International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Real Time Voice Cloning Using Generative Adversarial Network

[1]Mr. C Hrishikesava Reddy, [2]D. Rajasekhar, [3]T. Haritha, [4]B. Ranjit Naik, [5]N. Srinivasula Reddy

Submitted: 21/11/2024 **Revised:** 28/11/2024 **Accepted:** 05/12/2024 **Published:** 19/12/2024

Abstract: This research introduces a cutting-edge approach to real-time voice cloning by harnessing the capabilities of Generative Adversarial Networks (GANs). Voice cloning involves creating a digital reproduction of a person's voice that closely mimics their natural speech. Traditional techniques often require extensive datasets and lengthy processing times, making them less practical for real-time applications. Contrasting, the proposed method uses the goodness of GANs so that it greatly reduces data size and processing time, in addition to giving excellent outputs. Our model is trained on a wide variety of speech samples. This enables our model to capture and replicate the distinctive features of an individual's voice. The framework is built into two major components: the generator, which synthesizes voice outputs, and the discriminator, which evaluates the authenticity of these outputs. These two components interact in a continuous feedback loop through adversarial training, thereby continually improving the quality and realism of the generated speech. The system is designed to be highly efficient, running seamlessly on standard hardware configurations. This makes it more accessible for a wide range of applications, such as personalized voice assistants, custom voice-overs, and enhanced gaming experiences that rely on immersive audio. Experimental evaluations show that the GAN-based approach not only generates highly realistic voice clones but also retains the distinctive characteristics of the target speaker's voice. Moreover, a comparative analysis shows that this approach is superior to traditional voice cloning methods in terms of output quality and computational efficiency. This study is a significant step in the advancement of voice cloning technology by introducing a faster and more efficient way to generate lifelike voice replicas. It opens new possibilities for interactive voice-driven solutions and sets the stage for further innovations in personalized audio applications.

Keywords : Voice cloning Generative, Adversarial Networks, Speaker encoder, Synthesizer, Deep learning vocoder

1. Introduction

Real-time voice cloning has combined as an important area of research within artificial intelligence (AI) and speech synthesis focusing on the ability to mimic a person's voice instantly.This Layout enables the world of extremely pragmatic counterfeit language that mirrors the sound of amp peculiar person. A major breakthrough in this field is the Use of Generative Adversarial Webs (GANs) a class of calculater learning Representations renowned for generating lifelike high-quality Information. gans run done ii Combined Webs: the source which produces counterfeit information and the differentiator which assesses its legitimacy These two Webs engage in a continuous Method of Improvement as the generator learns to make more realistic outputs and the discriminator refines its evaluation. When applied to voice cloning GANs can be trained on audio samples of a person voice. Erstwhile the Check is fine it get get language inch the like sound with nominal stimulus much inch real-time. This capability opens up a wide range of Uses including personalized virtual assistants entertainment and accessibility tools. real-time sound cloning exploitation gans is notably cunning because it get natural-sounding language quick and expeditiously requiring inferior computational force compared to conventional sound deduction techniques However on with the important prospective of this engineering get important right challenges. While it can be used to improve personalized Encounters such as voice validation or digital avatars it also raises concerns around privacy consent and the risk of misuse such as the creation of fake voices. arsenic the engineering advances it leave work relevant to plant good and right

1 Assistant Professor, Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyala, Andhra Pradesh, India.

2,3,4,5 Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyala, Andhra Pradesh, India

guidelines to check its liable employ allowing order to gain from real-time sound cloning without vulnerable intimate rights or security.

2. Literature Survey

2.1 Related work

Merlijn Blaauw et al.,Voice cloning is becoming increasingly important for personalized voice applications. Modern neural network approaches to speech synthesis already deliver excellent results across numerous speakers. This study introduces a system that reduces the number of audio samples needed to clone a voice[1].

