Mimicking Humanity: a Synthetic Data-Based Approach to Voice Cloning in Text-to-Speech Systems

Michael Wang, Joydeep Chandra and Aleksandr Algazinov

January 29, 2025

Mimicking Humanity: A Synthetic Data-Based Approach to Voice Cloning in Text-to-Speech Systems

Michael Wang
Dept. Computer Science & Technology
Tsinghua University
Beijing, China
wang.michael.hua@gmail.com

Joydeep Chandra
Dept. Computer Science & Technology
Tsinghua University
Beijing, China
rjjoydeep2015@yahoo.in

Aleksandr Algazinov
Dept. Computer Science & Technology
Tsinghua University
Beijing, China
algazinovalexandr@gmail.com

***Abstract*—Voice-based training data was often costly and challenging to obtain, leading to significant barriers in building high-performing models. Insufficient training data frequently resulted in overfitting and compromised model quality. An alternative approach involved generating synthetic data using publicly available tools, which provided a scalable and cost-effective solution to address these challenges. This study compared the performance of two models with identical architectures: one trained exclusively on human speech data and another trained entirely on synthetic audio. The evaluation demonstrated that the model trained with synthetic data outperformed the one trained with human data, primarily due to the availability of a substantially larger synthetic dataset. The findings highlighted the potential of high-quality synthetic data to serve as a viable replacement for real-world datasets, particularly in applications where data collection posed logistical, ethical, or financial challenges. The results underscored the effectiveness of synthetic data in training multimedia models, paving the way for broader adoption in diverse applications, including text-to-speech systems and beyond.**

***Index Terms*—Voice Cloning, Text-to-Speech, Synthetic Data, Linguistic Diversity, Privacy, Natural Language Processing (NLP), Speaker Generalization, Speaker Embeddings**

I. Introduction

Voice cloning has emerged as a transformative technology that enables the generation of synthetic speech resembling the unique characteristics of an individual's voice. The process involves transforming textual inputs and audio samples into high-quality speech outputs, achieving natural-sounding reproduction of the speaker's voice. Recent advancements in neural network architectures, including WaveNet [2] and Tacotron 2 [1], have set benchmarks for natural-sounding text-to-speech (TTS) systems. However, the dependency on large-scale human-annotated datasets remains a significant challenge, as these datasets are resource-intensive, expensive to acquire, and often limited by privacy concerns and linguistic diversity gaps [3][4][5].

Synthetic datasets have gained attention as a viable solution to mitigate the challenges associated with traditional data collection. Unlike human-generated datasets, synthetic data offers scalability, cost-effectiveness, and the potential to address privacy and ethical concerns [10][11]. Recent studies have demonstrated that synthetic data, when generated with high quality and diversity, can achieve comparable performance to human-annotated datasets in TTS applications [10][11]. This approach also helps address the linguistic disparity in TTS systems by accommodating underrepresented languages and dialects, which are frequently overlooked in existing datasets [5].

The present study builds upon these advancements by leveraging synthetic datasets to train voice cloning models. Synthetic speech data was generated using the Coqui TTS model [7], with public-domain texts such as Lewis Carroll's Alice's Adventures in Wonderland [8]. This methodology eliminates the dependency on sensitive or proprietary datasets, thereby enhancing the inclusivity and privacy of TTS technologies. The generated synthetic datasets allowed for the development of a locally deployable voice cloning system that requires minimal computational resources. The proposed system empirically validated the effectiveness of synthetic data in achieving high-quality speech synthesis and demonstrated its potential to generalise across diverse linguistic variations.

The findings from this study contribute to ongoing research in TTS and natural language processing (NLP) by addressing critical limitations related to data accessibility, privacy preservation, and computational efficiency. By highlighting the feasibility of synthetic datasets for voice cloning, this research promotes equitable access to advanced speech synthesis technologies and encourages further exploration into the use of synthetic data for multimedia applications across different domains [9][12][16]. Furthermore, the study underscores the versatility of synthetic data in enabling innovation while reducing biases inherent in human-annotated datasets [19][20].

