A Study on Audio Input Representations on GANSynth MIDI Synthesizer for Bumbong Bamboo Instrument

Undergraduate Student Project

by

Ryle Emmanuel Raagas
2015-05329
B.S. Electronics and Communications Engineering

Advisers:

Crisron Rudolf Lucas
Jose Marie Mendoza
Carl Timothy Tolentino, PhD

University of the Philippines, Diliman
July 2023

UNIVERSITY OF THE PHILIPPINES

Bachelor of Science in Electronics and Communications Engineering

Ryle Emmanuel Raagas

A Study on Audio Input Representations on GANSynth MIDI Synthesizer for Bumbong Bamboo Instrument

Undergraduate Project Advisers:

Crisron Rudolf Lucas
Carl Timothy Tolentino, PhD
Jose Marie Mendoza

Electrical and Electronics Engineering Institute
University of the Philippines Diliman

Undergraduate Project Reader:

Alberto De Villa

Electrical and Electronics Engineering Institute
University of the Philippines Diliman

Date of Submission

July 2023

Permission is given for the following people to have access to this thesis:

Circle one or more concerns: I P C	
Available to the general public	Yes/No
Available only after consultation with author/thesis adviser	Yes/No
Available only to those bound by confidentiality agreement	Yes/No

Students' signature/s: 

Signature/s of undergraduate project advisers:

Approval Sheet

In partial fulfillment of the requirements for the degree of BS Electronics and Communications Engineering, this project entitled "A Study on Audio Input Representations on GANSynth MIDI Synthesizer for Bumbong Bamboo Instrument", prepared and submitted by Ryle Emmanuel Raagas, is hereby recommended for approval.

Crisron Rudolf Lucas	Date
Adviser	

Carl Timothy Tolentino	Date
Adviser	

Jose Marie Mendoza	Date
Adviser	

Accepted in partial fulfillment of the requirements for the degree of BS Electronics and Communications Engineering.

Alberto De Villa	Date
Panel Member	

Lew Andrew Tria	Date
Director, Electrical and Electronics Engineering Institute	

University Permission Page

I hereby grant the University of the Philippines non-exclusive worldwide, royalty-free license to reproduce, publish, and public distribute copies of this work in any form subject to the provisions of applicable laws, the provisions of the UP IPR policy and any contractual obligations, as well as more specific permission marking on the Title Page.

Specifically I grant the following rights to the University:

- to upload a copy of the work in the theses database of the college/school/institute/department and in any other databases available on the public internet;
- to publish the work in the college/school/institute/department journal, both in print and electronic or digital format and online; and
- to give open access to above-mentioned work, thus allowing “fair-use” of the work in accordance with the provisions of the Intellectual Property Code of the Philippines (Republic Act No. 8293), especially for teaching, scholarly, and research purposes.

Ryle Emmanuel Ragas

02 July 2023

Date

Acknowledgment

Finally, the wait is over. Ito na 'yun.

Sa loob ng walong taon, naipon ang magkahalong luha, puyat, pagod, saya, pawis, anxiety, at marami pang iba.

Sa mga Kisay friends ko (special mention: Angelooo, Theaaa, Remby, Miah, Erica, MM, Darluh, Alexia, Laurent, Nanami, GeAnne), maraming salamat sa pagsama sa akin sa lungkot at saya. Wala akong orgs pero naging sapat kayo para sa akin. Hindi ko naramdamang malungkot ako kahit lagi akong nag-iisa.

Sa mga naging UPD friends ko (special mention: Claire B, Kath, DaFi, Alex M, Bennet, Karl, Hannah, Ken, Cheska), maraming salamat! Kahit minsan lamang tayo nagkasama o mag-usap ay naappreciate ko yun lalo na sa mga group study, library hangouts, Discord kamustahan, atbp. Hindi ko kayo makakalimutan.

Sa mga naging customers ko sa aking munting business na KazutechPH, maraming salamat sa pagsuporta ninyo. Naging malaking bagay kayo para suportahan yung pag-aaral ko since 2017 hanggang sa ngayon. See you around, KazutechFam!

Sa mga nagdulot ng hinanakit sa puso ko at sa mga hindi naniwala, maraming salamat pa rin. Naging motivation ko kayo kahit papaano at masasabi kong kinaya ko itong lahat na hindi kayo kasama.

Sa mga naging professors at instructors ko, maraming salamat. Marami akong natutunan regardless kung anong grade ang naibigay ninyo sa akin. Malaking karangalan na makapag-enroll sa mga classes ninyo.

Sa mga advisers ko na sila Sir Crisron, Sir Carl, at Sir Jomari, maraming salamat sa pagtitiis ninyo sa akin hahaha. Maraming salamat sa pagpapatuloy ninyo sa akin sa DSP Family.

Sa mga magulang at mga kapatid ko, maraming salamat sa suporta at pagtitiwala na makakapagtapos ako. Mahal na mahal ko kayo.

Lastly, salamat din kay Lord. Natupad na rin yung wish ko.

Ryle Emmanuel Bayan Raagas (2015-05329)

Abstract

A Study on Audio Input Representations on GANSynth MIDI Synthesizer for Bumbong Bamboo Instrument

This paper introduces audio synthesis for Bumbong bamboo instrument using neural network particularly generative adversarial network (GAN). The Bumbong is a bamboo aerophone instrument composed of different length bamboo poles originated in the Philippines. This project implemented GANSynth audio synthesis with MEL and Constant-Q Transform (CQT) spectrograms as input representations with preprocessing through VAE-GAN timbre transfer. The MEL spectrogram was used to accommodate mel scale which is closer to human hearing perception while CQT spectrograms could create higher evaluation accuracy based on multi-class classification. The input representations trained separately on GANSynth shows that the synthesized Bumbong audio on the MEL group have a 6.08 percent better subjective test and 2.435 better difference in Frechet Audio Distance score than the CQT group.

