Text Echo Personalized TTS System

1st Dr. Pramod Patil
Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
pramod.patil@dypvp.edu.in

2nd Mrs. Vasudha Phaltankar
Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
vasudha.phalatankar@dypvp.edu.in

3rd Shivprasad Waghmare
Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
shivprasadwaghmare2003@gmail.com

4th Ankur Bombarde
Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
ankurbombarde@gmail.com

5th Chetan Lande
Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
chetanlande504@gmail.com

6th Omkar Raskar
Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
omkarraskar03@gmail.com

***Abstract*—Text Echo is a personalized text-to-speech (TTS) system that uses advanced deep learning models like Coqui XTTS v2 to generate natural-sounding speech in multiple languages, including Hindi and Marathi. Coqui XTTS v2 demonstrates a speech naturalness and intelligibility accuracy of around 86%–90 %, as reflected by its impressive Mean Opinion Score (MOS) ranging from 4.3 to 4.5 on a 5-point scale. It allows users to clone their voices for a customized listening experience and offers features such as real-time processing, text extraction through Optimal character recognition (OCR), and the ability to adjust emotional tones. The system outperforms existing TTS systems by providing user-specific voice synthesis, accommodating regional languages, and offering an intuitive interface. It employs artificial neural networks to identify vocal features and generate speech that maintains the user's distinct intonation and emotional expression. Ethical considerations like data security and misuse prevention are addressed through voice sample encryption, anonymization, and audio watermarking. Future work will focus on expanding language support, integrating real-time processing capabilities, and developing the system as an API for broader integration.**

Keywords— LSTM, OCR, Deep Learning algorithms, Vocalizations, Text-to-Speech (TTS), voice-based summarization, Indic languages, Coqui XTTS v2 (WaveNet, Tacotron), neural networks.

I. INTRODUCTION

The four-decade history of TTS technology, as it developed from the initial robot-like systems of the 1960s through to the later, more natural concatenative techniques, native and latent Markov model-based approaches of the 1990s. In the 2010s, deep Learning methodologies, including Coqui XTTS v2 (WaveNet, Tacotron), have transformed TTS technology by enabling more spontaneous and natural speech. 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 strides have been made, there are still challenges to processing underrepresented speech material and accents, notably assisting with regional languages such as Hindi and Marathi. Recent research attempts to enhance the voice-per- The expressiveness, personalization, and adaptability of artificial speech in these languages, allowing greater use and greater access by native speakers.

A. Importance of TTS Systems

The creation of a skilled TTS system is necessary in most tasks, especially for expanding accessibility, improving education, and improving entertainment. A better TTS engine can get over accessibility challenges by translating written text to natural speech, enabling the visually impaired or readers with barriers to freely access information. In schools, TTS technology can be an effective tool for a range of learners, through promoting understanding by audio story learning.

Interestingly, there is a large disparity in the quantity of competent TTS systems for Indian languages such as Marathi and Hindi. It would bridge this gap and make a more complete strategy to cater to several 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 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

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

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 technological engagement.

The adoption of TTS technology is growing across various sectors, driven by its ability to enhance accessibility, user experience, and information delivery. Table 1 shows TTS technology's market size and projected growth across various sectors.

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%

B. Existing State of Art

Despite progress, underrepresented speech data and accents, especially in real-time applications, are challenging. Voice personalization, expressiveness, and adaptability continue to advance through research for more accessibility. There are still constraints, however, such as limited capability in less prevalent languages such as Marathi and Hindi, robotic and emotionally neutral synthetic voice impacting user experience, and latency in processing that slows down real-time applications. Loss of context in verbal summary results in inaccuracy, and ethical and privacy issues are threats of misuse. Limited adaptability to user preferences and the lack of customizable voice options also slow down personalization and user interaction.

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 TTS, which provides multiple voices and languages to enhance user experience [1].

Vocals, an app aiding vocally impaired individuals in scheduling appointments, utilizes 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].

In contrast to [4], which addresses general TTS systems, Text Echo is tailored for personalized voice cloning and supports regional languages. While [5] and [6] delve into deep learning methods for TTS, our research combines OCR and NLP to improve text extraction and contextual comprehension. Although reference [8] covers preprocessing methods, our system advances further by including post-processing to adjust emotional tones.

An approach for identifying unauthorized outdoor advertising utilizes CLIP fine-tuning for image analysis and 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].