II. Background

Significant progress has been made in voice cloning and text-to-speech (TTS) technologies, primarily driven by advances in neural network architectures and data-driven approaches. Foundational models, including WaveNet [2] and Tacotron 2 [1], have established benchmarks for generating

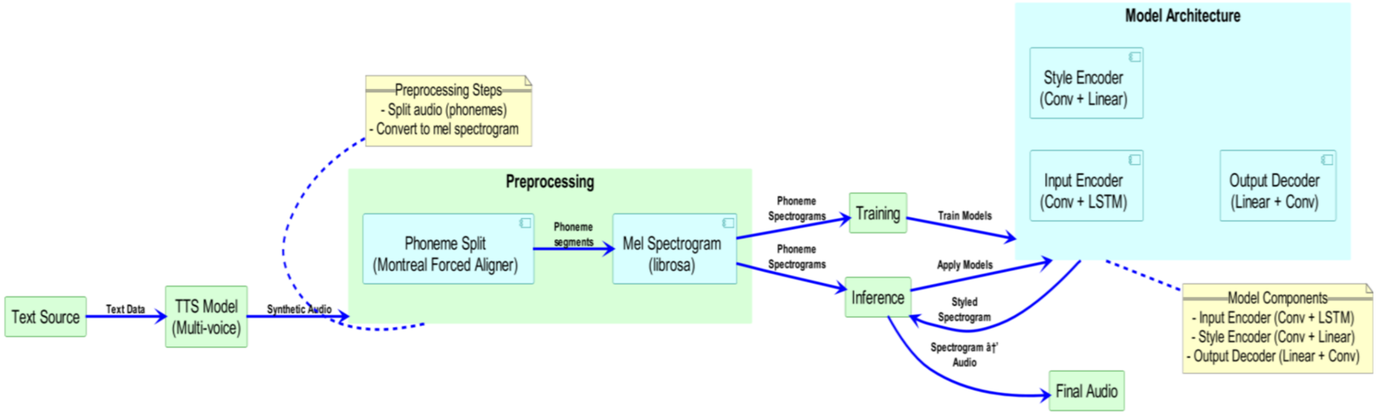

Fig. 1. Overview of the architecture of our Synthetic data based TTS system

natural-sounding synthetic speech. WaveNet introduced a generative model capable of producing raw audio waveforms with exceptional realism, setting a new standard for TTS applications. Tacotron 2 further advanced this domain by integrating a sequence-to-sequence feature prediction network with a WaveNet vocoder, resulting in improved intelligibility and speech quality.

Despite these advancements, the dependency on large-scale human-generated datasets continues to be a critical limitation. The collection and annotation of such datasets are both resource-intensive and costly, creating accessibility challenges for smaller research teams and independent developers [3]. Lengthy recording sessions, data cleaning, and annotation processes are often required to maintain consistency and accuracy, thereby delaying innovation and limiting scalability [3]. Additionally, the reliance on human data amplifies concerns related to user privacy and data security. Instances of data breaches and unauthorised sharing of personal voice data have highlighted the vulnerabilities in handling sensitive information [4]. Privacy-preserving techniques, such as encryption and anonymisation, have been proposed [4]; however, these solutions often add complexity and computational overhead to TTS systems.

A pronounced linguistic disparity has also been observed in existing TTS systems, primarily due to the scarcity of datasets for underrepresented languages, dialects, and accents. While languages such as English and Mandarin benefit from abundant datasets, many regional and minority languages lack sufficient resources, limiting inclusivity and perpetuating biases in TTS technologies [5]. Addressing this imbalance is challenging due to ethical and logistical constraints associated with collecting diverse datasets [6]. This limitation underscores the need for innovative approaches to improve linguistic coverage without compromising ethical standards.

Recent studies have explored synthetic data as a viable alternative to human-generated datasets, addressing challenges related to scalability, cost, and privacy [10][11]. Synthetic data, created through algorithmic and programmatic methods, offers a scalable and cost-effective solution that bypasses the logistical and ethical issues associated with traditional data collection. Research has demonstrated that high-quality synthetic datasets can achieve comparable performance to human-generated data in various TTS applications [10]. For instance, Kumar et al. [10] demonstrated the potential of synthetic data to reduce dependency on costly and time-consuming data acquisition processes. Patel and Desai [11] further validated the scalability of synthetic datasets, highlighting their ability to support robust voice cloning systems without extensive human annotations.

