Towards High Quality multi-speaker TTS of Chinese

Yi Li[1], **Xiaoyun Zheng**[2], **Lifan Chen**[2], **Xuewen Zhong**[2]

[1]AI class
[2]Information class
31520241154485, 31520241154539, 30920241154547, 31520241154487

Abstract

Text-to-Speech (TTS) systems have seen widespread application across various fields and have attracted increasing attention. However, traditional TTS models often focus on generating neutral speech, neglecting the personalized characteristics of the speaker. On the other hand, despite the emergence of many high-performance models in recent years, Chinese speech synthesis has not received sufficient attention in this development. To address this issue, we propose a two-stage model: (i)we use x-vectors to extract speaker-specific feature vectors from reference audio and then input these feature vectors, along with the paired text, into the pre-trained SpeechT5 model to generate the Mel spectrogram of the target audio. (ii) we employ HiFi-GAN to convert the generated Mel-spectrogram into high-fidelity Chinese audio signals. Experimental results show that our model achieves a Mean Opinion Score(MOS) of 3.17, with audio quality and fidelity approaching that of the current state-of-the-art multi-speaker TTS systems.

Introduction

Multi-speaker TTS, which is sometimes referred to as speech clone in specific contexts, is designed to synthesize speech from input text while retaining the voice characteristics of the speaker given in a reference speech. TTS models have fundamentally transformed the field of Natural Language Processing (NLP), enabling the conversion of written text into synthetic speech that closely resembles human voice[15, 17]. In this process, it is essential not only to ensure the fluency of the synthesized voice but also to make the generated speech as similar as possible in quality and style to the real voice [14]. In recent years, the TTS industry has made significant strides and now plays a critical role in various applications, including voice assistants, audiobooks, and accessibility features[8]. It is worth noting that multi-speaker TTS is not just a technology but also a new cultural and artistic practice. It provides users with personalized systems [16], helps language learners mimic specific voice styles to improve pronunciation and speaking skills [18], and can also be used to clone the voices of deceased relatives or historical figures. Additionally, multi-speaker TTS provides new data and methods for related research, contributing to the further development of speech recognition and synthesis technologies.

With the strengthening of deep learning and pre-training technologies in multi-speaker TTS [5, 9], various efficient synthesis algorithms have emerged, significantly advancing the progress of multi-speaker TTS. For instance, Tacotron, as an end-to-end model for generating text-to-speech, can directly convert characters into speech [28]. FastSpeech is a new feedforward network based on Transformer that can parallelly generate mel spectrograms, significantly enhancing the generation speed of mel spectrograms [21]. Moreover, Parallel WaveGAN, using generative adversarial networks (GANs), effectively captures the time-frequency distribution of real speech waveforms by jointly optimizing multi-resolution spectrograms and adversarial loss functions, allowing for real-time synthesis of high-quality speech [29]. It is worth mentioning that SPEAR-TTS [10] only requires a minimal amount of parallel data for training, and can synthesize speech using voice samples as short as 3 seconds, maintaining the voice characteristics of speakers not seen before. Vall-E [26] using discrete codes derived from an off-the-shelf neural audio codec model, and sets a new sota in speech naturalness and speaker similarity. These technological advancements have led to significant improvements in both the speed and quality of multi-speaker TTS. Recently, GPT-soVITS[22] has emerged as a significant milestone, facilitating the generation of high-quality and high-fidelity speech synthesis, while also showcasing few-shot learning capabilities. GPT-soVITS is a decoder-only model that starts with an auto-regressive process, generating reference audio from transcription while embedding speaker features into the latent space. The model then processes the target text and produces the corresponding target audio, reflecting the desired timbre.

Although previous researchers have contributed significantly to creating multi-speaker TTS models, they still faces several challenges and issues. For example, accurately extracting voice features such as timbre, pitch, and speech rate remains a difficult task [1]. Achieving high-quality multi-speaker TTS often requires a substantial amount of labeled data, and collecting this data is both time-consuming and costly [23, 4]. Additionally, there are limitations in emotional expression and voice consistency, and challenges remain in addressing diversity and naturalness [28, 21,

16].Furthermore, computational resources present another significant challenge that cannot be overlooked.

