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

TBA

ACKNOWLEDGEMENTS

TBA

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

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 a 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 Differentiable Digital Signal Processing (DDSP), Differential Wavetable Synthesis (DWTS), and Neural Instrument Cloning (NIC) in order to replicate instrumental tones, comparing them in terms of accuracy and efficiency. The sub-objectives are as follows:

- 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 how different audio synthesis methods handle replicating tones from instrumental domains not included during training.

1.3 Research Questions

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

- 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 models compare to the original audio input in terms of perceptual quality?

- How robust are different audio synthesis methods in accurately replicating tones from instrumental domains not included during training?

1.4 Scope and Limitations

To define the scope of this paper, the study will be limited to implementations of the concerned methods (DDSP, DWTS and NIC) as outlined in their respective studies. The architecture for the use of DDSP will follow that of Engel et al. in their introduction of the library [3]. For DWTS, architecture from Shan et al. [17] will be used. Lastly, the architecture of NIC will be based on its original paper by Jonason and Sturm [10].

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. [3]. 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 Fréchet Audio Distance (FAD), Number of Statistically Different Bins (NDB/k), Mean of Opinion Score (MOS), and Multiple Stimuli with Hidden Reference and Anchor (MUSHRA). 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 MUSHRA 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/k 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. The study aims to provide valuable insights by evaluating three of the most recent generative models in the field: Differentiable Digital Signal Processing, Differential Wavetable Synthesis, and Neural Instrument Cloning. Through comparative analysis, the study evaluates which model has the greatest potential to enable novel creative possibilities and improve workflows in music production, audio restoration, processing or enhancement, and sound design. By selecting these models, the research strives to determine which holds the best promise for generating more realistic and natural audio, expanding the sonic palette available to creators, and streamlining the process of finding the perfect sound.

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. [5], 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. 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, introduced by Engel et al. [3], uses a Progressive GAN architecture [11] 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. [4]. 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.

Through a combination of oscillators $x(n) = \sum_{k=1}^{K} A_k(n)\, sin(\phi_k(n))$ (with $A_k(n)$ as the time-varying amplitude of the *k*-th sinusoidal component), where instantaneous phase $\phi_k(n)$ is achieved through $\phi_k(n) = 2\pi \sum_{m=0}^{n} f_k(m) + \phi_{0,k}$. The DDSP autoencoder is built to learn parameters from the input audio combined with the filter and the reverb module.

The autoencoder implemented by Engel creates an encoder taking in an input tuple $(f(t),\ l(t),\ z(t))$: in a given number of timesteps t, $f(t)$ is the fundamental frequency achieved using Convolutional Representation for Pitch Estimation (CREPE), a pitch tracking algorithm [13] which extracts the fundamental frequency from a given audio, $l(t)$ is the loudness which is

computed through an A-weighted log power spectrum of the audio, and z(t): the smoothed spectral envelopes of the harmonics done through passing MFCC coefficients into a GRU layer. 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.) In summary, 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. Their results show that the autoencoder architecture was able to accurately resynthesize the solo datasets.

2.1.2.1 Differentiable Wavetable Synthesis (DWTS)

Another form of audio synthesis derived from DDSP was introduced by Shan et al., who used DDSP methods as a building block in order to create Differentiable Wavetable Synthesis (DWTS) methods [17]. Their implementation relies mainly on a list of wavetables taken from one-shot audio samples (i.e. the NSynth dataset) as a learnable dictionary using gradient descent. From this dictionary, each wavetable represents a wide variety of timbres sourced from the target audio. By morphing between the wavetables in this dictionary, their implementation makes it possible to change timbre over time, theoretically leading to a more dynamic output. Initialized with a zero-centered Gaussian distribution, the signal $x(n)$ is synthesized by combining wavetables in the dictionary through the formula

$$x(n) = A(n) \sum_{i=1}^{N} c_i(n) \cdot \Phi(w_i, \widetilde{j}(n), \kappa),$$

where "$A(n)$ is a time-varying amplitude controlling the signal's overall amplitude and c_i denotes the time-varying attention on w_i. $A(n)$ and $c_i(n)$ are constrained positive via a sigmoid. The

function $\Phi(wi\ , j(n), \kappa)$ is a fractional indexing operator that returns the $\tilde{j}$-th element of the vector w_i by using an interpolation kernel κ to approximate $w_i\ [\tilde{j}]$ when $\tilde{j}$ is non-integer."

