A COMPARATIVE EVALUATION OF TONAL REPLICATION TECHNIQUES FOR MUSIC COMPOSITION AND SOUND DESIGN THROUGH TONAL SYNTHESIS

A Thesis

Presented to the

Department of Information Systems

and Computer Science

Ateneo de Manila University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science in Computer Science

by

Angelo Joaquin B. Alvarez

John Aidan Vincent M. Ng

Justin Carlo J. Reyes

2024

ABSTRACT

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

CHAPTER 1

Introduction

1.1 Context of the Study

Music is an art that evolves as musicians search for new ways to express their messages, emotions, or even their identity. Throughout music history, there has been a consistent push towards innovation whether it be by composition, sound design, or other areas of study. One such example is the musical expression of jazz, considering its relatively recent rise in the early 1900's. In the past half century, synthesizers have led to the generation of audio signals and waveforms using additive, subtractive, or frequency modulation synthesis techniques.

The advent of synthesizers gave rise to a new era of music-making, giving birth to an array of new genres such as electronic, techno, ambient, and house music, among others. Oscillators, filters, envelopes and modulation sources came with the rise of synthesizers to empower musicians in sculpting sounds to the limits of their creative expression. Through the advancements of these features, synthesizers have broken the barriers of genres. Synthesizers have been integrated into various musical contexts, giving each genre the modern technological edge.

The dominance of synthesizers over the past few decades of the music industry cannot be understated. Nevertheless, new methods of music production have come to light due to advancements in synthesizer technology and artificial intelligence. Methods such as algorithmic composition, modular synthesis, sample-based production, and artificial intelligence-assisted composition have increased in usage over the past years.

This study aims to explore the intersection of artificial intelligence and synthesizer technology, focusing on the replication of instrumental tones through an audio synthesis from an instrumental melody. Through comparing the latest methods and techniques in the field of tone replication, this study will contribute to our understanding of how synthesizers can further shape the landscape of music composition and sound-design in the coming years.

1.2 Research Objectives

The study aims to utilize various techniques in audio synthesis and replication, namely Tone Transfer, GANs (Generative Adversarial Networks), and DSPGAN in order to replicate instrumental tones, comparing them in terms of accuracy and efficiency. The sub-objectives are as follows:

- To determine how the quality and choice of input audio encoding and representation affect sound replication efficiency and accuracy,
- To determine the stability of the outlined methods in handling various factors in the audio input, such as noise, frequency masks, low audio quality, and fragmented audio inputs,
- To determine the perceptual quality of synthesized audio as compared to the original audio input, and
- To determine the effect of contrasting feature selection and representation on the accuracy and efficiency of synthesized audio.

1.3 Research Questions

The study seeks to answer the question: How do methods in audio synthesis and replication such as Tone Transfer, GANs, and DSPGAN compare in their accuracy and efficiency in replicating instrumental tones? In answering this question, the following sub-questions can be answered:

- How does the quality and the choice of encoding and representation for the input audio impact the efficiency and accuracy of sound replication?
- How stable are different methods in audio synthesis and replication in handling noise and frequency masks in the audio input, low audio quality, and fragmented audio inputs?
- How does the synthesized output generated by the model compare to the original audio input in terms of perceptual quality?
- How do the contrasting feature selection and representation employed in different methods contribute to their accuracy and efficiency in replicating instrumental tones?

1.4 Scope and Limitations

To define the scope of this paper, the study will be limited to implementations of the concerned methods (Tone Transfer, GANs, and DSPGAN) as outlined in their respective studies. The architecture for the Tone Transfer method will follow that of [study]. For the utilization of GANs, architecture from Engel et al., known as GanSynth [cite], will be used. Lastly, the architecture of DSPGAN will be based on its original paper by Song et al. [cite]. Tentatively, this study will focus on the comparison of these architectures, refraining from delving into deeper techniques such as model or architecture creation.

In the context of the study, audio input will be defined as any melodic recording of an instrument– tentatively, the instruments to be used will be the violin, guitar, flute, trombone,

organ, and marimba. With regards to training data, these implementations will be trained on the NSynth dataset, a set of musical notes from various instruments created by Engel et al. [cite]. Notably, there is not yet any literature regarding the training of the concerned models on polyphonic data, such as audio with multiple instruments. Several metrics will be used in evaluating the synthesized output of the various models, namely Frechet Audio Distance (FAD), Number of Statistically Different Bins (NDB,) Mean of Opinion Score (MOS), and Subjective Preference Testing.