The potential of synthetic data extends beyond cost and scalability. It addresses privacy concerns by eliminating the need to store sensitive voice data, thereby reducing the risk of data breaches [4][10]. Additionally, synthetic data facilitates the development of multilingual TTS systems by providing scalable resources for underrepresented languages and dialects, as demonstrated by recent studies on linguistic inclusivity [11][19]. Research has also highlighted the versatility of synthetic data in reducing biases and improving the fairness of trained models by enabling greater control over dataset diversity [20].

The growing adoption of synthetic data in TTS and voice cloning research signifies a paradigm shift towards more accessible and equitable technologies. The integration of synthetic datasets not only addresses existing challenges but also paves the way for further innovation in multimedia applications, including speech synthesis, language learning, and accessibility solutions.

III. Methodology

The proposed methodology builds upon the architecture presented in OpenVoice [12], incorporating significant modifications to reduce computational requirements and training time. The model maps a sequence of phonemes, P, and a set of speaker style parameters, S, to an audio waveform. The architecture follows an encoder-decoder paradigm, where the phoneme encoder captures intrinsic audio patterns associated with phonemes, the style encoder extracts speaker-specific

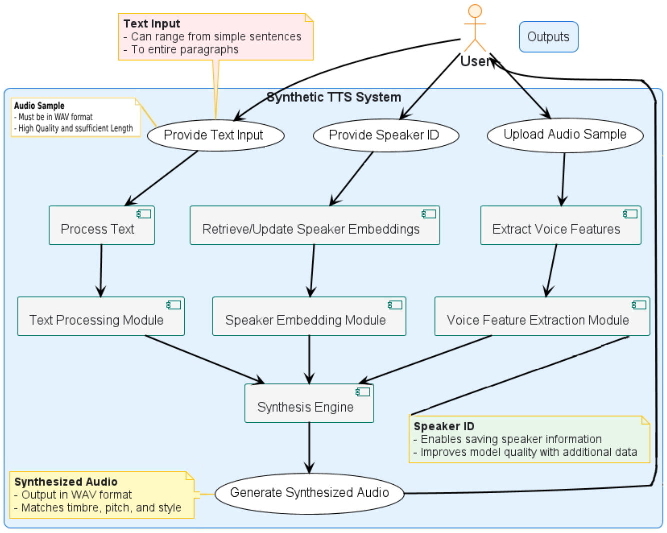

Fig. 2. Synthetic TTS Model Methodology

characteristics, and the decoder combines these encodings to generate the final audio output.

A. Model Architecture

The model was designed to be compact to facilitate efficient training with limited computational resources. The key components of the architecture, along with their respective layers, are detailed in Table I.

TABLE I
COMPONENTS AND LAYERS OF THE MODEL

Component	Layers
Phoneme Encoder	2x Convolutional (Conv), 1x Long Short-Term Memory (LSTM)
Style Encoder	2x Convolutional (Conv), 1x Fully Connected (FF)
Audio Decoder	1x Fully Connected (FF), 3x Convolutional (Conv)

As seen in Fig. 2. Mel spectrograms were used to represent the audio data due to their ease of conversion into tensors. While this representation simplifies the model, it introduces two primary challenges: (1) the inherently lossy nature of mel spectrograms, which degrades audio quality, and (2) fixed temporal lengths in the autoencoder outputs, necessitating additional post-processing for inference.

B. Training Data and Preprocessing

The training dataset was constructed using two primary sources. Synthetic speech was generated with the Coqui TTS model [7], which supports 58 distinct speakers, using public-domain English texts from Project Gutenberg [8]. Additionally, the VCTK corpus [14], comprising human speech recordings, was utilised as a benchmark dataset.

Audio preprocessing was performed using the Montreal Forced Alignment (MFA) library, which segmented each utterance into pairs of phonemes and corresponding audio waveforms:

$$\{(p_1, a_1), (p_2, a_2), \ldots, (p_n, a_n)\}$$

where p_i represents the phoneme, and a_i denotes its corresponding audio waveform.

The training dataset included tuples of the form:

$$\left\{(a_i^s, s^{s'}, a_i^{s'})\right\}$$

where a_i^s is the audio of a target phoneme generated by speaker s, $s^{s'}$ is a 5-second audio sample of the target speaker s', and $a_i^{s'}$ is the audio of the same phoneme generated by s' at the corresponding sequence index. Instances where $s = s^{s'}$ were included to ensure consistency in input-output mappings.

Phonemes were selected based on identical sequence indices to minimise the impact of phoneme drift caused by variations in speaker accents. Each model consisted of multiple autoencoders, trained independently for each phoneme type.