It was found that their implementation resulted in sounds that matched the physics of NSynth samples. The model also takes advantage of phase relationships within and between wavetables, while also opening up for possible polyphonic passages when an autoencoder model is initialized with pretrained wavetables and combined with multiple phase accumulators— however, polyphonic testing was not included in their dataset. Their model (tested on N = 5, 10, 20, 100 wavetables) resulted in a lower reconstruction error when compared to Engel's work [4] , with a value of 0.5712 when N = 20 compared to 0.5834 on DDSP Additive Synthesis.

2.1.2.2 Neural Instrument Cloning (NIC)

Further attempts to evolve from the DDSP methods include Jonason and Sturm's Neural Instrument Cloning from Few Samples: from usually requiring ten minutes of data from the instrument of interest, their methodology resulted in needing only four to 256 seconds by leveraging speech voice cloning techniques while also incorporating additional information in the autoencoder [10]. One such piece of information is the F0 confidence of the pitch estimation, an additional piece of data obtained through CREPE. With a total of 12,548,993 parameters, their model architecture involves a trainable recording-specific embedding notated Z and an entirely new two-part reverb design mixed with a "wet/dry" knob. Unlike other papers where one-shot audio was used in training, raw solos were obtained through YouTube queries and compiled into their own dataset. Artifacts included the synthesized audio sounding as if it was recorded from afar, as well as echoes present in the audio, which may have been caused by the dataset. They found that their models using both F0-confidence conditioning and the two-part reverb obtained the lowest losses. While pre-training their model did not improve quality, it provided a much faster turnaround time

for cloning instruments. Much like DWTS, the overall result made for a more natural-sounding synthesis. They evaluated these different configurations by looking at the multi-scale spectral reconstruction loss of test excerpts. It was also found that F0-confidence reduced artifacts, and that the two-part reverb design became less impactful as the sample size from the dataset grew bigger.

2.1.2.3 Other Applications for DDSP

In recent years, DDSP has been retrofitted into even more methodologies in the academic field, as discussed in a catalogue and survey into the use of said techniques in sound and music synthesis by Hayes et al. [8]. 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.'s neural waveshaping synthesis method [7] 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 [14] implemented a more precise version by including controllable parameters such as cutoff frequency, a differentiable subtractive synthesizer, and more.

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 which applies 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. In the paper regarding its creation, Song et al. discuss how pitch jitters and discontinuous harmonics would commonly be found in GAN-based vocoders, leading them to add a DSP module. They found that the resulting model could "generate high-fidelity speech for various TTS models trained using diverse data" [18].

Despite advancements in DDSP-based generative models, Tone Transfer, a model developed by Carney et al. [2], remains the most easily accessible 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.

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 [1]. Common evaluation metrics of generative models are often driven by intuition and may overlook certain features of the product [3]– these limitations are also present in audio synthesis. Vinay and Lerch [19] cite various metrics that can be used for evaluating audio generative models following methodologies of previous studies, which will be discussed in the following sections. It should be noted as well, however, that Betzalel et al. [1] recommend using a diverse set of evaluation metrics when comparing generative models to control the variability in scores. As such, not only are objective evaluations necessary for evaluating models– subjective evaluations, such as those requiring human discrimination, are also necessary to ensure 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. delineate six (6) evaluation metrics used for implicit generative models, such as KL-divergence, IS, and Fréchet Inception Distance [1]. The paper further recommends dropping the use of IS in favor of 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. [12] by replacing the inherent Inception network within FID with the VGGish model. The developed technique, called Fréchet 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, FAD will be helpful in objectively evaluating the quality of produced audio from the three generative models.