Contents

List of Figures iii

List of Tables iv

1 Introduction 1
1.1 Documentation Flow and Organization . . . 3

2 Related Works 4
2.1 Neural Networks . . . 4
2.2 Sound Synthesis of Indigenous Instruments . . . 5
2.3 Input Representations in Neural Networks . . . 6
2.3.1 Constant-Q Transform (CQT) Spectrograms . . . 7
2.3.2 MEL Spectrograms . . . 7
2.4 Timbre Transfer Pipeline . . . 8

3 Problem Statement and Objectives 10
3.1 Problem Statement . . . 10
3.2 Objectives . . . 10
3.3 Scope & Limitations . . . 11

4 Methodology 12
4.1 Audio Data Collection . . . 12
4.2 Data Preprocessing . . . 14
4.3 Model Training and Validation . . . 14
4.4 Evaluation & Analysis . . . 16
4.4.1 Objective Test (FAD) . . . 16
4.4.2 Subjective Test (Mean Opinion Score) . . . 16

5 Results and Discussion 18
5.1 GANSynth Output . . . 18
5.2 Evaluations . . . 18
5.2.1 Fréchet Audio Distance (FAD) . . . 20
5.2.2 Subjective Test . . . 20

6 Conclusions and Recommendations 22
6.1 Conclusions . . . 22

6.2 Recommendations 23
6.2.1 Computer Specifications 23
6.2.2 Future Work 23

Bibliography **25**

List of Figures

1.1 Bumbong Play Tutorial 2

2.1 CQT Formula 7
2.2 PCG Signal Performance Evaluation on Input Representations 8
2.3 VAE-GAN Model 9

4.1 Methodology Flowchart 13
4.2 A Bumbong Bamboo Instrument and Sample Note Waveform (C4) 13
4.3 Preprocessing Flow 14
4.4 Progressive GAN Training Model 15
4.5 GANSynth Training 15

5.1 Spectrogram View of Audio Tracks 19
5.2 FAD Formula 20

List of Tables

2.1 NSynth and GANSynth Comparison 5
2.2 Evaluation Results of Similar Projects 6
2.3 GANSynth Evaluation Results 7

4.1 Bumbong Dataset Information 12
4.2 MIDI File Details 16

5.1 FAD Results 20
5.2 Subjective Test Summary 21

6.1 PC Specifications Summary 23

Chapter 1

Introduction

In this time of COVID-19 pandemic in the Philippines, the music industry was hardly impacted as community quarantine was implemented. The concerts and choirs are prohibited as number of people that perform together are highly limited. Musicians do also have a problem in borrowing instruments as it may concern hygiene-related reasons especially on aerophones such as flutes and trumpets. In the Philippines, an aerophone called Bumbong is a bamboo instrument made from the bamboo culm and represent a different note depending on the culm size [6]. The Bumbong player must hold the instrument on face level and blow wind on its mouth as shown in Figure 1.1. This tutorial can be accessed via Pacita M. Narzo's YouTube Channel.

The current pandemic also boosted the online activity of people to able to stay connected for school, work, and various reasons. This project will help mainly on the promoting and preservation of an indigenous instrument such as Bumbong using modern methods of audio synthesis with help of neural networks and later on with MIDI implementation. There are currently various efforts in the preservation of indigenous instruments including collecting audio data and DOST's Bamboo Musical Instruments Innovation Research & Development program improving production of Bumbong to make it resistant to insect attacks, stains, and weather conditions. [8] These efforts can help increase the listener's appreciation of indigenous instruments and hopefully to be recognized more in the Philippines.

In this project, neural networks will be implemented to produce digital signals based on instrument audio recordings. Indigenous instruments such as the Bumbong do not have MIDI ports that common instruments such as electric pianos and guitars have. Neural networks may help generate digital signals based on the Bumbong's audio database. This project will use GANSynth neural network architecture for training which were used to produce high fidelity audio with the exploration of two (2) input representations for improving audio accuracy and human perception.

Figure 1.1: Bumbong Play Tutorial

1.1 Documentation Flow and Organization

The flow of the project shows that Chapter 2 discusses the related works of other neural network-based audio syntheses. Chapter 3 discusses about the problem statement, objectives, scope, and limitations of the project. Chapter 4 discusses the methodology of the project. Chapter 5 includes the results and discussion of the project. Chapter 6 includes the conclusion and recommendations.

Chapter 2

Related Works

2.1 Neural Networks

In [17]. Natsiou and O'Leary indicated that researchers have withdrawn from using classic signal processing methods for sound generation and is now focusing on deep learning which tackles neural networks. Choosing a neural network architecture along with the input representation can be difficult as it may depend on the audio such as instruments and human speech.

Neural Synthesizer (NSynth) and Generative Adversarial Networks Synthesizer (GANSynth) are both collaboration work of Google Brain team and DeepMind using TensorFlow Python module [1, 2, 3]. A NSynth dataset containing a large collection of annotated musical notes sampled from each of around 1000 instruments was produced to explicitly factorize the generation of music into notes and other qualities such as timbre and dynamics.

NSynth is based on WaveNet-style autoencoders that conditions an autoregressive decoder on temporal codes on trained raw audio waveforms that demonstrates a well-tuned spectral autoencoder baseline [2]. GANSynth uses the NSynth database with the difference of using Progressive GAN as its neural network model. GANs are known for generating high-quality images but was considered to generate high-fidelity audio in parallel and is significantly faster than standard WaveNet by around 50,000 times on a modern graphics card (GPU).

GANSynth was developed in 2019 while NSynth was in 2017. GANSynth is still under consideration as audio artifacts are more present than NSynth but its researchers mentioned that GANSynth can still be improved and can open up avenues for domain transfer and other applications.

In this project, GANSynth is chosen over NSynth as it significantly needs less training time and lower computer hardware specifications as shown on Table 2.1. The listed graphics card

Neural Network	Graphics Cards Used	Training Duration Approx.
NSynth (WaveNet)	10x NVIDIA Tesla K40 (Synchronous)	10 Days
GANSynth (Prog. GAN)	1x NVIDIA Tesla V100	~3-4 Days

Table 2.1: NSynth and GANSynth Comparison

models are not the listed minimum requirement but can be used for roughly estimating training time when using a different graphics card model. The listed training durations are based on the event that both NSynth and GANSynth default setups are reproduced. The training duration depends mainly on the scale of the audio datasets to be trained and of the graphics card (GPU) used.

2.2 Sound Synthesis of Indigenous Instruments

In [4], Laguna, Valdez, and Guevara, used an instrument called Kulintang, typically made up of five to seven gongs and is played by striking with a stick, for MIDI Implementation using the VST2.4 standard. They have used the Kulintang to give musicians access to the sounds of an indigenous Philippine instrument. The project produced a Virtual Studio Technology (VST) instrument plugin with twenty-three (23) tuning presets with eight (8) gongs each that can be controlled by a Digital Audio Workstation (DAW) which is a common tool used today. Similar to [9] Bagaforo and Gayo, indigenous intruments, Tongali and Kolintong, were used for digital audio synthesis as a VST plugin. In [11], Castillo, and Domingo also implemented a VSTi plugin for Bamboo instruments such as Angklung, Bumbong, and Marimba and were synthesized using Sum of Sinusoids and Digital Waveguide which were both not using neural networks.

