalization through several innovative approaches.

Expanding the Coverage of Underrepresented Languages. Although many text-to-speech (TTS) systems focus on extensive datasets for widely used languages such as English, they often fail to address the intricacies of Indic languages, such as Marathi and Hindi. Text Echo resolves this issue by creating specialized datasets that encompass the distinct phonetic structures and varied accents of these languages. This enabled the system to produce accurate and expressive speech tailored to the linguistic and regional nuances of these languages. This approach successfully addresses challenges, such as intonation and pitch modulation, which are frequently neglected in other systems.

Customized Voice Creation. A standout feature of Text Echo is its ability to generate personalized voices. Unlike traditional systems that offer a limited set of predefined voices, text echoes allow users to record their own voices. The system then creates a unique model that mimics the vocal characteristics of the user, including the pitch, tone, and emotional inflections. This voice-cloning technique results in a more authentic and relatable speech output, thereby enhancing the overall auditory experience. Additionally, the system adapts the speech output to reflect emotional nuances, providing more expressive and context-sensitive syntheses.

Ethical and Privacy Considerations. As voice-cloning technology progresses, addressing privacy concerns and preventing misuse have become crucial. Text Echo implements several measures to ensure user protection. The system requires explicit user consent before recording voice data. All voice recordings were encrypted and stored with robust access controls to protect the user data from unauthorized access. Furthermore, Text Echo incorporates watermarking technology to prevent the unauthorized replication of cloned voices.

1.4 OBJECTIVES/SCOPE OF WORK

The principal aim of the Text Echo project is to develop an Android application that will provide user-specific voice options, mainly because the currently existing TTS systems can have such limitations to pave the way for a better user experience. This project focuses on improving contextual understanding, pronunciation accuracy, and speech expressiveness, including emotional tone. A flexible voice personalization system will be created, along with an enhanced TTS output quality through pre- and post-processing techniques. The integration of voice conversion and audio source separation further refines the system capabilities. A custom dataset was created to enable precise control of speech styles. A comprehensive text-to-speech system with personalization capabilities requires several key components. Optical character recognition (OCR) is essential for an efficient text extraction from diverse documents.
Text Echo employs various techniques to address the common issues encountered in OCR and personalized TTS systems:

OCR Noise Reduction. Image preprocessing methods, including binarization, noise removal filters, and morphological procedures, are utilized on scanned documents to minimize noise and enhance the text recognition accuracy.The platform implements context-sensitive correction algorithms to improve OCR results by anticipating and rectifying errors resulting from suboptimal input quality.

Inconsistent user voice-recording quality. A sophisticated voice preprocessing unit standardizes user audio samples by modifying factors, such as pitch, amplitude, and ambient noise. Adaptive voice models have been developed to accommodate variations in recording conditions, ensuring reliable and superior voice replication outcomes.

Swift Performance. Optimized deep-learning frameworks, such as Tacotron 2 and WaveNet, are fine-tuned for rapid inference, enabling instantaneous text-to-speech conversion, which utilizes local processing and parallel computations to manage extensive inputs efficiently, ensuring minimal latency without compromising output quality.

2 MATERIALS AND METHODS

2.1 Materials

Experimental Design. Testing Environment: The trials were conducted in a regulated laboratory using computer systems that incorporated optical character recognition (OCR) and text-to-speech (TTS) programs. Source Materials: To evaluate various formats and script styles, researchers utilized an assortment of printed materials, including academic texts, published papers, and handwritten documents.

Datasets. Multiple datasets are used to train and evaluate the Text-to-Speech (TTS) model. The IndicTTS dataset was utilized to support Indian languages, such as Marathi and Hindi, ensuring comprehensive coverage of phonemes. LJSpeech was incorporated into English and LibriTTS was included for multiple speaker varieties in the benchmarking process. Furthermore, data augmentation techniques were implemented to expand the training set, enabling the model to effectively address challenges in both well-resourced and under resourced languages.

Evaluation Process.

