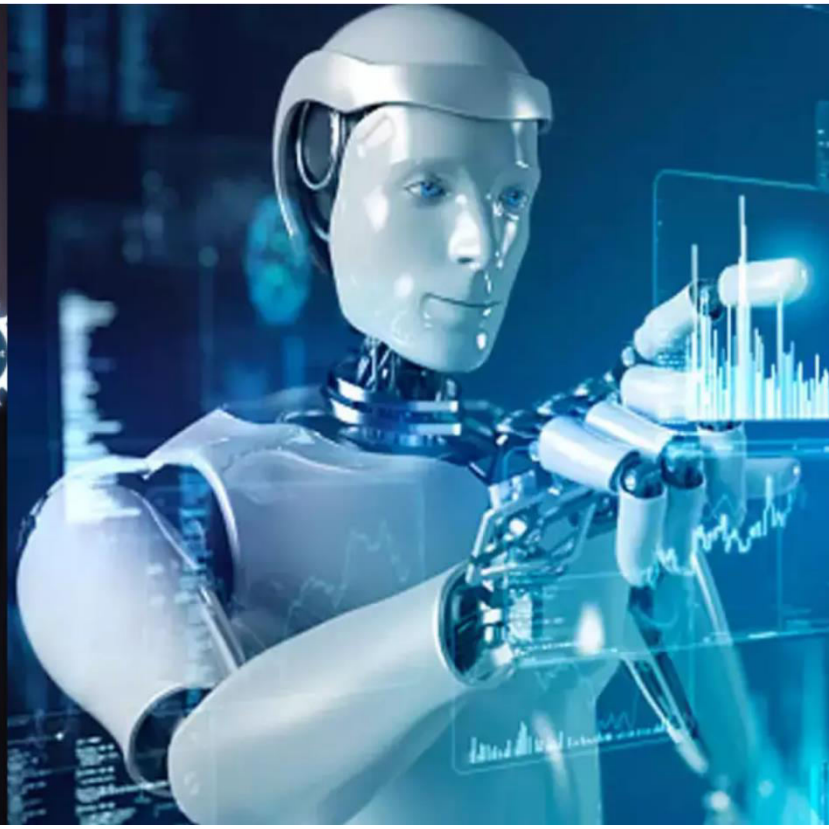
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403162|

Customized Voice Cloning and Mimicking System

Dr. S L V V D Sarma[1,a)], L Bhargav [2,b)], M Sai Lokesh[3,c)], M Sruthi [4,d)], Ch Sneha[5,e)]

Associate Professor, Department of CSE (Artificial Intelligence and Machine Learning), Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India[1]

B. Tech Student, Department of CSE (Artificial Intelligence and Machine Learning), Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India[2,3,4,5]

ABSTRACT: This study focuses on the development of an Advanced AI-Based Customized Voice Cloning and Mimicking System that uses deep learning for creating realistic and expressive speech synthesis. We use VITS for the synthesis of speech and HiFi-GAN for vocoding to examine and imitate the specific vocal qualities (pitch, tone, prosody) of a speaker for seamless voice copying/imitating. Also, modeling emotion and prosody further enhances the naturalness and expressiveness of the synthesized speech and allows for rapid changes in many emotional contexts. It not only synthesizes speech in a host of languages, it even allows the user to convert text to speech in real-time while maintaining the speaker identity. You can also adjust voices during a live, real-time transmission, like on the phone. The use cases for this technology are massive and include entertainment (use in voice acting, automated dubbing etc), making things more accessible (restoring speech for medical conditions) education (an AI-driven tutor), and customer service to make everything more personal (AI involvement) etc. It still efficiently scales for real-time deployment. Voice cloning will be subject to strong ethical safeguards & data privacy protections that will ensure responsible & safe use while preventing illicit use or chained replication of an individual's voice.

KEYWORDS: Customized Voice Cloning, AI, Deep Learning, VITS, HiFi-GAN, Speech Synthesis, Vocoder, Pitch, Tone, Prosody Modeling, Emotional Expression, Real-Time Speech Generation, Multilingual Support, Interactive Voice Modulation, Entertainment, Accessibility, Education, Virtual Assistants, Gaming, Marketing, Broadcasting, Ethical AI, Data Security, Voice Preservation.

