International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

AI Based Voice Cloning and Generation for Vocally Challenged

Mr.Dhanush B
Department of
Artificial Intelligence and
Machine Learning
Sri Shakthi Institute of
Engineering and Technology
Coimbatore,India

Mr.Dhanush K
Department of
Artificial Intelligence and
Machine Learning
Sri Shakthi Institute of
Engineering and Technology
Coimbatore,India

Ms.Narmadha M
Department of
Artificial Intelligence and
Machine Learning
Sri Shakthi Institute of
Engineering and Technology
Coimbatore,India

Mr.Vijay S
Department of
Artificial Intelligence and
Machine Learning
Sri Shakthi Institute of
Engineering and Technology
Coimbatore,India

Mr.Varshik Daniel L
Department of
Artificial Intelligence and
Machine Learning
Sri Shakthi Institute of
Engineering and Technology
Coimbatore,India

Abstract- Voice cloning, the ability to replicate a person's voice with high fidelity, has significant applications in entertainment, accessibility, virtual assistants, and forensic sciences. Recent advancements in deep learning have demonstrated the potential of Generative Adversarial Networks (GANs) to generate realistic synthetic audio. This study explores the application of GANs for voice cloning, leveraging their adversarial training mechanism to achieve high-quality and natural-sounding voice synthesis.

The proposed framework integrates a generator network designed to produce realistic audio waveforms and a discriminator network tasked with distinguishing between real and synthetic samples. The system is trained on a dataset of diverse voice recordings, focusing on capturing both prosody and speaker-specific features. To enhance the cloning process, the model employs auxiliary losses, such as mel-spectrogram reconstruction and perceptual loss, ensuring the generated audio aligns closely with human perception.

Experimental results demonstrate that GAN-based voice cloning outperforms traditional methods in both audio quality and speaker similarity, even with limited data. The research also discusses ethical considerations, including misuse risks and countermeasures, emphasizing the importance of responsible deployment. The findings establish GANs as a promising approach for advancing voice cloning technology while highlighting avenues for future research in robustness and real-time applications.

INTRODUCTION

Voice cloning is a speech synthesis method that allows machines to synthesize the speech of a specific target speaker. It also provides a critical technical means for

generating personalized speech. In personalized human-computer interaction scenarios, voice cloning technology possesses a wide range of applications in intelligent electronic terminal equipment like autonomous robots, Internet of Vehicles, Internet of Things, etc.

This technology can be divided into two categories based on the amount of target speaker corpus used. One method is a speech synthesis method that is based on a large amount of the target speaker corpus. The fundamental principle of this method is to train a speech synthesis system with a large amount of the target speaker's speech and synthesize the voice of the target speaker. The main disadvantage is that it needs to collect a large number of speech samples of a specific person, which is a tedious job in many cases. Hence, this method is rarely used. The second approach is based on a small number of samples. There are two ways to implement this method. One method involves using the speaker adaptation method. The basic idea is to obtain a more matchable acoustic model by finetuning the parameters of the trained multispeaker generation model through an adaptive algorithm. Speaker adaptation entirely depends on adaptation parameters and leads to an increase in memory storage and serving costs. The second approach is to use the speaker encoding method. The basic idea is to select an independent speaker encoder to extract the embedding vector of the target speaker and then splice the speaker embedding vector into the multispeaker speech generation model for controlling the speech. Finally, the voice of the target speaker is synthesized by the vocoder. The advantage of this method is that the trained speaker encoding model does not require any finetuning, and the speaker embedding vector can be directly inferred from only a couple of speech samples of the target speaker, so the cloning speed is fast. The key part of this method is to design a good speaker encoder that has the capacity to extract the features that characterize the target speaker from a small number of speech fragments. The similarity of the cloned speech can be determined by the quality of extracted speaker features. Arik et al. made a detailed comparison between the speaker encoding and speaker adaption methods and both methods performed well in voice cloning tasks. The speaker adaptation method requires thousands of finetuned steps to achieve a high-quality adaptive effect, making it more difficult for deployment in mobile devices without real-time synthesis. The cloning time and necessary memory for the speaker encoding method are really less, which is crucial for practical applications.