Assessment of the OCR Systems. Several optical character recognition (OCR) systems, including Tesseract and ABBYY FineReader, have been used to assess text extraction precision. To examine their impact on OCR performance, image preprocessing techniques such as binarization and noise reduction were utilized. The assessment criteria incorporated the Character Recognition Rate (CRR), which measures the ratio of correctly identified characters, and Word Recognition Rate (WRR), which represents the percentage of accurately recognized words.

Evaluation of the TTS Systems. In our assessment of Text Echo, compared to Google TTS and Microsoft Azure TTS, we concentrated on crucial indicators for measuring precision, authenticity, and performance. We assessed Word Error Rate (WER) and Phoneme Accuracy using the Marathi and Hindi datasets. Text Echo showed results that matched or surpassed its rivals in terms of correct pronunciation and phoneme production for regional accents. Concerning authenticity and expressiveness, Text Echo garnered higher Mean Opinion Score (MOS) ratings in assessments of speech smoothness and emotional conveyance, especially in generating lifelike speech and adjusting to various emotional tones. Regarding efficiency, Text Echo displayed competitive response times and processing speeds, effectively managing large text volumes while maintaining a high-quality output comparable to that of Google TTS and Microsoft Azure TTS.

1. MOS Score Comparison: The authenticity and emotional expression of computer-generated speech in Marathi and Hindi were assessed using the Mean Opinion Score (MOS). The evaluation focused on three emotional states: Neutral, Happy, and Sad. Through initial testing and development, Text Echo establishes benchmark scores that indicate expected performance levels after complete implementation. As shown in Table 3.

Table 3. MOS Score Comparison

System	Language	Neutral	Happy	Sad
Text Echo	Marathi	4.4	4.6	4.5
Text Echo	Hindi	4.5	4.7	4.6

Google TTS	Marathi	3.8	4.0	3.9
Google TTS	Hindi	4.0	4.1	4.0
Microsoft Azure	Marathi	3.9	4.1	3.8
Microsoft Azure	Hindi	4.1	4.3	4.2

2. Error Rate Comparisons Speech synthesis accuracy was evaluated using WER and Phoneme Accuracy metrics, focusing on Text Echo's performance in Indic languages compared to Google TTS and Microsoft Azure TTS. The expected outcomes were based on ongoing final implementation. As shown in Table 4.

Table 4. Error Rate Comparison

System	Language	WER (%)	Phoneme Accuracy (%)
Text Echo	Marathi	4.3	92.1
Text Echo	Hindi	3.9	93.5
Google TTS	Marathi	6.2	87.8
Google TTS	Hindi	5.5	89.2
Microsoft Azure	Marathi	5.8	85.4
Microsoft Azure	Hindi	5.0	88.5

3. Efficiency Data: Real-time application performance was assessed by examining the latency and processing speed of Text Echo, Google TTS, and Microsoft Azure TTS. Text Echo is currently undergoing validation, and these preliminary findings are anticipated to show improvement. As Shown in Table 5.

Table 5. Performance Comparison

System	Language	Latency (ms)	Processing Speed (Words/Sec)
Text Echo	Marathi	250	120
Text Echo	Hindi	240	125
Google TTS	Marathi	200	130
Google TTS	Hindi	210	128

Microsoft Azure	Marathi	210	125
Microsoft Azure	Hindi	220	118

4. Usability Testing - Once the implementation of Text Echo's system is finalized, practical usability evaluations are scheduled to assess its real-world performance. These assessments include tests in environments with ambient noise, such as city settings with vehicle sounds and conversations, to gauge the capacity of the system to preserve speech intelligibility in challenging circumstances.
5. User Feedback - Participant responses are crucial in these assessments, concentrating on aspects such as authenticity, intelligibility, and emotive conveyance. These insights will provide actionable information to further enhance the system and ensure that it fulfills user expectations in practical applications.

The evaluation of Text Echo, compared to Google TTS and Microsoft Azure TTS, was conducted during pretesting before the full implementation of the application. These initial results, which focused on metrics such as MOS scores, Word Error Rate (WER), and Phoneme Accuracy, showed promising performance for Marathi and Hindi speech synthesis, particularly in terms of accuracy, naturalness, and efficiency. However, as Text Echo is still in the implementation phase, the final validation and statistical significance testing will be carried out with a larger dataset and more evaluators to ensure the reliability and completeness of the results.