It should be noted that in the field of audio synthesis, none of the metrics currently used in evaluating synthesis techniques are individually sufficient to determine the accuracy of a model. For example, MOS and Subjective Preference testing can be subjective, given that the two metrics are based on scoring from human subjects. As such, they can produce varying results. The objective metrics are similarly inconclusive: FAD only measures the distribution of generated audio, while NDB ignores the temporal aspect of audio in favor of spectral content. It is hoped that utilizing these metrics in tandem could aid in mitigating their individual shortcomings.

1.5 Significance of the Study

This study lies at the intersection of computer science and computational musicology, focusing on advancing music sound synthesis through generative artificial intelligence. Its aim is to provide valuable insights by evaluating the three most prominent generative models in the field: Tone Transfer, GANs, and DSPGAN. Through comparative analysis, the study seeks to determine which model holds the greatest potential for further advancement in the field. By selecting and scrutinizing these models, the research strives to shed light on which of the following models will be better suited in the use of developing new tools in the realms of music production, audio restoration, processing or enhancement, and sound design.

CHAPTER 2

Review of Related Literature

2.1 Previous Methods for Audio Synthesis

2.1.1 WaveNet and GANs (Generative Adversarial Networks)

The use of WaveNet autoencoders for neural audio synthesis by Engel et al. (2017), one of the most relevant technological advancements regarding audio synthesis of the past decade, paved the way for computational musicology by using conditional autoencoders learned from raw audio waveforms [cite]. Another contribution from the same study is the NSynth dataset, a "large-scale dataset for exploring neural audio synthesis of musical notes," which was composed of over 300,000 notes belonging to instruments of different families (strings, vocals, wind instruments, etc.) It was found in their study that playing styles such as vibrato could be replicated when looking at instantaneous frequencies in a spectrogram, and that harmonic structures and overtones blended more smoothly. While revolutionary, this method was only able to recreate sample-based audio, and was not able to capture the full global context. Regardless, the paper showed that creating sample-by-sample audio signals was possible through the use of deep learning.

Instead of using vocoders, frequency modulation, MIDI synthesizers, or any combination of the three and other possible methods, deep learning for audio synthesis is headed toward directly replicating the waveform of audio samples. GANSynth, a study conducted by Engel et al. (2019), uses a Progressive GAN architecture [cite] combined with conditioning of an additional feature: a one-hot representation of musical pitch. GANSynth uses Short Time Fourier Transforms (STFT) and IF-Mel (log magnitudes and mel frequency scales) variants in order to generate samples over 50,000x faster than WaveNet autoencoders. Aside from faster sample-generation, GANSynth utilizes information from the latent features and musical pitch of the training dataset to generate audio exhibiting smooth timbral interpolation and timbral consistency across different pitches. The

introduction of GANSynth marks a significant development in the use of GANs for audio generation, synthesizing audio with superior quality when compared to the previous WaveNet autoencoder.

2.1.2 DDSP (Differentiable Digital Signal Processing)

In 2019, the term differentiable digital signal processing (DDSP) was introduced by Engel et al. (2020). In their study, the pitfalls of various previous methods were discussed, such as strided convolution models (WaveGAN, SING) giving only general representations that can present any waveform, Fourier-based models (GANSynth) not being able to assert the issue of spectral leakage leading to phase misalignments, and autoregressive models (WaveNet, RNNs) being able to hold their own but being prone to exposure bias and incompatibility with other spectral audio features.

The DDSP library introduced by Engel et al. offers a suite of differentiable components, including Spectral Modeling Synthesis, Harmonic Oscillators, overlapping Hamming window-based amplitude envelopes, a time-varying FIR filter, a subtractive synthesizer, and a reverb module. Evaluated with both supervised (NSynth) and unsupervised (solo violin) datasets, DDSP exhibits high-fidelity synthesis capabilities. Crucially, it enables independent control of loudness and pitch while facilitating timbral transfer (e.g., transforming a singing voice into violin-like sounds). The entire architecture of the autoencoder essentially channels multiple audio signals through different audio synthesis algorithms and combines each signal into one using a reverberation module.