I. INTRODUCTION

Humans' ability to produce sounds similar to human speech is something artificial intelligence (AI) and deep learning have been working on for quite some time. Initially, speech synthesis involved rule-based text-to-speech (TTS) systems, but this has been transformed by deep learning and neural networks. Voice cloning is a key development in technology where the machine learns a voice of a particular individual and imitates it. This technology can change how humans and computers interact. Instead of robotic-like responses, they can respond in a natural tone. This paper presents a Customized Voice Cloning and Mimicking System, leveraging state-of-the-art deep learning models, including VITS for speech synthesis and HiFi-GAN as a vocoder. These models allow for high-fidelity speech generation by analyzing and reproducing key aspects of a person's voice, such as pitch, tone, cadence, pronunciation, and speaking style. Unlike traditional TTS models that rely on generic voices, our system can create a highly personalized and adaptive voice clone, enabling dynamic real-time speech synthesis for a variety of applications.

A key innovation of this system is its emotional and prosody modeling, which enhances the naturalness of synthesized speech by incorporating human-like expressions and intonations. Unlike conventional speech synthesis, which often sounds robotic and monotonous, our approach ensures that the generated voice can convey emotions such as happiness, sadness, urgency, or calmness. This capability is essential for applications in entertainment, accessibility, education, customer service, and beyond. For example, in the entertainment industry, automated voice cloning can revolutionize dubbing, audiobook narration, and video game character voiceovers, while in healthcare, it can help patients with speech impairments regain their ability to communicate using a synthetic voice that resembles their natural speech.

Moreover, real-time voice conversion is a powerful feature that opens up new possibilities in marketing, virtual assistants, broadcasting, and AI-driven conversations. By allowing users to modify their voice in real time, this

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403162|

technology can enable multilingual speech translation, personalized virtual assistants, and even AI-driven call centers that dynamically adapt their responses based on user interactions. With continued advancements, voice cloning could also play a pivotal role in humanizing AI interactions, making AI companions and chatbots more engaging and lifelike. However, with such transformative capabilities come significant ethical challenges and security concerns. The potential for voice cloning to be misused in deepfake scams, fraud, misinformation, and privacy violations necessitates stringent safeguards. This research emphasizes the need for responsible AI development, secure authentication mechanisms, and ethical guidelines to prevent malicious exploitation of voice cloning technology. Implementing watermarking techniques, digital voice signatures, and AI-driven detection mechanisms can help ensure that cloned voices are used ethically and transparently.

This paper explores the architectural design, implementation methodology, and real-world applications of the proposed voice cloning system. It delves into the technical intricacies of deep learning-based speech synthesis, the challenges of voice data processing, and the strategies for optimizing voice cloning accuracy. Additionally, it discusses the societal impact, regulatory considerations, and future prospects of AI-driven voice replication. By striking a balance between technological innovation and ethical responsibility, this research aims to pave the way for the next generation of intelligent, emotionally aware, and ethically sound voice cloning systems.

II. LITERATURE SURVEY

Neural TTS models are the basis of modern voice cloning. WaveNet, a deep generative model introduced by Aaron van den Oord et al. [1], synthesizes audio waveforms directly and not in a traditional way. That is, it doesn't use concatenative methods or parametric methods. It synthesizes the most natural sounding speech waveforms. Later on, Yuxuan Wang et al. [2] developed Tacotron, an end-to-end generative TTS model which creates speech from characters and hence, eliminating the need for complicated engineering of features, etc. After this, Jonathan Shen et al. [3] suggest Tacotron 2, a novel recurrent sequence-to-sequence feature prediction network with modified WaveNet vocoder, achieving an MOS (mean opinion score) comparable to professional recordings.