Although previous works in voice cloning have appropriately considered the limited speech samples in personalized voice, they have not completely addressed the key issues. They finetune the whole model or the decoder part, achieving good quality but leading to too many adaptation parameters. Reducing the number of adaptation parameters is crucial for the practical application of voice cloning tasks. Also, the memory storage can explode due to the increase in the number of users. Some works only finetune the speaker embedding or train the speaker encoder part, which does not require any fine-tuning during voice cloning. Although these approaches lead to a lightweight and efficient adaptation, they provide poor cloning quality.

LITERATURE REVIEW

So far, Tacotron 2, based on sequence-to-sequence architecture, has been a very popular model in the field of speech synthesis, possessing great development prospects and significant versatility. Tacotron 2 can be divided into two submodules: the acoustic feature prediction module and the vocoder module. At the time of training, the acoustic feature prediction module usually inputs the text sequence and outputs the acoustic features, while the vocoder module restores the predicted acoustic features to speech waveforms. However, Tacotron 2 cannot precisely control and synthesize diverse speech samples. To synthesize sounds closer to human beings, Tacotron 2 is usually extended and applied in several other tasks such as voice cloning, speech style control, speech prosody control, code-switching, etc. Wang et al. achieved prosody transfer and enhanced the emotional information of synthesized speech by adding a prosody encoder in the acoustic feature prediction module to model and learn the prosody features in both supervised as well as unsupervised ways. The speaker encoder can be simply regarded as a text-independent speaker recognition model. Kinnunen et al. completed speech conversion based on *i*-vector in speaker recognition . It is difficult to retain the nonlinear features in the original data when the *i*-vector uses a linear transformation to reduce the dimensions, which have

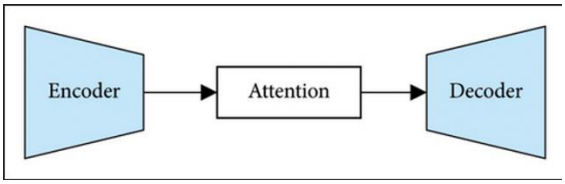
been replaced by a *d*-vector with strong antinoise ability. By adding the speaker encoder and extracting the *d*-vector with the target speaker features, the multispeaker TTS system realizes the preliminary voice cloning. However, *d*-vector does not fully consider the context information of the entire utterance, which leads to the omission of speaker information. Snyder et al. first proposed a framework for extracting speaker embedding features based on the time-delay neural network (TDNN) and successfully obtained the *x*-vector, which was applied to the speaker verification task and outperformed the traditional speaker vector. The application of the *x*-vector in the speaker encoder can effectively guide the prediction of acoustic features, which can significantly improve the similarity of the cloned speech. By combining WaveNet based on the autoregressive model as a vocoder, Tacotron 2 can generate high-quality speech, but the sequential reasoning process of the autoregressive model makes it sluggish and inefficient to generate speech, which cannot meet the requirements of real-time applications. To address the limitations of autoregressive models, more and more researchers began focusing on nonautoregressive models based on generative adversarial networks (GAN). Parallel WaveGAN and MelGAN are the early attempts of GAN on vocoder. Although the model reasoning speed can be significantly improved, the speech they generated is not satisfactory in terms of quality. The appearance of HiFi-GAN breaks the shackles not only by effectively modeling the long-term correlation of the speech waveform but, more importantly, by effectively modeling the periodic mode of the speech waveform. Besides, it achieves real-time and high-fidelity speech waveform generation. Moreover, as one of the most advanced vocoder networks, the HiFi-GAN model is used as the backend by many end-to-end speech synthesis systems to restore the predicted Mel spectrum to speech waveforms. However, there are still some shortcomings in HiFi-GAN, which fail to balance the speech quality with model parameters and inference speed. Therefore, the HiFi-GAN model must be improved for better application in the voice cloning task.

Although the multispeaker TTS model based on Tacotron 2 can expand the system architecture along with supporting the voice cloning function of multiple speakers, it is still slightly insufficient in terms of speaker feature extraction and synthesis speed. The speech information of the entire sentence is not taken into consideration by the d-vector, which affects the similarity of the cloned voice. The WaveNet vocoder can have a severe impact on the speed of speech generation. To better balance the relationship between speech quality, model parameters, and inference speed of the voice cloning system, the speaker features are extracted based on TDNN, and the output of each speech segment is aggregated after passing through the model through statistical pooling. This represents the feature vector of the target speaker and improves the quality of the generated speech. In this paper, a competitive multiscale convolution (CMSC) strategy and a depth-wise separable convolution (DSC) strategy are introduced to improve the HiFi-GAN model, which replaces the WaveNet vocoder, to significantly reduce the number of model parameters and further enhance the inference speed.