Unfortunately, Fréchet 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. To address this, one could utilize another metric known as Number of Statistically Different Bins (NDB/k,) which was introduced by Richardson and Weiss to measure the diversity of generated audio from the original dataset [15]. 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 may gain a more comprehensive understanding of an audio generative model's strengths and weaknesses.

2.2.2 Subjective 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 a random sample from an audio pool on a 5-point Likert scale across a set of questions that pertain to the quality, loudness, and preference of the audio [9]. The MOS has been used as an evaluation metric for audio generative models such as WaveGAN, DSPGAN and generative neural speech models [18]. Tables 2.1, 2.2, and 2.3 show details of the rating criteria of the MOS in aspects of quality, effort, and loudness preference, respectively, according to the recommendations set by the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T) [9]. By aggregating scores from a diverse group of listeners, MOS provides a quantitative measure of how humans perceive the generated audio.

Table 2.1. Mean Opinion Score on the Listening-Quality Scale

Score	Quality of the Audio
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2.2. Mean Opinion Score on the Listening-Effort Scale

Score	Effort to Recognize the Instrument of the Audio
5	Complete relaxation possible; no effort required
4	Attention necessary; no appreciable effort required
3	Moderate effort required
2	Considerable effort required
1	No meaning understood with any feasible effort

Table 2.3. Mean Opinion Score on the Loudness-Preference Scale

Score	Loudness Preference
5	Much louder than preferred
4	Louder than preferred
3	Preferred
2	Quieter than preferred
1	Much quieter than preferred

To supplement the evaluation with human perception, Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) builds upon the Mean Opinion Score (MOS) by providing a more nuanced assessment suitable for comparing the perceptual quality of audio codecs, particularly those introducing distortions [7]. It introduces hidden reference and anchor audio samples, with the reference being the original, uncompressed audio providing a benchmark for the highest quality and the anchors being the manipulated versions with low quality to serve as a comparison point. One of the samples is intentionally manipulated to be at the lowest quality to provide a reference point. Listeners rate the perceived quality of each test sample compared to these references. This approach goes beyond MOS by establishing benchmarks and allowing for a more detailed analysis of strengths and weaknesses across different instruments and timbres in these generative models. Figure 2.1 exhibits webMUSHRA, a program specifically designed for

conducting MUSHRA tests [16], which allows the subject to listen to the reference and rate each sample in relation to the reference. Overall, combining these objective and subjective evaluations provides a well-rounded picture of the strengths and weaknesses of different audio generative models.

Figure 2.1. webMUSHRA Interface.

CHAPTER 3

METHODOLOGY

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 (DDSP) by Engel et al. [4],
2. Differentiable Wavetable Synthesis (DWTS) by Shan et al. [17], and
3. Neural Instrument Cloning From Very Few Samples (NIC) by Jonason and Sturm [10].

In the interest of feasibility, DDSP-based methods which used external components such as synthesizers or those with more focus on learning parameters were not chosen. The ones above also use similar audio features in order to generate synthesized audio. For consistency, the researchers will retrain each architecture (DDSP, DWTS, and NIC) using the recommended hyperparameters and training configurations from the respective papers.