To dive into the aforementioned issues, this paper proposes a voice cloning model consisting of three modules: (i) the x-vectors technology, used to extract speaker-specific features from reference audio; (ii) a pre-trained SpeechT5 model for generating spectrograms from text; (iii) a HiFi-GAN model for converting the generated spectrograms into natural waveforms. The x-vector method, as a feature extraction approach, can effectively extract speaker characteristics, maintaining good generalization ability even with insufficient sample sizes, and showing excellent performance in handling noise [23]. Simultaneously, HiFi-GAN offers fast synthesis speeds and allows for controlling the style and emotion of synthesized speech by adjusting input parameters, thus generating high-quality, natural-sounding speech [13]. SpeechT5 utilizes unlabeled speech and text data for pre-training, significantly enhancing the model's performance in various application scenarios, accurately generating speech content and improving the quality of speech synthesis [2].

Specifically, the proposed voice cloning model consists of three steps: First, to synthesize the voice of a specific speaker, the speaker's reference speech must be encoded separately to provide prior information. The embedding feature representation of the speaker can be obtained using the x-vectors technology, which utilizes a Time-Delay Neural Network(TDNN) to map the extracted acoustic features into a high-dimensional space, generating fixed-length feature vectors that effectively represent the speaker's voice characteristics. Second, the paired input text and the speaker feature representation obtained in the first step are jointly fed into the SpeechT5 model, which learns the complex relationships between linguistic content and acoustic features, and generates a spectrogram containing both textual information and the speaker's voice characteristics. Third, the generated spectrogram is input into the vocoder — the HiFi-GAN model, which reconstructs a high-quality speech waveform from the Mel spectrogram, while preserving the target speaker's voice features.Finally, we evaluate the performance of the proposed model using the MOS and compare our model with GPT-soVITS.

The main contributions of this paper can be summarized as follows:

- A two-stage model constructed using a cascade approach is proposed, combining multiple technologies such as x-vectors, SpeechT5, and HiFi-GAN, to enhance the quality of multi-speaker TTS and achieve high-quality, personalized, and natural speech synthesis.
- By fine-tuning the English model on a Chinese dataset, the model expands its knowledge scope and enhances its ability to accommodate speech with varying languages, pronunciations, and speaking speeds, thus improving its application in Chinese speech synthesis.
- Our model achieves a MOS of 3.17, which approaches the performance of the current state-of-the-art multi-speaker TTS models in Chinese tasks of 3.92.

Related Work

In this study, we try to improve the effectiveness of voice cloning of Chinese from multiple perspectives. In the following, we review relevant research and previous efforts in these areas.

Cross-modal Mapping Between Speech and Text

Cross-modal mapping between speech and text is a critical component in speech synthesis, particularly in the context of voice cloning. While traditional text-to-speech (TTS) systems aim to generate natural-sounding speech from textual input, voice cloning involves a more complex mapping. It not only synthesizes speech but also replicates the unique voice characteristics of the speaker.

Several studies have explored this mapping to improve the accuracy and naturalness of synthesized speech. For instance, Tacotron [28] introduced a sequence-to-sequence model that converts linguistic features, such as phonemes, into mel-spectrograms, which are subsequently used to generate speech waveforms. Subsequent research, including techniques like pretraining the speech decoder in TTS via autoregressive mel-spectrogram prediction [7] and methods for masking and reconstructing mel-spectrograms [6], has further refined these approaches to improve the alignment between text and speech.

Although the above studies have made a series of progress, the inherently complex tones and syntactic structures of Chinese make cross-modal mapping more complicated, and there is still a long way to go. In our work, we addressed this issue by leveraging the tokenizer from a large-scale pre-trained model that is natively designed to support Chinese.

Disentangling speech content and timbre

Voice cloning systems typically aim to synthesize both intelligible speech content and the unique voice characteristics (timbre) of a specific individual. In traditional speech synthesis models, these two elements are often fused, resulting in a less flexible and rigid voice output. The lack of clear separation between content and timbre can limit the expressiveness and personalization of the synthesized voice.

In this work, we explore the disentangling of speech content and timbre in the context of Chinese voice cloning. Our approach seeks to ensure that the timbre remains faithful to the original speaker, while allowing for flexible manipulation of the speech content. This separation enhances the flexibility of voice cloning systems and is particularly valuable for applications such as personalized speech assistants and dubbing, where it is crucial to preserve the speaker's identity across diverse speech contexts.

Multi-speaker Pre-trained Model

In recent years, pre-trained models have achieved significant success in various speech processing tasks, including voice cloning. These models leverage large-scale, diverse datasets to capture a wide range of speaker-specific features, enabling them to generalize effectively to new, unseen speakers with minimal fine-tuning.