.

EXISTING SYSTEM

The feature prediction network in this paper is based on the encoder-decoder model. Its primary function is to direct the conversion of the input text into the Mel spectrum with the target speaker's characteristics after splicing with the *x*-vector vector output by the speaker encoder that describes the speaker's characteristics, so that the vocoder can restore waveforms. Figure depicts its fundamental architectural principle.



The encoder first models the contextual information of the input text sequence with a 3-layer convolutional network, and the output of the final convolutional layer is fed into a bidirectional long-short-term memory network with 512 units to convert the input text sequence into a high-level feature sequence. The attention mechanism computes the weight of each element in the high-level feature

sequence, assigns different weights to the encoder output, performs weighted summation, and then feeds it into the decoder. In this case, the attention network employs the location-sensitive attention mechanism, which extends the additional attention mechanism, alleviating potential error patterns caused by the decoder repeating or ignoring some subsequences. The decoder is a 5-layer convolution postprocessing network with a 2-layer fully connected preprocessing network, a 2-layer unidirectional long short-term memory network, two linear mapping layers, and a 2-layer fully connected preprocessing network. The posterior probability of the output sequence and the output Mel spectrum are computed.

PROPOSED SYSTEM

HiFi-GAN uses GAN as the basic generative model and includes a generator and two discriminators, which can efficiently convert the spectrum generated by the acoustic model into high-quality audio. HiFi-GAN is a vocoder commonly used in both academia and industry in recent years, but it still has some shortcomings. In order to reduce the model parameters of HiFi-GAN and improve the inference speed without sacrificing the speech quality, we use CMSC and DSC strategies to improve the HiFi-GAN model, and the details are described in the following submodules.

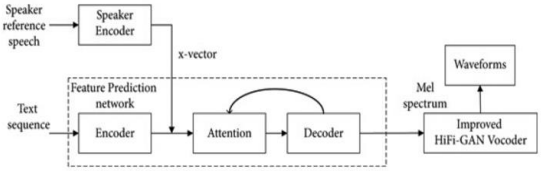
Generative Adversal Network

Synthetic voices are becoming more realistic and versatile, thanks to the advances in artificial intelligence (AI). One of the techniques that enables this is generative adversarial networks (GANs), which are composed of two competing neural networks that learn from each other. In this article, you will learn how GANs work, what are some of the applications and challenges of synthetic voice generation, and how you can experiment with GANs yourself.

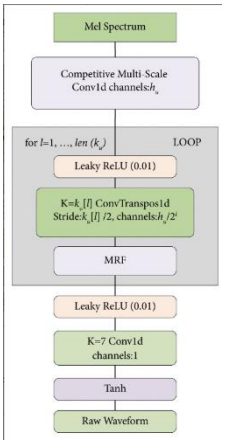
GANs are a type of AI model that can generate new data that resembles the original data. For example, a GAN can create realistic images of faces, animals, or landscapes that never existed before. A GAN consists of two neural networks: a generator and a discriminator. The generator tries to create fake data that can fool the discriminator, while the discriminator tries to distinguish between real and fake data. The two networks train each other in a game-like scenario, improving their performance over time.

To generate synthetic voices, a GAN needs to learn from a large dataset of human speech samples. The generator can then produce new speech samples that mimic the characteristics of the original voice, such as pitch, tone, accent, and emotion. The discriminator can then evaluate how realistic and natural the generated voice sounds, and provide feedback to the generator. By repeating this process, the GAN can create synthetic voices that are indistinguishable from human voices.