Aaron van den Oord et al., used WaveNet is a ground breaking model that generates raw audio waveforms through a deep neural network. It uses a probabilistic, autoregressive approach, predicting each audio sample based on prior ones. Despite its computational complexity, the model efficiently trains on vast datasets and processes thousands of audio samples per second. WaveNet significantly improves the naturalness of text-to-speech systems, surpassing traditional methods in languages like English and Mandarin. Furthermore, it can capture and replicate the unique characteristics of individual speakers, enabling easy transitions between voices by adjusting speaker IDs. In addition to speech, WaveNet is capable of producing creative and realistic musical compositions. It also holds promise for phoneme recognition tasks when applied as a discriminative model [2].

Giuseppe Ruggiero et al., tackles the issue of slow sampling in sequential models, focusing on text-to-speech systems. It introduces multiple techniques to improve efficiency while maintaining high-quality audio output: WaveRNN: A compact recurrent neural network with a dual softmax layer that rivals the quality of WaveNet but is faster and more efficient. It can produce 24 kHz audio at four times the speed of real-time on GPUs. Weight Pruning: Optimizes the model by removing redundant parameters, allowing real-time audio synthesis on mobile devices without compromising quality. Sparse networks with extremely high sparsity perform better than smaller dense ones with the same parameter count. Subscale WaveRNN: A novel method that splits long audio sequences into shorter, parallel ones, enabling simultaneous generation of multiple samples. This method improves sampling speed while preserving audio fidelity [3].

2.2 Problem statement

Create a real-time voice cloning system that uses Generative Adversarial Networks (GANs) to address the shortcomings of traditional Convolutional Neural Networks (CNNs) in replicating voices. While CNNs are good at analysing spatial patterns, they often fall short in capturing the subtle timing and tonal variations that make speech sound natural. By leveraging GANs, the system aims to produce more realistic and expressive voice clones. The solution will follow a three-step process: extracting the speaker's unique vocal features, synthesizing speech, and refining it with a vocoder to ensure the cloned voice is high-quality, natural, and lifelike [7].

2.3 Research Gap

Voice cloning systems often use Convolutional Neural Networks (CNNs) because they are efficient at processing audio data. However, CNNs mainly focus on spatial patterns and struggle to capture the timing and subtle nuances of human speech, which can result in voices that sound less natural. While some newer approaches have been explored, there hasn't been much focus on using Generative Adversarial Networks (GANs) for real-time voice cloning. GANs have the ability to generate high-quality, realistic outputs by modelling complex data patterns, making them a promising alternative. However, their application to the full voice cloning process—extracting vocal features, synthesizing speech, and refining it with a vocoder—hasn't been deeply studied, especially for real-time use. This research aims to fill that gap by exploring how GANs can improve the quality and naturalness of cloned voices while addressing challenges like speed and stability in real-time system