Multi-speaker pre-trained models typically utilize speaker embeddings that encode the unique characteristics of individual speakers. These embeddings are learned during the pre-training phase and can be fine-tuned on smaller, domain-specific datasets to adapt the model to new speakers. This approach greatly reduces the amount of training data required for voice cloning, making it feasible to generate high-quality synthetic voices for a broad array of speakers, even when data is limited.

Our work leverages this concept to enhance multi-speaker TTS for Chinese speakers.

Multi-speaker TTS with Fine-tuning Techniques

Multi-speaker text-to-speech (TTS) systems leverage speaker embeddings to condition the speech generation process, enabling the model to learn the distinctive characteristics of each speaker's voice without the need for large amounts of speaker-specific data. This is particularly advantageous in scenarios where training data for individual speakers is limited or when the goal is to develop a system capable of synthesizing a wide variety of voices from a relatively small set of training examples.

Fine-tuning techniques, where pre-trained models are adapted to new speakers or new conditions, have proven to be highly effective in this domain. One such example is the FastSpeech 2[20] with speaker conditioning, which has demonstrated that a fine-tuning strategy can allow a model to efficiently adapt to unseen speakers while retaining high synthesis quality.

This is particularly useful for languages such as Chinese where high-quality speech data is scarce, as collecting enough speaker-specific data can be very resource-intensive. In our work, we explore the potential of fine-tuning a voice cloning model on a multi-speaker Chinese dataset, enabling it to accurately capture the nuances of Chinese pronunciation, rather than synthesizing accented Chinese.

Method

Problem Definition

The voice cloning task can be viewed as a mapping between the text and the corresponding speech signals of the target speaker. Let:

- $T = \{t_1, t_2, ..., t_n\}$ represent the sequence of text to be synthesized.
- $S = \{s_1, s_2, ..., s_m\}$ represent speech waveform in the desired timbre.
- The objective of the speech cloning model is to learn a mapping f that generates speech sequences S from the text sequence T, i.e.,

$$S = f(T; \theta)$$

where θ represents the reference audio of the target speaker.

In practice, this task involves both linguistic and speaker-specific information. While traditional TTS models focus on generating neutral speech, voice cloning models need to incorporate speaker-specific features to produce personalized speech. These challenges can be addressed through the use of advanced neural architectures and multi-modal learning.

Speaker Embedding Extraction To synthesize speech that sounds like a specific target speaker, it is essential to encode the speaker's unique vocal traits. This is typically achieved through the extraction of speaker embeddings—fixed-length vectors that capture the speaker's identity. Speaker embeddings can be obtained using pre-trained models such as Speaker Recognition Networks or autoencoders. The embedding vector e_s is derived from a small amount of reference speech from the target speaker, and it represents key features such as voice timbre, pitch, and other distinctive acoustic patterns.

Formally, given a reference audio clip from the target speaker, the speaker embedding extraction process can be defined as:

$$e_s = g(\text{audio\_clip}; \phi)$$

where g is an embedding function, ϕ represents its parameters, and e_s is the resulting speaker embedding.

Once the speaker embedding e_s is computed, it is used during the synthesis process to condition the generation of the target speaker's speech, alongside the text input T. This allows the model to adjust the prosody, pitch, and other speaker-specific features, ensuring that the output speech matches the target speaker's voice.

Mel Spectrogram Generation The next step in voice cloning involves generating a time-frequency representation of the speech, typically in the form of a Mel spectrogram. The Mel spectrogram is a widely used feature in speech synthesis tasks, as it effectively represents the speech signal while reducing dimensionality and preserving perceptually relevant information.

To generate the Mel spectrogram, the model takes the paired input T (text or phoneme sequence) and the speaker embedding e_s as inputs and processes them through a sequence of neural network layers. The model outputs a sequence of Mel spectrograms $\hat{S} = \{\hat{s}_1, \hat{s}_2, ..., \hat{s}_m\}$ that represent the target speech signal in a time-frequency domain.

This step can be modeled as:

$$\hat{S} = h(T, e_s; \theta)$$

where h represents a neural network mapping that generates the Mel spectrogram, and θ represents the model parameters.

To ensure that the generated spectrogram preserves both the linguistic content of the input text and the speaker-specific features, the network typically incorporates both a sequence-to-sequence architecture and a speaker conditioning mechanism. Sequence-to-sequence architectures, such as Tacotron or Transformer-based models, are commonly used to map the text input to the Mel spectrogram, while speaker embeddings are fed into the model through attention mechanisms or concatenation layers.

