Audio Cloning on Historical Figures

Wenhao He
School of Engineering and Applied Science
State University of New York at Buffalo
Buffalo, New York 14261
Email: wenhaohe@buffalo.edu

Dr. David Doermann
Empire Innovation, Interim Chair
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, New York 14261
Email: doermann@buffalo.edu

***Abstract*—In the upcoming year, 2024, we will commemorate the 70th anniversary of the landmark civil rights case, Brown v. Board of Education. This civil rights case decided that separating students by race in public schools violated the Fourteenth Amendment to the Constitution. The decision in the court case led to significant changes in the laws, not only fighting against the harmful separation of races in schools but also creating an example that would help guide future efforts and legal actions for civil rights. During that time, various key decisions were made, but many from this era were never publicly read by the court. Acknowledging this gap and under the supervision of Dr. David Doermann, this project aims to develop an audio cloning application on historical figures using Deep Learning techniques. By collecting the recordings of the voices of the judges, the project intends to revisit and reenact scenarios from decades ago, bringing Justices' Voices to life and recreating the genuine sound of history. Looking forward, the project envisions extending its scope to include ambient court sounds and the unique acoustics of historical courtrooms for cases in history. This pursuit not only commemorates the past but also opens new avenues for educational purposes for students and teachers when they are taught about the history of the United States at school. This pays homage to a significant milestone in civil rights cases and demonstrates the potential of advanced audio technology in historical preservation and education purpose. These forthcoming improvements will help merge technological innovation with the rich legacies of the past of the United States.**

1. Introduction

The objective of this project is to explore the use of Audio Cloning [1], also referred to as digital cloning, in the reproduction of voices of famous individuals and to utilize these voices in the recreation of history. The project aims to use AI (Artificial Intelligence) to closely mimic human voice. This audio cloning project has interest in the AI community, particularly for its application in recreating the voices of historical figures. Utilizing these voices in educational and documentary contexts to bring history to life, play a transformative role, and provide a more immersive learning experience for future generations. One of the ethical challenges we face involves the consideration and the need for responsible use guidelines, which are crucial to prevent misuse, such as creating Deepfake voices to any famous individuals.

In addressing the state of the art in this audio cloning project, referenced the work on the Google paper titled 'Transfer Learning from Speaker Verification to Multi-speaker Text-To-Speech Synthesis,' published by Google researchers [2]. The paper introduces a new method, a three-stage pipeline deep learning framework called SV2TTS. This model, equipped with a voice encoder, can generate the voice of anyone in real-time using only 5 seconds of tuning data. Our approach to achieving the goal of this audio cloning project includes implementing the model introduced in the paper, which creates a numerical representation of a voice. This representation is then used to condition a text-to-speech model to generate new audio.

Although the paper was previously introduced, it lacked open-source accessibility for public viewing. Consequently, the implementation of the model was facilitated through a GitHub repository [3] implemented the model from the paper. The work on the model was significantly important, and it performed well in audio cloning with training. The amount of work was manageable, the toolbox design from the original work was easy to understand, and there was extensive training on genders, male and female. However, due to the rapidly fast updates in the field of AI, this implemented work is quickly becoming outdated. It also lacked support for punctuation in the implementation of the model and could not support long scripts to output at once, such as the script [8] we used in this civil rights case, Brown v. Board of Education. There was also no support for the most commonly used audio file type, like MP3, from the original work to convert. Based on the experiment and the use of data relating to historical figures for audio cloning, make changes to the original work, adapting it to use historical figures and thereby generating better quality voice outputs of them. Implement a system for accurately simulating the Supreme Court environment for future work on this audio cloning project.

2. Experiment Context

The SV2TTS model introduces a three-stage pipeline that enables the cloning of a voice unseen during the training process with just a few seconds of a speech audio clip, eliminating the need for retraining the model as shown in Figure 1, based on the Multispeaker speech synthesis model from the paper [2]. The process is separated into three distinct stages. The first stage, the speaker encoder, derives an embedding from a short utterance of a speaker. This embedding represents the voice of the speaker meaningfully, capturing the characteristic of speaker and identifying the characteristic. The second stage, the synthesizer, uses a sequence of graphemes or phonemes as input, conditioned on the embedding of the speaker, and generates a Mel spectrogram from the provided text. The third stage, the voice encoder, infers an audio waveform from the spectrograms generated by the synthesizer, resulting in the output voice.