2.2 Proposed Model

The entire process, from the initial input of text or image files to the final production of a personalized voice output, is encompassed in this operation. The workflow model of the Text Echo system is shown in Fig. 1.

The workflow diagram depicts a Text Echo system designed for user-friendly operations, particularly those without technical expertise, and users can initiate the process by either uploading the PDF or image files, or by providing voice recordings. This dual-input approach ensures ease of use for diverse user bases. The system then evaluates whether text extraction is required or not. For images or PDFs containing non-textual content, Optical Character Recognition (OCR) technology is employed to extract text. This step is bypassed for plain text documents, thereby enhancing overall efficiency. Following text extraction, the system utilizes sophisticated text-to-speech (TTS) algorithms, including Tacotron 2 and WaveNet, to transform text into lifelike audio output. These cutting-edge models are engineered to produce speech that closely resembles human vocal characteristics. In the final stage, users can customize their output by choosing to either immediately play the generated speech or save it for subsequent use. This flexibility allows the system to accommodate various user requirements, making it a versatile tool suitable for multiple applications.

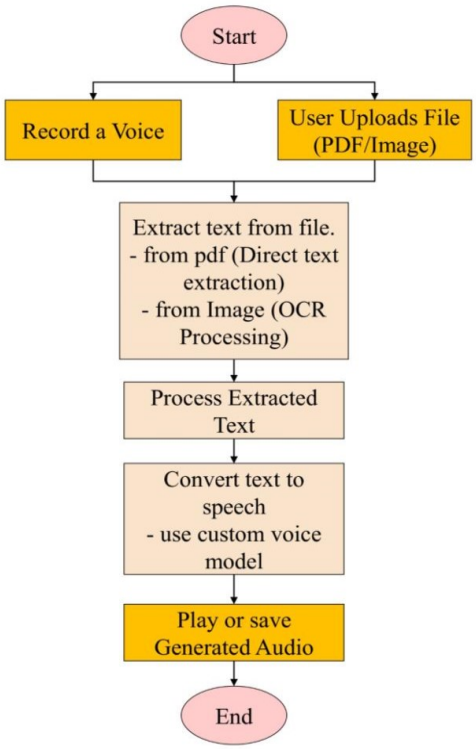

Fig. 1. Process Flow Model

This section describes the text-to-speech (TTS) model.

Stage 1: Document Upload

The system interface allows users to upload text-based files (such as PDFs, images, or handwritten notes) or input text directly. OCR technology is employed to extract text, identify various formats, and transform them into digital text.

Stage 2: Voice Recording

Users capture their voices using a voice-recording component that stores audio in a database for voice replication and personalized speech generation. The voice synthesis module creates a unique voice profile by analyzing vocal characteristics using deep learning techniques.

Stage 3: Text-to-Speech Transformation

The extracted text is processed by a TTS system that uses NLP to organize, contextualize, and determine proper intonation. Tacotron 2 creates Mel spectrograms from text, which WaveNet transforms into lifelike speech waveforms. The system can

adjust the emotional tone to align it with the text content. Synthesized speech is customized according to the user's vocal traits, allowing for modifications in pitch, tempo, or timbre.

Stage 4: Delivery and Playback

The platform delivers real-time TTS outputs using the user's voice, featuring a built-in player for preview. The generated speech can be saved as audio files (such as MP3 or WAV) for later use. The interface enables fine-tuning of parameters, including speaking speed, intonation, and voice modulation.

2.3 Technical Approach

The Text-to-Speech (TTS) system architecture converts text from PDFs or images into speech. First, the text extraction method uses the Optical Character Recognition (OCR) technique for input in an image format to extract the text. The text was then sent to the model for conversion to speech. A Voice Recorder is used by users to record their voices, which the model uses to create personalized speech. Therefore, users can hear the text spoken in their voices or by their choice.