Audio Generation Once the Mel spectrogram is generated, the next step is to convert it back into a speech waveform that can be played by a speaker. This step is commonly referred to as waveform generation or vocoder. The goal is to reconstruct a high-quality speech waveform from the Mel

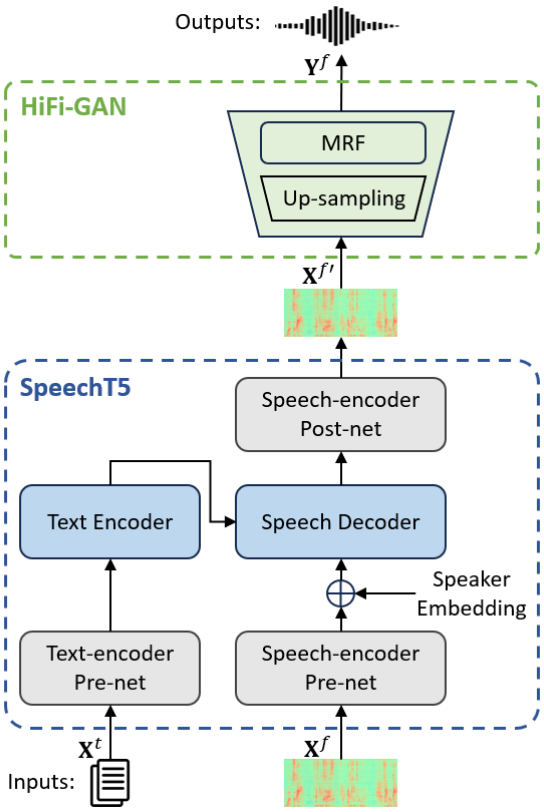

Figure 1: Architecture of Our Model

spectrogram while preserving the characteristics of the target speaker's voice.

A popular approach for this task is to use a neural vocoder, such as WaveNet, HiFi-GAN, or WaveGlow. These models are trained to convert spectrograms into raw audio by learning the complex mapping between the frequency-domain representation and the time-domain waveform. In the context of voice cloning, the vocoder takes the Mel spectrogram $\hat{S}$ and outputs the corresponding speech waveform x, i.e.,

$$x = v(\hat{S}; \phi)$$

where v is the vocoder model, and ϕ represents its parameters. The vocoder is typically trained on a large dataset of paired Mel spectrograms and waveforms.In modern systems, the vocoder may also take speaker embeddings as additional input to ensure that the output waveform retains the unique characteristics of the target speaker's voice. By combining speaker embeddings with the Mel spectrogram, the vocoder is able to generate a speech waveform that sounds not only linguistically accurate but also natural and personalized.

The overall voice cloning pipeline can therefore be summarized as a multi-stage process, where each stage focuses on one aspect of speech generation, from text processing and speaker embedding extraction to Mel spectrogram generation and waveform synthesis. By jointly optimizing these components, the system is able to produce high-quality synthetic speech that closely resembles the voice of the target speaker.

Model Architecture

Our voice cloning model is constructed using a cascade approach with two key components (Fig 1): a text-to-spectrogram model and a vocoder that converts the generated spectrogram into an audible waveform. The two components work in tandem to synthesize high-quality, speaker-specific speech from text input.

Spectrogram Model In the text-to-spectrogram module, we leverage the SpeechT5, which is different from the SOTA models [27, 11]. It is a powerful pre-trained multi-modal model designed for a range of text-to-speech and speech-related tasks. SpeechT5 is a versatile architecture that learns joint representations of text and speech, allowing it to better capture the complex relationships between linguistic content and speech signals. By pre-training the model on large-scale multi-modal datasets, SpeechT5 learns to generate high-quality spectrograms that can be used as an intermediate representation for speech synthesis.

The choice of SpeechT5 offers several advantages:

- Generalization across tasks: SpeechT5 is pre-trained on a variety of speech synthesis tasks, allowing it to generalize well to different languages, accents, and voice styles. In particular, SpeechT5 is shown to outperform other models in the field of voice conversion, which is similar to voice cloning.
- Robustness to text input variations: The model is capable of handling diverse input representations, including raw text, phonetic transcriptions, or even text mixed with prosody markers (such as pitch or duration).
- Integration of speaker-specific features: By fine-tuning SpeechT5 on target speaker data, the model can adapt to generate speech in the target voice, preserving the speaker's characteristics such as tone, pitch, and speech patterns.