C. Contributions Of Proposed System Text Echo

Text Echo is a personalized TTS system that uses advanced deep learning models like Coqui XTTS v2 to generate natural-sounding speech in multiple languages, including Hindi and

Marathi. It allows users to clone their voices for a customized listening experience and offers features such as real-time processing, text extraction through OCR, and the ability to adjust emotional tones. The system outperforms existing TTS systems by providing user-specific voice synthesis, accommodating regional languages, and offering an intuitive interface. Ethical considerations like data security and misuse prevention are addressed through voice sample encryption, anonymization, and audio watermarking. Future work will focus on expanding language support, integrating real-time processing capabilities, and developing the system as an API for broader integration.

D. OBJECTIVES

Text Echo is a project to create an Android application that provides user-dependent voice options, transcending current limitations of TTS systems to enhance user experience. It aims to add contextual knowledge, pronunciation, and speech expressiveness. The project will yield a flexible vocal sound personalization system and improve TTS output quality using pre- and post-processing techniques. Voice conversion and audio source separation are combined to adjust system parameters. A new dataset was built with deliberate control. Creating a deep text-to-speech model with personalization requires key features, such as core OCR, for effective text extraction from documents.

II. ALGORITHMS AND METHODS

A. METHODS

1) ALGORITHMS & MODELS

MODEL COQUI XTTS V2

We used the pre-trained model Coqui XTTS v2 for TTS configuration in our application. Coqui XTTS v2 is a powerful voice generation model enabling personalized voice cloning with a short audio sample. It is designed for multilingual speech synthesis and cross-language voice cloning, making it advanced for voice-based applications. We integrated Coqui XTTS v2 to facilitate personalized voice cloning for text-to-speech conversion, selected for its ability to generate high-quality, natural-sounding speech while maintaining unique vocal characteristics. Coqui XTTS v2 supports 17 languages, but our implementation focuses on three: English, Marathi, and Hindi.

ADVANTAGES OF COQUI XTTS V2 OVER EXISTING SYSTEMS

The Coqui XTTS v2 model, utilized in our project for TTS conversion, introduces notable improvements over earlier models in areas such as voice cloning, phoneme accuracy, emotional expressiveness, and processing efficiency. A key strength of Coqui XTTS v2 lies in its ability to clone voices with just a short reference sample, addressing a major limitation of older models like VQ-VAE Voice Conversion, which required a substantial amount of speaker data and often encountered challenges related to speaker identity retention [11].

Furthermore, our model enhances speech naturalness and expressiveness by efficiently transferring emotion and speaking style from a reference voice. Earlier TTS models, such as Tacotron 2 combined with WaveGlow, often produced speech that sounded monotone or robotic, struggling to replicate the subtle emotional variations present in natural conversations [8]. By contrast, Coqui XTTS v2 effectively captures vocal nuances, ensuring a more lifelike and dynamic speech synthesis experience.

Our model focuses on English, Marathi, and Hindi, ensuring high phoneme accuracy across these languages. Many previous models, such as Star GAN-VC, often faced challenges in accent retention and speaker identity consistency when applied to multilingual speech synthesis [11]. Coqui XTTS v2 mitigates these limitations by preserving speaker characteristics effectively across languages, leading to improved pronunciation accuracy and speech consistency.

In terms of efficiency, Coqui XTTS v2 operates at a 24 kHz sampling rate, generating high-fidelity speech with minimal delay. Older architectures like FastSpeech-2 demanded extensive computational resources and prolonged training times to achieve comparable quality [8]. Additionally, handling regional accents and phoneme variations in Indian languages has been a persistent challenge for many TTS models. For instance, StarGAN-VC struggled with phoneme accuracy in non-English languages, whereas Coqui XTTS v2 demonstrated notable improvements in articulation and fluency for Marathi and Hindi [11].

III. 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 proposed system is shown in Fig. 1.

The workflow diagram depicts a text-echo system for user-friendly operation. Users can upload PDF or image files or provide voice recordings. This system evaluates whether text extraction is required. OCR technology extracts text from images or PDFs with non-textual content. This phase was skipped for unformatted text files. The system uses TTS models, like Coqui XTTS v2, to convert text into lifelike audio. These models generate speech akin to human vocal traits. Clients can then play or save the speech, meeting the user's needs.