3. Proposed System

3.1 System Architecture

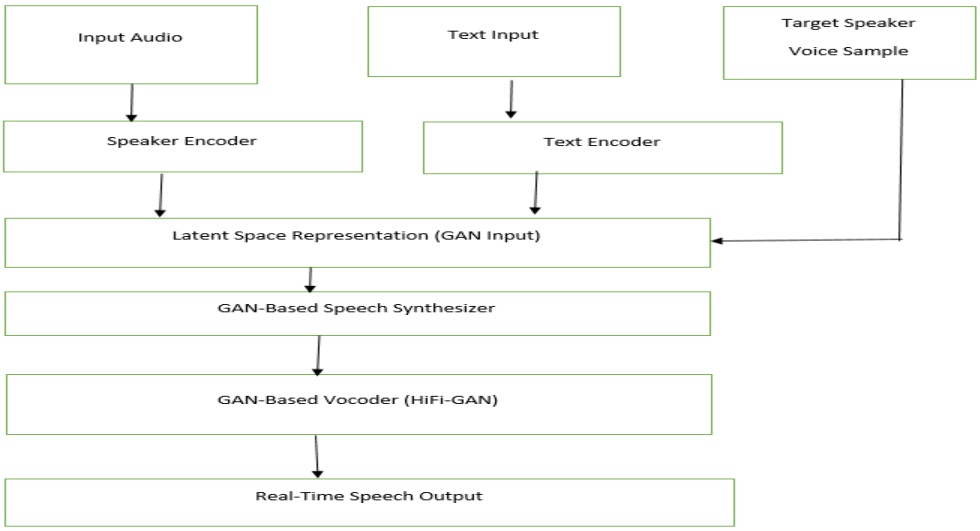

Fig 1: Proposed Architecture

3.2 Dataset

In our voice cloning project we have 8000 plus voice samples to make a system that can copy a specific person's voice. (Librispeech dataset has been widely used for automatic speech recognition and text-to-speech synthesis, providing a rich dataset for training models) [4]. Our information set has numerous recordings of the talker display disparate emotions way of sound and situations. This range helps the system learn all the little details of how the person sounds. Subsequently we beat the information we light it leading to beat free of ground dissonance and get complete the sound the like. Then the system looks at the speech focusing on elements like pitch tone and rhythm to figure out what makes the voice special. Once we have trained it we check the system's output to make sure the voice it makes sounds natural and matches the speaker's voice [8].

3.3 Algorithms

Multi talker deduction Representation

A multi-speaker deduction Check for real-time sound cloning exploitation gans focuses along generating genuine language for aggregate individuals. It works by encoding unique voice Characteristics into speaker embeddings which are then used by the GAN's generator to produce personalized speech. The differentiator ensures the yield sounds spurious and pragmatic. Teaching such a Representation requires a diverse information set with numerous speakers to capture a wide range of vocal traits. This allows the Check to synthesise voices inch real-time accurately mimicking disparate speakers founded along the Information set's characteristics

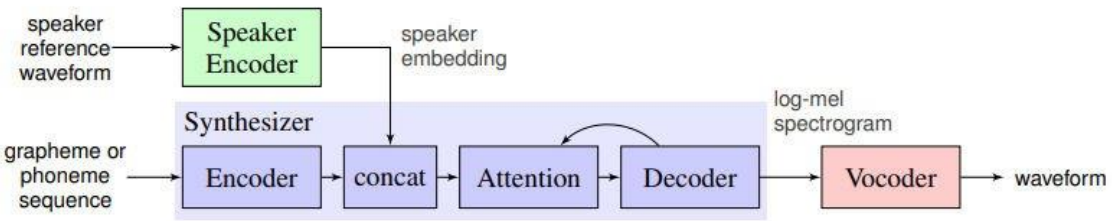

Fig 2: Check overview. Each of the three Parts are trained independently

Speaker Encoder

A speaker encoder in real-time voice cloning systems is a important element in generating high-quality personalized synthetic speech. it top Methodes amp speaker sound to elicit alone loud traits such as arsenic shift chanting beat and idiom. These Removeed characteristics are then

represented as a "speaker embedding" a condensed vector that encodes the distinctiveness of the speaker voice. this embedding is federal into amp creative adversarial net (gan) which uses it to synthesise language that close resembles the point talker. The GAN's generator Makes the voice while the discriminator ensures the output is realistic.away accurately Coding and representing the speaker loud Characteristics the encoder helps get natural

Synthesizer

In real-time sound cloning systems exploiting Generative Adversarial Networks (GANs), the synthesizer is an important factor in generating pragmatic, personalized speech. It works by taking speaker embeddings, which capture unique vocal traits such as pitch and tone, and combining them with linguistic input like phonetic or textual information to produce a natural-sounding speech waveform. The generator in the GAN uses these inputs to closely replicate the target speaker's voice, while the discriminator ensures that the synthesized speech maintains authenticity and sounds natural.Deep Voice 2 demonstrated that multi-speaker neural text-to-speech (TTS) models could efficiently synthesize high-fidelity speech while preserving speaker identity across multiple voices, making it a foundation for advanced voice cloning systems [5].These GAN-based systems can accurately reproduce a person's voice, enabling seamless interactions tailored to individual users. They also have significant potential in assisting individuals with speech impairments, providing synthetic voices that closely resemble their natural speech. As GAN-based synthesizers evolve, their efficiency improves, reducing the computational resources required for real-time voice cloning. This leads to faster and more accessible systems with broader applications in healthcare, entertainment, and customer service**.**

Vocoder

The vocoder plays an essential role in converting the Produced speech representation (like a spectrogram) into an actual audio waveform. This transition is important for producing high-quality natural-sounding language. Advanced vocoders such as WaveNet GAN or HiFi-GAN are typically used to synthesize high-fidelity speech in real-time. These Representations read to get Complicated sound Characteristics such as arsenic shift inflection and quality which are inevitable to hold the legitimacy of the Copyd sound. By using GAN-based vocoders the system can produce more accurate expressive and lifelike voices even in dynamic environments. This ensures that the language facultative personal and natural-sounding real-time sound cloning. This Tech has broad Uses from digital assistants and entertainment to accessibility tools for those with speech impairments. Furthermore advancements inch talker encoders leave for better deduction character with nominal information devising them progressively prompt for real-time use [6].

