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 [smth].

In the past decades, the dominance of synthesizers in 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 Machine Learning 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 aims to 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

I will not Paragraph Muna:

- This study is focused on the specific implementations of:

    - Tone transfer architecture as delineated in [Synthesis] FM Tone Transfer with Envelope Learning.pdf (Caspe et. al, 2023)
    - A Generative Adversarial Network architecture as specified in [Synthesis] GanSynth - Adversarial Neural Audio Synthesis.pdf (Engel et. al, 2019)
    - Variational Autoencoders architecture as specified in [MAY STUDY BA].
- The Audio Input is further specified as:
    - Any melodic recording of an instrument.
    - TENTATIVE Instrumental recordings will use the following instruments:
        - Violin
        - Guitar
        - Flute
        - Trombone
        - Organ
        - Marimba
    - Quality of audio varies to determine the stability of each architecture.
    - Note Qualities also vary to determine which features of each architecture affect the audio output
- The dataset to be used will be the NSynth Dataset consisting of 305,979 musical notes with unique pitch, timbre and envelope played from 1,006 instruments.
- Metrics to be used to determine the accuracy of the output from the input are:
    - Human Evaluation (tentative)
    - Inception Score (IS)
    - Pitch Accuracy (PA) and Pitch Entropy (PE)
    - Frechet Inception Distance (FID)
    - Number of Statistically-Different Bins (NDB)
- Limitations include
    - There is no single metric that could effectively determine the accuracy of the models. As such these scores may not capture all aspects of accuracy within tonal replication.
        - Human evaluation to determine output quality can be subjective and produce unexpected or varied results

        - Inception Score does not capture perceptual accuracy
        - Pitch Metrics ignores timbre and nuanced musical characteristics
        - NDB ignores temporal aspects and focuses on spectral content
    - Results may be overfit to the NSynth Dataset (tbd by the models itself)
        - Unsure cause replication nga dba so dat magoverfit talaga 😕
    - This study focuses on the comparison of specific implementations of architecture in each neural network and does not delve into deeper and more advanced techniques such as unique model and architecture creation.

1.5 Significance of Study

In the intersection of computer science and the area of computational musicology, this study aims to further the field of music sound synthesis using machine learning by figuring out how

CHAPTER 2

Review of Related Literature

Important?

- https://research.facebook.com/publications/sing-symbol-to-instrument-neural-generator/
- https://dl.acm.org/doi/pdf/10.1145/3616195.3616196
- https://openreview.net/attachment?id=B1x1ma4tDr&name=original_pdf
- https://openreview.net/pdf?id=H1xQVn09FX ← Metrics

Possible sections:

- Previous models
- CNNs, RNNs
- Explanation for FM
- Methods of evaluation

U think for this part we should just say the study and what they are for muna?

CHAPTER 3

Methodology

- Nsynth dataset
- Envelope Learning:
    - Instead of using oscillators from generated patches, we could use:
        - Chroma Features?
    - Model Inference (3.3)
        - Feature extraction....
            - **WHERE THE FUCK IS THIS IN THE CODE??????**

(2) Control Prediction, we use our neural network g_ϕ to infer a set of frame-wise FM synthesis controls, the oscillator output levels $\hat{ol}_k$, from the conditioning signals $\hat{a_k}$ and $\hat{f_k}$.

$$\hat{ol}_k = g_\phi(\hat{a_k}, \hat{f_k}) \quad (3)$$

-
            - This is from Tone Transfer with Envelope Learning: however, it uses **generated synth patches** while we are working with raw data...
            - Possible alternative:
                - a^k <- loudness
                    - A^k could be a tuple in itself (or is it called a latent vector)
                    - Ex. overall RMS first value
                    - Other values: 0-500hz, 500-2000hz, etc. GETS BA AWESOME LOW LEVEL MID LEVEL HIGH LEVEL BASS MID TREBLE GANON
                - f^k <- f0
                - Extra parameter: velocity? (is that relevant)

(3) An FM oscillator bank $S_p(.)$ renders a window of N audio samples from output levels $\hat{ol}_k$, fundamental frequency f_{0k}. We configure the bank with the oscillator routing and frequency ratios of the patch p used to train g_ϕ, although this can be changed during inference.

$$s_{Nk}, ...s_{N(k+1)} = S_p(\hat{ol}_k, f_{0k}) \quad (4)$$

-
            - Where DDSP synthesizer can be used?

- Harmonic synthesizer ← implemented by DDSP
    - Idk

We use Pytorch as a training framework. The process takes about four hours per model using a single NVIDIA GeForce RTX 2080 Ti GPU.

Mas maayos na methodology

1. Dataset ← **NSynth → Training Tuples**
    a. F0 ← Fundamental frequency using CREPE
    b. A-weighted Log Amplitude Loudness
    c. Getting “pseudo-envelope” through Hilbert transform (di ko pa sure to)
    d. Z-encoder: using CQT ← constant q transform
2. Data Processing
    a. GAN to refine pitch / F0 for micro-pitch variations???????? (ewan din kasi ginagamit sa vocoder)
    b. Essentially same as DDSP
    c. So many different DDSP variations...
        i. Waveshaping (prediction)
        ii. Subtractive synthesizers...
        iii. FM tone transfer
        iv. Wavetables
3. Metrics
    a. Subjective:
        i. MOS (Mean Opinion Score) ← Rating 1-5
        ii. ABX Testing
    b. Quantitative:
        i. Frechet Audio Distance
        ii. Nearest Neighbor Comparison
        iii. Ewan ko pa pero meron dito https://pypi.org/project/Audio-Similarity/