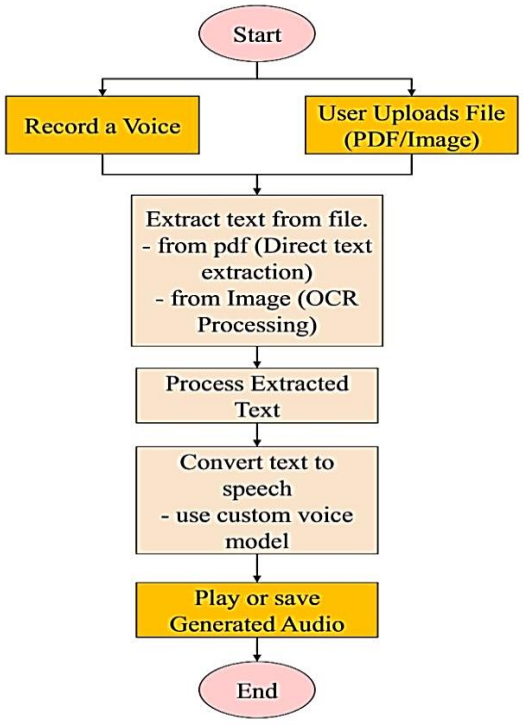

Fig. 1. Process Flow of Proposed Model

This section describes the proposed 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. Coqui XTTS v2 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.

Ultimately, the system utilizes Coqui XTTS v2, an advanced text-to-speech model, for replicating voices. Users can provide a brief audio recording, which the model processes to create a unique voice profile. This model captures emotional nuances and regional dialects, delivering natural-sounding speech in various languages, such as Hindi and Marathi.

A. Technical Approach

The designed TTS system architecture converts text from PDFs or images into speech. First, the text extraction method uses OCR for input in image format to extract text. The text is then sent to the model for speech conversion. Users can record their voices, which the model uses to create personalized speech. Users can hear the text spoken in their own voices or by their choice.

Fig. 2 depicts the operational process of the Text Echo system. Initially, voice recording is processed by an Artificial Neural Network (ANN) to identify vocal features. These features are utilized by the Coqui XTTS v2 model to produce customized speech. The ANN ensures that the generated speech maintains the user's distinct intonation and emotional expression. This details the modular design of the text-echo system, starting with a user-friendly interface for uploading PDF/image files or voice recordings. 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. All extracted or input data are stored in a central database for scalability and easy retrieval. The text is processed using advanced algorithms such as Coqui XTTS v2, known for high-quality, realistic voice outputs. The system concludes by playing or saving synthesized speech and offering customization options to complete the process for each user.

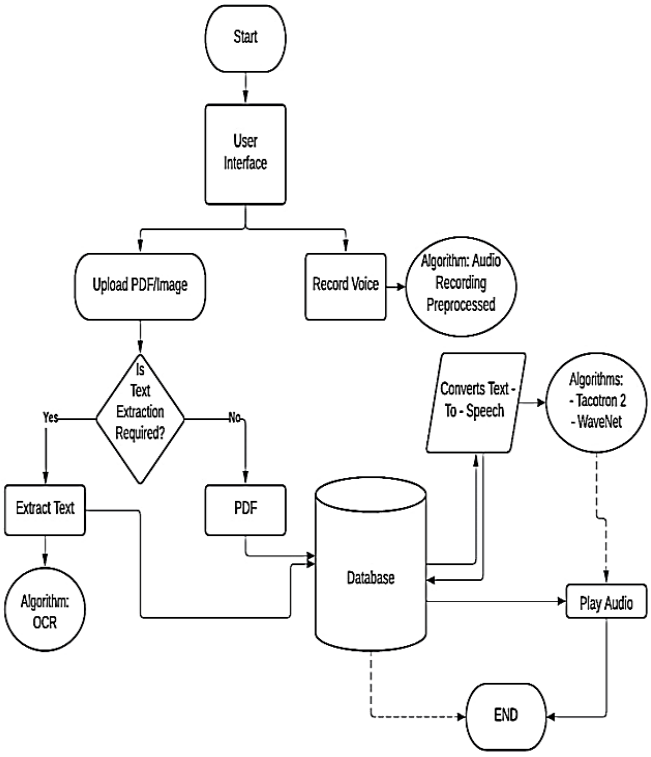

Fig. 2. Architectural Overview of the Designed Text Echo System

The system comprises OCR, NLP, and machine-learning models to achieve a natural, personalized speech output.