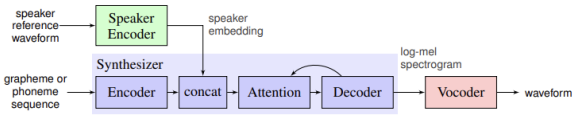

Figure 1. The general SV2TTS architecture, a three-stage pipeline, allows you to clone a voice unseen during the training process. Figure extracted from (Jia et al., 2018).

A spectrogram visually shows the different frequencies of sound signals as they vary with time. The spectrograms shown in Figure 2 serve as references to generate speaker embeddings and their corresponding outputs. The alignment between the text and the spectrogram, indicated in red, guides the synthesis process for easier understanding.

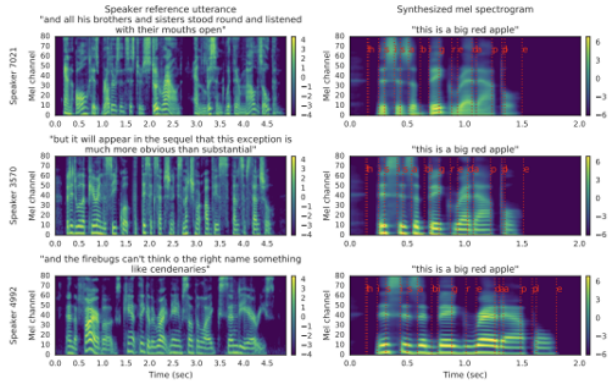

Figure 2. Comparative spectrograms visualized of Reference Speech and Synthesized Outputs respectively. The alignment between the text and the spectrogram is indicated in red for differences in spectrograms. Figure extracted from (Jia et al., 2018).

The interface of the model, as illustrated in Figure 3, was developed by the project [3]; it utilizes Python with the Qt5 graphical interface as a Python module. The embedding vector with a heatmap plot of the synthesized utterance is drawing on the left side of the figure upon completion of the generation process. After the embedding is computed, it can generate a spectrogram. This is subsequently projected using UMAP [4](Uniform Manifold Approximation and Projection), a technique focused on dimensionality reduction, similar to PCA [5](Principal Component Analysis) and other methods for reduction.

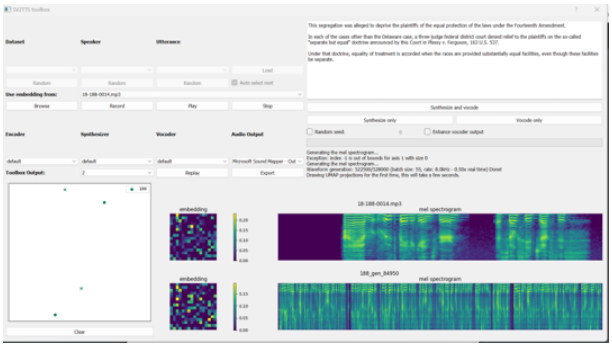

Figure 3. The interface of the model.

At the beginning stage of the experiment, various dimensionality reduction methods for embeddings were examined, including Principal Component Analysis (PCA) and others, notably contrasting PCA with Uniform Manifold Approximation and Projection (UMAP). After thorough evaluation and comparison, UMAP was found to be the more helpful technique in dimensionality reduction compared to other techniques for projections, effectively considering the overall aspects.

UMAP stands for Uniform Manifold Approximation and Projection and is a technique utilized in machine learning for dimensionality reduction. It is adept at preserving the essential characteristics of voice data while reducing its dimensionality. As a non-linear algorithm, UMAP can capture complex, nonlinear relationships within audio data. Compared to other techniques, UMAP often yields more interpretable and insightful visualizations of high-dimensional audio data, especially in contrast to linear techniques such as PCA.