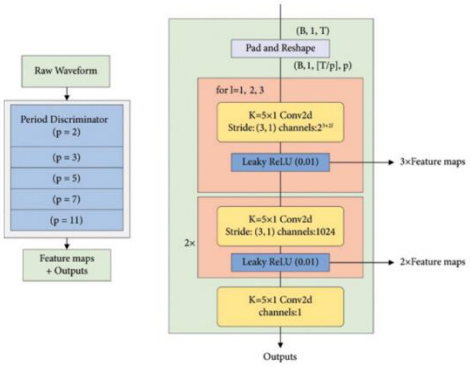
A key element responsible for creating fresh, accurate data in a Generative Adversarial Network (GAN) is the generator model. The generator takes random noise as input and converts it into complex data samples, such text or images. It is commonly depicted as a deep neural network.The training data's underlying distribution is captured by layers of learnable parameters in its design through training. The generator adjusts its output to produce samples that closely mimic real data as it is being trained by using backpropagation to fine-tune its parameters.



FLOW DIAGRAM



Generator



Discriminator

DESIGN

DATASET

With the rise of AI applications offering voiceovers and human-like conversational voices, there's a growing interest in building custom text-to-speech models. Many developers and companies seek to avoid the costs of paid voiceover services by fine tuning their own models. However, the first and most critical step in creating a high-quality text-to-speech system is acquiring a rich, well-prepared dataset. This guide walks through a comprehensive process to build such a dataset, focusing on extracting clear and accurate vocal samples essential for effective voice cloning.As an intermediate step of my project, I'm working on setting up an automated pipeline that can seamlessly perform each of these steps. This guide provides a detailed walkthrough on creating a high-quality dataset, covering everything from video downloading to audio transcription. Toward the end, I'll discuss some of the challenges encountered along the way..

GENERATOR

The sum of the outputs of multiple residual blocks (ResBlock) is accumulated by the MRF module. Each residual block is composed of a series of one-dimensional convolutions. These convolutions have different convolution kernels and dilation rates that form different sized receptive fields, effectively modeling the long-term correlations of speech waveforms.Unlike the original generator network, a CMSC strategy is used to extract the features from the input Mel spectrum. The multisized convolution kernels are used to process the input Mel spectrum, and the sum of these processed results is returned. Compared with the original convolutional layer with a fixed convolution kernel size to extract features from the Mel spectrum, CMSC can better capture the local features between different frames and interframe correlations of the Mel spectrum and express the feature information extracted from the Mel spectrum in a better way while providing sufficient information for the subsequent network learning. It thus improves the learning ability of the model. Besides, the original generators are composed of standard 1D convolutional layers except for a few transposed convolutional layers for upsampling. Inspired by the DSCs in images , in this paper, these standard 1D convolutions are replaced with 1D DSCs, which is expected to further compress the model size and speed the model inference without compromising the quality of the generated speech; making it significant for applications with limited hardware. It must be noted that DSC using weight normalization is equivalent to the depth-wise convolution and the pointwise convolution, which adopt weight normalization

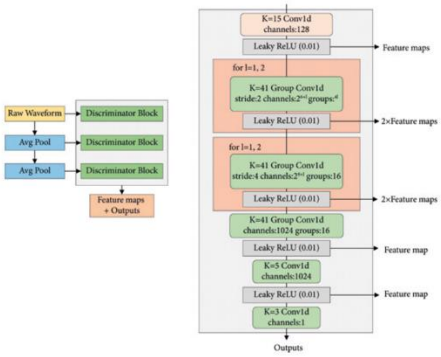
MULTIPERIOD DISCRIMINATOR

To realize that the sub discriminator captures the periodic pattern in the speech signal, the sub discriminators do not directly process the speech waveform but pad and reshape the speech waveform. Figure highlights the case when the period parameter p is 3. For ensuring that each sub discriminator only accepts equally spaced sampling points of the input speech waveform, the interval is represented by p. Thus, the original one-dimensional speech of length T is processed into two-dimensional data with height T/p and width p. Therefore, the MPD needs to use a two-dimensional convolutional neural network to process these data. Other than the last network layer, the other layers use two-dimensional stride convolutions, which only stride in height, and each convolution layer uses weight normalization. In each convolutional layer of MPD, the size of the width axis of the convolution kernel is limited to 1. This leads to an independent processing of the periodic speech samples in the width axis direction. Thus, each sub