In [5], Sunjankhom, Chivapreecha, Chitanont, and Kato produced neural network-based sound synthesis for a Thai Duct Flute instrument called Khlui. Khlui is a reedless wooden wind instrument that is played by blowing air into its mouthpiece. The neural networks used for the sound synthesis are combined Multilayer perceptron (MLP) and Recurrent Neural Network (RNN). The audio data used was sampled at 16kHz and was extracted into harmonic and stochastic signals for Spectral Modeling Synthesis (SMS). The output sound of the trained model were compared to the original recorded sound and were then classified as natural and realistic by cross correlation function.

In Table 2.2, the evaluation method and results are shown for discussed papers. The subjective evaluation method such as the Two-Interval Forced Choice (2IFC) is considered as it requires human hearing tests that can perceive better of the irregularities or difference between the

Instruments Used	Eval. Methods	Evaluation Results
Kulintang [5]	None	N/A
Khlui [6]	Cross-correlation	Strong Correlation (0.9397-0.9907)
Tongali, Kolintong [10]	2IFC, FFT	2IFC (43%-53% correct), FFT (below 30% error)
Angklung, etc. [12]	2IFC	62.5% up to 93.7% correct on 3 instruments

Table 2.2: Evaluation Results of Similar Projects

synthesized and raw audio. In this project, the objective tests to be used are considered depending on the synthesis method used. The GANSynth project used objective tests that are specialized in evaluating GAN-based neural networks.

2.3 Input Representations in Neural Networks

In [16]. Natsiou and O'Leary indicated that numerous input representations have proven beneficial for audio synthesis. Input representations can be raw audio, spectrograms, acoustic features, and more. Using raw audio as an input representation may take longer training time and may miss some sound features that only spectrograms can represent by using Short Time Fourier Transform (STFT). In this project, spectrograms will be used as input representations which are usually used as inputs to neural network systems and are usually tailored depending on the data used such as images and audio [12]. According to [13], choosing the appropriate input representation (spectrograms) can increase transcription accuracy by 8.33% and reduction in error by 9.39%. Although some neural network models such as WaveNet can input raw audio data, the use of input representations is more preferred as it is also common in projects related but not limited to speech recognition and music transcription.

In [1], GANSynth used Instantaneuos Frequency (IF) + MEL + High Frequency (IF-MEL + H) for the representation of its input data as produced the best win in human evaluation and Number of Statistically-Different Bins (**NDB**) tests as shown in Table 2.3. **FID** stands for Fréchet Inception Distance, **IS** for Inception Score, **PA** for Pitch Accuracy, and **PE** for Pitch Entropy. The **bolded data** on Table 2.3 represents the best result for each category or column except for the real data row which is the best result of mostly all example rows.

There are various input representations available such as different variations of Short-time Fourier Transform (STFT), Mel Frequency Cepstral Coefficients (MFCC), and Constant-Q Transform (CQT). In GANSynth [1], they have tried to use MEL & Phase input representations but not of any Constant-Q Transforms.

Examples	Human Evaluation	NDB	FID	IS	PA	PE
Real Data	549	2.2	13	47.1	98.2	0.22
IF-MEL + H	**485**	**29.3**	167	38.1	97.9	0.40
IF + H	308	36.0	**104**	**41.6**	**98.3**	**0.32**
Phase + H	225	37.6	592	36.2	97.6	0.44
IF-MEL	479	37.0	600	29.6	94.1	0.63
IF	283	37.0	708	36.3	96.8	0.44
Phase	203	41.4	687	24.4	94.4	0.77
WaveNet	359	45.9	320	29.1	92.7	0.70
WaveGAN	216	43.0	461	13.7	82.7	1.40

Table 2.3: GANSynth Evaluation Results

$$X^{cq}[k_{cq}, \tau] = \sum_{n=0}^{N_{k_{cq}}-1} x[n + h\tau] \cdot e^{-2\pi i Q \frac{n}{N_{k_{cq}}}}$$

Figure 2.1: CQT Formula

2.3.1 Constant-Q Transform (CQT) Spectrograms

Constant-Q Transform (CQT) is a modified version of Short Term Fourier Transform (STFT) with the notable difference using factor Q instead of k in STFT [13]. In [14], Tiwari, Jain, Sharma, and Almustafa used a variation of Constant-Q Transform (CQT), hybrid constant-Q transform (HCQT), for a phonocardiogram (PCG) signal. This signal represents murmurs and sounds from the cardiovascular system and has reached 96% accuracy in multi-class classification compared to other input representations (MFCC, VQT, CQT) using the Convolutional Neural Network (CNN) architecture as shown in Figure 2.2.

2.3.2 MEL Spectrograms

MEL Spectrograms remaps the value in hertz to the mel logarithmic scale. The mel scale, derived from melody, is what human hearing generally perceives [17]. Humans can find it difficult to differentiate higher frequencies than lower frequencies. MEL Spectrograms are better suited to human hearing perception-related applications than linear audio spectrograms. The MEL Spectrograms can be found useful for instruments that humans likes to play and hear.

In [10], there was an increase in performance for identification of Aras bird sounds when