3. Experiment Overview

Judge Earl Warren [6], who was Supreme Court Chief Justice during the civil rights case Brown v. Board of Education, along with Judge Ruth Bader Ginsburg [7], an Associate Justice of the Supreme Court of the United States from 1993 until her passing in 2020, were the focus of our audio cloning project. The project aimed to preserve the “noise” and mimic human voice to avoid a machine-like output despite the limited data from an era before widespread audio recordings. Our approach involved obtaining audio data for these target individuals, exploring the impact of the original voice of a speaker quality on the model, experimenting with the effects of voice clip length on the outputs, and enhancing audio fidelity by tuning the embedding for a better Mel spectrogram.

Upon examination of the original work, the initially generated voice outputs were unsatisfied, primarily due to the input voice as source recordings from decades ago; the source recordings were not captured in controlled, disturbance-free environments as heard; it was not made at a quiet place with none disturbed. This made improvements and modifications to the original work a must for the voice generation process.

This semester, four members of our research group conducted a series of experiments with original audio lengths, various text lengths for audio generation, and both seen and unseen text, evaluating the quality across several parameters. These revealed that slurred speech occurred with numerical numbers, and varying the length of input voice and text resulted in different quality outputs. Notably, trimming the same voice input at different lengths resulted in inconsistencies. We determined that a text input length not exceeding approximately 2300 words resulted in no slurred speech.

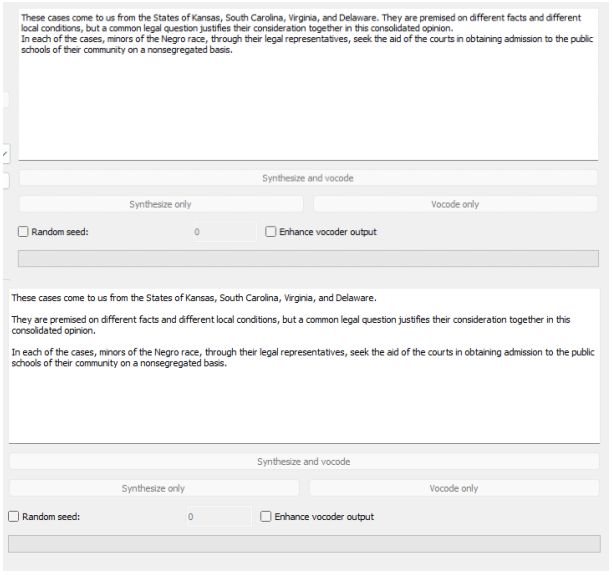

Figure 4. Comparison of the same text with different ordering of sentences.

In addition to the results obtained from the experiments during this project, several challenges were encountered. The first challenge was the inability to process long sequences of text, which we addressed by separating long sequences by sentence, generating output for individual sentences, and then merging the results into one voice output file for the limitation on the result found on the model for no more than a specific range of length in words for the best quality of output voice; the second was slurred speech in the output since the original work does not support any punctuation. Therefore, modifying how to handle sentences in a long script was challenging. As illustrated in Figure 4, the order of sentences on how to place them in the input text field matters for better sounding improvement. We added pauses, achieved through the insertion of commas or tabs between sentences, which would efficiently maintain quality output voice on historical figures even with old recordings for input voice. The last was the variation based on the length of tuning speech, where, for short inputs, averaging embeddings to generate spectrograms in smaller numbers. However, for long inputs, this approach could dilute the distinctive voice features of the speaker. As a solution to this challenge, utilizing the voice input within ten seconds would provide better quality of the voice output.

4. Experiment Summary

The project works in a collaborative, supportive environment. Done the work within a peer group of four, fostering cooperative work and data sharing. Nevertheless, the coding changes, analysis, and model adaptation within this project were conducted independently, with great support from advisor Dr. David Doermann. Peers provided syntax assistance within the code and aided in debugging; however, developed the vast majority of this model by myself.