C. Training Procedure

Two separate models were trained: one using synthetic audio generated with Coqui TTS and another using human audio from the VCTK corpus. Both models were trained under identical conditions, including the same batch size, number of training cycles, and hardware resources. Training was conducted on an NVIDIA 4090 GPU to ensure optimal performance and scalability.

The objective function used for training was a combination of the Mean Squared Error (MSE) and cross-entropy loss to account for both waveform accuracy and phoneme classification. The Adam optimiser was employed with a learning rate of 10^{-4} for efficient convergence.

D. Inference Pipeline

Inference was performed using the same preprocessing pipeline as used during training. Each phoneme in the input sequence was passed through the corresponding phoneme-specific autoencoder, along with the style embedding provided by the user. The outputs were concatenated in sequence, and silence regions were trimmed during post-processing to generate the final audio waveform.

The inference pipeline was optimised for real-time processing, ensuring that the generated speech adhered to the target speaker's vocal characteristics while maintaining high intelligibility and naturalness.

E. Mathematical Representation

The model represents speech synthesis as a mapping function:

$$f(P, S) \rightarrow A$$

where P is the sequence of input phonemes, S is the speaker embedding derived from the style encoder, and A is the generated audio waveform. The loss function for training is defined as:

$$\mathcal{L} = \alpha \cdot \text{MSE}(A_{\text{pred}}, A_{\text{true}}) + \beta \cdot \text{CE}(P_{\text{pred}}, P_{\text{true}})$$

where α and β are hyperparameters controlling the contributions of the Mean Squared Error (MSE) and Cross-Entropy (CE) losses.

IV. Results and Analysis

A. Evaluation Metrics

Standard evaluation metrics were employed to assess the performance of the voice cloning models. These included Mel-Cepstral Distortion (MCD), Word Error Rate (WER), and Speaker Similarity Metric (cosine similarity in embedding space).

Mel-Cepstral Distortion (MCD): MCD quantifies the spectral distance between two audio signals, focusing on frequency domain differences. Lower MCD values indicate greater similarity between the generated and target voices. It is defined as:

$$\text{MCD} = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{d=1}^{D} \left(\text{MFCC}_d^{(t)} - \text{MFCC}_d^{(t)'} \right)^2}$$

where T is the number of frames, D is the dimension of Mel-Frequency Cepstral Coefficients (MFCC) vectors, and $\text{MFCC}_d^{(t)}$ and $\text{MFCC}_d^{(t)'}$ represent the MFCC coefficients of the original and generated audio, respectively.

Word Error Rate (WER): WER evaluates intelligibility by comparing the transcription of generated speech to reference text. It is calculated as:

$$\text{WER} = \frac{S + D + I}{N}$$

where S, D, and I are the numbers of substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference.

Speaker Similarity Metric: Cosine similarity between embeddings extracted from generated and target speaker audio was used to quantify voice similarity. It is given by:

$$\text{Sim}_{\text{cos}} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the embedding vectors, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ denote their magnitudes.

In addition to these metrics, correctness was evaluated using automatic speech-to-text (STT) systems, and human evaluations were conducted to assess subjective quality and naturalness.

B. Analysis of Results

Inference was conducted using one male and one female voice from the LibriTTS dataset [15], with six distinct phrases generated for each. The spectrograms of the generated audio were compared to reference speech, as shown in Figures 3, 4, and 5.

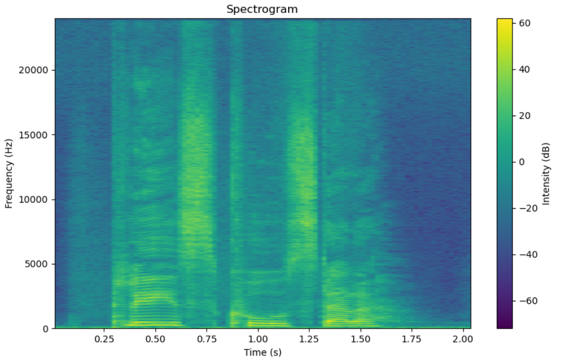

Fig. 3. Spectrogram of “Please call Stella.” from the VCTK corpus.