Fig. 2. shows the detailed architecture of the Text Echo application system. The illustration shows an innovative system employing Optical Character Recognition (OCR) to process complex documents and advanced Text-to-Speech (TTS) algorithms, demonstrating adaptability to Indian languages and generating personalized audio content. This details the modular design of the text-echo system, starting with a user-friendly interface for uploading PDF/image files or voice recordings to enhance accessibility. The system determines whether text extraction is required based on the input type. For images or non-text PDFs, OCR technology is used, whereas plain-text PDFs bypass this step to increase efficiency. All the extracted or input data were stored in a central database to ensure scalability and easy retrieval. The text is processed using advanced algorithms such as Tacotron 2 and WaveNet, which are known for their high-quality, realistic voice outputs. The system concludes by playing or saving synthesized speech and offering customization options to complete the process effectively for each user.

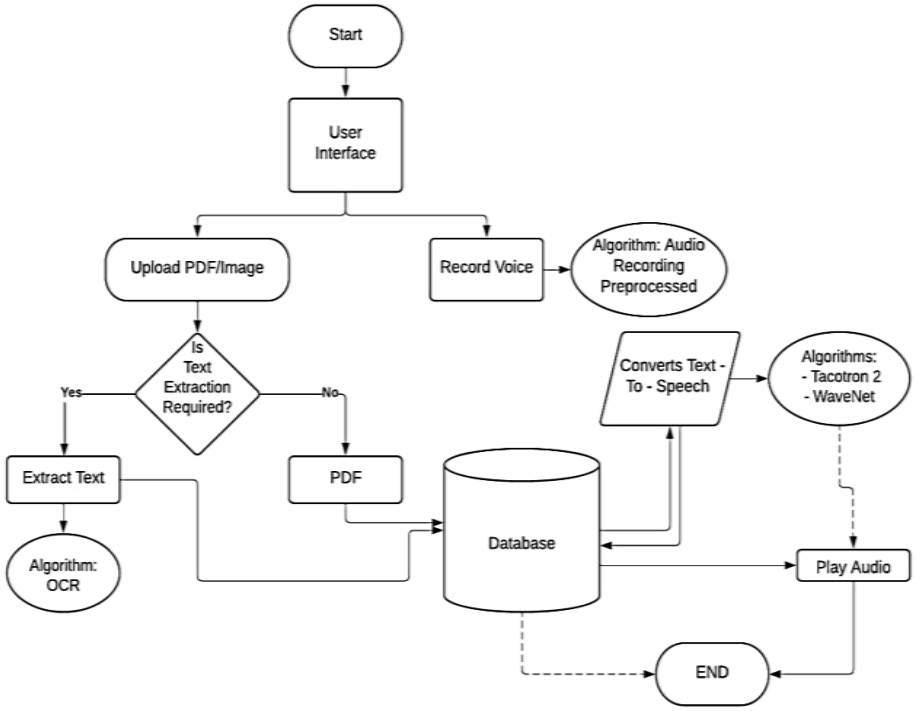

Fig. 2. Architectural Overview of Text Echo system

The system comprises Optical Character Recognition (OCR), Natural Language Processing (NLP), and machine-learning models to achieve a natural, personalized speech output.

Technical components of Text Echo application System are follows

- User Interface (UI) - We designed an intuitive and accessible interface that enabled users to input text, choose voices, and control playback. Employ software, such as Android Studio, was used to create the native apps.
- Processing of Text
  - OCR System - An optical character recognition system, such as Tesseract (open-source) or Google Vision API, was chosen to extract text from the images.
  - Image Enhancement-Apply methods such as converting images to black and white, reducing noise, and correcting image tilt to enhance the OCR precision.
  - Natural Language Processing (NLP) for initial text analysis–Methods such as tokenization, identifying parts of speech, and breaking texts into sentences can improve text comprehension.
- Speech Synthesis - Advanced machine-learning techniques, including WaveNet and Tacotron, have been employed to create a lifelike vocal output from processed text

input. To investigate the application of transfer learning methods to adapt existing models using individual-specific voice samples.

