World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

AI powered voice synthesizer

V. Vanaja, Venkatesham Tunge, Nithin Kumar Kanagala *, Harsha Vardhan Bhumandla and Shruti Kana

Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 663-671

Publication history: Received on 22 March 2025; revised on 02 May 2025; accepted on 04 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0590

Abstract

The AI Voice Synthesizer is an advanced real-time, multilingual voice cloning system that utilizes state-of-the-art deep learning techniques to generate personalized speech with high naturalness and accuracy. Built on the open-source Coqui.ai's XTTSv2 framework, the system enables users to synthesize speech using their own voice—or any voice sample—by analyzing just a few seconds of audio. It then uses this voice profile to generate natural-sounding speech in multiple languages, even those the original speaker has never spoken, offering a revolutionary leap in the field of synthetic speech and human-computer interaction.

Traditional text-to-speech (TTS) systems often suffer from robotic tone, lack of personalization, limited language support, and high latency. In contrast, this project provides a lightweight, low-latency (<200 ms), and user-friendly platform that supports cross-lingual, few-shot voice cloning. Designed with modularity in mind, the system consists of several independent components: speaker embedding extraction, multilingual text processing, real-time speech synthesis, and a web-based front end. These components are integrated into a seamless workflow that is intuitive and accessible for non-technical users, while also being scalable and customizable for developers and researchers.

Keywords: Real-Time Speech Synthesis; Few-Shot Text-To-Speech; Multilingual TTS; Coqui.AI Speaker Embedding; Personalized Synthetic Voice; Real-Time Voice Cloning System

1. Introduction

Artificial Intelligence (AI) has rapidly transformed numerous domains, and among the most impactful areas is speech synthesis. One of the most advanced innovations in this field is **AI Voice Cloning**—a technique that allows machines to replicate a person's voice using deep learning models. Unlike traditional text-to-speech (TTS) systems, which often produce robotic and monotonous output, voice cloning captures the unique tonal qualities, pitch, and speaking style of an individual. This project focuses on building a real-time, multilingual voice cloning system that offers fast, high-quality, and personalized voice synthesis through open-source AI models.

Voice synthesis has been part of human-computer interaction for decades. Early systems relied on concatenative synthesis, combining pre-recorded segments of speech to produce output. While functional, these systems lacked flexibility and naturalness. With the emergence of machine learning, particularly **deep learning**, speech synthesis has reached new levels of realism and adaptability. Technologies like Google's Tacotron, Mozilla's TTS, and now Coqui.ai's TTS have revolutionized the domain, allowing systems to generate speech that is virtually indistinguishable from human speech.Voice cloning builds upon these foundations by not just synthesizing generic speech, but by mimicking specific voices. By analyzing a short audio sample, advanced models can capture and replicate a speaker's vocal signature. This opens up a range of possibilities—from personal voice assistants that sound like their users, to localized narration in multiple voices, to accessibility tools that give users their own synthetic voice.

* Corresponding author: Nithin kumar Kanagala.

Voice is a fundamental element of human communication. It conveys not only linguistic content but also emotional tone, personality, and intent. In recent years, the development of artificial intelligence (AI) has enabled machines to understand, generate, and even replicate human speech. This advancement has culminated in the emergence of **voice cloning**, a subfield of AI that focuses on synthesizing human-like voices based on real-world samples. This project, **AI Powered Voice Synthesizer**, builds upon these innovations to provide a real-time, multilingual, speaker-specific speech generation system.

2. Literature review

The domain of text-to-speech (TTS) and voice cloning has witnessed rapid advancements over the past decade, driven largely by the evolution of deep learning models. Early speech synthesis systems relied on rule-based or concatenative techniques, which generated intelligible but robotic-sounding speech. These methods were rigid, lacked personalization, and were unsuitable for multilingual or expressive applications.

A significant leap occurred with the development of statistical parametric speech synthesis, particularly using Hidden Markov Models (HMMs). However, true naturalness was achieved only after the introduction of deep neural networks. Models such as Tacotron and Tacotron 2 introduced sequence-to-sequence architectures with attention mechanisms, allowing for end-to-end training and high-quality mel-spectrogram generation. These models were further improved through vocoders like WaveNet, which synthesized waveforms from spectrograms with exceptional fidelity.