The project has demonstrated that audio cloning is not just a technical challenge but also an exploration of the ethical and educational implications of AI. The knowledge I have acquired from courses such as CSE 676 Deep Learning, CSE 574 Machine Learning, and EAS 595 Fundamental of AI was instrumental in this endeavor. The coding expertise greatly supported the project, essential in developing the cloning process and integrating various files using contemporary techniques. Furthermore, soft skills such as critical thinking, problem-solving, and ethical reasoning played a significant role in this project.

The model could be better, and there is room to develop a system for accurately simulating the environment of the Supreme Court, including its unique acoustics and layout, the project has the potential to incorporate advanced features. These include modulating the emotional tone of the voices and their intonation and capturing regional speech variations in future work. And perhaps the implementation of noise reduction algorithms and audio processing techniques. Such advancement aims to increase the realism of voice reproductions and transform our approach in class to historical learning, significantly revolutionizing our interaction with historical content and making it more interactive, offering more dynamic and engaging educational experiences. The long-term goal involves advancing AI, particularly in ethical AI development. Paving the way towards achieving this goal is essential for the project improvement.

As advice to more junior members of the Engineering Science in Artificial Intelligence program, encourage focusing on both the technical and ethical aspects of AI after completing project work. It is essential to develop strong coding and machine learning skills and understand the broader implications of work.

5. Conclusion

Throughout this project, understanding the text-to-speech process and recognizing its broader implications has been deepened. The significance of the civil rights

case Brown v. Board of Education also depends on each American, highlighting that ethical issues will challenging on historical contexts. An instance of such ethical dilemmas is the misuse of deepfake technique, where criminals manipulate the voices of individuals never actually made by the person before, resulting in potential blackmail or fraudulent account transfers to their friends or families.

It is an important step to approach these techniques with a balanced and responsible mindset, especially using historical figures or histories. It is vital to avoid distortion of historical facts, clear of any deception or misrepresentation regarding history or historical figures. It is crucial to prohibit using techniques, particularly in any manner, that could deceive or mislead people by believing they are interacting with historical figures. Respecting individual rights is of the most important thing in honoring the rights and legacies of historical figures, ensuring the factual accuracy of their contributions to history is educated correctly to everyone. Never distort the role of historical figures in the past, work with techniques that avoid defraudation or lying to people, and ensure that relentless work is prohibited, the individual rights of historical figures are respected.

To conclude, I sincerely thank Dr. David Doermann for providing invaluable guidance. Additionally, I would like to acknowledge my peer group members for their collaborative efforts on data and code modifications that contributed to the progress of this project.

6. Acknowledgment

My gratitude to my advisor, Dr. David Doermann, for his wise advice and his patience during this semester.

I would like to thank my peers Ashish Pyla, Tushar Goyal and Tanamair for their collaboration on data and help with code modifications of my work throughout this project.

I wish to thank my family for their great support in all aspects during the work of my study of master degree at the University at Buffalo.

References

[1] Jon Truby and Rafael Brown, *Human digital thought clones: the Holy Grail of artificial intelligence for big data*. Information & Communications Technology Law, 2021.

[2] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu, *Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis*, 2018. [Online]. Available: https://arxiv.org/pdf/1806.04558.pdf

[3] CorentinJ, *Real Time Voice Cloning*. GitHub Repository. [Online]. Available: https://github.com/CorentinJ/Real-Time-Voice-Cloning

[4] Leland McInnes, John Healy and James Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2020. [online]. Available: https://arxiv.org/pdf/1802.03426.pdf

[5] JOLLIFFE, I.T, *Principal Component Analysis*, 2002. second edition, New York: Springer-Verlag New York, Inc.

[6] Wikipedia contributors, "Earl Warren," *Wikipedia, The Free Encyclopedia*. [Online]. Available: https://en.wikipedia.org/wiki/Earl_Warren

[7] Wikipedia contributors, "Ruth Bader Ginsburg," *Wikipedia, The Free Encyclopedia*. [Online]. Available: https://en.wikipedia.org/wiki/Ruth_Bader_Ginsburg

[8] Supreme Court of the United States, *Brown V. Board of Education*. U.S. National Archives and Records Administration. [Online]. Available: https://www.archives.gov/milestone-documents/brown-v-board-of-education