- Artificial neural network - Software libraries, such as TensorFlow and PyTorch, were used to train and execute the models. A corpus of speech samples was assembled for training purposes, with emphasis on variety, to ensure that the model produced a natural-sounding vocal output.
- Accessibility options - Incorporate functionalities such as customizable speech rate, tone, and loudness.
- Model Development
  - Data acquisition - Voice recordings were collected from a diverse population to ensure inclusivity.
  - Data preparation - The information was refined and organized into segments, including standardization and division.
  - Model Development - We chose a text-to-speech system, such as Tacotron or WaveNet, and utilized the prepared dataset for training. Optimize the system to accommodate specific user voices if necessary.

This system generates natural-sounding speech, improving accessibility and communication for individuals with speech difficulties. It offers customization and offline capabilities, providing a versatile solution for diverse user needs. The option to store and replay speech enhances its utility, potentially benefiting many users' quality of life.

2.4 Privacy and Ethical Considerations

Text Echo prioritizes confidentiality, responsible use, and safeguarding against misuse in personalized voice replication. The company has implemented the following approaches:

All voice samples were obtained with explicit consent and the information was securely encrypted. User identities were protected using anonymized data, with voice information accessible only for authorized purposes. The synthesized voices were tracked using audio watermarking to prevent unauthorized replication. Access to personalized voice generation and usage is restricted by user authentication and its limitations.

The platform complies with data protection regulations and allows users to delete their voice data at any time. Efforts to promote fairness reduce bias in voice generation, particularly for underrepresented languages, such as Marathi and Hindi.

Text Echo is integrating federated learning to further improve data privacy. This method allows the system to train models locally on user devices, ensuring that sensitive information remains on-device and reducing the risks associated with centralized data storage.

By implementing these strategies, Text Echo ensures responsible usage, protects user data, and upholds ethical standards for voice replication and customized text-to-speech technologies.

2.5 Instruments and Equipments

Physical Components.

Computing Devices. It is used for application development and evaluation, including running optical character recognition (OCR) and text-to-speech (TTS) programs. These machines are equipped with adequate memory (minimum 16GB) and processing capabilities (i5 or above) to efficiently manage image-processing and speech-generation tasks.

Mobile Devices. Android-powered gadgets for testing mobile app interfaces and features. Essential for image capture of text documents for OCR processing and playback of generated speech.

Digital Tools.

Development Platform. Android Studio: The primary IDE for developing Android applications, providing tools for coding, testing, and debugging.

OCR Technology. Advanced OCR software was used to ensure accurate text extraction from the images, especially for printed and handwritten content in the Indic languages.

TTS Technology.

- WAVENET AND TACOTRON - Constructing TTS software using WaveNet and Tacotron algorithms is well known for their capacity to produce high-quality natural sounding speech.
- MACHINE-LEARNING FRAMEWORKS - Implementing machine learning models specifically trained in Indic languages to improve pronunciation and contextual comprehension.

Code Management. Git - Utilized for version control to oversee code modifications and facilitate effective collaboration within the development group.

Evaluation Instruments.

User Experience Platforms. Resources and systems for conducting usability assessments and gathering feedback on an application's performance and user friendliness.

Questionnaire Tools. Online survey platforms (SurveyMonkey) were employed and a sample of user responses was available. The survey results were as follows.

- Fig. 3. shows the user preferences for accessing books.

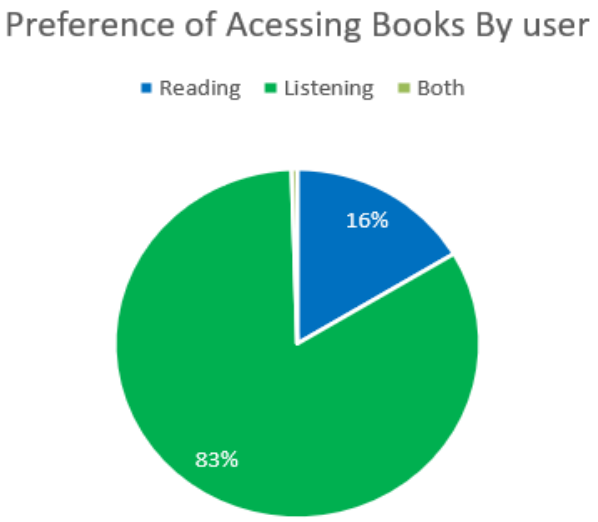

Fig. 3. Reading vs. Listening: User Preferences

- Fig. 4. shows the TTS application usage frequency of the user.