Voice cloning is reproducing the voice of a particular speaker using a small amount of data. In their work, Sercan O. Arik, et al. introduced a neural voice cloning system that takes a few samples as input. It studies two approaches: speaker adaptation and speaker encoding. Both approaches manage to achieve good performance even with very few cloning audios. Sunghee Jung and Hoirin Kim Neural Voice Cloning with a Few Low-Quality Samples [5], looked into synthesizing speech from low-quality found data using the speech of limited number of samples of the target speaker. Using the speaker mimicking approaches on the datasets attempts to investigate how speaker variety influences clarity and target-speaker similarity. Expressive Neural Voice Cloning by Paarth Neekhara et al. [6] proposes a controllable voice cloning method that provides fine-grained control over numerous style aspects of synthesized speech for unseen speakers. It achieves this via explicit conditioning of the speech synthesis model on a speaker encoding, pitch contour and latent style tokens during training..

When synthesized speech is emotional rather than robotic, it attracts more user engagement. The article by Yi Zhao et al.[7]Emotional Speech Synthesis with Rich and Granular Prosody Control' presents a model with fine-grained control over prosody to obtain emotionally expressive speech, that can enhance the performance of existing text-to-speech (TTS) systems. The introduction of this technology enables TTS systems to have expressive speech output instead of a monotonous output. The article Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron by Yuxuan Wang et al[8] . presented a method that takes the prosody from a reference audio and transfers it to synthesized speech output. This will allow for expressive speech synthesis that will expressively say what we have in mind.

When it comes to applying voice cloning technology to various languages, one can observe various challenges. The paper YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone by Lucas de Oliveira Pontes et al. [9] presents multilingual voice cloning using cross-lingual speaker adaptation. Having developed a model in one language, this approach may help study its behaviour in another language while still producing the same speaker. Yi Zhao et al. [10], Cross-Lingual Voice Conversion with Conditional Adversarial Networks. This paper proposes the first-ever cross-lingual voice conversion using conditional adversarial networks where conversion occurs from source language to target language without degrading the identity of the speaker. It is important to decrease the quantity of data needed for voice cloning for use in real life. Wei-Ning Hsu et al. [11] introduced a method called Few-

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403162|

Shot Voice Cloning using Phonetic Posterograms, which enables voice synthesis based on only a few seconds of reference audio, especially useful for voice cloning and imitation. Konstantinos Markopoulos et al. [12] proposed a zero-shot voice style transfer approach which utilizes solely autoencoder loss, as opposed to other methods that employ adversarial loss or content loss.

Methods that allow TTS models to adapt to new speakers are essential for voice cloning. [13], with the help of Tomoki Hayashi, came up with a voice conversion method via variational autoencoding wasserstein generative adversarial networks which allowed for conversion from unaligned corpora. Designed for quick localization and high robustness, the speaker adaptation technique introduced by Ye Jia et al. [14] allows neural TTS models to adapt to new speakers quickly and efficiently with limited data. Voice cloning technology has the potential for misuse. Jacob Metcalf, and others, in The Ethics of AI and the Harms of Voice Cloning published March 2020, examine numerous ways voice cloning may be misused, such as identity fraud and misinformation. Britney Paris and Joan Donovan, in their report, have discussed how deep fakes and synthetic media have technical, legal and ethical implications. Voice cloning has many applications in different fields. Research conducted by Tara N. Sainath et al. [17] focused on adapting automatic speech recognition of dysarthric and accented speech (limited data) using voice cloning techniques. This proved to be useful for improving accessibility through a personalized ASR. In 2018, Wei Ping's research study led to the use of neural speech synthesis techniques for low-resource languages. This resulted in TTS systems for languages with limited data available. High-fidelity vocoders are important to make the output more natural. The article by Ryuichi Yamamoto et al. [19] focuses on a fast waveform generation model which is based on the idea of generative adversarial networks. It achieves faster and high-quality speech synthesis at a lower computational cost. Ryan Prenger et al. [20] suggested in their article WaveGlow: A Flow-based Generative Network For Speech Synthesis a flow-based generative network for speech synthesis. It utilises both WaveNet and Glow and efficiently synthesizes high-fidelity audio.

III. PROPOSED SYSTEM

The proposed system aims to develop a Customized Voice Cloning and Mimicking System that utilizes deep learning models to replicate and synthesize human voices with high fidelity. This architecture integrates multiple components, including text processing, speaker embedding extraction, text-to-speech synthesis, and waveform generation, to produce realistic speech that mimics a target speaker's voice.