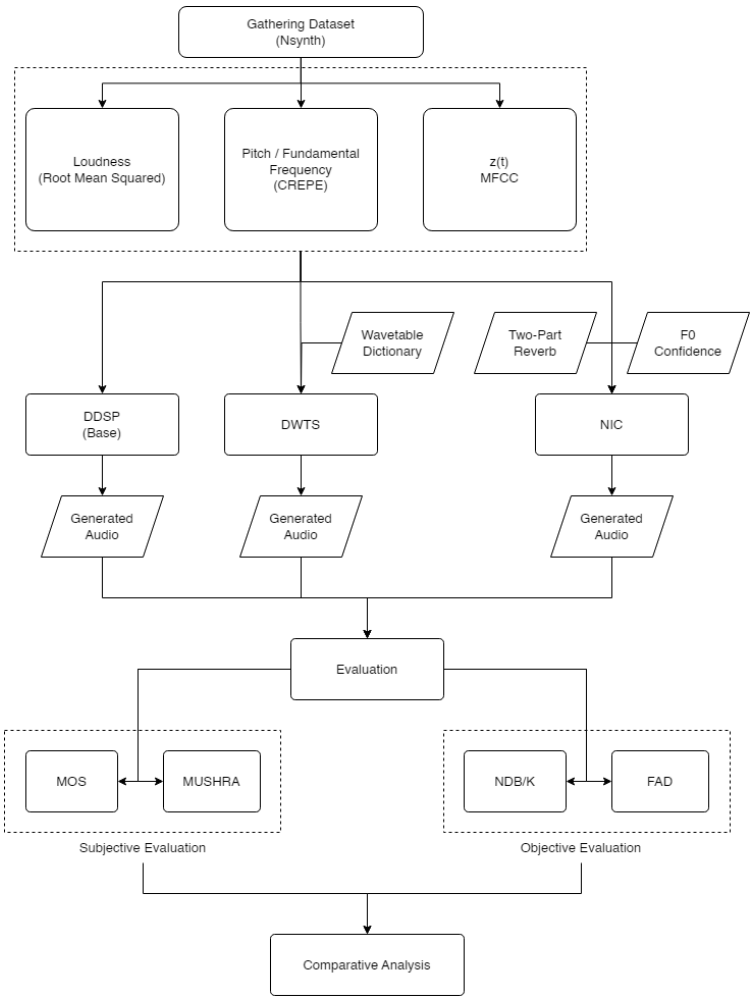

Figure 3.1. Flowchart of general methodology.

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.

In training the models, the features that will be used are the fundamental frequency, or F0. Pre-trained models of CREPE, which have proven to be accurate within a strict evaluation threshold of 10 cents [13], will be used for obtaining F0. Additionally, in order to analyze loudness, all of the papers above use an A-weighting of the power spectrum, which 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, taking F0 and loudness into account allows for providing the instantaneous fundamental frequency and intensity of a note sequence at a constant frame rate.

3.2.1 Audio Input Set-ups

In order to obtain a comprehensive set of synthesized audio from each of the studied models, a base set of audio inputs from various instruments will be procured. Based on this set, several set-ups differing in audio configuration will be used in order to observe how noise and other artifacts in audio input affect the synthesis process of the tested methods:

1. The control set-up, which is the unaltered base set,
2. The noise set-up, which will consist of instances from the base set with a variable amount of white noise added,
3. The mask set-up, which will contain audio that has been passed through a notch filter set to various bandwidths, removing certain parts of the frequency, and
4. The fragmented set-up, which will consist of audio interspersed with variable lengths of silence.

Samples from these set-ups will be passed through the trained models and the resulting synthesized audio will be subject to the mentioned evaluation metrics.

Another set of audio inputs will be obtained as well, consisting of instruments that were not present in the NSynth dataset– this will be used in order to determine how the models handle

audio inputs from instruments that are not included in the training data. The resulting synthesized audio from this set will also be put through the evaluation metrics.

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. [17] that the succeeding formula,

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

would work better (in terms of training stability) than the one made by Engel et al. where

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

$$L_{reconstruction} = \sum_i L_i.$$

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). It was found that this change did not affect the quality of synthesized audio.

3.3.2 Dataset Use

While the NSynth dataset provides a vast amount of data, subsequent studies utilizing the dataset for GANSynth and DDSP used smaller subsets totalling "70,379 examples composed mostly of strings, brass, woodwinds and mallets with pitch labels within MIDI pitch range 24-84" [3-4, 6]. A notable factor to consider between the three papers is that from DDSP to DWTS to NIC, the sample size for the inputs shrink, i.e. training data for NIC is sourced from mere YouTube

queries. Like these papers, our methodology will be limited to the same families considering the training size and availability of resources to train the model.