Technical components of the Text Echo application System are as follows.

1. 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.
2. Processing of Text
    a. OCR System - An optical character recognition system, like Tesseract (open-source) or Google Vision API, was chosen to extract text from images.
    b. Image Enhancement - Convert images to black and white, reduce noise, and correct image tilt to enhance OCR precision.
    c. NLP for initial text analysis – Methods like tokenization, identifying parts of speech, and sentence segmentation can improve text comprehension.

3. Speech Synthesis - Advanced machine-learning techniques, including Coqui XTTS v2, 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.

The proposed system creates speech that sounds natural, enhancing accessibility and communication for those with speech impairments. It includes options for customization and offline use, making it a flexible solution for various user requirements. The ability to store and replay speech increases its usefulness, potentially improving the quality of life for many users. Text Echo is a sophisticated personalized TTS system that employs deep learning models like Coqui XTTS v2 to produce clear and natural speech. With an accuracy rate of 86%-90 % and a high Mean Opinion Score (MOS) of 4.3 - 4.5, it can replicate user voices and supports multiple languages, such as Hindi and Marathi.

IV. Instruments and Equipment

A. Physical Components

1) Physical Hardware

Computing systems with a minimum of 16GB RAM and an i5 processor are used for processing OCR and TTS, while Android smartphones are employed for testing, image capture of text, and speech playing.

B. Digital Tools

Android Studio is used for app development, Coqui XTTS v2 enables natural-sounding speech synthesis, assisted by machine learning frameworks for enhanced pronunciation and context awareness. Git is used for version control.

C. Evaluation Instruments

1) User Experience Platforms

Resources and systems for conducting usability assessments and gathering feedback on an application's performance and user-friendliness. Ex. SurveyMonkey, Google Feedback forms.

V. Results and Discussion

A. Differences from Google Translate

This invention surpasses Google Speak, particularly with its user-specific voice synthesis features. Unlike Google Speak's predefined voices, it allows users to create speech output in their voice, offering a personalized experience. It employs advanced OCR technology to extract text from varied sources, including handwritten materials and images, functions not present in Google Speak. Additionally, it accommodates regional languages, including Hindi and Marathi, making it more useful for Indic language speakers. The innovation provides an easy-to-use interface that allows for vocal documentation, file submissions, and voice personalization. This system ensures users of varying technical expertise can utilize it effectively without extensive preparation or guidance.

B. *Designed Application component*

1) *Home Screen*

Fig. 3 shows the home screen of the Text Echo app, with AI-powered voice cloning for text-to-speech. It highlights key features like multiple language support, superior sound, ease of sharing, demo samples, and a how-it-works section.

2) *Gradio model without input and output*

Fig. 4 displays the interface of voice cloning software that enables users to upload an audio file and a PDF or image to create synthesized speech resembling the original speaker. The software, developed using Gradio, supports text extraction from uploaded documents to produce the final speech output.

3) *Gradio model with input and output*

Fig. 5 presents a screenshot illustrating a voice cloning application where an individual has submitted an audio file and an image. The user has produced speech and is playing the resultant audio. The application extracts text from PDF documents or images, using this text to synthesize speech in the uploaded speaker's voice.

