STYLE BASED SINGING SYNTHESIS

CULTURAL AND SOCIAL IMPACT OF MUSIC SYNTHESIS SYSTEMS

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Liam Brennan

November 1, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Professor Joshua Earle, Department of Engineering and Society

Professor Yuan Tian, Department of Computer Science

Introduction

For both my STS paper, as well as my Technical Thesis, I am going to be exploring Artificial Intelligence (AI) based music synthesis systems.

Throughout history, technology has helped facilitate the ability of humans to express themselves through music. From the first flutes over 40,000 years ago to modern music creation software, technology continuously redefines and revolutionizes music. However, with the advent of Deep Learning, a subset of AI categorized by extremely complex and powerful models, we are rapidly approaching an inflection point that will change the relationship between music and technology. Recent projects have shown the capabilities of Deep Learning based systems to synthesize art comparable to human professionals in a completely unsupervised manner (Dhariwal et al., 2020)(Pal*, Saha, & Anita, 2020)(Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022). If the current trend continues, these systems will begin to monopolize the creative process, instead of supplementing it.

For my STS paper, my main research question will be, "Who benefits, and who is disadvantaged, from the continued development of AI-based music synthesis systems?". Part of this project will be analyzing current music synthesis systems, and understanding the strengths and weaknesses of these systems compared to traditional methods of music. In addition, because AI music synthesis is a relatively new and growing subject, understanding not just where the technology is at, but where it's headed, is instrumental to my analysis. After this technical research, I will be exploring the diversity of music across different cultural, geographical, and socioeconomic backgrounds. Combining these two types of research, I will analyze who stands to benefit from the development of AI-based music systems, and propose solutions to make the current, and future, state of artificial music more ethical and equitable.

This work is important for a couple of reasons. For better or for worse, AI is becoming a dominant technological force. In 2022, the global AI industry is estimated at over 100 Billion dollars and is expected to grow by almost 20x to nearly 2 Trillion dollars by 2030 (Technologies, 2022). However, work on AI is disproportionately focused on the technical aspects, instead of the wide-ranging societal impacts. AI art, particularly music, is a new frontier in AI that has the capacity to change our perspective on what makes us inherently human. A discussion of how AI music synthesis might change society on a global scale, and whom this technology impacts, is of utmost importance.

For my technical project, I will be building my own music synthesis project, using machine learning. In particular, I will be focusing on building synthesis models for one of the most popular, and complex, instruments: the human voice. Stemming from my previous work on generative machine learning models, and my own research projects on voice authentication, I will build a human voice synthesis model. Leveraging my experience with deep learning projects, my model will build off the latest works in generative deep learning. As this project involves building an AI music synthesis system, it is directly related to my STS project. My STS project will discuss the social implications of AI music synthesis systems, while my technical project will let me get hands-on experience building my own AI music synthesis systems.

This prospectus is organized as follows. In the first section, I will discuss the specifics of my technical project. In the second section, I'll discuss my STS project, including my

research plan. In the third section, I'll go over key texts that have large implications for my project.

Technical Project

For my technical project, I will be building a voice synthesis project using generative deep learning methods. Generative models train on a large number of samples (music, in this case), and learn to produce their own samples that sound similar. In particular, I will be leveraging Generative Adaptive Networks (GANs) (Goodfellow et al., 2014) as my main generative architecture. GANs work by building a system of 2 models in competition. One model, the Generator, produces samples, and the other, the Discriminator, tries to figure out if the samples are "real" or from the Generator. Over time, the Generator gets better at making realistic samples by trying to trick the Discriminator. In addition, I will build on top of recent "style" based generative methods such as StyleGAN (Karras, Laine, & Aila, 2018) that allow parameters (such as tone or tempo) to be tweaked at synthesis to generate pieces of different "styles". I will be building and training my system using the open source python deep learning framework PyTorch (Paszke et al., 2019).

First, I will build my own dataset. My plan for this is to scrape music off of various websites, including YouTube, Soundcloud, and Spotify. I will then process the music to extract only the appropriate voice samples using various music processing techniques. In addition to extracting the voice samples, I will be collecting various metadata about the voice samples and the songs that they sampled from.

To train the model, I will begin by pre-training on large spoken word datasets. Then, I will then use my collected data to fine-tune the model, which will learn how to create its own singing samples. For my system itself, users will be able to use a reference voice sample and edit the voice sample using a variety of metadata. I plan to release this model online through a web application.

STS Project

My main research question for this project is "Who benefits, and who is disadvantaged, from the continued development of AI-based music synthesis systems?". This is an important subject for a variety of reasons. As mentioned in the introduction, AI technology is quickly becoming a global force. In particular, AI art synthesis, such as music synthesis, is completely uncharted territory. Never in recorded history has technology had the capacity to surpass humans at something so fundamental to our identity as a species. Because of the potential of this technology to shape many of the fundamental aspects of life, a detailed social analysis of the technology is important. Understanding whom this emerging technology affects and how it affects them will be important social commentary. This is especially true due to music existing as a core part of many different cultures and identities.

While AI music synthesis has broad implications in every nook and cranny of society, due to every group and individual having a unique musical identity, it is not feasible to discuss every societal implication. Consequently, for this work, I will be focusing on how the emer-

gence of AI-based music synthesis affects two different types of hierarchies: socioeconomic and cultural.