3.3.3 Model-Specific Implementations

3.3.3.1 DDSP

The supervised variant from the original DDSP paper is to be followed.

3.3.3.2 DWTS

Since DWTS includes learnable wavetables, it is represented as a learnable dictionary $D = \{w_i\}^{N_i}$. Using gradient descent, D is learned during training. A phase accumulator will be used for synthesis with an input sequence of time-varying fundamental frequency $f_0(n)$ over time steps n, through the formulas used in the original paper.

3.3.3.3 NIC

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 as well. This design models the room impulse response (IR) as a combination of two components.

3.4 Evaluation

To comprehensively assess the performance of the compared DDSP-based audio synthesis methods (DDSP, DWTS, NIC) and the proposed hybrid model with GAN architecture, we will employ a combination of objective and subjective evaluation metrics.

3.4.1 Objective Evaluation

We will utilize objective metrics to assess the audio quality and diversity of the generated outputs. The Fréchet Audio Distance (FAD) will measure audio quality by detecting distortions in the generated audio compared to a clean reference set. This allows us to directly compare the three existing methods and the new model's ability to recreate high-fidelity audio. A model with a lower FAD score indicates fewer distortions and potentially higher audio quality [12].

Multivariate Gaussians are computed on the evaluation set embeddings $N_e(\mu_e, \Sigma_e)$ sourced from an enhanced evaluation set (in this case, the synthesized audio) and a VGGish model consisting of 96 consecutive frames of 64 dimensional log-mel features extracted from the audio's magnitude spectrogram. From these embeddings, the FAD is computed as

$$F(N_b, N_e) = ||\mu_b - \mu_e||^2 + tr(\Sigma_b + \Sigma_e - 2\sqrt{\Sigma_b \Sigma_e}),$$

where tr is the trace of a matrix.

Furthermore, the Number of Statistically Different Bins (NDB/k) metric will be used alongside FAD to assess the diversity of the generated audio. This metric identifies models that go beyond simply replicating the training data by analyzing the statistical uniqueness of the generated audio samples. Ideally, a high NDB/k score alongside a good FAD score would indicate a model capable of generating high-fidelity audio with a variety of unique timbres and characteristics [14].

3.4.2 Subjective Evaluation

To incorporate human perception into the evaluation and assess aspects beyond objective metrics, we will conduct subjective evaluations using both the Mean Opinion Score (MOS) and the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test.

The MOS evaluation will involve participants rating the generated audio samples from all models on a standard Likert scale based on predefined questions about perceptual quality [9]. By aggregating scores from a diverse listening group, MOS provides a quantitative measure of how natural and high-quality the generated audio is perceived by humans. Comparing MOS scores across the models will help us identify which method, including the new hybrid model, produces audio that is subjectively perceived as the most natural-sounding.

For a more nuanced human evaluation, we will also employ the MUSHRA test alongside MOS. MUSHRA builds upon MOS by introducing hidden reference and anchor audio samples. Listeners will rate the perceived quality of each generated audio sample in relation to these references. This approach offers several advantages. It establishes benchmarks for listeners, allowing for more detailed and comparative judgments of audio quality across different instruments and timbres within the models. By identifying how well each model performs in replicating specific instruments or timbres compared to the reference and anchors, MUSHRA can guide the development of the new model by focusing on areas where improvement is needed to achieve a wider range of high-fidelity and natural-sounding audio generation [7].

The MUSHRA evaluations may be done using webMUSHRA, a web framework developed in PHP by Schoeffler et al [16]. A web-based MUSHRA listening test, webMUSHRA will allow for the ITU-R BS.1534 listening method as it allows for easy configuration of comparisons between reference and target audio samples. Like in an ITU-R BS.1116 test, listeners

can instantaneously switch between the reference and the conditions when listening, and the order is randomized.