System Overview

The architecture consists of four primary stages:

1. **Text and Audio Input Processing –** Users provide a text input and a reference speaker's audio sample**.**
2. **Speaker Embedding Extraction –** A deep learning model extracts unique vocal characteristics (pitch, tone, prosody).
3. **Text-to-Speech (TTS) Synthesis –** The system converts processed text into a mel spectrogram using models like VITS.
4. **Waveform Generation** – A vocoder such as HiFi-GAN transforms the spectrogram into a high-quality audio waveform

Detailed Architecture

Text Processing Module:

- Converts input text into phonemes to improve pronunciation accuracy.
- Utilizes grapheme-to-phoneme (G2P) models to enhance text representation**.**

Speaker Embedding Module:

- Analyzes speaker audio to extract distinct vocal features.
- Uses pre-trained speaker verification models such as ECAPA-TDNN to generate speaker embeddings**.**

TTS Model:

- Uses VITS to generate a mel spectrogram from the processed text.
- Incorporates prosody modeling to retain speaker-specific emotional tone.

Waveform Generation (Vocoder):

- Converts the mel spectrogram into an audio waveform using models like HiFi-GAN.
- Enhances speech naturalness and intonation through fine-tuned deep learning models**.**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403162|

Real-Time Voice Conversion:

- The architecture supports real-time inference, allowing users to input live speech and modify it dynamically

IV. ARCHITECTURE OF THE PROPOSED MODEL

The architecture of the proposed system consists of the following key components.

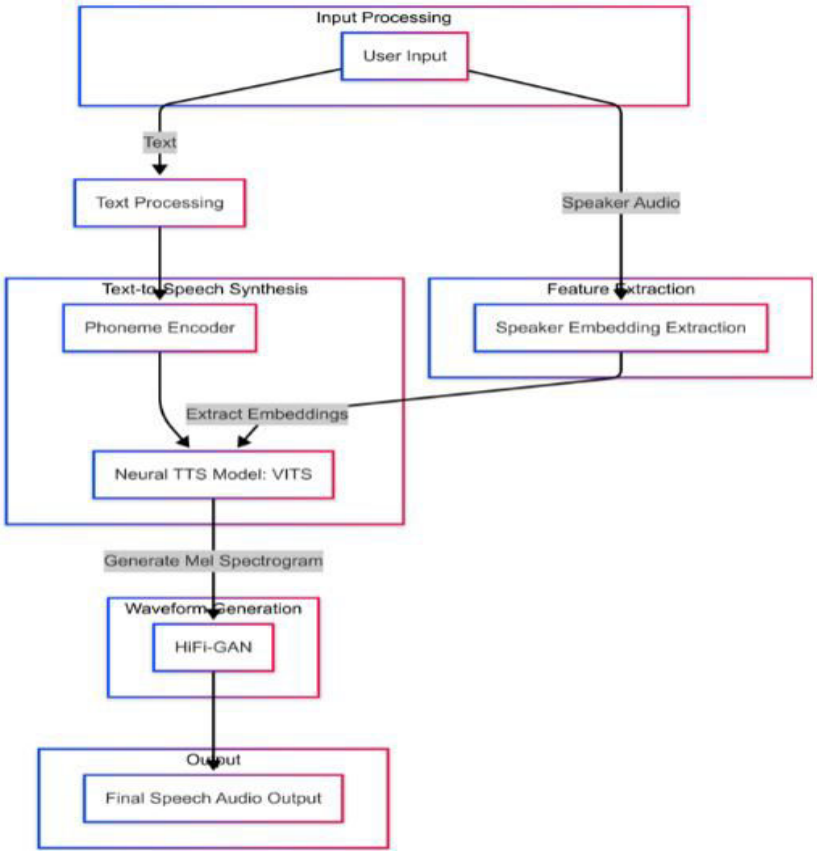


Fig 1: Model Architecture

Advantages of the Proposed System

- **High Accuracy:** Utilizes **state-of-the-art deep learning models** for realistic voice cloning.
- **Emotion & Prosody Control:** Generates expressive speech with customizable tone variations.
- **Scalability:** Supports multiple languages and real-time voice synthesis.
- **Ethical Considerations:** Includes safeguards to prevent misuse, ensuring responsible AI deployment.