For the socioeconomic hierarchy, I am going to explore how AI-based music synthesis plays into the working and ruling class struggle. While AI music synthesis has the potential to democratize music creation, allowing even non-musicians to produce professional quality music, the reliance on big data and computing power means that the powerful and wealthy could monopolize AI music synthesis. This exclusivity could have long-reaching societal implications.

For the cultural hierarchy, I will explore the uneven extent to which AI music-based synthesis will lend itself to different cultural backgrounds. Music is an essential part of nearly every cultural identity, and every culture has different styles of music that are important to that identity. Historically, music synthesis has been focused on western classical and pop music. As the scope of music synthesis expands, exploring which styles of music these synthesis systems lend themselves toward is an important topic of conversation.

One of the main STS frameworks that I will use to answer my research question is the Social Construction of Technology (SCOT). SCOT posits that technological advancement is usually a reflection of society, and not vice versa. This framework lends itself almost perfectly to AI-based music synthesis. In contrast to most types of technology, music synthesis has almost no practical application. Art is, almost by definition, a societal construct. Art is used to express emotion and identity for groups and individuals, not to complete a task or solve a problem. Consequently, AI music synthesis is a reflection of society. The types of music people like to listen to, or the styles of music people want to make, will completely drive the technology. Thus, the SCOT framework is an excellent way to analyze AI music synthesis.

The other main framework that I will use to conduct my analysis will be Race Critique. As mentioned above, music is a large part of different cultural and ethnic identities. Additionally, music synthesis has been used historically for white, western audiences. By using Race Critique, I will look into how music synthesis currently, and in the future, is disproportionately built for different types of audiences, and the implications that that has.

I will conduct my research in three main steps. Firstly, I will dive into the technicalities of current, and possibly future, AI-based music synthesis systems. Doing this will allow me to understand the strengths and weaknesses of these systems are, and where the technology is heading. Next, I will explore the diversity of musical styles across different cultural and socioeconomic groups. Through this, I'll be able to contrast the trajectory of the current technology with the diversity of music throughout these different groups. Thirdly, I'm going to look into the impacts of the music industry in general, and how prejudice plays a key role. This will allow me to bring my previous research together and analyze the effect of AI-based music synthesis systems across different demographics.

Key Texts

There are a few key texts that I analyzed to get a better understanding of AI music synthesis and its social implications.

As mentioned in the previous section, my first research objective is to understand the

state of the art in AI music synthesis. To do this, I looked into technical research papers that outline some of the best recent works. The first work I looked into is MuseNet (Pal* et al., 2020), released by OpenAI in 2020. This work was one of the first front-to-back music synthesis platforms, creating entire midi compositions emanating different musical styles, such as country, classical, and pop. The authors of this paper use a generative model called a Transformer, which is very good at modeling long sequences. To train the model, they used a large database of midi notes and asked the model to predict the next sequence of midi notes. Another recent AI music synthesis work is Jukebox (Dhariwal et al., 2020), released by OpenAI in 2021. In a similar vein, this work is a front-to-back music synthesis platform that can produce songs of different musical styles. Instead of midi notes, this work trains on raw audio from songs and uses a different architecture called a Variational Autoencoder (VAE). Autoencoders work by "compressing" samples into a small set of normally distributed features called *embeddings*. Then, samples are generated by asking the model to produce a sample by "decompressing" some randomly generated embedding.

In regards to STS texts, I first looked into "The Sonic Episteme" by Robin James (James, 2019). This book looks into the relationship between music in the 21st-Century and white supremacy, capitalism, and patriarchy. James argues that many of the seemingly harmless trends in 21st century music perpetuate some of the inequalities and injustices that exist within our society. Additionally, I looked into "Artificial Unintelligence" by Meredith Broussard (Broussard, 2019). This text looks into the strengths and weaknesses of AI, and what things AI is poor at compared to other, more traditional systems. Broussard argues against the idea of *techno chauvinism*, the idea that technology is the solution to any problem. In Broussard's eyes there are some problems AI cannot solve, and expecting AI to solve everything is just as dangerous as it is unrealistic.

References

Broussard, M. (2019). *Artificial unintelligence: How computers misunderstand the world.* The MIT Press.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). *Jukebox: A generative model for music.* arXiv. Retrieved from https://arxiv.org/abs/2005.00341 doi: 10.48550/ARXIV.2005.00341

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). *Generative adversarial networks.* arXiv. Retrieved from https://arxiv.org/abs/1406.2661 doi: 10.48550/ARXIV.1406.2661

James, R. (2019). *The sonic episteme: Acoustic resonance, neoliberalism, and biopolitics.* University Press.

Karras, T., Laine, S., & Aila, T. (2018). *A style-based generator architecture for generative adversarial networks.* arXiv. Retrieved from https://arxiv.org/abs/1812.04948 doi: 10.48550/ARXIV.1812.04948

Pal*, A., Saha, S., & Anita. (2020). Musenet : Music generation using abstractive and generative methods. *International Journal of Innovative Technology and Exploring Engineering*, *9*(6), 784–788. doi: 10.35940/ijitee.f3580.049620

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with clip latents.* arXiv. Retrieved from https://arxiv.org/abs/2204.06125 doi: 10.48550/ARXIV.2204.06125

Technologies, N. G. (2022). *Artificial intelligence market size report, 2022-2030.*