The output of this model is a sequence of mel spectrogram frames, which serves as an intermediate representation of the speech signal, encoding both the content and the speech signals relevant to the target speaker.

Vocoder The vocoder is used to convert the generated spectrogram into a natural waveform. Earlier approaches, such as Griffin-Lim and WaveNet [25], faced limitations in terms of speed and quality. Compared to others recent models like WaveGlow [19] and Parallel WaveGAN [29], HiFi-GAN [13] excels at generating natural and expressive speech, particularly with low computational costs, making it ideal for real-time applications.

HiFi-GAN generates highly realistic speech waveforms that closely resemble human speech, reducing artifacts and improving perceptual quality, and for which we apply it for our task.

The vocoder takes the spectrograms generated by the Speecht5 model and produces the final audio waveform. This step ensures that the synthesized speech retains the target speaker's voice while sounding natural and fluid.

Experiment

Dataset Details

In adapting our model for Chinese voice cloning task while enhancing its performance in complex scenarios, we utilized the Chinese Internet Celebrity Speech Dataset, which provides a rich and diverse source of voice samples, and is crucial for training our model to accurately capture the nuances and characteristics of Chinese speech. By leveraging this dataset, we aim to improve the model's ability to generate natural-sounding and contextually appropriate Chinese cloned speech, thereby ensuring its effectiveness in various applications.

The dataset contains text-speech pairs of several internet celebrities, including but not limited to Ding Zhen, Sun Xiaochuan, Dian Gun, and Cai Xukun. The audio samples are sourced from social media, short video platforms, and diverse online live streaming services. To enhance the diversity of pronunciations and vocal dynamics, we have also incorporated data from Kobe into the training set.

Evaluation Metrics

We adopted the mean opinion score (MOS) to evaluate our model by native speakers on the randomly selected 20 sentences without overlapping with training data. MOS is a widely used subjective scoring method in speech synthesis tasks. In our experiments, evaluators score audio samples based on aspects such as clarity, naturalness, and speaker reducibility, with higher scores indicating better speech quality and similarity with the reference speaker. For a certain model, we average scores on all generated samples to get the MOS.

Model Implementation

We borrowed the configuration of speechT5[3]. The encoder-decoder backbone network consists of 12 Transformer encoder blocks and 6 Transformer decoder blocks, where the model dimension is 768, the dimension of FFN is 3,072, and the number of attention heads is 12. The speech-encoder pre-net includes 7 temporal convolution blocks, each of which consists of 512 channels, with strides (5, 2, 2, 2, 2, 2, 2) and kernel sizes (10, 3, 3, 3, 3, 2, 2). For text-encoder pre-net, we use a embedding layer with a dimension of 768. In addition, we optimized the model with the Adam optimizer [12] and applied a learning rate warm-up during the first 8% of updates, gradually increasing it to a maximum value of 2×10^{-4}. After reaching the peak, the learning rate was then linearly decayed for the remaining updates. On this basis, we fine-tuned the pre-trained model on the Chinese Internet Celebrity Speech Dataset with cross-entropy loss and attention loss[24]. We used HiFiGAN [13] as the vocoder to convert the Mel spectrogram into high-quality speech waveform.

Results and Analysis

We compare our model with GPT-soVITS, the decoder-only sota method. GPT-soVITS initiates an auto-regressive model by generating reference audio from transcription while incorporating speaker features into the latent space. Subsequently, the model consumes the target text and completes the following target audio that embodies the desired timbre.

Model	MOS
GPT-soVITS	3.92 ± 0.02
Ours	3.17 ± 0.03

Table 1: Results of multi-speaker TTS on the Chinese Internet Celebrity Speech Dataset.

Results shown in Table1 indicate that our model is capable of generating speech while effectively preserving the reference speaker's timbral characteristics. Our model also achieved advancements in several aspects, including delving into the potential of encoder-decoder architectures for voice cloning and demonstrating the model's capability for cross-lingual transfer, specifically from English to Chinese.

However, out model is still inferior to GPT-SoVITS, several observations and conclusions from the experiments are listed below:

- decoder-only TTS model such as GPT-soVITS tends to yield better performance over encoder-decoder models. A reasonable explanation is that in the decoder-only paradigm, the text is incorporated into computation via self-attention, preserving more information compared to the semantic embeddings generated by the encoder. This approach maintains the integrity of the original text, leading to a richer representation retaining nuanced details of the input.
- The baselines in TTS typically apply large-scale, high-quality datasets covering various domains and containing millions of training samples. However, due to limited computational resources, we failed to train our model on large-scale Chinese TTS corpus such as Common Voice. Consequently, we can only apply a relatively small dataset that primarily focuses on speech from specific internet personalities. While this dataset featuring high personalization and cultural traits, its scale and diversity still limits the model performance, and affects model convergence occasionally.