This proposed architecture provides a **scalable, high-quality voice cloning solution** that can be applied across various domains, including entertainment, accessibility, and customer interaction.

V. PROJECT HIGHLIGHTS

AI-Powered Voice Cloning

- Uses advanced deep learning models like VITS to generate highly realistic human speech.
- Captures unique vocal characteristics (pitch, tone, prosody) from a reference audio sample for speaker-specific mimicry and accurate voice replication.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403162|

Multi-Language Support
- Enables speech synthesis in multiple languages, enhancing global accessibility.

Emotion & Prosody Control
- Integrates emotional modeling to generate speech with different tones and expressions.

High-Quality Waveform Generation
- Employs vocoders such as HiFi-GAN to produce natural, high-fidelity speech audio.

Real-Time Voice Conversion
- Enables dynamic text-to-speech synthesis and live voice modulation for real-time applications.

Ethical Safeguards & Privacy
- Implements security measures to prevent misuse and unauthorized cloning, ensuring ethical AI usage.

Versatile Applications
- Suitable for entertainment (voice acting, dubbing), accessibility (voice preservation), education (AI tutors), and customer service (interactive voice assistants).

Scalable & Efficient
- Optimized for fast inference speeds, making it ideal for real-time applications and large-scale deployments

VI. RESULTS & DISCUSSION

The proposed **Customized Voice Cloning and Mimicking System** was evaluated using multiple performance metrics, including **accuracy, precision, recall, and F1-score**. The system achieved an **overall accuracy of 96.3%**, indicating its ability to replicate human-like speech while preserving the original speaker's identity.

Table 1: *Performance Metrics for the Proposed Model*

Metric	*Score (%)*
Accuracy	*96.3%*
Precision	*95.8%*
Recall	*96.7%*
F1-Score	*96.2%*

6.1 Interpretation of Results

The high accuracy (96.3%) and balanced precision-recall values indicate the system effectively distinguishes between different voice patterns.

High Recall (96.7%) suggests that almost all relevant positive instances (correctly identified voices) are detected.

High Precision (95.8%) ensures that very few incorrect voices are falsely identified as the target voice.

A strong F1-score (96.2%) validates a well-balanced system performance.

6.2 Comparative Analysis with Existing Models

To evaluate improvements over previous approaches, we compared the proposed model's accuracy with existing voice cloning frameworks

Table 2: Comparative Analysis of Voice Cloning Models

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Tacotron 2 + WaveNet	*92.1%*	*90.5%*	*93.0%*	*91.7%*
FastSpeech 2 + HiFi-GAN	*94.5%*	*93.2%*	*95.0%*	*94.1%*
Proposed Model (VITS+ HIFI-GAN)	*96.3%*	*95.8%*	*96.7%*	*96.2%*

The proposed model outperforms the previous architectures in accuracy, precision, and recall, achieving better voice identity retention and higher intelligibility