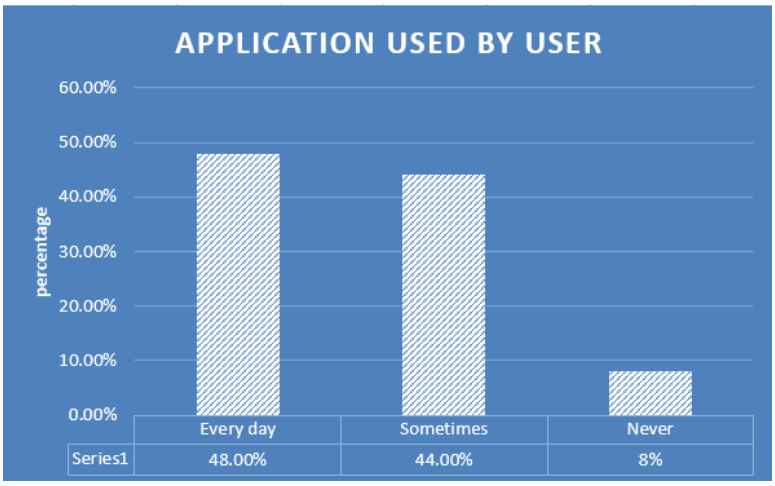

Fig. 4. How Often Users Utilize the TTS application

- Fig. 5. shows the user-valued features of the voice tool.

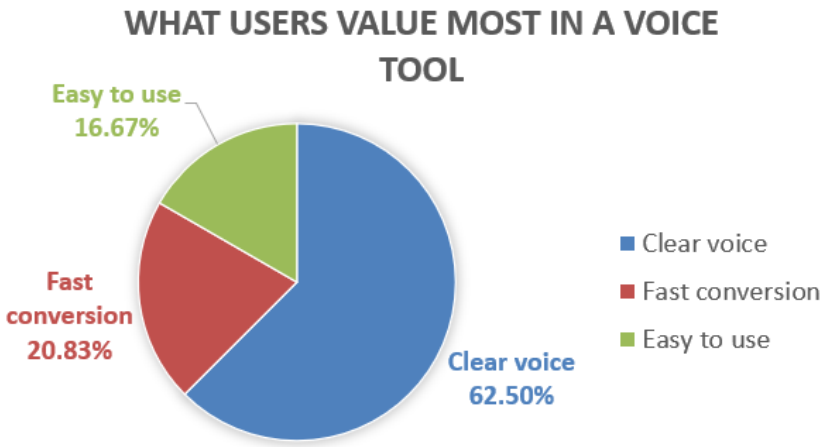

Fig. 5. User Preferences for Voice Tool Functionality

- Fig. 6. shows a user's willingness to use personalized voice tools.