Voice cloning emerged as a critical subfield, aiming to reproduce a specific speaker's voice using limited training data. The introduction of speaker embeddings and few-shot learning techniques, as seen in SV2TTS and AutoVC, enabled the cloning of voices using only a few seconds of audio. Recently, Coqui.ai's XTTS and XTTSv2 frameworks have made real-time, multilingual voice cloning accessible through open-source libraries, providing state-of-the-art results in voice personalization and cross-lingual synthesis.

3. Existing system

Existing voice synthesis and cloning systems, while increasingly advanced, are still limited in several critical areas. Commercial platforms like **Google Cloud Text-to-Speech**, **Amazon Polly**, and **Microsoft Azure Speech Services** deliver high-quality speech output through cloud-based APIs but restrict users in terms of customization, personalization, and accessibility. These systems typically offer a fixed set of pre-trained voices with limited options for emotional tone, prosody variation, or speaker-specific customization. As a result, users must choose from a pre-defined library of synthetic voices that may not align with their personal or cultural needs.

Furthermore, while some commercial platforms support a handful of languages, true multilingual capability—particularly for low-resource or regional languages—is rarely implemented with consistent quality. Most services excel in English but provide suboptimal results in languages with fewer available datasets or complex phonetic structures. Additionally, many systems operate with high latency, often taking several seconds or more to synthesize a short utterance. This makes them impractical for real-time or interactive applications such as voice chatbots, live presentations, or accessible communication tools for individuals with speech impairments.

Another major shortcoming of many existing systems is their **lack of real-time performance and offline deployment**. Most commercial TTS solutions are cloud-dependent, meaning they require constant internet access and incur usage-based charges. For developers or institutions with privacy or cost concerns, this dependency poses a significant barrier. Moreover, the proprietary nature of these platforms means that developers cannot modify the model architecture, fine-tune on custom datasets, or implement experimental features like emotion modeling or speaker verification.

4. Proposed system

The **AI Voice Synthesizer** proposed in this study addresses the limitations of existing voice cloning and text-to-speech (TTS) systems by offering a real-time, few-shot, multilingual solution that is open-source, modular, and ethically designed. Built upon **Coqui.ai's XTTSv2** architecture, the system allows users to synthesize speech in their own voice—or a cloned voice—across multiple languages using just a few seconds of audio input. This represents a significant shift from traditional TTS models that require extensive training data and operate primarily in batch processing mode.

The primary innovation of the proposed system lies in its **few-shot speaker cloning** capability. By leveraging a powerful speaker embedding network, the system extracts unique voice characteristics from as little as 3 to 10 seconds

of user-provided speech. These embeddings are then used to condition a multilingual TTS model, enabling the synthesis of personalized speech outputs in over 16 languages, even those the original speaker has never spoken. This is made possible through a shared phoneme tokenization strategy and language conditioning tags supported by XTTSv2.

In terms of architecture, the system is modular, consisting of several independent components: a **Speaker Embedding Module**, a **Text Preprocessing Module**, a **Multilingual XTTSv2 TTS Engine**, and an **Audio Output Interface**. Each component can be updated or replaced without affecting the rest of the system, which allows for flexible experimentation and easy future upgrades. The user interface is developed using **Gradio** or **Streamlit**, providing a simple and accessible platform for end users. It allows voice upload, text input, language selection, and audio preview or download in real time.

5. Methodology

The development of the AI Voice Synthesizer involves a well-defined, six-step pipeline designed to support real-time, speaker-specific, multilingual voice cloning. These steps encompass both signal processing and deep learning procedures to convert a short voice sample and a block of text into natural-sounding speech. The following stages outline the system's methodology in detail:

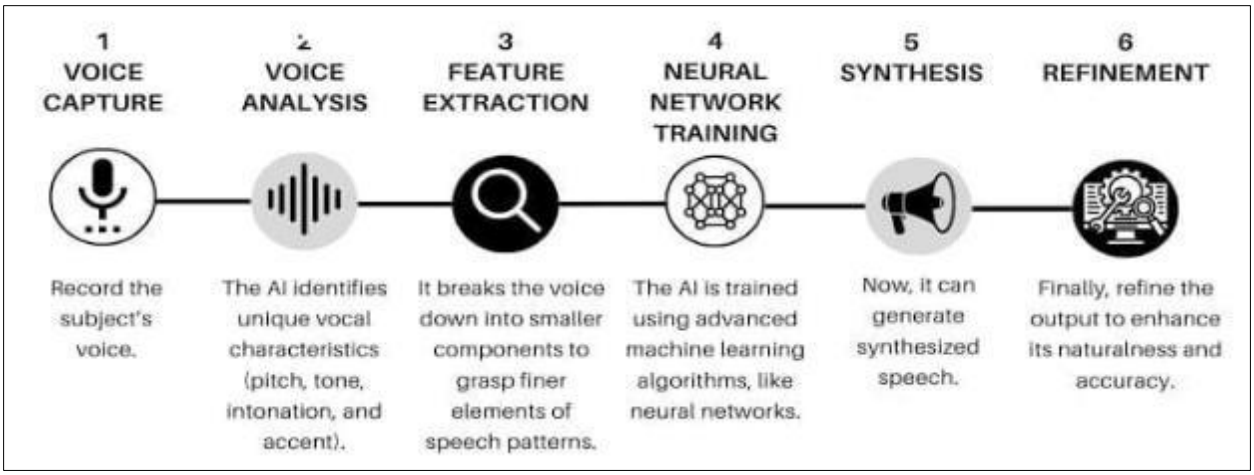

Figure 1 Methodology

5.1. Voice Capture

The process begins with capturing a short voice sample from the user. This sample—typically ranging from 3 to 10 seconds—is uploaded via the user interface in a common audio format such as WAV or MP3. It should be clean, noise-free, and consist of neutral speech to ensure optimal cloning accuracy. The captured sample forms the basis for cloning the user's unique vocal identity.

5.2. Voice Analysis

Once the voice sample is acquired, it undergoes a basic preprocessing phase. This involves audio normalization, noise suppression, and silence trimming to ensure a consistent and clean signal. The system analyzes the sample for pitch, tone, speed, and phoneme clarity. These aspects are crucial for accurate modeling of speaker identity in the later stages.

5.3. Feature Extraction

The refined audio sample is passed through a **speaker encoder**, which uses a deep neural network to extract a **speaker embedding**—a compact, high-dimensional vector that represents the unique vocal features of the speaker. These embedding captures timbre, cadence, accent, and other vocal traits. It is used to condition the synthesis model, ensuring that all future speech outputs retain the speaker's identity.

5.4. Neural Network Training

Although the AI Voice Synthesizer uses pre-trained models for deployment, this stage involves understanding the underlying training that powers XTTSv2. The TTS model is trained on large multilingual datasets using **sequence-to-**

sequence learning. During training, the model learns to map tokenized text and speaker embeddings to mel-spectrograms. These spectrograms are further converted into audio waveforms via a vocoder. The training process equips the model to generalize across unseen text and voices.

5.5. Synthesis

During inference, the user provides text input alongside the previously extracted speaker embedding. The TTS engine (XTTSv2) uses both inputs to generate a corresponding mel-spectrogram. The vocoder then transforms the spectrogram into a high-quality audio waveform that reflects the speaker's voice while articulating the new textual content. This process takes place in real-time with minimal latency on GPU systems.

5.6. Refinement

Finally, the synthesized audio undergoes optional post-processing for clarity and realism. Techniques such as dynamic range compression, denoising, and silence removal may be applied. Optional watermarking can be embedded for content verification. The final result is presented to the user for playback or download.

5.7. System Architecture

The architecture of the AI Voice Synthesizer has been designed with modularity, scalability, and real-time performance in mind. The system follows a layered client-server model, allowing seamless interaction between the user interface, processing modules, and synthesis engine. Its components are loosely coupled, enabling independent updates, replacements, and easy integration with third-party tools. This modular structure makes the system suitable for both personal and enterprise-level deployments.

At the highest level, the User Interface Layer handles interaction with users. Developed using web-based frameworks like Gradio or Streamlit, this layer provides an accessible platform where users can upload voice samples, input or paste text, select output languages, and play or download the synthesized speech. The UI also offers real-time feedback during synthesis and includes privacy controls for data handling.

Beneath the interface lies the Voice Analysis and Embedding Layer. This module is responsible for preprocessing the uploaded audio, applying noise reduction and trimming, and extracting a speaker embedding using a pre-trained encoder. The embedding is a numerical representation of the speaker's unique vocal features, including tone, accent, and pitch. This vector is stored temporarily for use in conditioning the synthesis engine.