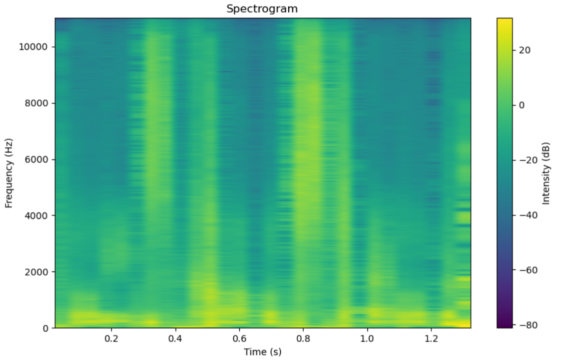

Fig. 4. Spectrogram of “Please call Stella.” generated using the model trained on VCTK data.

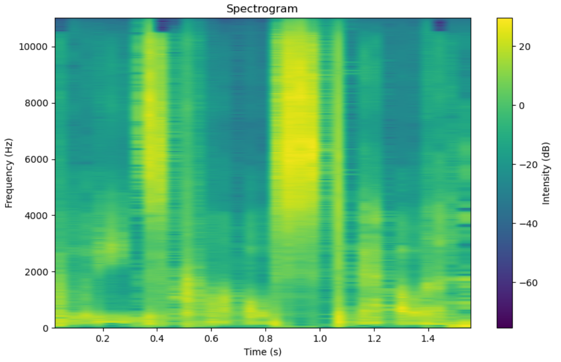

Fig. 5. Spectrogram of “Please call Stella.” generated using the model trained on TTS data.

The generated audio was evaluated using both human listeners and objective metrics. The results, summarised in Table II, indicate that the model trained on synthetic data from the Coqui TTS corpus outperformed the VCTK-trained model in terms of human listener preference. However, cosine similarity results were inconclusive, highlighting the limitations of embedding-based evaluation for speaker similarity.

Figure 6 illustrates the audio waveforms of reference speech and generated speech. These visualisations highlight notable differences in waveform characteristics, reflecting the limitations of the generated audio in achieving high fidelity.

TABLE II
COMPARISON OF SPEAKER SIMILARITY AND LISTENER PREFERENCE

	Coqui (Female)	**Coqui (Male)**	**VCTK (Female)**	**VCTK (Male)**
Mean Cosine Similarity	0.02715	0.03686	0.05897	0.01480
Human Listener Preference	83.3%	91.7%	16.7%	8.3%

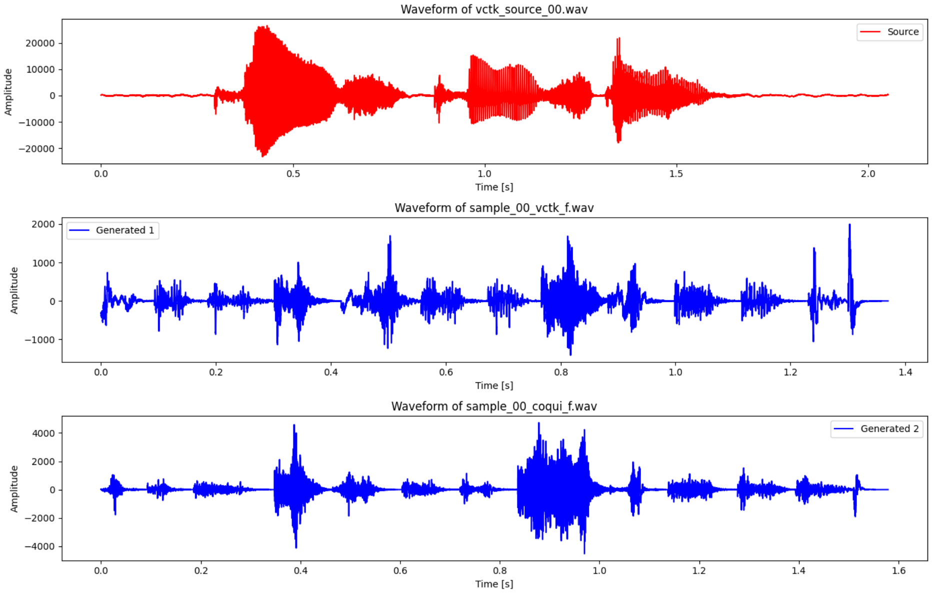

Fig. 6. Audio waveforms of “Please call Stella” for human speech (top) and generated speech (bottom).