The combined application of FAD, NDB/*k*, MOS, and MUSHRA evaluations will provide a thorough analysis of the strengths and weaknesses of each audio synthesis method. By objectively measuring audio quality, diversity, and subjectively assessing human perception across various timbres, this evaluation will not only distinguish the most effective method among the existing DDSP approaches but also guide the further development of the proposed hybrid model. This will ensure the new model achieves its goal of generating high-fidelity and perceptually natural-sounding audio that surpasses the capabilities of the compared methods.

BIBLIOGRAPHY

[1] Betzalel, E., Penso, C., Navon, A., and Fetaya, E. 2022. A Study on the Evaluation of Generative Models. (June 2022). Retrieved from http://arxiv.org/abs/2206.10935

[2] Carney, M., Li, C., Toh, E., Zada, N., Yu, P., and Engel, J. 2021. Tone Transfer: In-Browser Interactive Neural Audio Synthesis. Retrieved from http://ceur-ws.org

[3] Engel, J., Agrawal, K.K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. 2019. GANSynth: Adversarial Neural Audio Synthesis. (February 2019). Retrieved from http://arxiv.org/abs/1902.08710

[4] Engel, J., Hantrakul, L., Gu, C., and Roberts, A. 2020. DDSP: Differentiable Digital Signal Processing. Retrieved April 1, 2024 from http://arxiv.org/abs/2001.04643

[5] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. Retrieved from https://arxiv.org/abs/1704.01279

[6] Hantrakul, L., Engel, J., Roberts, A., and Gu, C. 2019. Fast and Flexible Neural Audio Synthesis. (2019). Retrieved from https://archives.ismir.net/ismir2019/paper/000063.pdf

[7] Hayes, B., Saitis, C., and Fazekas, G. 2021. Neural Waveshaping Synthesis. (July 2021). Retrieved from http://arxiv.org/abs/2107.05050

[8] Hayes, B., Shier, J., Fazekas, G., McPherson, A., and Saitis, C. 2023. A Review of Differentiable Digital Signal Processing for Music & Speech Synthesis. (August 2023). Retrieved from http://arxiv.org/abs/2308.15422

[9] ITU-T P. 800. Methods for Subjective Determination of Transmission Quality. (1996). Retrieved from https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.800-199608-I!!PDF-E&type=items

[10] Jonason, N. and Sturm, B.L.T. 2022. Neural Music Instrument Cloning From Few Samples. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-326017

[11] Karras, T., Aila, T., Laine, S., and Lehtinen, J. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. (October 2017). Retrieved from http://arxiv.org/abs/1710.10196

[12] Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. 2018. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. (December 2018). Retrieved from http://arxiv.org/abs/1812.08466

[13] Kim, J.W., Salamon, J., Li, P., and Bello, J.P. 2018. CREPE: A Convolutional Representation for Pitch Estimation. Retrieved April 1, 2024 from http://arxiv.org/abs/1802.06182

[14] Masuda, N. and Saito, D. 2021. Synthesizer Sound Matching With Differentiable DSP. Retrieved from https://archives.ismir.net/ismir2021/paper/000053.pdf

[15] Richardson, E. and Weiss, Y. 2018. On GANs and GMMs. Retrieved from https://github.com/eitanrich/gans-n-gmms

[16] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J. 2018. webMUSHRA — A Comprehensive Framework for Web-based Listening Tests. *JORS* 6, 1 (February 2018), 8. https://doi.org/10.5334/jors.187

[17] Shan, S., Hantrakul, L., Chen, J., Avent, M., and Trevelyan, D. 2021. Differentiable Wavetable Synthesis. (November 2021). Retrieved from http://arxiv.org/abs/2111.10003

[18] Song, K., Zhang, Y., Lei, Y., Cong, J., Li, H., Xie, L., He, G., and Bai, J. 2022. DSPGAN: a GAN-based universal vocoder for high-fidelity TTS by time-frequency domain supervision from DSP. (November 2022). Retrieved from http://arxiv.org/abs/2211.01087

[19] Vinay, A. and Lerch, A. 2022. Evaluating generative audio systems and their metrics. (August 2022). Retrieved from http://arxiv.org/abs/2209.00130