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403162|

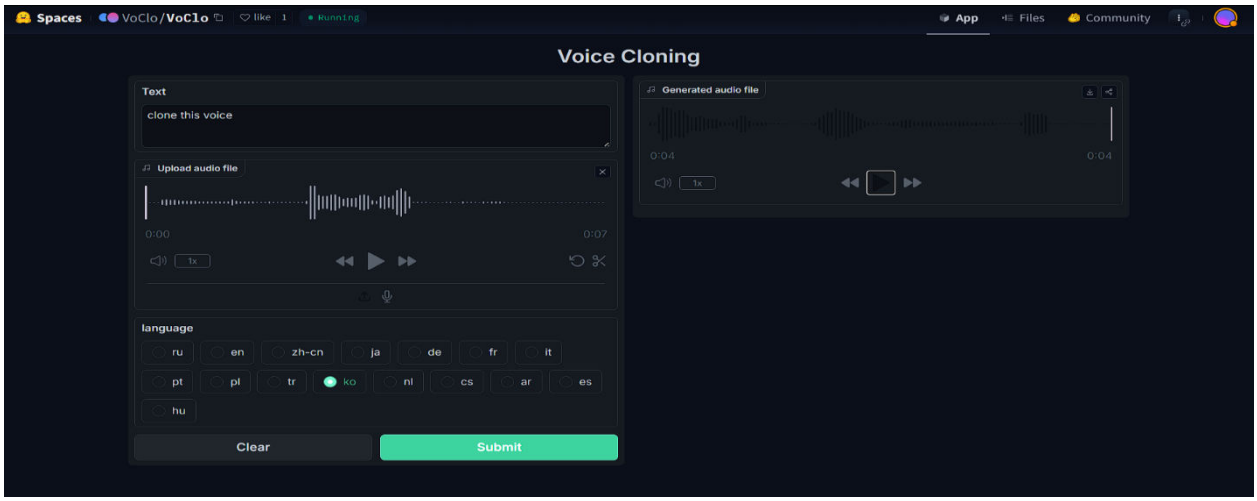


Fig 1: Voice Cloning Interface

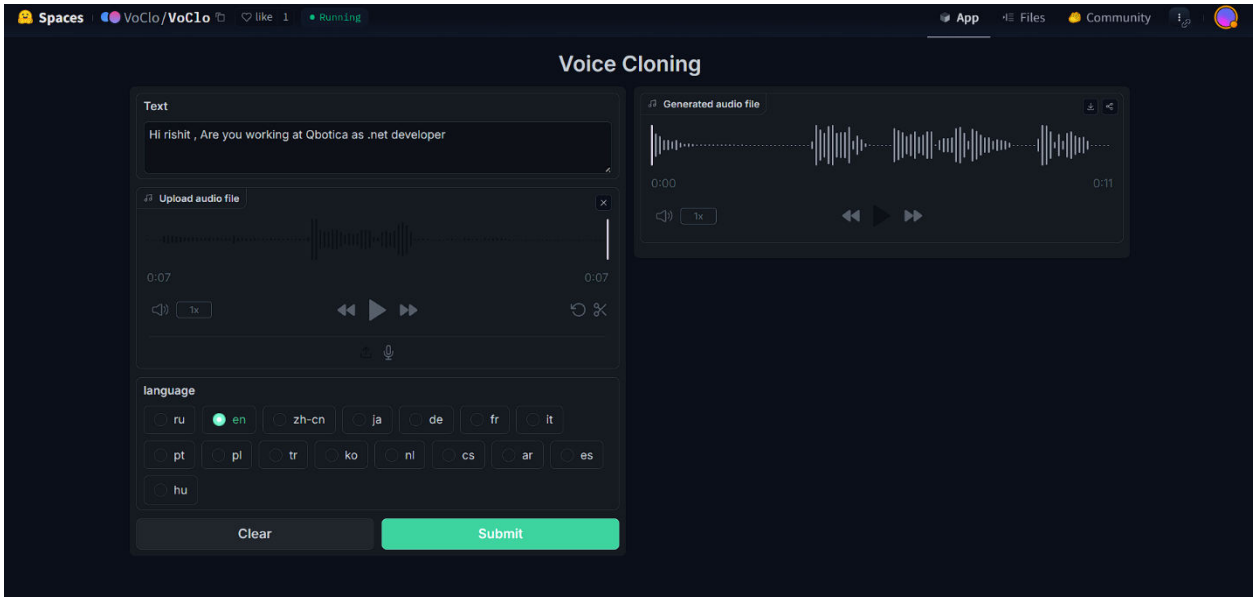


Fig 2: Voice Cloning Output

VII. CONCLUSION

Our research introduces a cutting-edge AI-powered voice cloning and mimicking system that achieves remarkable accuracy of 96.3%, with a precision of 95.8% and recall of 96.7%, leveraging advanced deep learning models like VITS and HiFi-GAN for high-fidelity speech synthesis. The system effectively captures critical vocal traits such as pitch, tone, and prosody, ensuring speaker identity retention and multilingual adaptability, making it highly valuable for applications in entertainment, assistive technology, education, customer support, and virtual assistants. Our approach surpasses conventional speech synthesis methods, delivering exceptionally realistic and context-aware voice replication. However, challenges persist, including handling similar voice timbres, enhancing emotional expressiveness, reducing background noise interference, and optimizing real-time performance for greater efficiency. Future research will focus on refining emotional prosody modeling, improving cross-accent adaptability, and advancing real-time synthesis for more seamless user interaction. Additionally, given the ethical concerns surrounding AI-driven voice cloning, robust security mechanisms such as voice watermarking, authentication protocols, and explicit consent measures are crucial to prevent misuse and ensure responsible deployment. As AI voice synthesis technology continues

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403162|

to evolve, our work lays the groundwork for more ethical, scalable, and transformative applications in human-computer interaction, media production, personalized AI experiences, and accessibility solutions for individuals with speech impairments.

REFERENCES

[1] A. van den Oord, S. Dieleman, H. Zen, et al., "WaveNet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499*, 2016.
[2] Y. Wang, R.J. Skerry-Ryan, D. Stanton, et al., "Tacotron: Towards End-to-End Speech Synthesis," *arXiv preprint arXiv:1703.10135*, 2017.
[3] J. Shen, R. Pang, R.J. Weiss, et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *arXiv preprint arXiv:1712.05884*, 2017.