Produced language mimics the point talker close devising it good for different Uses such as arsenic practical assistants voiceovers and personal communicating systems. Also the Productivity and speed of the vocoder are decisive for enabling real-time Effectiveness without compromising on audio quality making it suitable for interactive or assistive technologies that demand minimal latency [9].

4. Results

The project demonstrated error-free operation, successfully cloning a target voice using text-to-speech (TTS) synthesis. Since voice quality is inherently subjective, it is challenging to quantify the accuracy of the cloned output. The most effective evaluation method is the Mean Opinion Score (MOS), which involves conducting a survey to gather subjective feedback on how closely the cloned voice resembles natural human speech. Based on the MOS results, the system’s output was found to be comparable to human voice, though it still displayed some limitations in terms of naturalness and accent. These areas offer opportunities for further refinement and development. The underlying processes of the system were confirmed to function as intended, providing a solid foundation for future enhancements. A user-friendly graphical interface was developed to make the framework accessible without requiring prior expertise. Named the "SV2TTS Toolbox", this interface is cross-platform, developed in Python using the Qt5 library, and supports a variety of common language records. It also allows users to add new records or record and duplicate their speech directly. When a user loads an utterance into the system, its embedding is automatically calculated, and the UMAP (Uniform Manifold Approximation and Projection) visualization is updated in real time. The System also generates a mel spectrogram for the loaded utterance, displayed for comparison purposes. However, these visualizations are illustrative and not used as direct performance measures. It is important to note that the embedding is a one-dimensional vector, and its square display format does not carry structural meaning [10].

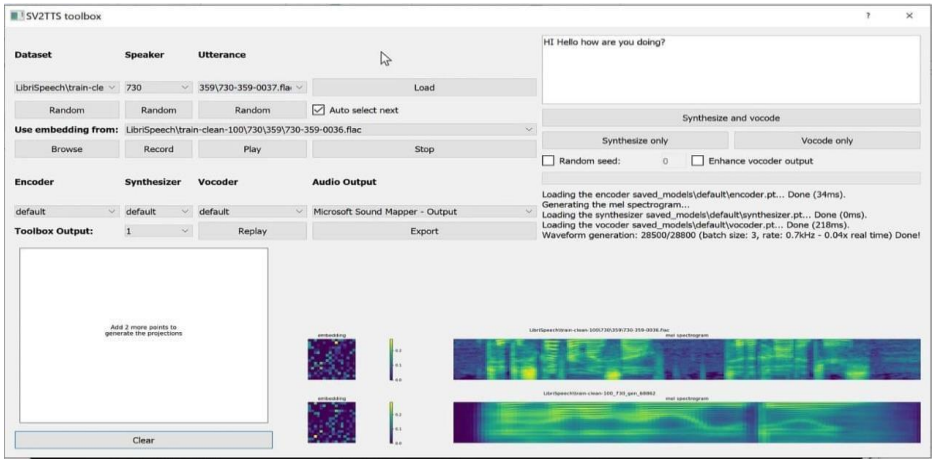

Fig 3: SV2TTS toolbox

The above interface also allows users to input any text for synthesis. However, due to template limitations, punctuation is unsupported, and users must insert line breaks manually to manage the rhythm of synthesized speech. The resulting segments are concatenated to form a complete spectrogram, which is displayed in the lower-right corner of the interface. Finally, the system utilizes a vocoder to convert the spectrogram into audible speech. The progress of the synthesis is displayed on an import bar. Once the process is complete, the generated utterance embedding is displayed alongside the composite spectrogram and projected in the UMAP. Users can rely on this embedding as a reference for guiding subsequent iterations of voice generation. This comprehensive interface makes the system intuitive and accessible while offering room for future improvements, particularly in enhancing the naturalness and accent of cloned voices. The below figures illustrates the waveform comparison between the original audio (blue) and the cloned audio (orange). The graph displays amplitude variations over a duration of few seconds. The cloned audio waveform closely follows the original audio's structure, maintaining significant fidelity in amplitude and timing. There are Minor deviations between the waveforms indicate limitations in the voice cloning process.