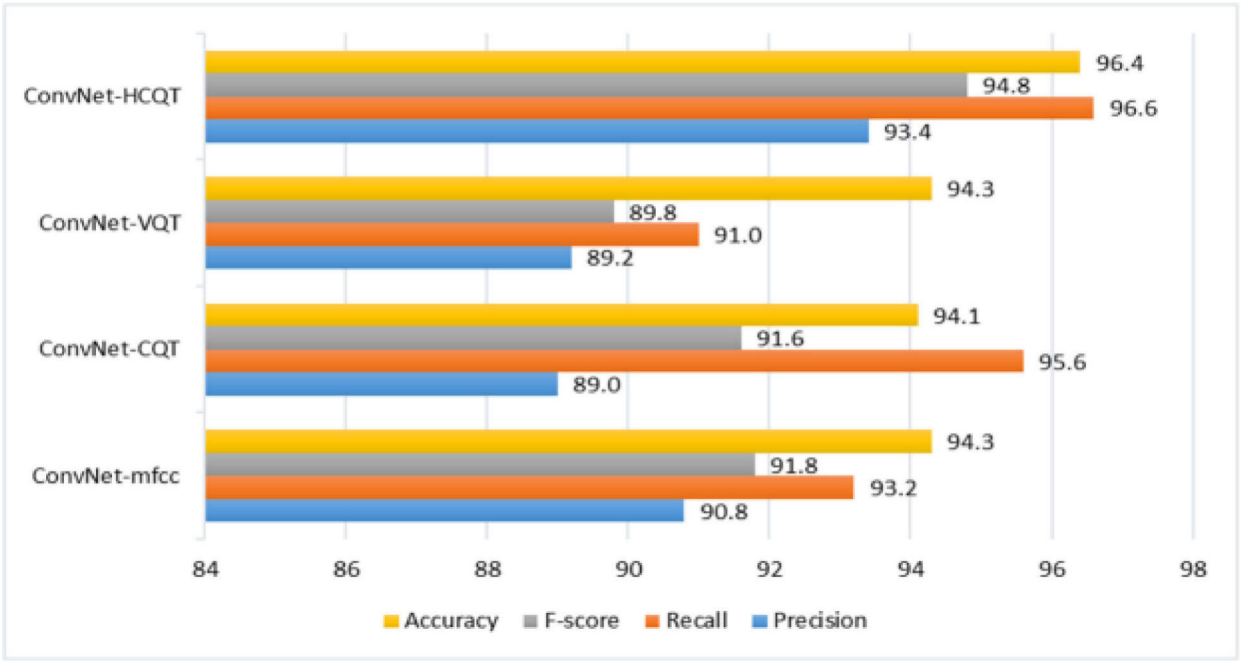

Figure 2.2: PCG Signal Performance Evaluation on Input Representations

using MEL spectrograms versus raw audio for training using CNN. Processing the raw audio to MEL spectrogram increased the model performance by 4%.

2.4 Timbre Transfer Pipeline

In [18], Cua, Sta. Maria, and Tumanut utilized Variational Autoencoder-Generative Adversarial Network (VAE-GAN) of Bonnici et al.[19] to perform timbre transfer of stringed bass instruments. VAE-GAN's timbre transfer converted source input audio to MEL spectrograms, trained using a Variational Autoencoder (VAE), infered the trained model to a playable audio file (.wav format) using Griffin Lim algorithm, and then trained using WaveNet.

In [18], VAE-GAN was modified to cater MEL and Constant-Q Transform (CQT) spectrograms for the timbre transfer pipeline. The VAE-GAN model (single path) is shown in Figure 2.3.

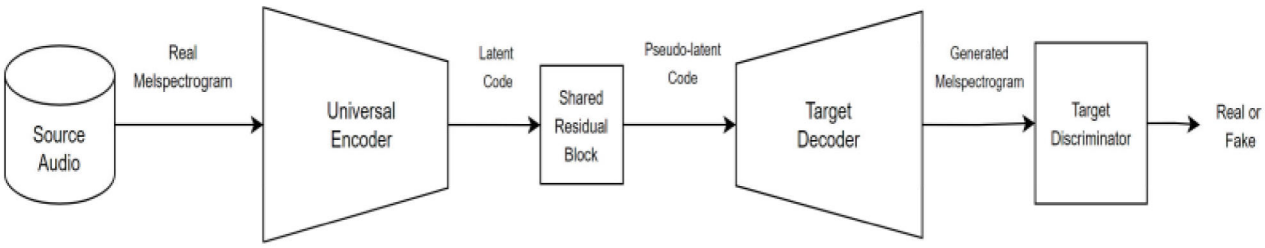

Figure 2.3: VAE-GAN Model

Chapter 3

Problem Statement and Objectives

3.1 Problem Statement

Indigenous Music in the Philippines is not common due to inaccessibility to the instruments and lack of recognition. Due to the COVID-19 pandemic, accessibility to instruments whether indigenous or common becomes more limited as performers can't perform freely on their respective homes or even physically come due to community quarantine restrictions. This project aims to use GANSynth Neural Network model, based on Progressive GAN known for producing state-of-the-art images, to synthesize Bumbong audio to generate high-fidelity audio with significantly less training time than of known models based on AutoRegressive (AR) models like WaveNet.

To improve GANSynth audio generation, this project will use and analyze two (2) input representation namely Constant-Q Transform (CQT) and MEL spectrograms. CQT Spectrograms are used to help generate an accurate synthesized audio on multi-class classification and MEL Spectrograms are used to improve human hearing perception of audio with respect to the mel scale.

This project also aims to be relevant to the current music industry where high fidelity audio is becoming popular and to also promote Bumbong as an instrument that interested people may try on using a MIDI file without initially needing to physically play the instrument.

3.2 Objectives

The main objective of this project is to synthesize audio of the Bumbong instrument using neural networks. For this project, Progressive GAN (GANSynth) will be used as the neural network model and CQT and MEL as input representations. The specific objectives are as follows:

1. Implement GANSynth audio synthesis with Constant-Q Transform (CQT) spectrograms for Bumbong bamboo instrument.

2. Implement GANSynth audio synthesis with MEL spectrograms for Bumbong bamboo instrument.

3. Evaluate the performance of the systems in terms of both subjective and objective tests.

3.3 Scope & Limitations

The project will only focus on evaluating two separate input representations (CQT and MEL Spectrograms) for the GANSynth audio synthesis. The limitations are listed below.

1. The provided Bumbong audio dataset are composed of audio recordings from three (3) different individuals with 1-2 Bumbong sets used. There are no physically available Bumbong instrument for recording.

2. Training GANSynth is recommended using any modern NVIDIA GPU with CUDA processing cores. The available GPU for training are of consumer-level only (e.g. NVIDIA GeForce) which would take longer training time compared to the data center-level GPU used by GANSynth (e.g. NVIDIA Tesla GPU). There are no listed minimum GPU requirement for GANSynth.

3. The computer to be used for training and evaluation has significantly lower specifications that the machines used by the GANSynth developers.

Chapter 4

Methodology

This project will be done in five main parts namely audio data collection, data preprocessing, model training and validation, evaluation and comparison, and results and discussion. The general flowchart is shown in Fugure 4.1 and each parts are discussed.

4.1 Audio Data Collection

The researcher has decided to choose Bumbong bamboo instrument for this project with Table 4.1 showing 3 recorded Bumbong players and its information. A sample Bumbong C4 note stereo recording with its spectrogram view as seen in Figure 4.2. The Bumbong audio data are acquired through Philippine Bamboo Musical Instruments by the Department of Science and Technology Forest Products Research and Development Institute (DOST - FPRDI) [6, 7]. This study only includes the legato part of the dataset as the GANSynth subset from NSynth is composed of legato recordings.