Parallel to the embedding process, the Text Processing Layer prepares the user's input text. It performs tokenization, language tagging, and phoneme conversion where necessary, depending on the selected language. This ensures that the text is appropriately structured for input into the multilingual speech model.

The core of the system is the XTTSv2 Synthesis Engine, located in the Model Inference Layer. It takes as input the speaker embedding and processed text, then generates a mel-spectrogram. This is further converted into an audio waveform using an integrated neural vocoder. XTTSv2 supports multilingual synthesis and can clone voices across more than 16 languages with impressive fidelity and low latency, typically under 200 milliseconds on GPU-enabled machines.

Following synthesis, the Audio Output Layer handles the rendering and post-processing of the generated audio. Features such as dynamic gain adjustment, silence trimming, and optional watermarking are included here. The output is then presented to the user for real-time playback or download in formats like WAV or MP3.

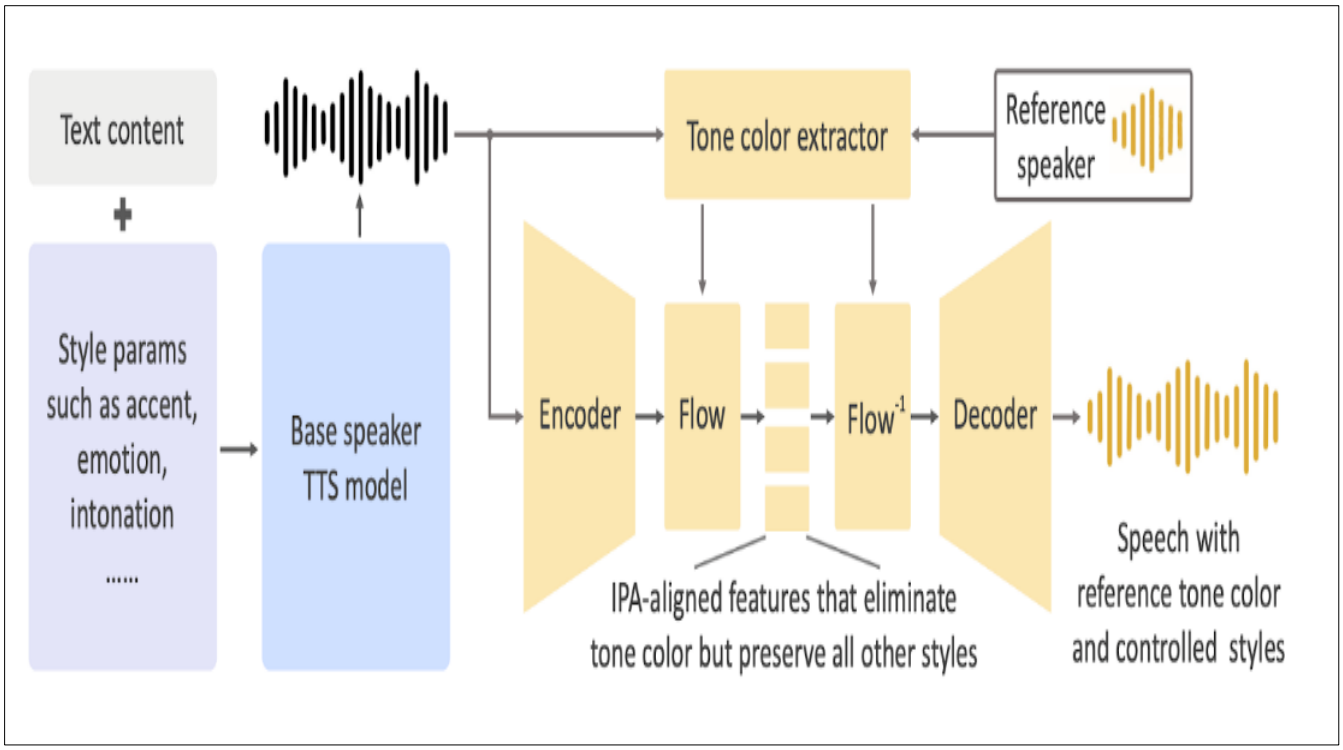

Figure 2 System Architecture

6. Results and Discussion

Figure 3 User Interface

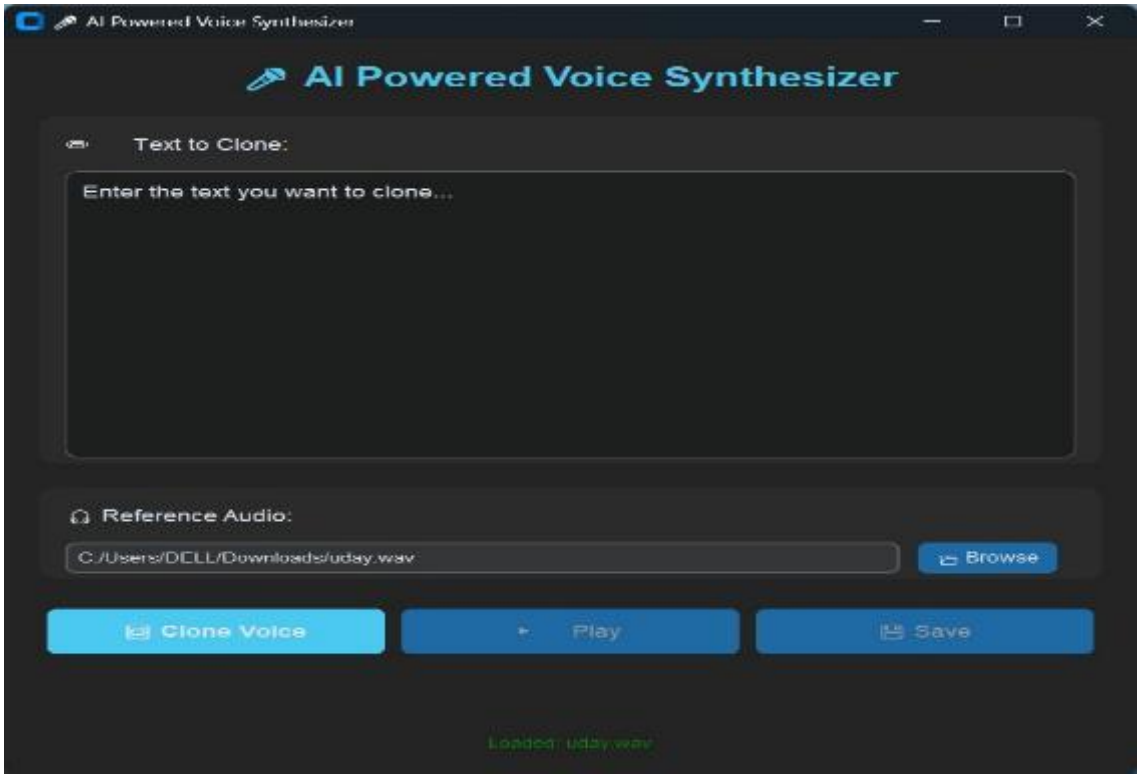

Figure 4 Loading Audio

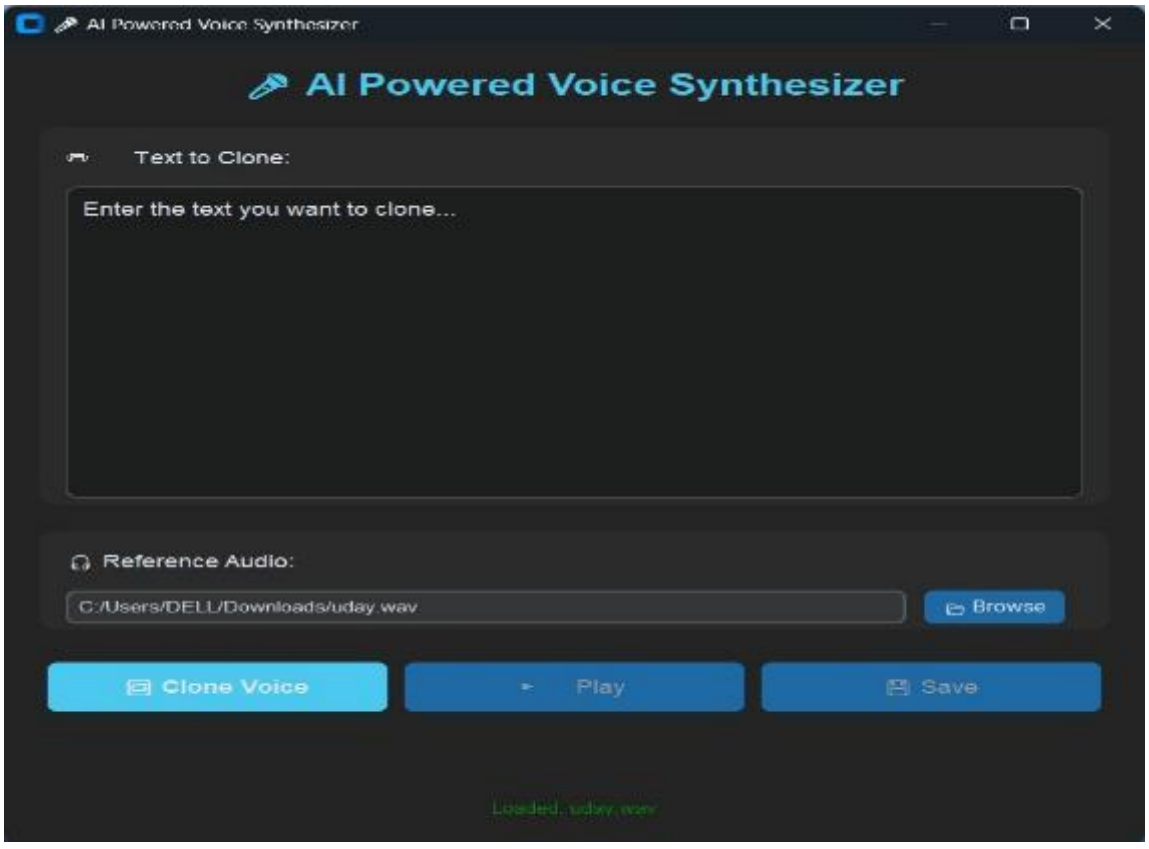

Figure 5 Cloning Audio.wav