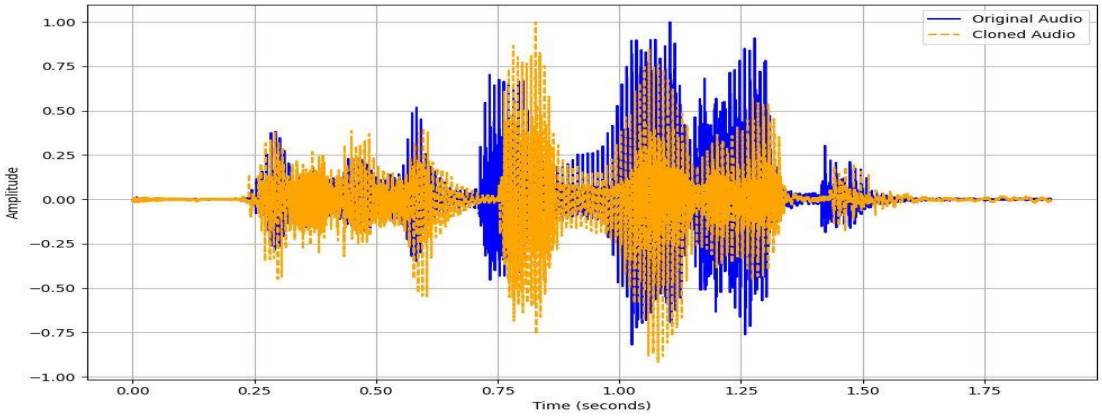

Fig 4: Waveform Comparison: Original vs Cloned Audio

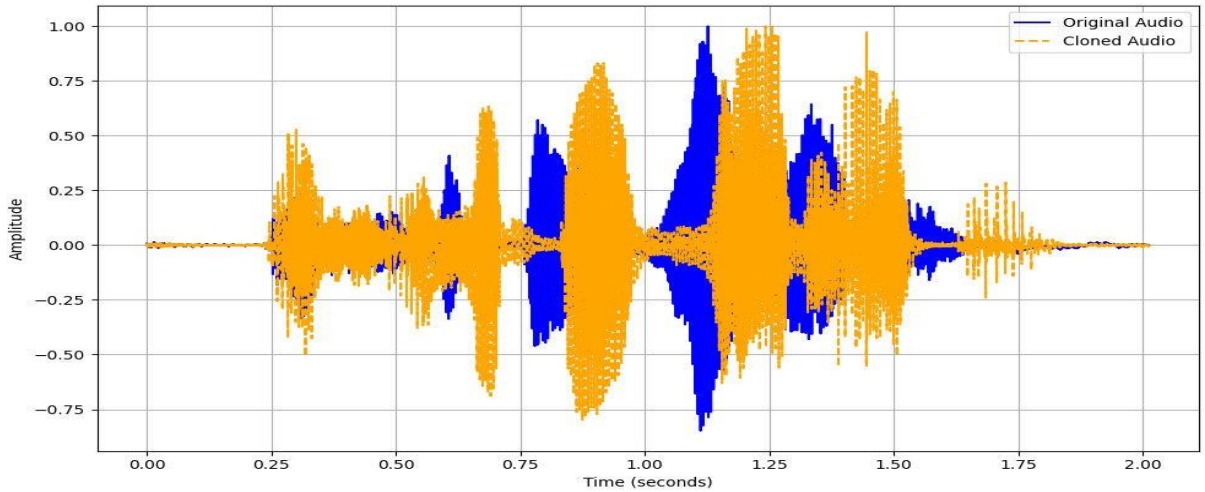

Fig 5: Waveform Comparison: Original vs Cloned Audio

5. Overall comparison analysis on Suggested model

Table 1. Comparison Analysis

Aspect	Real-Time Voice Cloning	Tortoise and Bark	Piper
Efficiency	Clones voice with just a few seconds of input audio.	Requires less training data but may not be as fast.	Requires ~3 hours of voice data for training.
Fidelity	May exhibit artifacts or slight deviations from the original.	Higher fidelity cloning compared to earlier models.	Quality can vary depending on dataset and tuning.
Processing Time	Real-time synthesis capability.	Likely slower than Real-Time Voice Cloning.	Processing time depends on the dataset size.
Data Requirements	Minimal—only a few seconds of audio input required.	Minimal training data needed for effective cloning.	Needs significant data (3+ hours) for training.
Accessibility	Open-source and highly accessible for experimentation.	Limited by implementation or resource availability.	Open-source but resource-intensive for high quality.
Applications	Immediate speech synthesis for real-time applications.	Suitable for higher-quality offline applications.	Suitable for custom voice cloning projects.
Limitations	Quality variations and artifacts; struggles with unique voices.	Relatively new; capabilities not fully established.	Resource-intensive; may not achieve optimal results.

6. Conclusion

It dived into the intricate and very dynamic area of real-time voice cloning using Generative Adversarial Networks (GANs). The central goal was to design a system that may generate synthetic speech pretty close to the voice of the target speaker, with minimal latency requirements for real-time applications. The project was successful, demonstrating the feasibility of this approach: a system producing speech with a high degree of fidelity to the original voice was produced. This achievement is one of the major milestones in the development of speech synthesis technology and opens a very wide range of possible applications.

This approach to the problem is based on training a GAN model using a dataset of recorded speech samples. In summary, the generator network produced raw audio waveforms as output, and the network of discriminator made it real or generated the sound. Very crucial in generating speech would be the type of training called adversarial where each network tries to outpace the other in which outputs have been realized and produced very accurately with natural sound in its generation. Optimizing good real-time performance has been great significance in this project. In fact, very careful designs of the network architecture combined with