Conclusion

In this study, we introduce a two-phase cascaded framework designed to tackle the challenges associated with high-quality Chinese multi-speaker text-to-speech (TTS) synthesis. In pushing the limits of capturing the differences between various languages and cultural preferences, we applied Chinese Internet Celebrity Speech Dataset to train and evaluate our model. Experimental results show that the proposed model delivers excellent performance in audio quality and fidelity, approaching the state-of-the-art of multi-speaker TTS.

References

[1] Almutairi, Z.; and Elgibreen, H. 2022. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, 15(5): 155.

[2] Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; Wei, Z.; Qian, Y.; Li, J.; and Wei, F. 2022. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5723–5738. Dublin, Ireland: Association for Computational Linguistics.

[3] Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

[4] Arik, S.; Chen, J.; Peng, K.; Ping, W.; and Zhou, Y. 2018. Neural voice cloning with a few samples. *Advances in neural information processing systems*, 31.

[5] Arık, S. Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Ng, A.; Raiman, J.; et al. 2017. Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*, 195–204. PMLR.

[6] Bai, H.; Zheng, R.; Chen, J.; Ma, M.; Li, X.; and Huang, L. 2022. A 3 T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing. In *International Conference on Machine Learning*, 1399–1411. PMLR.

[7] Chung, Y.-A.; Wang, Y.; Hsu, W.-N.; Zhang, Y.; and Skerry-Ryan, R. 2019. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6940–6944. IEEE.

[8] Dutoit, T. 2004. High-quality text-to-speech synthesis : an overview.

[9] Huang, R.; Wang, Y.; Hu, R.; Xu, X.; Hong, Z.; Yang, D.; Cheng, X.; Wang, Z.; Jiang, Z.; Ye, Z.; et al. 2024. VoiceTuner: Self-Supervised Pre-training and Efficient Fine-tuning For Voice Generation. In *ACM Multimedia 2024*.

[10] Kharitonov, E.; Vincent, D.; Borsos, Z.; Marinier, R.; Girgin, S.; Pietquin, O.; Sharifi, M.; Tagliasacchi, M.; and Zeghidour, N. 2023. Speak, Read and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision. *Transactions of the Association for Computational Linguistics*, 11: 1703–1718.

[11] Kharitonov, E.; Vincent, D.; Borsos, Z.; Marinier, R.; Girgin, S.; Pietquin, O.; Sharifi, M.; Tagliasacchi, M.; and Zeghidour, N. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11: 1703–1718.

[12] Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[13] Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33: 17022–17033.

[14] Luong, H.-T.; and Yamagishi, J. 2020. NAUTILUS: A Versatile Voice Cloning System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2967–2981.

[15] Nair, J.; Krishnan, A.; and S, V. 2022. Indian Text to Speech Systems: A Short Survey. In *2022 International Conference on Connected Systems Intelligence (CSI)*, 1–8.

[16] Neekhara, P.; Hussain, S.; Dubnov, S.; Koushanfar, F.; and McAuley, J. 2021. Expressive neural voice cloning. In *Asian Conference on Machine Learning*, 252–267. PMLR.

[17] Padmesh, M. L.; and Kumar, P. S. 2015. Implementation of Viterbi coder for text to speech synthesis. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1–5.

[18] Pérez, A.; Diaz-Munio, G. G.; Gimenez, A.; Silvestre-Cerda, J. A.; Sanchis, A.; Civera, J.; Jiménez, M.; Turro, C.; and Juan, A. 2021. Towards cross-lingual voice cloning in higher education. *Engineering Applications of Artificial Intelligence*, 105: 104413.

[19] Prenger, R.; Valle, R.; and Catanzaro, B. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621. IEEE.

[20] Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

[21] Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

[22] RVC-BOSS. 2024. GPT-SoVITS. Accessed 21 August 2024. https://github.com/RVC-Boss/GPT-SoVITS/.

[23] Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5329–5333. IEEE.

[24] Tachibana, H.; Uenoyama, K.; and Aihara, S. 2018. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4784–4788.

[25] Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K.; et al. 2016. Wavenet:

A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

[26] Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; He, L.; Zhao, S.; and Wei, F. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv:2301.02111.

[27] Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

[28] Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

[29] Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6199–6203. IEEE.