In recent years, DDSP has been retrofitted into various methodologies in the academic field, as discussed by Hayes et al. (2023)'s catalogue and survey into the use of said techniques in sound and music synthesis. According to their review, DDSP-based audio synthesis in the field of musical audio synthesis can be classified into four distinct areas: (i) Musical Instrument Synthesis,

(ii) Performance Rendering, (iii) Timbre Transfer, and (iv) Sound Matching. For example, Hayes et al. (2021)'s neural waveshaping synthesis method uses NEWTs (neural waveshaping unit, learning shaping functions from unlabelled audio) fed into a harmonic-plus-noise synthesizer, much like DDSP. There are also several implementations of the library using PyTorch– Masuta and Saito (2021) implemented a similar-yet-controllable version by having controllable parameters such as cutoff frequency and a differentiable subtractive synthesizer, and more. Additionally, Shan et al. used DDSP methods as a building block, combining it with differentiable wavetable synthesis.

Despite advancements in DDSP-based generative models, Tone Transfer, a model developed by Carney et al., remains the most prominent model using the DDSP architecture. Tone Transfer employs a modified DDSP architecture to enable a small, fast, efficient, and computationally inexpensive web application. It substitutes the DDSP encoder's weighted spectrograms for loudness with root-mean-squared power for the loudness waveform, while adding a few more layers to the DDSP decoder to achieve an application condensed enough to work efficiently on the web.

Another significant development is a model that combines two different kinds of generative models. DSPGAN, a GAN-based universal vocoder for high-fidelity speech synthesis by applying the time-frequency domain supervision from digital signal processing (DSP), is one of the most recent developments combining both GANs and methods from Engel's methodology. The paper regarding its creation by Song et al. discussed how pitch jitters and discontinuous harmonics would commonly be found in GAN-based vocoders, and so they combined a DSP module, finding that "it can generate high-fidelity speech for various TTS models trained using diverse data."

2.2 Audio Synthesis Evaluation Metrics

Despite the development of generative models in audio, video, and image synthesis, finding evaluation methods still pose a challenge, particularly for implicit models that do not generate quantifiable values. Although these models can easily be judged through perceptual evaluation, objective metrics are still significant in the comparison of models, architectures, and hyperparameters. Common evaluation metrics of generative models are often driven by intuition and may overlook certain features of the product. These limitations are also present in audio synthesis. Vinay and Lerch cite various metrics that can be used for evaluating audio generative models following methodologies of previous studies. Moreover, Betzalel et al. recommends using a diverse set of evaluation metrics when comparing generative models to control the variability in scores. As such, not only objective evaluations are necessary, subjective evaluations, such as those requiring human discrimination, are also necessary for ensuring that the produced audio will cater to the perception of humans.

2.2.1 Objective Evaluation Metrics

Numerous metrics have been developed to evaluate generative models, with Inception Score (IS) being the most commonly used. Betzalel et al. delineates six (6) evaluation metrics used for implicit generative models. The paper further recommends dropping the use of IS in favor of Frechet's Inception Distance (FID), as IS performs relatively worse among other evaluation metrics. However, it should be noted that this study is primarily focused on image generative models. IS, FID, and other metrics that utilize Inception measure the quality and diversity of an image. Using metrics such as IS and FID for other generative models, such as audio and video, requires replacing the Inception component of these metrics with features of the intended input data.

Despite being originally developed for image evaluation, these metrics have been modified to adapt to audio-based generative models. A workaround to adapt FID to the audio domain has been developed by Kilgour et al. by replacing the inherent Inception network within FID with the VGGish model. The developed technique, called Frechet Audio Distance (FAD), works by generating embedding statistics with the VGGish model on the evaluation set and compares it to the embedding statistics generated on a reference set of clean music, which is usually the training set. This metric is used to measure audio quality by analyzing distortions within the audio. Thus, this metric will be helpful in objectively evaluating the quality of produced audio from the three generative models.