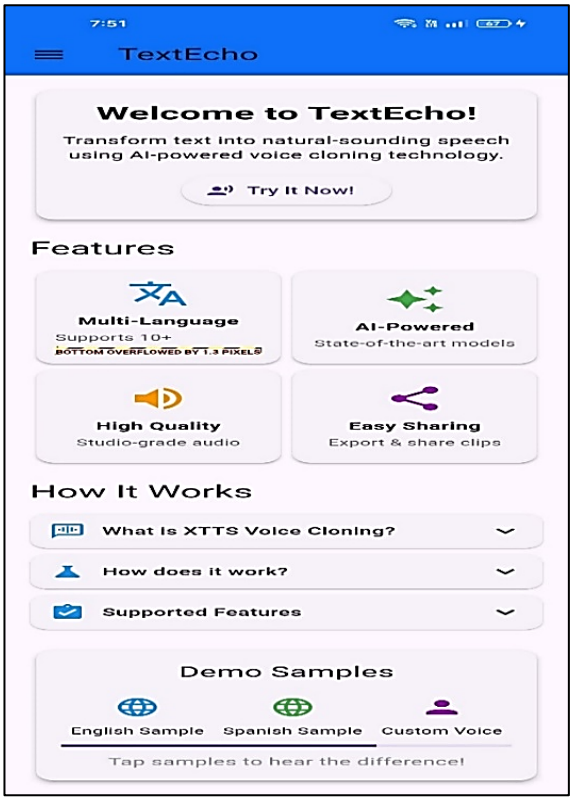

Fig. 3. Home Screen page of Text Echo

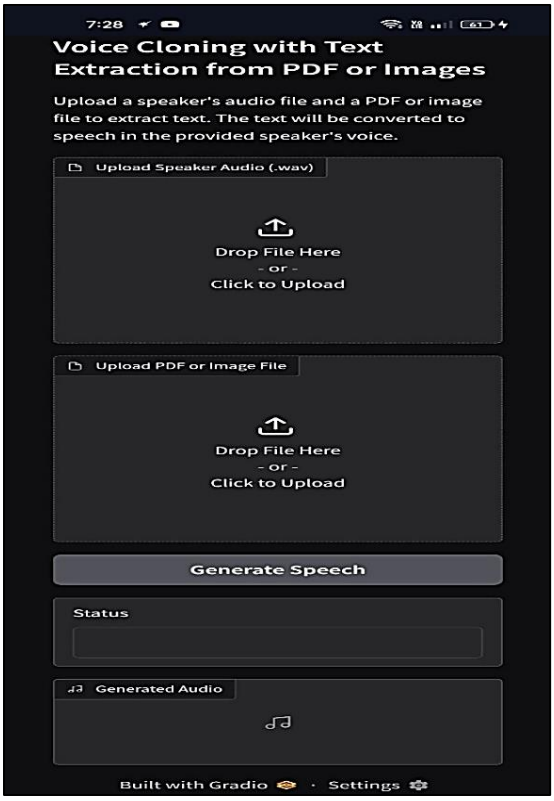

Fig. 4. Gradio model of Text Echo

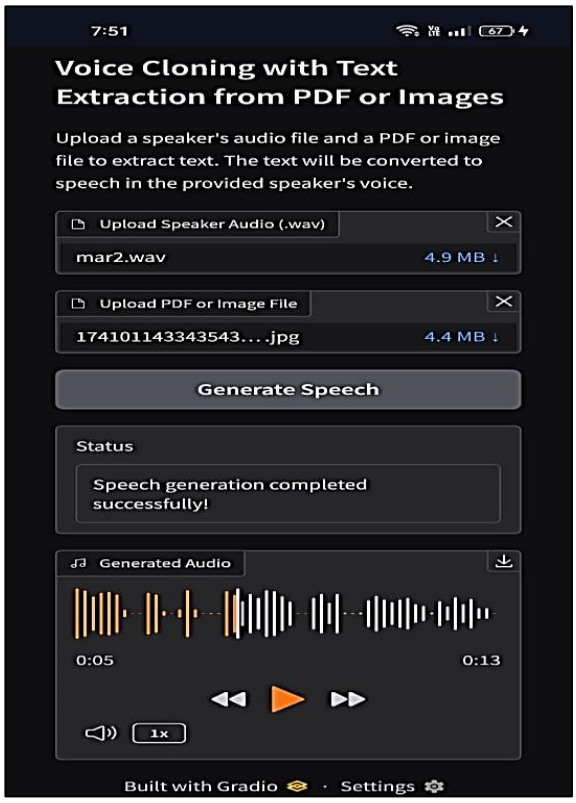

Fig. 5. Gradio model with input and output provided

C. Evaluation of Proposed TTS System

1) OCR Tools

Both Tesseract OCR and EasyOCR are extensively utilized OCR software. EasyOCR is noted for its superior accuracy and user-friendliness. Tesseract, an OCR engine based on long short-term memory (LSTM) networks, supports over 100 languages but necessitates substantial preprocessing to achieve optimal results. It encounters difficulties with complex fonts, handwritten text, and noisy images. Conversely, EasyOCR, which employs a deep learning-based OCR engine incorporating convolutional neural networks (CNN), LSTM, and connectionist temporal classification (CTC), offers enhanced accuracy, particularly for handwritten, stylized, and multi-script text. It requires minimal preprocessing and benefits from GPU acceleration, thereby improving efficiency.