Player	Data Sample	Composition
Calabig	58 Recordings	Various notes A2 to G4
Ramos	120 Recordings	Various notes A3 to G5
Toledo	62 Recordings	Various notes A2 to G#4

Table 4.1: Bumbong Dataset Information

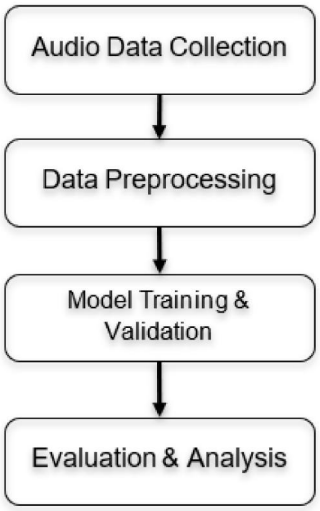

Figure 4.1: Methodology Flowchart

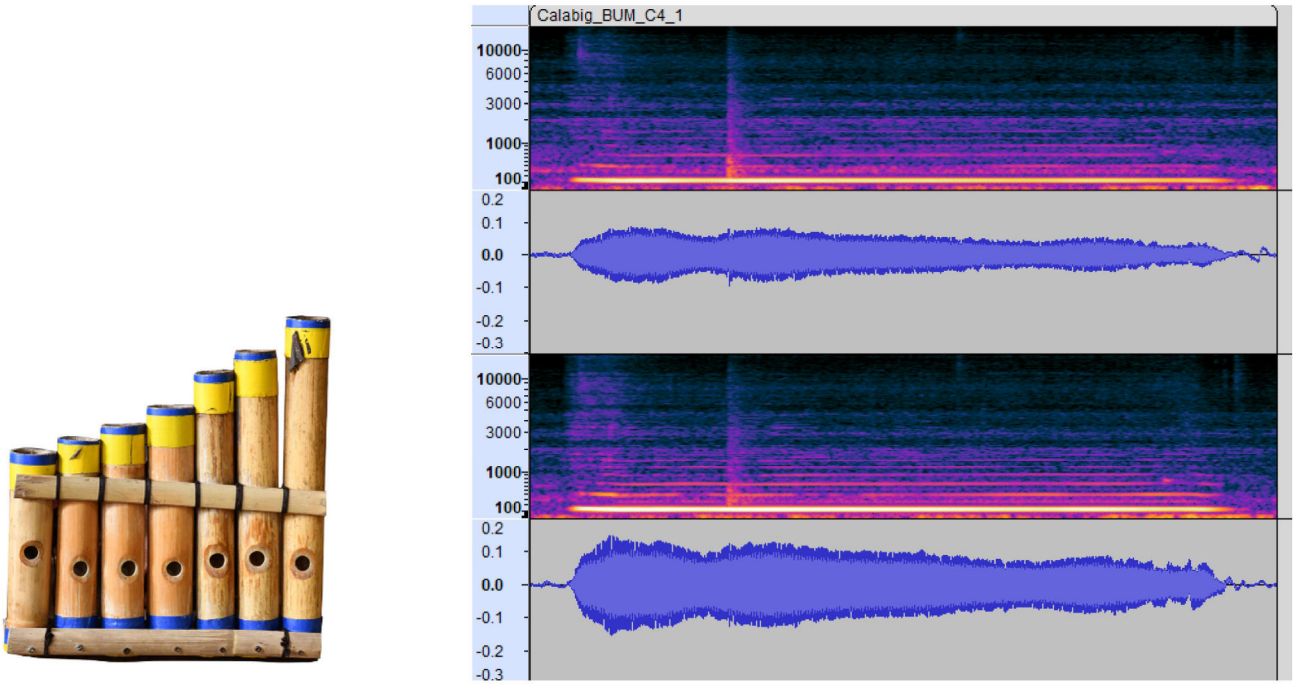

Figure 4.2: A Bumbong Bamboo Instrument and Sample Note Waveform (C4)

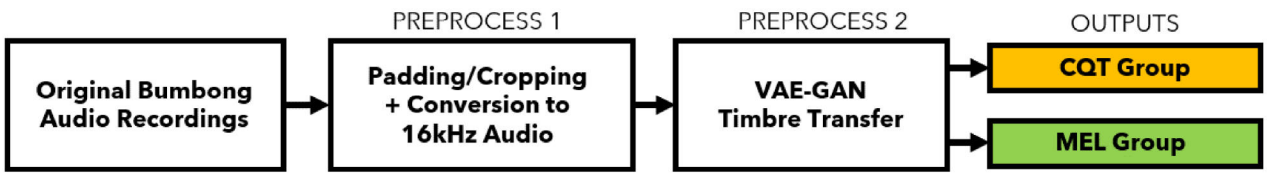

Figure 4.3: Preprocessing Flow

4.2 Data Preprocessing

Based on the implementation of GANSynth to the NSynth database, the audio samples were set to the following specifications below and will generate 64,000 dimensions and is intended to sound natural to an average listener. [1] Audio files with longer duration than the NSynth specification were cropped with mostly quiet portions removed while some audio files with slightly lower duration than 4 seconds were padded with quiet portions.

- Duration: **4 seconds each**
- Sampling Frequency: **16kHz**
- Fundamental Pitch Range: **MIDI 24-84** (~32Hz-1kHz)

The audio samples will also be converted to specific spectrograms as it may improve accuracy, reduce error rate, or decrease training time. In this project, the input representations used are CQT and MEL Spectrograms through the VAE-GAN pipeline [18][19] as seen in Figure 4.3. The output of the VAE-GAN (CQT) and VAE-GAN (MEL) pipelines in audio (.wav) format were later trained using GANSynth. VAE-GAN was used for preprocessing than directly using CQT or MEL spectrograms as GANSynth recognizes playable audio (.wav) files for its input. The output audio files from VAE-GAN (both MEL and CQT) was not converted to back to its original waveform but there were observed cropping and looping difference when played which were expected due to most spectrograms with hopping bins and window size as parameters.