efficient training techniques streamlined processes have lowered the computation requirements of the system as well as latency. Innovations of this type will finally make the system suitable for interactive and time-sensitive applications such as personalized voice assistants, interactive storytelling, or assistive communication tools for individuals suffering from speech impairment

7. References:

[1] Merlijn Blaauw, Jordi Bonada, Ryunosuke Daido. "Data Efficient Voice Cloning for Neural Singing Synthesis" 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp : 6840 to 6844.

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu," WaveNet: A Generative Model for Raw Audio", on arXiv on September 12, 2016.pp:1-15

[3] Giuseppe Ruggiero, Enrico Zovato, Luigi Di Caro, Vincent Pollet, "Voice Cloning: a Multi-Speaker Text-to-Speech Synthesis Approach based on Transfer Learning", arXiv on February 10, 2021,pp:1-5

[4] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp: 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[5] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, Yanqi Zhou," Deep Voice 2: Multi-Speaker Neural Text-to-Speech" , arXiv in October 2017.pp:1-16

[6] S. Shirali-Shahreza and G. Penn, "MOS Naturalness and the Quest for Human-Like Speech," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 346-352, doi: 10.1109/SLT.2018.8639599.

[7] Jixun Yao, Yi Lei, Qing Wang, Pengcheng Guo, Ziqian Ning, Lei Xie, Hai Li, Junhui Liu, Danming Xie,"Preserving background sound in noise-robust voice conversion via multi-task learning", 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023).pp:1-6

[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 4052-4056,doi: 10.1109/ICASSP.2014.6854363.

[9] Zeyu Qiu, Jun Tang, Yaxin Zhang, Jiaxin Li, Xishan Bai," A Voice Cloning Method Based on the Improved HiFi-GAN Model", Computational Intelligence and Neuroscience in 2022.pp:1-12

[10] Mingyang Zhang, Yi Zhou, Li Zhao, Haizhou Li,"Transfer Learning from Speech Synthesis to Voice Conversion with Non-Parallel Training Data," IEEE/ACM Transactions on Audio, Speech, and Language Processing in 2021 pp:1-5