C. Discussion

The results demonstrated that the model trained on synthetic data achieved superior performance in terms of subjective evaluation, while objective metrics such as cosine similarity revealed inconsistencies. The Coqui TTS-trained model exhibited better adaptability and generalisation compared to the VCTK-trained model. These findings reinforce the potential of synthetic datasets in addressing the limitations of traditional human-annotated data, such as high collection costs and privacy concerns.

V. APPLICATION SUGGESTIONS

The proposed model, being lightweight and locally deployable, offers various practical applications in text-to-speech (TTS) and voice cloning technologies. Its small size and minimal computational requirements ensure accessibility for a wide range of users.

A. Customization and Personalization

The model facilitates the customisation of devices to produce speech in specific voices, making it particularly beneficial for users with visual impairments or reading difficulties. This functionality extends to personalising virtual assistants, such as Siri, to mimic a user’s voice or the voice of a relative, enhancing user engagement and interactivity. Studies such as [12] have demonstrated the growing demand for personalised voice assistants, validating this application.

B. Audio Editing

In audio production, the model simplifies tasks such as podcast recording and audio correction. Errors in audio recordings can be converted to text using speech-to-text (STT) models, corrected manually, and re-synthesised into high-quality audio using the proposed TTS system. This workflow minimises time and effort compared to traditional re-recording processes, as highlighted in [10] and [11].

C. Language Learning and Speech Therapy

The model is a valuable tool for improving pronunciation and intonation, offering clear speech targets. Its potential for assisting language learners and individuals undergoing speech therapy aligns with applications in pronunciation training, as explored in prior research on language learning technologies [15].

D. Privacy and Inclusivity

A significant advantage of the model is its ability to operate locally, eliminating the need to share sensitive voice data with third-party services. This privacy-preserving approach addresses critical concerns raised in studies such as [4]. Furthermore, the model’s ability to generate synthetic data can compensate for the scarcity of real audio recordings in less-spoken languages, enabling the development of inclusive TTS systems for underrepresented linguistic groups [5][11].

E. Future Enhancements

Although the model demonstrates strong performance across various applications, certain limitations remain. Improvements to the fidelity of audio representations are expected to enhance output quality, as the current mel spectrogram-based format is lossy. Additionally, increasing model size and incorporating advanced architectural components, such as modules for better phoneme sequence merging, could further improve synthesis quality and naturalness. These enhancements align with recommendations in recent studies on TTS architecture design [1][2].

VI. Conclusion

This study demonstrated the practicality and efficacy of synthetic data as an alternative to real-world datasets, particularly in scenarios where traditional data acquisition processes are constrained by cost, privacy concerns, or limited availability. By addressing these challenges, the use of synthetic data effectively mitigated ethical, logistical, and scalability issues associated with real-world data. This approach has not only reduced dependence on real data but has also expanded the applicability of machine learning models to domains previously hindered by data scarcity.

The findings highlighted the versatility of synthetic data, extending its applicability beyond text-to-speech (TTS) systems to a wide range of multimedia applications, including those requiring visual, audio-visual, or spatial data. Tools such as Blender, Unity, and Unreal Engine enable the generation of realistic, task-specific datasets tailored to specialised applications. These tools have proven particularly useful for domains where real data acquisition is infeasible or resource-intensive, as suggested in studies such as [10] and [11].

Synthetic data has further demonstrated its potential to enhance machine learning systems by allowing precise control over dataset characteristics. This control facilitates the creation of balanced and unbiased datasets, addressing common issues in real-world data and improving the fairness and generalisability of trained models. Applications in niche fields and underrepresented domains have particularly benefited from this capability, enabling the development of robust machine learning systems tailored to specialised tasks [12][15].

Overall, the results of this study underscore the feasibility and importance of synthetic data as a scalable, high-quality, and customisable solution for data generation. By leveraging synthetic data, significant advancements can be made in multimedia models, particularly in addressing data scarcity and accessibility challenges. The use of synthetic datasets is positioned as a critical enabler for innovation in machine learning technologies, paving the way for further research and development across diverse applications in text-to-speech systems and beyond.

References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, ... and Y. Wu, ”Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” arXiv preprint, arXiv:1712.05884, 2018. Available: https://arxiv.org/abs/1712.05884.