4.3 Model Training and Validation

GANSynth is based on Progressive GAN with additional method of appending a one-hot representation of musical pitch to the latent vector to be able to independently control pitch and

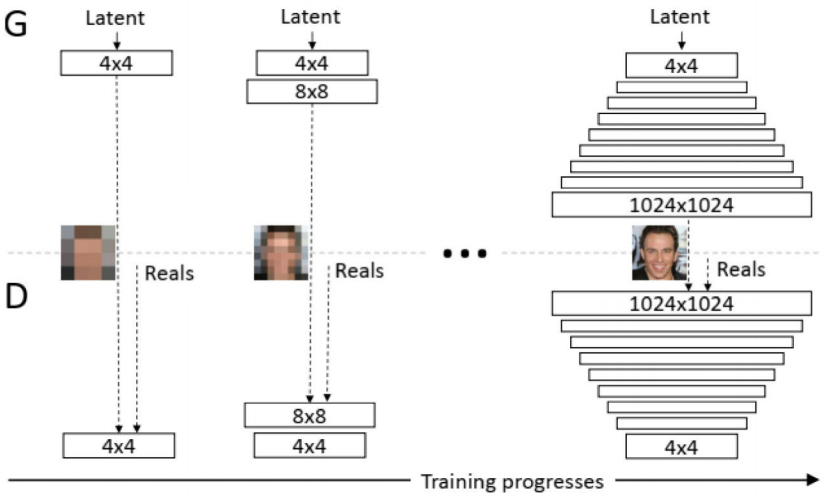

Figure 4.4: Progressive GAN Training Model

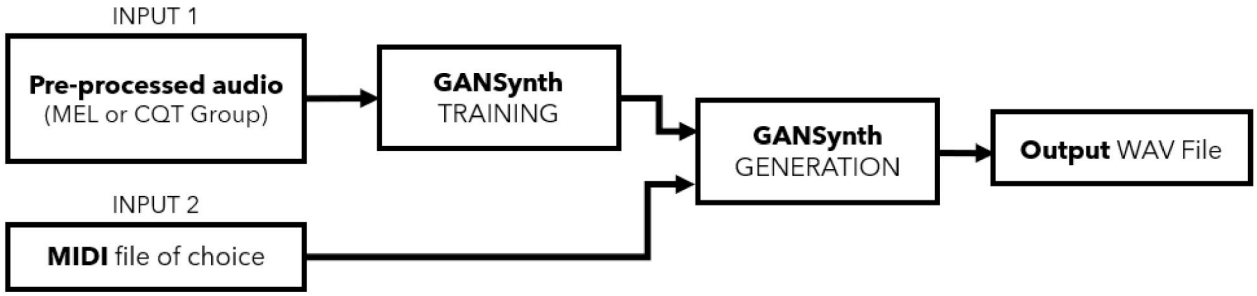

Figure 4.5: GANSynth Training

timbre. The representation of Progressive GAN training model with human photos is shown in Figure 4.4 which starts with both Generator (**G**) and Discriminator (**D**) [1]. The representation of images on Figure 4.3 were associated with preprocessed audio data as stated in Section 4.2.

The preprocessed audio file groups (CQT and MEL) were trained separately using GANSynth along with an input MIDI file of choice as seen in Figure 4.5. The training of a model in GANSynth can be done once and the generation with any MIDI (.mid) files can be done repeatedly. In this project, the training duration for both CQT and MEL groups were below 30 minutes only while the generation duration varies from 1-5 minutes depending on the length and complexity of the input MIDI file.

Filenames	Composition	Duration (m:s)
London.mid	London Bridge (Nursery Rhyme)	0:46
Mary.mid	Mary Had a Little Lamb (Nursery Rhyme)	0:21
NGGYU.mid	Never Gonna Give You Up (by R. Astley)	3:36
522.mid	Custom MIDI file with repeating notes with octaves from 5 to 2.	1:00
allreg.mid	Custom MIDI file with simple notes only (no #).	0:55

Table 4.2: MIDI File Details

4.4 Evaluation & Analysis

Sound evaluation is a challenge as it is hard to formalize, and people can have different perspective on audio hearing. Based on GANSynth metric evaluations, this project will adopt the Fréchet Audio Distance (**FAD**) evaluation to cater outputs on GAN-based neural networks [10].

4.4.1 Objective Test (FAD)

Fréchet Audio Distance (FAD) adapted the Fréchet Inception Distance (FID), which usually evaluates images from GAN, to evaluate audio generated from GAN [14]. FAD helps evaluate the difference of the original and synthesized audio of the Bumbong instrument. In this project, the FAD evaluation used is obtained from VAE-GAN's GitHub repository. Five (5) MIDI files with more simple compositions (in terms of layer of instruments used) were prepared as shown in Table 4.2.

To prepare the FAD setup, three (3) groups were prepared namely the MEL, CQT, and the reference dataset. The reference dataset consists of five (5) audio (.wav) files composed of MIDI songs similar to Table 4.1 merged with a custom Bumbong soundfont (.SF2) file. The generated soundfont file from a software called Polyphone was composed of all Bumbong unprocessed audio recordings separated into three (3) groups of instruments according to the three individuals that was recorded for the dataset. Both MEL and CQT groups also used the same MIDI file from Table 4.1 for GANSynth generation for consistency.

To compute the FAD score, all three groups were processed to first create their respective statistic files for comparison. Both MEL and CQT groups were both compared to the reference dataset and the scores were recorded.

4.4.2 Subjective Test (Mean Opinion Score)

Thirty (30) participants were asked to rate the similarity of the synthesized and referenced Bumbong audio. The rating is set from 1 (not similar) to 10 (very similar) for this evaluation.

Three (3) spliced audio sample groups with all less than 15 seconds were prepared which utilized NGGYU.mid as Track01, 522.mid as Track02, and Mary.mid as Track03 to have a variation in terms of music simplicity. Total of nine (9) audio samples for CQT, MEL, and reference groups were posted for evaluation. Other information such as the audio device used (wired, wireless, loudspeaker) were recorded for reference.

Chapter 5

Results and Discussion

5.1 GANSynth Output

The spectrogram view of all audio groups are shown respectively below for spliced tracks, Track01 (NGGYU.mid), Track02 (522.mid), and Track03 (Mary.mid). It can be observed that the reference audio using a normal soundfount has a cleaner and less distorted spectrogram view compared to both CQT and MEL groups. However, CQT and MEL groups possesses similarity of patterns in the spectrogram view which suggests retention of features. Track01 shows a more distorted spectrogram view on CQT and MEL when it reaches the more complex part of the song. It can also be observed that prominent horizontal lines of the reference audio can also be observed on the CQT and MEL groups. The Track02 with repeating and downsizing notes shows a consistent distortion on the spectrogram views of CQT and MEL groups while the reference does not have much distortions. The possible reason for the distortion on the CQT and MEL groups of Track02 suggests that GANSynth tries to recreate the longer notes of Bumbong and its spectral features. Track03 with simple notes of a nursery rhyme shows that there are similarities in terms of collective area of distortions and has consistency on every note played.