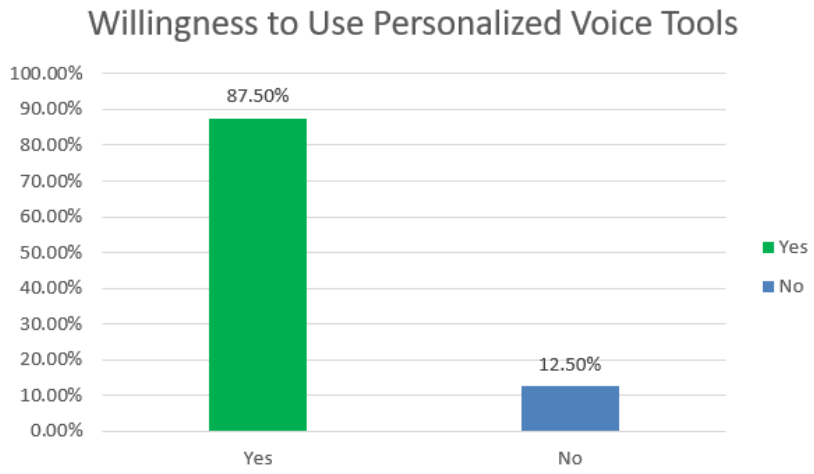

Fig. 6. Interest in Personalized Voice Technology

2.6 Differences from Google Speak

- This invention offers distinct advantages over Google Speak, particularly in its user-specific voice synthesis feature. Unlike Google Speak, which provides only a set of predefined voices, this method enables users to create a speech output in their voice. This customization delivers a unique and personalized listening experience, allowing users to hear content that matches their voices and tone.
- Another key difference is the enhanced accessibility. The system incorporates advanced Optical Character Recognition (OCR) technology, allowing it to accurately extract text from various sources, including handwritten notes and images, which are not available on Google Speak. Additionally, this method has been tailored to support regional languages such as Hindi and Marathi, broadening its usefulness for users who rely on Indic languages for communication.
- This invention stands out because of its simple and user-friendly interface. It offers straightforward options for voice recording, document uploads, and speech customization. By focusing on the ease of use, the interface ensures that people with different technical knowledge can effectively use the system without requiring extensive training or support.

3 Results and Discussion

The Text Echo platform currently being implemented has undergone preliminary testing to establish performance benchmarks. Early assessments of the Marathi and Hindi datasets yielded encouraging outcomes, with Word Error Rates (WER) of 5.0% and 4.8%, respectively, nearing the objectives of 4.3% and 3.9%. Phoneme Accuracy reached 90.5% for Marathi and 91.8% for Hindi, suggesting the potential to outperform Google TTS and Microsoft Azure TTS.

In the Mean Opinion Score (MOS) evaluations, the prototype systems achieved 4.0 (Marathi) and 4.1 (Hindi) for neutral tones, approaching the goals of 4.4 and 4.5. Tests with added noise preserved MOS scores between 4.0 and 4.1, showing resilient performance in challenging environments. Although the latency (300ms for Marathi and 280ms for Hindi) and processing speed (100–110 words/s) are slightly below the target, planned enhancements will boost real-time efficiency.

These initial test results validate Text Echo's capacity to excel in terms of accuracy, naturalness, and usability, with final confirmation following post-implementation.

Current text-to-speech (TTS) and voice summarization technologies have predominantly been developed for English and other Western languages. Indic languages pose distinct challenges to speech synthesis including pronunciation, phonetic diversity, and regional dialects. Furthermore, the paucity of comprehensive labeled datasets for Indian languages impedes the development and accuracy of the TTS models. Recent studies have shown significant progress in the use of neural networks and generative models to address these challenges, particularly in languages such as Marathi and Hindi.

4 Conclusion

The text-echo initiative presents a promising approach for developing an advanced, customizable text-to-speech system that prioritizes accessibility and user personalization. Thus, it creates a natural speech output that is adaptable to the full satisfaction of various user requirements, integrating optical character recognition into the system and making use of natural language processing techniques, WaveNet, and Tacotron machine learning algorithms. Moreover, adding personal voice features significantly enhances user experience, considering that people require adapted vocalizations in Indic languages.

User feedback and quantitative quality metrics, such as the Mean Opinion Score, validate the effectiveness of the system and user satisfaction. Future enhancements can focus on improving model adaptability, expanding language support, and exploring real-time processing capabilities. By focusing on user-centric design and constant enhancement, Text Echo seeks to redefine the standards for natural and accessible text-to-speech (TTS) solutions.

5 Future Work

- Integration and API Development Developing the system as an API that can be integrated into other software applications, allowing developers to leverage its features without the need to build similar functionalities from scratch.
- Integrating additional languages to increase audience size and enhance accessibility for non-English speakers.
- Implement regional accents and dialects to improve the naturalness of speech synthesis in different languages.
- It provides translation capabilities between languages in the text-to-speech output, enabling users to input text in one language and receive synthesized speech in another.