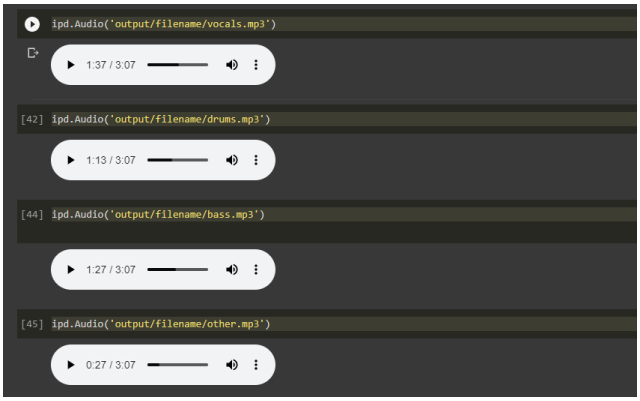
discriminator can capture the underlying periodic patterns that differ from each other in the speech by observing different parts of the speech waveforms.

MULTISCALE DESCRIMINATOR

Figure below shows the structure of the multi-scale discriminator (MSD). The left represents the overall structure, and the right represents the network structure of the subdiscriminator. The MSD is a combination of three discriminators with the same network structure but working at different scales: processing the raw speech, ×2 average-pooled audio, and ×4 average-pooled audio. To allow the use of larger-sized kernels while keeping a smaller number of parameters, the subdiscriminator employs grouped convolutions. Apart from applying spectral normalization in the first subdiscriminator for raw speech processing, which is used here to help stabilize training, the other two sub-discriminators apply weight normalization.



RESULT



CONCLUSION

In this paper, a voice cloning method with fewer parameters, faster inference speed, and higher voice quality is proposed based on the multispeaker TTS model. First, to improve the similarity of cloning speech, the x-vector feature vector that can better represent the characteristics of the target speaker is extracted based on TDNN. Then, the HiFi-GAN vocoder is improved to effectively characterize the input Mel spectrum through a competitive multiscale convolution strategy, providing sufficient feature information for the subsequent network to generate a higher-quality speech signal. Finally, the model parameters are effectively reduced by the use of depth-wise separable convolution, and the inference speed is improved without degrading the quality of the generated speech. According to the experimental results, the method in this paper effectively reduces the parameters of the HiFi-GAN model and improves the generated speech quality (MOS increased by 0.13, PESQ increased by 0.11), and the model inference speed on GPU and CPU is increased by about 11.84% and 30.99%, respectively. This proves to be very meaningful for deploying the model to application scenarios with insufficient hardware conditions and limited memory and for improving the adaptability of the model. The improved HiFi-GAN model has remarkable performance and good compatibility on the voice cloning task and achieves the highest combined score combined with x-vector embedding in all the tests.

FUTURE ENHANCEMENT

The future of voice cloning using GANs lies in improving model architectures, such as multi-scale GANs, hybrid approaches with transformers or diffusion models, and lightweight designs for real-time applications. Enhanced training techniques, like few-shot learning, self-supervised methods, and advanced speaker embeddings, will reduce data requirements while capturing nuanced voice features like emotion and accent. Ethical safeguards, including synthetic speech watermarking and detection tools, will mitigate misuse. Multimodal integration, combining voice cloning with visual inputs, will enable lifelike digital avatars and virtual assistants. Additionally, real-time cloning and cross-lingual capabilities will expand

applications in accessibility, education, entertainment, and healthcare, making GAN-based voice cloning more robust, expressive, and inclusive.

REFERENCES

[1] Ian J. Goodfellow et al, Generative Adversarial Nets, This is the foundational paper that introduced GANs. It describes the adversarial framework

[2] Alec Radford, Luke Metz, Soumith Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (DCGAN),Introduces Deep Convolutional GANs, a popular architecture that uses convolutional layers, making GANs more stable and effective for image generation.

[3] Yaniv Taigman, Lior Wolf, Adam Polyak, Eliya Nachmani,Tacotron: Towards End-to-End Speech Synthesis,Proposes an end-to-end TTS model that generates speech directly from text, producing intelligible and natural-sounding speech.

[4] Aaron van den Oord et al,WaveNet: A Generative Model for Raw Audio,introduces WaveNet, a powerful generative model for producing high-quality, human-like speech and audio from raw waveforms.

[5] Lifa Sun, Hao Wang, Shiyin Kang, Kun Li, Helen Meng,GAN-TTS: A Generative Adversarial Text-to-Speech Model,introduces GAN-TTS, which uses a GAN-based architecture for text-to-speech synthesis.