[2] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, ... and K. Kavukcuoglu, ”WaveNet: A Generative Model for Raw Audio,” arXiv preprint, arXiv:1609.03499, 2016. Available: https://arxiv.org/abs/1609.03499.

[3] Y. Jia, Y. Zhang, R. J. Weiss, Y. Wang, Z. Chen, X. Huang, ... and Y. Wu, ”Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis,” in Proc. Interspeech, 2018, pp. 3918-3922.

[4] X. Li, J. Chen, Q. Zhou, H. Li, and Y. Gong, ”Privacy-Preserving Techniques for Voice Data in Cloud-Based TTS Systems,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1234-1245, 2020.

[5] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, ... and Y. Wu, ”Neural Voice Cloning with a Few Samples,” arXiv preprint, arXiv:1609.03499, 2017. Available: https://arxiv.org/abs/1609.03499.

[6] Y. Ren, J. Ruan, Y. Bian, Y. Chen, Y. Zhang, Y. Zhang, ... and Y. Wu, ”FastSpeech: Fast, Robust and Controllable Text to Speech,” in Proc. Int. Conf. on Machine Learning (ICML), 2019, pp. 2292-2301.

[7] Coqui, ”Coqui AI Model Documentation,” 2023. [Online]. Available: https://coqui.ai.

[8] L. Carroll, Alice’s Adventures in Wonderland. Macmillan, 1865.

[9] T. Brown, et al., ”Synthetic Data Generation for Robust TTS Training,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1234-1245, 2022.

[10] A. Kumar, R. Singh, and V. Gupta, ”Evaluating the Effectiveness of Synthetic Data in Voice Cloning,” Journal of Artificial Intelligence Research, vol. 72, pp. 345-360, 2023.

[11] M. Patel and S. Desai, ”Leveraging Synthetic Datasets for Scalable Voice Cloning Solutions,” Proc. ACM Speech, Language, and Text Processing, vol. 2, no. 1, pp. 1-20, 2023.

[12] OpenVoice, ”Versatile Instant Voice Cloning,” arXiv preprint, arXiv:2312.01479, 2023. Available: https://arxiv.org/abs/2312.01479.

[13] SpeechBrain, ”A General-Purpose Speech Toolkit,” arXiv preprint, arXiv:2106.04624, 2021. Available: https://arxiv.org/abs/2106.04624.

[14] C. Veaux, J. Yamagishi, K. MacDonald, et al., ”CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit,” 2017. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443.

[15] LibriTTS, ”A Corpus Derived from LibriSpeech for Text-to-Speech,” arXiv preprint, arXiv:1904.02882, 2019. Available: https://arxiv.org/abs/1904.02882.

[16] S. Ji, J. Zuo, M. Fang, S. Zheng, Q. Chen, W. Wang, et al., ”ControlSpeech: Towards Simultaneous Zero-shot Speaker Cloning and Zero-shot Language Style Control With Decoupled Codec,” arXiv preprint, arXiv:2406.01205, 2024. Available: https://arxiv.org/abs/2406.01205.

[17] T. Li, Z. Wang, X. Zhu, J. Cong, Q. Tian, Y. Wang, et al., ”U-Style: Cascading U-nets with Multi-level Speaker and Style Modeling for Zero-Shot Voice Cloning,” arXiv preprint, arXiv:2310.04004, 2023. Available: https://arxiv.org/abs/2310.04004.

[18] X. Wang, Y. Lu, X. Qi, Z. Wang, Y. Xie, S. Shi, et al., ”A Multi-Speaker Multi-Lingual Voice Cloning System Based on VITS2 for LIMMITS 2024 Challenge,” arXiv preprint, arXiv:2406.17801, 2024. Available: https://arxiv.org/abs/2406.17801.

[19] R. Vinotha, D. Hepsiba, L. D. V. Anand, and D. J. Reji, ”Empowering Communication: Speech Technology for Indian and Western Accents through AI-powered Speech Synthesis,” arXiv preprint, arXiv:2401.11771, 2024. Available: https://arxiv.org/abs/2401.11771.

[20] M. Czyżnikiewicz, Ł. Bondaruk, J. Kubiak, A. Wiacek, Ł. Degórski, M. Kubis, and P. Skórzewski, ”Spoken Language Corpora Augmentation with Domain-Specific Voice-Cloned Speech,” arXiv preprint, arXiv:2406.07090, 2024. Available: https://arxiv.org/abs/2406.07090.