5.2 Evaluations

The evaluation for this project is separated into objective (FAD) and subjective test (Mean Opinion Score). The summary of results is shown the the tables below.

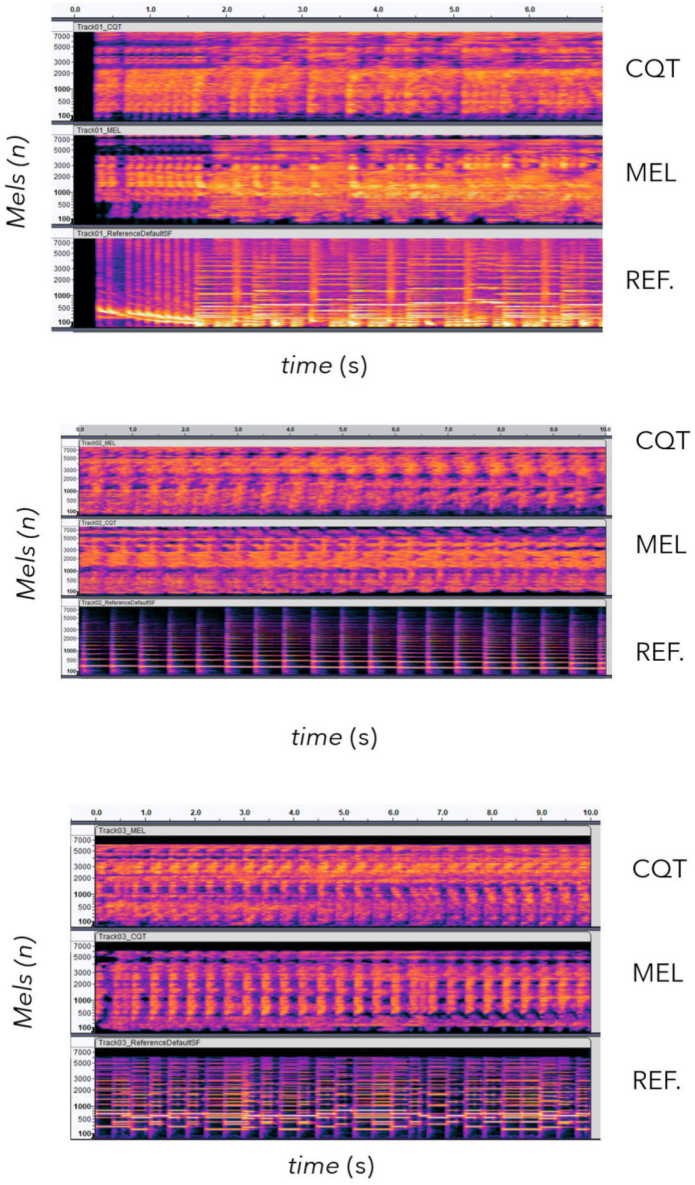

Figure 5.1: Spectrogram View of Audio Tracks

$$FAD = ||\mu_r - \mu_g||^2 + tr(\Sigma_r + \mu_g - 2\sqrt{\Sigma_r \Sigma_g})$$

Figure 5.2: FAD Formula

Audio Group	FAD Score
CQT Spectrogram	13.559
MEL Spectrogram	**11.124**

Table 5.1: FAD Results

5.2.1 Fréchet Audio Distance (FAD)

FAD scores show the intensity of distortion or difference of a target audio file compared to a clean/reference audio. In this project, the target audio file groups (MEL and CQT) were compared to the reference group (composed from recorded audio via soundfont). The formula for FAD is shown below.

As seen in Table 5.1, the MEL spectrogram group scored lowest and has a difference of ~2.435 to the CQT spectrogram group. Lower FAD (**bolded**) score means a better result. This evaluation shows that CQT did not achieve a more accurate result on FAD compared to MEL.

As an additional reference, the FAD score on both CQT and MEL groups compared to a portion of a different dataset (MagnaTagATune) which resulted 20.03 and 18.29 respectively. This additional result suggests that MEL still achieved a better FAD score than CQT even on different reference datasets.

5.2.2 Subjective Test

Based on the audio listening test with thirty (30) participants, the MEL group shows a more similarity to the reference audio by 6.08% than the CQT group as shown in Table 5.2. The audio devices used for the test were composed of 19 wired earphones (63.3%), 5 wireless earphones (16.7%), and 6 loudspeakers (20%) with all participants using a desktop/laptop PC for evaluation. The summary shows a low similarity rating (with 10 as perfect score) as both CQT and MEL groups have various distortions when played. The participants were able to tell that MEL sounds better as both Track02 and Track03 received the highest score while Track01 only had a 0.067 difference.

Audio Group	Track01	Track02	Track03	Mean Score
CQT Spectrogram	**4.367**	3.900	4.100	4.122
MEL Spectrogram	4.300	**4.200**	**4.667**	**4.389**

Table 5.2: Subjective Test Summary

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

This project implemented GANSynth audio synthesis for the Bumbong audio dataset played and recored by three players using both CQT and MEL spectrograms as the input representations. Both CQT and MEL spectrograms were converted back to an audio format using VAE-GAN timbre transfer pipeline retaining the features of a spectrogram as input for GANSynth neural network training. The 240 Bumbong audio dataset recordings were preprocessed thru VAE-GAN with batch size of 2 and epoch setting of 50. The MEL and CQT output groups of VAE-GAN were trained separately in GANSynth and an audio file can be produced using the trained groups using a MIDI file.

Based on the objective test (FAD), the MEL group showed a better result with a lower score than CQT. This lower score defines that the MEL group has produced lower distortions when trained in GANSynth. The GANSynth paper itself produced significantly higher Frechet Inception Score (FID) of 104-708, a metric which FAD is based, on all spectrogram groups tested with 13 as the best score on the reference data [1].

Whereas on the subjective test, thirty (30) participants scored the MEL group as the most similar based on the reference audio with 6.08% higher rating than the CQT group. Both CQT and MEL groups produced a lower collective mean score of 4.256 out of 10 as all audio outputs were observed to have distortions which greatly affects the melody and notes that makes a good-to-hear audio.

In this study, it is to be noted that MEL group yielded better results on both objective and subjective tests. The MEL spectrogram was known for its accommodation to human hearing perception which can be one reason that it scored higher than CQT.