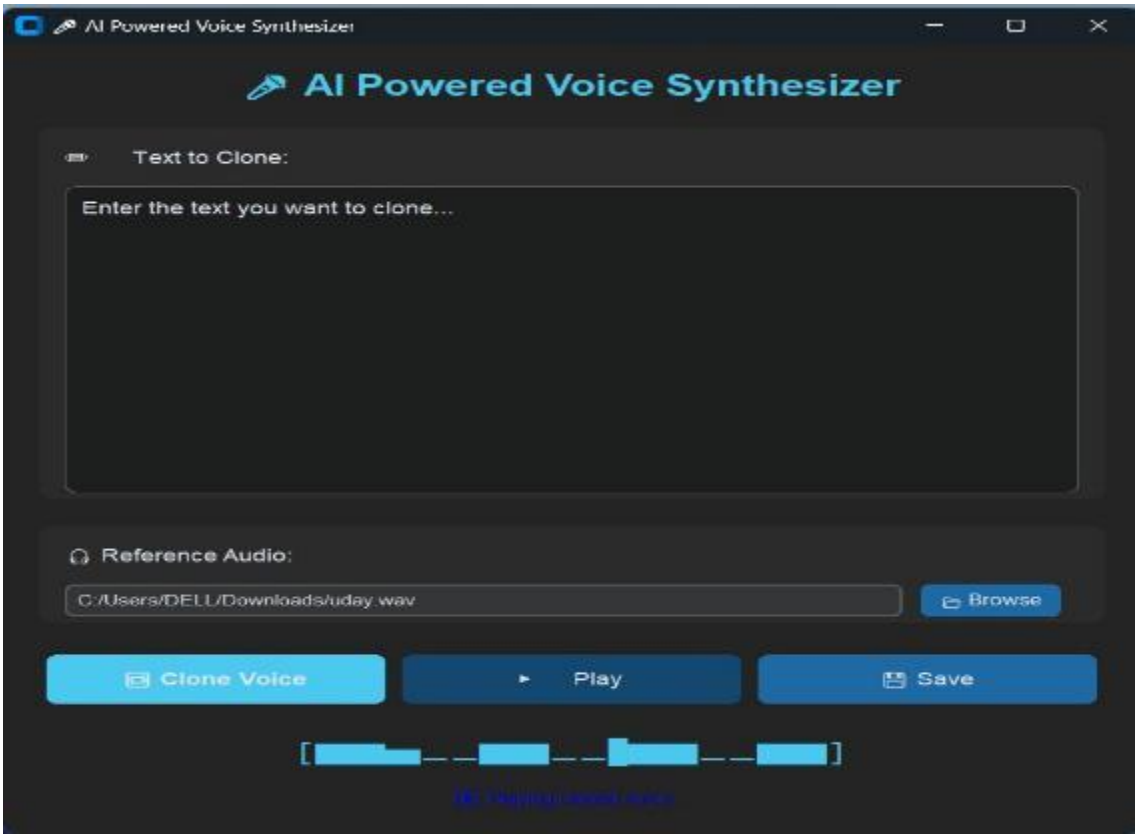

Figure 6 Playing Cloned Audio

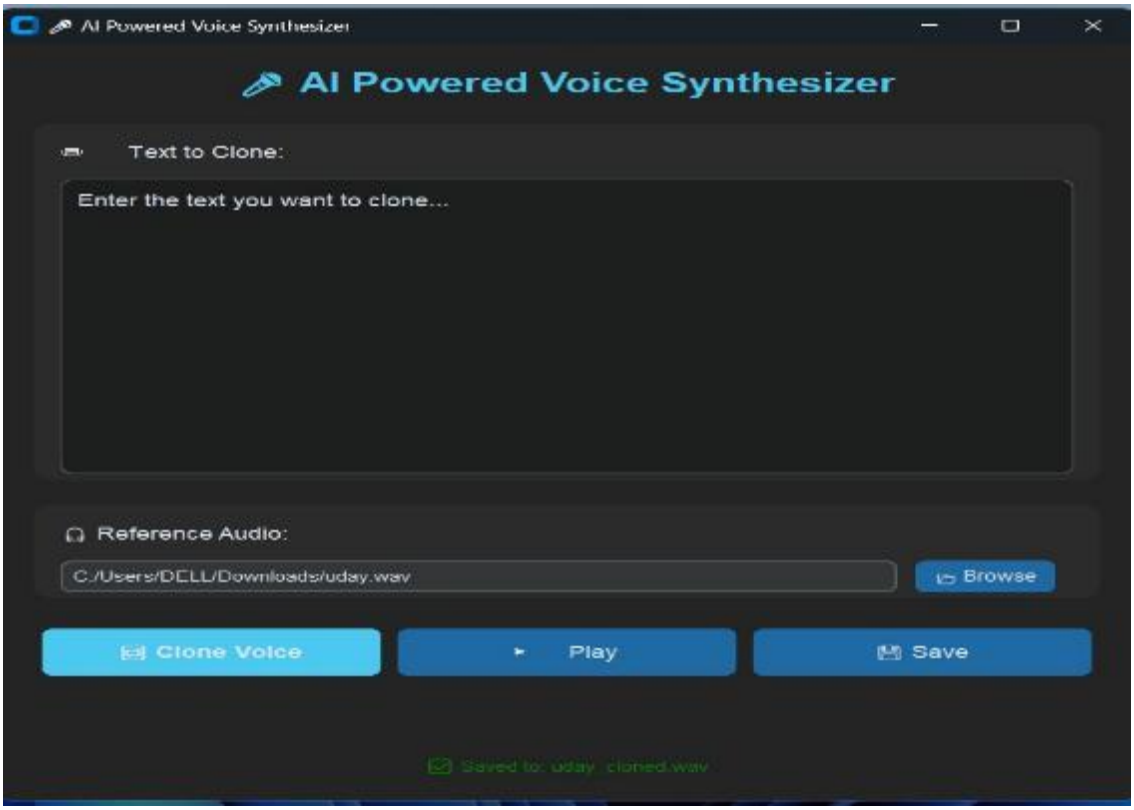

Figure 7 Saving Audio

7. Conclusion

The AI Voice Synthesizer represents a meaningful advancement in the field of voice synthesis and artificial intelligence by addressing the limitations of existing systems and offering a solution that is real-time, multilingual, speaker-specific, and ethically conscious. Unlike traditional text-to-speech (TTS) platforms that rely on static, pre-trained voices or require vast amounts of training data, the system presented in this work is capable of cloning a speaker's voice from just a few seconds of audio and reproducing that voice across multiple languages with remarkable accuracy and speed.

At the core of the system lies the XTTSv2 model, developed by Coqui.ai, which enables few-shot speaker adaptation and cross-lingual synthesis using shared phoneme representations and speaker embeddings. This approach empowers users to generate personalized speech with minimal effort, offering applications in accessibility, education, entertainment, content creation, and interactive voice-based systems. The system's ability to synthesize speech in real time—under 200 milliseconds on suitable hardware—makes it practical for use in live applications such as voice assistants, translation tools, or digital storytelling platforms.