The Frechet Audio Distance is not sufficient to fully evaluate the chosen generative models, as it only compares the distribution of the generated audio to the distribution of the original audio dataset. Another metric called Number of Statistically Different Bins (NDB/k) that was introduced by Richardson and Weiss to measure the diversity of generated audio from the original dataset. NDB/k works by distributing test samples into k clusters using an *L2* distance measure and performing two-sample t-tests between each cluster pair, with the NDB score being the proportion of statistically different clusters to the amount of clusters . This metric measures that the produced audio is uniquely generated and not replicated from the original dataset.

The use of these objective evaluations allows for a nuanced comparison between generative models. While FAD provides valuable insight into audio quality by measuring distortion from the training data, NDB/k offers a complementary perspective by assessing the diversity of generated audio, measuring its similarity to the training data. By considering both metrics, we gain a more comprehensive understanding of an audio generative model's strengths and weaknesses.

2.2.2 Human Evaluation Metrics

Since the generated audio will be made for consumption and use for music creation, human evaluation is necessary to ensure accurate perceptual quality. A commonly used subjective metric is the Mean Opinion Score (MOS) where participants are asked to rate the sound they hear in a 1 to 5 Likert Scale across a set of questions that pertain to the quality of the audio. These questions might focus on aspects like clarity, naturalness, pleasantness, and overall fidelity. By aggregating the scores from a diverse group of listeners, MOS provides a quantitative measure of how humans perceive the generated audio. Combining these objective and subjective evaluations provides a well-rounded picture of the strengths and weaknesses of different audio generative models.

[OTHER PART OF TESTING]

CHAPTER 3

Methodology

https://lucid.app/lucidchart/39cf1906-a300-44a6-b6e8-2be9975e22a6/edit?viewport_loc=-11%2C-11%2C2032%2C1289%2C0_0&invitationId=inv_03d5bddd-6d99-4076-8577-79b10c15d05d

3.1 Comparative Analysis

The researchers will examine the existing DDSP-based methods focusing on one-shot neural audio synthesis, such as:

1. Differentiable Digital Signal Processing by Engel et al. (2020)
2. Differentiable Wavetable Synthesis by Shan et al. (2021)
3. Neural Instrument Cloning From Very Few Samples by Jonason and Sturm (2022)

DDSP-based methods using external components like synthesizers and more on learning parameters were not chosen. The ones above also use similar audio features in order to generate synthesized audio. In order to stay consistent, the researchers will retrain each architecture (DDSP, Differentiable Wavetable Synthesis, and Neural Instrument Cloning) using the recommended hyperparameters and training configurations from the respective papers.

3.2 Dataset Preparation

The NSynth dataset will be used for its' large scale and annotations. NSynth contains over 300,000 musical notes spanning a wide range of pitches, timbres, and envelopes. This diversity allows for comprehensive evaluation of how well different DDSP methods can model various aspects of musical instrument sounds.

Additionally, each note in NSynth is monophonic and generated in isolation, allowing for focused analysis of timbre and synthesis quality without the complexities of polyphony or musical context. This controlled setting facilitates direct comparison of different DDSP models on their ability to accurately reconstruct individual notes.

The following audio features will be used:

1. F0 using CREPE (Convolutional Representation for Pitch Estimation)
2. Loudness using STFT and A-weighting

Pre-trained models of CREPE, a data-driven pitch tracking algorithm will be used for getting the fundamental frequency. It has been proven to be accurate within a strict evaluation threshold of 10 cents. Additionally, in order to analyze loudness, all of the papers above use an A-weighting of the power spectrum. A-weighting is applied to instrument-measured sound levels in an effort to account for the relative loudness perceived by the human ear, as the ear is less sensitive to low audio frequencies. In summary, F0 and Loudness allows for providing the instantaneous fundamental frequency and intensity of a note sequence at a constant frame rate.

3.3 Model Training

3.3.1 Loss Function

The multi-scale spectral loss, an objective function of the L1 difference for the Fast Fourier Transform, is used to compare the generated audio with the ground truth audio. It was found by Shan et al. that the preceding formula,

$$L_{\text{reconstruction}} = ||S_i - \hat{S}_i||_1$$

Would work better than the one made by Engel et al. where

$$L_i = ||S_i - \hat{S}_i||_1 + \alpha||\log S_i - \log \hat{S}_i||_1, \text{ and } L_{\text{reconstruction}} = \sum_i L_i.$$

S_i and $\hat{S}_i$ respectively denote magnitude spectrums of target and synthesized audio, and i denotes different FFT sizes. FFT sizes used in the papers were FFT sizes (2048,
1024, 512, 256, 128, 64).