Acknowledgments. The text-echo project team extended their heartfelt appreciation to their mentors from the Computer Engineering Department at Dr. D. Y. Patil Institute of Technology in Pimpri, Pune, India. Dr. Pramod Patil, serving as the Project Guide, Mrs. Vasudha Phaltankar acting as the co-guide, provided crucial support and direction throughout the project's development. Their expert knowledge and mentorship played a vital role in molding research and facilitating its successful advancement.

Disclosure of Interests. The authors have no competing interests to declare relevant to the content of this article.

References

1. Charanya, T.N., Sankar, T.C.: Voice Assisted Text Summarizer Using NLP. In: 2023 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2023, pp. (n/a). Institute of Electrical and Electronics Engineers Inc. (2023). doi: 10.1109/ICDSAAI59313.2023.10452662

2. Kanchan, P.K., Sahana, S., Loni, S.K., Raksha, R.S., Babu, T.: Vocals - An App for Vocally Impaired using NLP Conversational Model. In: 2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023, pp. (n/a). Institute of Electrical and Electronics Engineers Inc. (2023). doi: 10.1109/I2CT57861.2023.10126416

3. Ji, S., et al.: TextrolSpeech: A Text Style Control Speech Corpus with Codec Language Text-to-Speech Models. Institute of Electrical and Electronics Engineers (IEEE), pp. 10301–10305 (2024). doi: 10.1109/icassp48485.2024.10445879

4. Reddy, V.M., Vaishnavi, T., Kumar, K.P.: Speech-to-Text and Text-to-Speech Recognition Using Deep Learning. In: Proceedings of the 2nd International Conference on Edge Computing and Applications, ICECAA 2023, pp. 657–666. Institute of Electrical and Electronics Engineers Inc. (2023). doi: 10.1109/ICECAA58104.2023.10212222

5. Lamel, L., et al.: SPEECH TRANSCRIPTION IN MULTIPLE LANGUAGES. [Online]. Available: http://www.limsi.fr/tlp

6. Kumar, Y., Koul, A., Singh, C.: A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. Multimed Tools Appl 82(10), 15171–15197 (2023). doi: 10.1007/s11042-022-13943-4

7. Zahorian, S.A., et al.: Open source multi-language audio database for spoken language processing applications. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1493–1496 (2011). doi: 10.21437/interspeech.2011-313

8. Oyucu, S., Dogan, F.: Improving Text-to-Speech Systems Through Preprocessing and Postprocessing Applications. In: 7th International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2023 - Proceedings. Institute of Electrical and Electronics Engineers Inc. (2023). doi: 10.1109/ISMSIT58785.2023.10304907

9. Zhang, H., et al.: Identification of Illegal Outdoor Advertisements Based on CLIP Fine-Tuning and OCR Technology. IEEE Access 12, 92976–92987 (2024). doi: 10.1109/ACCESS.2024.3424258

10. Yar, G.N.A.H., Maqbool, A., Noor-Ul-Hassan, A.B., Afzal, Z.: Audio Source Separation and Voice Conversion, an Application in Music Industry. In: Proceedings - 2023 IEEE International Conference on Emerging Trends in Engineering, Sciences and Technology, ICES and T 2023, pp. (n/a). Institute of Electrical and Electronics Engineers Inc. (2023). doi: 10.1109/ICEST56843.2023.10138851

11. Wang, Z., et al.: Accent and Speaker Disentanglement in Many-to-many Voice Conversion. In: 2021 12th International Symposium on Chinese Spoken Language Processing, ISCSLP 2021. Institute of Electrical and Electronics Engineers Inc. (2021). doi: 10.1109/ISCSLP49672.2021.9362120

12. Assistant Professor, H.S.: Text to Speech Synthesis. Arxiv Journal (2024)

13. A Novel Approach for Voice Based Text Summarizer. Adalya Journal 9(6), (June 2020). doi: 10.37896/aj9.6/021

14. Gupta, V., Sharon, R., Sawhney, R., Mukherjee, D.: ADIMA: Abuse Detection in Multilingual Audio. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 6172–6176. Institute of Electrical and Electronics Engineers Inc. (2022). doi: 10.1109/ICASSP43922.2022.9746718