In addition to its technical achievements, the AI Voice Synthesizer emphasizes ethical AI practices. Voice data is handled responsibly, with temporary storage and options for immediate deletion, ensuring user privacy and data control. Optional watermarking allows synthesized audio to be marked for authenticity, which can help deter misuse such as voice impersonation or deepfake audio creation. These safeguards promote responsible innovation and align the project with broader calls for transparency and ethical standards in AI development.

The modular design of the system contributes to its scalability and extensibility. Individual modules—such as speaker embedding, text preprocessing, the TTS engine, and the user interface—can be independently modified or upgraded. This design allows the system to evolve over time, potentially integrating future capabilities like emotion modeling, voice-to-voice translation, or mobile deployment for edge devices.

Compliance with ethical standards

Disclosure of conflict of interest

There is no conflict of interest.

References

[1] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135. –Introduced a sequence-to-sequence neural network for converting text to mel-spectrograms, a key inspiration for modern TTS pipelines.

[2] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing. – Presented a system that improved speech naturalness by integrating Tacotron and WaveNet.Lee, M. Y. (2023). Building Multimodal AI Chatbots. arXiv preprint arXiv:2305.03512. https://arxiv.org/abs/2305.03512

[3] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., ... & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Advances in Neural Information Processing Systems (NeurIPS). – Introduced few-shot speaker adaptation via speaker embeddings, paving the way for voice cloning.

[4] Coqui.ai. (2024). XTTSv2: Cross-lingual real-time voice cloning. Retrieved from https://coqui.ai The official documentation and model used for the AI Voice Synthesizer project.

[5] Mozilla Foundation. (2024). Common Voice Dataset. Retrieved from https://voice.mozilla.org An open-source dataset of multilingual voice recordings used in TTS model training and evaluation.

[6] Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... & Miller, J. (2017). Deep Voice: Real-time neural text-to-speech. In Proceedings of the 34th International Conference on Machine Learning (ICML). – One of the early efforts toward real-time neural TTS.

[7] Descript. (2024). Overdub: Voice cloning for podcasters. Retrieved from https://www.descript.com/overdub A commercial voice cloning platform illustrating proprietary approaches to TTS.

Author's short biography

Mrs. V. Vanaja Mrs. V. Vanaja is working as an Assistant Professor in the Department of CSE (DATA SCIENCE) at ACE Engineering College, Hyderabad (India). She had completed M.Tech(CSE) at JNTUK University at Kakinada (India). She is in software Industry for more than 8 years. She is in teaching profession for more than 8 years. Her main area of interest includes Data Mining, Sentiment Analysis and Computer networks.	
Venkatesham Tunge I am T. Venkatesham, a Final-Year B. Tech Student at ACE Engineering College, specializing in CSE (Data Science). I am passionate about coding and problem-solving in Python and Flutter Developer. I strive to improve myself continuously and innovate new things in the tech world. My goal is to explore industry work cultures and contribute to impactful technological advancements. I complated my internship in health letic lifestyle startup company in 3 months.	
Nithinkumar Kanagala I am K. Nithinkumar, a Final-Year B. Tech Student at ACE Engineering College, specializing in CSE (Data Science). I am passionate about coding and problem-solving In Java, C Programming and Web Developer. I strive to improve myself continuously and innovate new things in the tech world. My goal is to explore industry work cultures and contribute to impactful technological advancements and becoming as a front end developer and java developer.	
Harsha Verdhan Bhumandla I am B Harsha Vardhan is currently pursuing a B.Tech in Computer Science and Engineering (Data Science). His research interests include Deep Learning, with a focus on leveraging advanced computational techniques for data-driven applications. As an undergraduate researcher, he is passionate about exploring machine learning models to solve real-world challenges, particularly in intelligent automation and pattern recognition.	
Shruti Kana I am K Shruti, currently pursuing a B.Tech in Computer Science and Engineering with a specialization in Data Science. My academic journey has been driven by a deep interest in computer science, particularly in the field of machine learning. I have gained valuable experience. As an undergraduate, I am passionate about using data science to tackle real-world problems. I look forward to continuing to explore and contribute to this rapidly evolving field.