As presented in Table 2, EasyOCR outperforms Tesseract in English, Marathi, and Hindi, achieving lower Character Error Rate (CER) and Word Error Rate (WER), ensuring enhanced text recognition accuracy. The higher Bilingual Evaluation Understudy (BLEU) scores confirm better text structure retention, which aligns with OCR advancements using CLIP fine-tuning techniques for improved accuracy [9].

Overall, EasyOCR emerges as the preferred OCR solution, delivering superior accuracy and efficiency.

Table 2. OCR Accuracy Comparison: Tesseract OCR and EasyOCR

Language	Metric	Tesseract [Existing Model] [9]	EasyOCR [Proposed Model]
English	CER	0.3683	0.0271
	WER	1.4313	0.1294
	BLEU	0.6369	0.8787
Marathi	CER	0.5788	0.3529
	WER	3.4345	2.0762
	BLEU	0.1907	0.3325
Hindi	CER	0.0874	0.1542
	WER	0.3572	0.6870
	BLEU	0.6204	0.3034

2) PDF Tools

PyPDF2 and PDFplumber are Python libraries for PDF manipulation. PDFplumber is more precise for text extraction, preserving text structure, tables, and multi-column layouts, whereas PyPDF2 extracts raw text but struggles with structure and table extraction. PDFplumber excels in structured table extraction and manages images better, maintaining precise positioning and metadata. It handles multi-column PDFs well, maintaining text alignment. PDFplumber is primarily used for precise text parsing, structured data recovery, and layout preservation, minimizing post-processing.

Table 3 shows a comparison of text extraction accuracies of PyPDF2 and pdfplumber libraries for PDFs, focusing on word- and character-level accuracies as percentages. The findings showed each tool's effectiveness with English, Marathi, and Hindi PDFs. Data indicate pdfplumber consistently surpasses PyPDF2 in preserving text structure, managing multi-column formats, and maintaining alignment, especially in non-English languages. This analysis emphasizes choosing the right tool based on the PDF content complexity and language.

Finally, our model Coqui XTTS v2 generates natural speech with an approximate 86%-90 % accuracy.

Table 3. Comparative Analysis of Text Extraction: PyPDF2 and PDF Plumber

Language	Metric	PyPDF2 [Existing Model]	PDF Plumber [Proposed Model]
English	Word-Level Accuracy (%)	96.31	96.31
	Character-Level Accuracy (%)	97.67	97.67
Marathi	Word-Level Accuracy (%)	18.55	85.30
	Character-Level Accuracy (%)	72.04	94.29
Hindi	Word-Level Accuracy (%)	84.04	81.73
	Character-Level Accuracy (%)	77.84	86.67

VI. Conclusion

Text Echo is an advanced TTS system that employs sophisticated deep learning models to produce natural-sounding speech across multiple languages, including Hindi and Marathi. The system allows users to clone their voices, thereby offering a personalized auditory experience. Key features of Text Echo include real-time processing, text extraction via OCR, and the ability to adjust emotional tones. The system demonstrates superior performance compared to

existing TTS systems, particularly in user-specific voice synthesis and support for regional languages.

Text Echo is an advanced personalized TTS system using deep learning models like Coqui XTTS v2 to generate natural and intelligible speech. With 86%-90 % accuracy and a MOS of 4.3 - 4.5, it clones user voices and supports multiple languages, including Hindi and Marathi. Its real-time processing, OCR-based text extraction, and emotional tone adjustments enhance user experience.

The system will be developed as an API for seamless integration with software applications, allowing efficient use by developers. Expanding language support, including regional accents and dialects, will enhance accessibility and speech authenticity. Additionally, translation features within the TTS output will improve cross-linguistic communication, making Text Echo a comprehensive solution for personalized speech synthesis across diverse linguistic and technological landscapes.

VII. Acknowledgment

We extend our gratitude to Dr. Pramod Patil and Mrs. Vasudha Phaltankar from the Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, for their guidance and support.

Disclosure of Interests

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

VIII. 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 Kumar, Y., Koul, A., Singh, C.: A deep learning approach 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

[6] Kumar, Y., Koul, A., Singh, C.: A deep learning approach 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/ISMSIT58785.2023.10304907S. Liu, "Wi-Fi Energy Detection Testbed (12MTC)," 2023, GitHub repository. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC

[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 the 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