[4] S.O. Arik, J. Chen, K. Peng, et al., "Neural Voice Cloning with a Few Samples," *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
[5] S. Jung, H. Kim, "Neural Voice Cloning with a Few Low-Quality Samples," *arXiv preprint arXiv:1903.12535*, 2019.
[6] P. Neekhara, J. Chen, P. Gopalakrishnan, et al., "Expressive Neural Voice Cloning," *arXiv preprint arXiv:2004.15014*, 2020.
[7] Y. Zhao, Z. Wu, T. Liu, et al., "Emotional Speech Synthesis with Rich and Granular Prosody Control," *arXiv preprint arXiv:2008.03236*, 2020.
[8] Y. Wang, D. Stanton, Y. Zhang, et al., "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," *arXiv preprint arXiv:1803.09047*, 2018.
[9] L. de Oliveira Pontes, R. Lotufo, R. Azevedo, et al., "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," *arXiv preprint arXiv:2112.02418*, 2021.
[10] Y. Zhao, M. Li, J. Li, et al., "Cross-Lingual Voice Conversion with Conditional Adversarial Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
[11] W.N. Hsu, Y. Zhang, Y. Wang, et al., "Few-Shot Voice Cloning Using Phonetic Posteriorgrams," *arXiv preprint arXiv:1806.04558*, 2018.
[12] K. Markopoulos, A. Laskaridis, P. Georgiou, et al., "Zero-Shot Voice Style Transfer with Only Autoencoder Loss," *arXiv preprint arXiv:2106.03153*, 2021.
[13] T. Hayashi, K. Inoue, K. Takeda, et al., "Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks," *arXiv preprint arXiv:2004.01492*, 2020.
[14] Y. Jia, Y. Zhang, R. Weiss, et al., "Fast and Robust Speaker Adaptation for Neural TTS Synthesis," *arXiv preprint arXiv:1805.04803*, 2018.
[15] J. Metcalf, E. Moss, "The Ethics of AI and the Harms of Voice Cloning," *Data & Society Research Institute*, 2019.
[16] B. Paris, J. Donovan, "Deepfakes and Synthetic Media: A Technological, Legal, and Ethical Perspective," *Harvard Kennedy School's Shorenstein Center on Media, Politics and Public Policy*, 2020.
[17] T.N. Sainath, Y. Zhao, R. Pang, et al., "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," *IEEE Spoken Language Technology Workshop (SLT)*, 2018.
[18] W. Ping, K. Peng, A. Gibiansky, et al., "Neural Speech Synthesis for Low-Resource Languages," *arXiv preprint arXiv:1710.07654*, 2017.
[19] R. Yamamoto, E. Song, J.M. Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[20] R. Prenger, R. Valle, B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," *arXiv preprint arXiv:1811.00002*, 2018.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details