PC Parts	Used in this project	Minimum	Recommended
CPU	Intel Core i3-10100F	Any with AVX instruction set	Intel Core i7/AMD Ryzen 7
RAM	16GB DDR4 total	8GB total	32GB-128GB
Graphics	NVIDIA GTX1660S 6GB	NVIDIA 600 or later	NVIDIA RTX/Quadro GPU
OS	Windows 11 (WSL)	Windows 10 22H2 (WSL)	Windows 11 or later (WSL)

Table 6.1: PC Specifications Summary

6.2 Recommendations

The following are the recommendations for the future work of this study and other similar study involving musical intruments, spectrograms, and neural network architectures.

6.2.1 Computer Specifications

Due to the computer specifications used in this project, training limitations are placed. The compiled PC specifications are listed in Table 6.1.

GANSynth's training duration depends on the number of audio files used and is optimized depending on the graphics card/s present. FAD Evaluation [19] requires more memory (RAM) as memory leak can occur on reference dataset with number of audio files that the code cannot accommodate. VAE-GAN was also observed to perform well depending on the CPU cores/threads set on the parameters to perform training in one of its pipelines (see train.py). The listed recommended specifications are based on the consumer products available as industrial components are highly expensive and harder to acquire.

6.2.2 Future Work

- Training Recommendation
  - Due to the current PC specification used, the number of epochs used in this project was limited to 50 as the training loss for the GAN discriminator reaches below 0.45 mark from approximate 1.02 mark on epoch 0 while the GAN generator reached a below 25 mark on approximate 57.0 mark on epoch 0. Depending on the input representation used in this study, the training duration per epoch varies from 3-10 minutes in VAE-GAN. Setting a higher epoch in the parameters may produce better and noticeable outputs but with possible exponentially or logarithmically increased training durations.
  - Due to GANSynth's specification, the training part is consistent with reaching 12 training stages regardless of number of files and filetypes. There were no available record

logs for training loss for reference. As GANSynth has a high possibilty of updating its research, its GitHub repository may be updated with more features and more specific documentation in near future.

- GANSynth Hyperparameters
  - GANSynth hyperparameters are utilized in a sample generation code snippet in GANSynth's GitHub repository but there were no documented hyperparameters that can be used. There were MEL-related hyperparameters present but other spectrograms were neither mentioned nor documented. The GANSynth code can be installed straight from the Python "pip install" command for Magenta which can be updated once available. The use of hyperparameters may help produce better results.
- Instrument Consideration
  - Other types of indigenous instruments such as stringed or bass can be considered as these types were commonly used for various audio synthesis models including GAN-based architectures.
- Timbre Transfers and Input Representations
  - There are other timbre transfer techniques and input representations (including variations and addition of audio features) that can be considered for future work.

Bibliography

[1] J. Engel, K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial Neural Audio Synthesis," arXiv:1902.08710, 2019. 1902.08710.

[2] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," arXiv:1704.0127, April 2017, 1704.0127

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," arXiv:1710.10196, February 2018. 1710.10196.

[4] A. F. B. Laguna, N. M. P. Valdez and R. C. L. Guevara, "MIDI implementation of a kulintang modal synthesizer using the VST2.4 standard," TENCON 2012 IEEE Region 10 Conference, 2012, pp. 1-5, doi: 10.1109/TENCON.2012.6412221.

[5] T. Sinjankhom, S. Chivapreecha, N. Chitanont and T. Kato, "Deep Neural Networks for Sound Synthesis of Thai Duct F1ute, Khlui," 2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST), 2021, pp. 63-67, doi: 10.1109/ICEAST52143.2021.9426294.

[6] Philippine Bamboo Musical Instruments. "Bumbong - Philippine Bamboo Musical Instruments," [Online]. Available: https://phbmi.com/bamboo-musical-instruments/bumbong-2/

[7] Philippine Bamboo Musical Instruments. "About the Program - Philippine Bamboo Musical Instruments," [Online]. Available: https://phbmi.com/about-the-program/

[8] M. C. Arayata, Uniting science, culture thru bamboo musical instruments." Novemeber 30, 2020. [Online]. Available: https://www.pna.gov.ph/articles/1123329

[9] C. A. Bagaforo and J. Gayo, "Digital sound synthesis of the Tongali and Kolitong implemented as a virtual instrument plugin," UP Electrical and Electronics Engineering Institute, June 2020.

[10] S. Bayat and G. Işık, "Identification of Aras Birds with Convolutional Neural Networks," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 2020, pp. 1-5, doi: 10.1109/ISMSIT50672.2020.9255205.

[11] J. P. Castillo and J. A. Domingo, "VSTi Implementation of Angklung, Bumbong, and Marimba," UP Electrical and Electronics Engineering Institute, December 2019.

[12] K. W. Cheuk, H. Anderson, K. Agres and D. Herremans, "nnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks," in IEEE Access, vol. 8, pp. 161981-162003, 2020, doi: 10.1109/ACCESS.2020.3019084.

[13] K. W. Cheuk, K. Agres and D. Herremans, "The Impact of Audio Input Representations on Neural Network based Music Transcription," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-6, doi: 10.1109/IJCNN48605.2020.9207605.

[14] S. Tiwari, A. Jain, A. K. Sharma and K. Mohamad Almustafa, "Phonocardiogram Signal Based Multi-Class Cardiac Diagnostic Decision Support System," in IEEE Access, vol. 9, pp. 110710-110722, 2021, doi: 10.1109/ACCESS.2021.3103316

[15] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms," arXiv:1812.08466, January 2019, 1812.08466

[16] A. Natsiou and S. O'Leary, "Audio representations for deep learning in sound synthesis: A review," arXiv:2201.02490, January 2022, 2201.02490

[17] Apple Developer. "Computing the Mel Spectrum Using Linear Algebra," [Online]. Available: https://developer.apple.com/documentation/accelerate/computing_the_mel_spectrum_using_linear_algeb

[18] T. M. Cua, Y. Sta. Maria, and D. N. Tumanut, "Timbre Transfer of Monophonic Stringed Bass Instruments Samples Using Variational Autoencoder-Generative Adversarial Network," UP Electrical and Electronics Engineering Institute, March 2023.

[19] R. S. Bonnici, C. Saitis, M. Benning, "Timbre Transfer with Variational Auto Encoding and Cycle-Consistent Adversarial Networks". arXiV:2109.02096, October 2021, 2109.02096

File Resources

To access the codes used in this project, you may check out the researcher's GitHub repository here. The evaluation files and other resources are also linked in the GitHub repository page.

GitHub Repository: https://github.com/upd-kazutoph/ECE198X