3.3.2 Dataset Use

While the NSynth dataset provides a vast amount of data, a smaller subset was used totalling "70,379 examples composed mostly of strings, brass, woodwinds and mallets with pitch labels within MIDI pitch range 24-84." A notable factor to consider between the three papers is that from DDSP to DWTS to NIC, the sample size for the inputs are able to shrink.

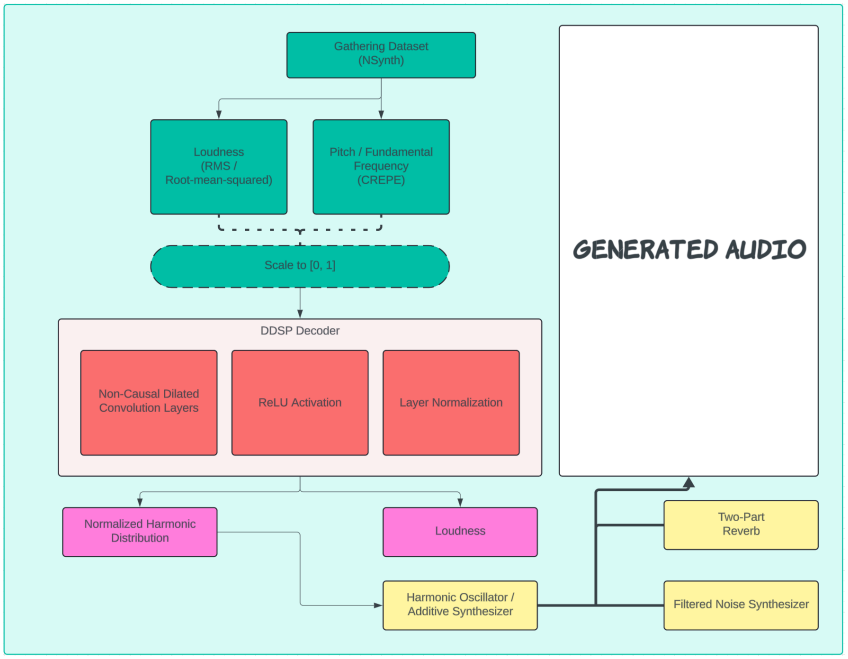
3.3.3 Key Differences

NIC introduces a trainable recording-specific embedding called Z. This embedding is not generated by an encoder, but rather trained jointly with the decoder for each individual recording. In addition to F0 and loudness contours, NIC also conditions the decoder on the F0 confidence score provided

by the CREPE pitch estimator. This score indicates how confident the pitch estimator is in its F0 prediction. By including this information, the model can better distinguish between pitched and unpitched sounds (e.g., breath noises, key clicks), leading to fewer artifacts in the synthesized audio, especially when training data is limited. NIC proposes a novel two-part reverb design to address the issue of excessive reverberation when training on small datasets. This design models the room impulse response (IR) as a combination of two components.

3.4 Hybrid Autoencoder-GAN Architecture

Incorporating a GAN into the DDSP autoencoder architecture and creating a hybrid approach would allow the researchers to produce outputs with finer details and textures that could possibly result in a more "realistic" and characterized waveform generation. Combining the strengths of the three methods, with (i) high audio quality and controllability from DDSP, (ii) compact and efficient timbre representation from DWTS, and (iii) data efficiency and reduced artifacts from NIC may further encourage the model to generate realistic and high-fidelity audio when combined with a GAN. By integrating the output of the hybrid model as a generator for the GAN, the discriminator network tries to distinguish between real and generated audio samples. As mentioned, DSPGAN is one of the papers that have examined the possibility of a similar architecture but more so in the field of vocoders instead of music synthesis, using mel-spectrograms generated by a DSP module.



Proposed Generation Model

The DDSP Decoder taken from this paper (Tone Transfer) uses two stacks of non-causal dilated convolution layers as the decoder.

non-causal dilated convolution layers

- nondilated input convolution layer
- Dilation factor
- Kernel
- Layer normalization

3.5 Metric