IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

Mimic Mania – AI Text-to-Speech (TTS) Generator

Megha Karmakar[1] and Prof . Ashwini Mahajan[2]
Student, Department of Computer Science and Engineering[1]
Co-Guide, Department of Computer Science and Engineering[2]
Abha Gaikwad Patil College of Engineering, Nagpur, Maharashtra
meghakarmakar236@gmail.com, ashwinimahajan@gpg.in

Abstract: *Artificial voice cloning replicates human speech characteristics using deep learning. This study explores neural architectures that synthesize lifelike audio with limited voice samples. A zero-shot approach is employed, leveraging a speaker recognition model transferred to a TTS setting. Our system effectively clones voices, even those it hasn't been trained on, by generating natural-sounding speech with minimal input.*

Keywords: Synthetic Speech, Voice Cloning, Deep Learning, Zero-Shot Learning, Tacotron, Neural TTS, Speaker Embedding

I. INTRODUCTION

Voice cloning involves creating a digital voice replica of an individual using artificial intelligence. This technology finds applications in various sectors—from healthcare (restoring lost voices) to entertainment, customer service, and smart assistants. With increasing demand for personalized and accessible interfaces, voice cloning is poised to become a cornerstone in the evolution of human-computer interaction.

II. LITERATURE REVIEW

TTS systems have progressed from rudimentary rule-based algorithms to highly accurate deep learning models. Early methods were limited in expressiveness and required extensive datasets. Modern systems like Tacotron and WaveNet utilize neural networks to deliver high-fidelity, expressive speech. These advances integrate linguistic modelling, acoustic processing, and neural synthesis, enabling multi-speaker support and real-time voice cloning.

With AI-driven solutions gaining traction, the need for personalized, adaptable, and realistic voice synthesis has become increasingly vital.

Several researchers have also highlighted the importance of inclusivity in TTS. Studies point out that many models are biased toward dominant languages and accents, often underrepresenting regional dialects and minority voices (Tatman, 2017). There is ongoing research to ensure fairer and more diverse TTS datasets.

Overall, the literature reveals a consistent trend toward realism, adaptability, and efficiency in TTS systems, driven by advances in AI and neural networks. However, the technology's rapid development demands a parallel emphasis on ethical frameworks and responsible use.

III. METHODOLOGY

Our approach utilizes a deep learning pipeline comprising:

- Speaker Encoder: Extracts speaker- specific features into a fixed-size vector from short audio input.
- Synthesizer: Uses a sequence-to- sequence model (e.g., Tacotron) to convert text into a Mel spectrogram conditioned on the speaker embedding.
- Vocoder: Reconstructs time-domain audio from the spectrogram using a neural vocoder like WaveNet.

This pipeline enables efficient voice synthesis with limited data. Python-based frameworks and pretrained models were utilized to streamline development and ensure reproducibility.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

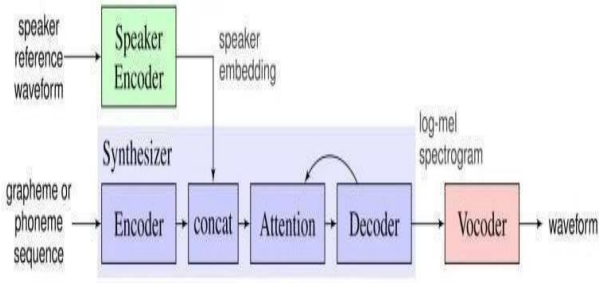

Figure: System Block Diagram

Applications of TTS

- Accessibility: TTS helps visually impaired individuals and those with reading disorders access content.
- Virtual Assistants: Tools like Siri, Alexa, and Google Assistant rely heavily on TTS.
- E-learning: TTS enhances learning by reading aloud educational material.
- Entertainment & Media: Voiceovers, audiobooks, and gaming now use AI- generated voices.
- Multilingual Communication: TTS enables real-time translation and localization.

IV. RESULTS AND DISCUSSION

The developed system effectively synthesized speech resembling the reference voices. Given the subjective nature of speech quality, we employed the Mean Opinion Score (MOS) technique to gather user feedback on intelligibility and naturalness.

The system successfully generated speech samples similar to the reference voices. Evaluation through the Mean Opinion Score (MOS) revealed that listeners found the cloned voices close to natural in terms of tone and clarity, though minor issues in prosody and background noise were noted. These findings affirm the model's potential and highlight areas for future refinement, such as emotion rendering and noise reduction.

Real Image of My App With full of working:

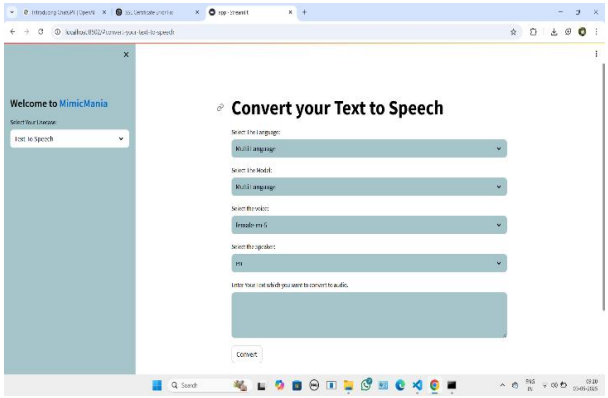

Figure: Text to Speech Conversion

How Modern TTS work

Modern TTS systems operate in two key stages:

- Text Processing: Input text is linguistically analysed—punctuation, context, and syntax are considered.
- Acoustic Modelling and Vocoding: Deep learning models predict how the text should sound, generating spectrograms. These are then converted into waveform audio using vocoders like Wave Glow or HiFi-GAN.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, May 2025
Impact Factor: 7.67

The integration of Natural Language Processing (NLP) and neural networks ensures nuanced, expressive speech generation.

V. CONCLUSIONS

This research demonstrates a viable method for low-resource voice cloning using a zero-shot neural TTS system. By extending speaker verification models to TTS synthesis, we achieved competent voice replication with minimal data. Future enhancements will aim to improve voice expressiveness and naturalness, bringing AI-generated speech even closer to human standards.

ACKNOWLEDGEMENT

We extend heartfelt thanks to the Department of Computer Science at Gaikwad Patil College for infrastructure support. Special appreciation is due to the Environmental Tech Lab for insightful feedback and assistance during the system's development and testing phases.

REFERENCES

[1]. Loupe, G., & Jemine, C. "Neural MultiSpeaker Voice Emulation Using Deep Learning." University of Liege.
[2]. Dieleman, S., et al. "WaveNet: Audio Sample Generation via Deep Neural Networks." arXiv:1609.03499
[3]. Simonyan, K., et al. "Accelerated Audio Synthesis with Neural Architectures."
[4]. Skerry-Ryan, R. J., et al. "Enhanced TTS via Spectrogram-Based Conditioning in WaveNet." arXiv:1712.05884
[5]. James, C. "Real-Time Speaker-Specific Voice Cloning Using Open-Source Tools."
[6]. Gilles Loupe, Corentin Jemine. Master Thesis: Automatic Multispeaker Voice Cloning. Faculty of Science Applications, University of Liege. URL http://hdl.handle.net/2268.2/6801
[7]. Sander Dieleman, Heiga Zen, Aaron van den Oord, Karen Simonyan, Nal Kalchbrenner, Andrew W. Senior, Oriol Vinyals, Alex Graves and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. CoRR, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499.
[8]. Karen Simonyan, Seb Noury, Norman Casagrande, Nal Kalchbrenner, Erich Elsen, Edward Lockhart, Florian Stimberg, Sander Dieleman, and Koray Kavukcuoglu, Aaron van den Oord. Efficient neural audio synthesis, 2018.