ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

A Comparative Evaluation of Tonal Replication Techniques for Music Composition and Sound Design Through Tonal Synthesis

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Introduction

Context of the Study

- Music has constantly evolved through innovation, whether in sound design or composition
- One good example: the use of synthesizers starting in the 1960s
  - The use of synthesizers led to genres such as techno & house
- Even more recently, we've seen more advancements in music
  - Algorithmic composition, sample-based production, AI and ML-assisted music production
- This study aims to explore an intersection **between AI and advancements in synthesizer technology: the replication of instrumental tones**

Research Objectives

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.

ACSML

Research Objectives

The study aims to utilize various techniques in audio synthesis and replication, namely DDSP, DWTS, and NIC in order to replicate instrumental tones, comparing them in terms of accuracy and efficiency.

Sub-objective #1:

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.

ACSML

Research Objectives

The study aims to utilize various techniques in audio synthesis and replication, namely DDSP, DWTS, and NIC in order to replicate instrumental tones, comparing them in terms of accuracy and efficiency.

Sub-objective #2:

Determine the perceptual quality of synthesized audio as compared to the original audio input.

ACSML

Research Objectives

The study aims to utilize various techniques in audio synthesis and replication, namely DDSP, DWTS, and NIC in order to replicate instrumental tones, comparing them in terms of accuracy and efficiency.

Sub-objective #3:

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

ACSML

Research Questions

Research Questions

How do methods in audio synthesis and replication such as **Differentiable Digital Signal Processing** (DDSP), **Differential Wavetable Synthesis** (DWTS), and **Neural Instrument Cloning** (NIC) compare in their accuracy and efficiency in replicating instrumental tones?

ACSML

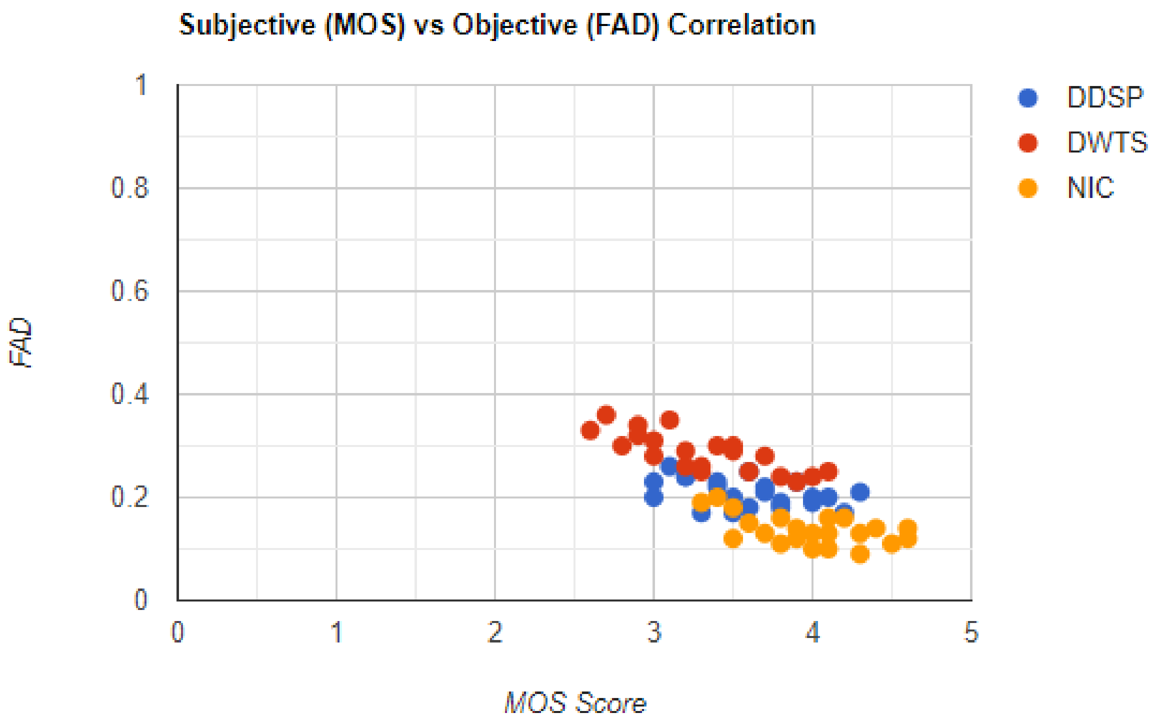
Example Visualizations

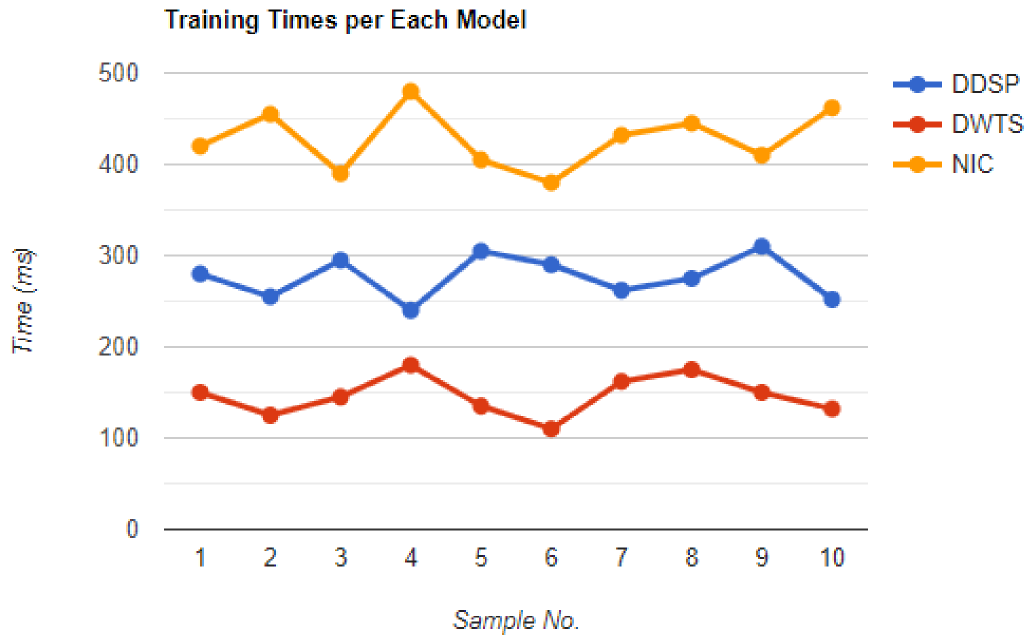
Accuracy

Average Accuracy Scores for Each Model

Metric	DDSP	DWTS	NIC
FAD	0.18	0.25	0.15
NDB/k	120	180	105
MOS	3.8	3.2	4.1
MUSHRA	75	60	82

Efficiency





*These are all fake data for the purpose of visualizations

Research Questions

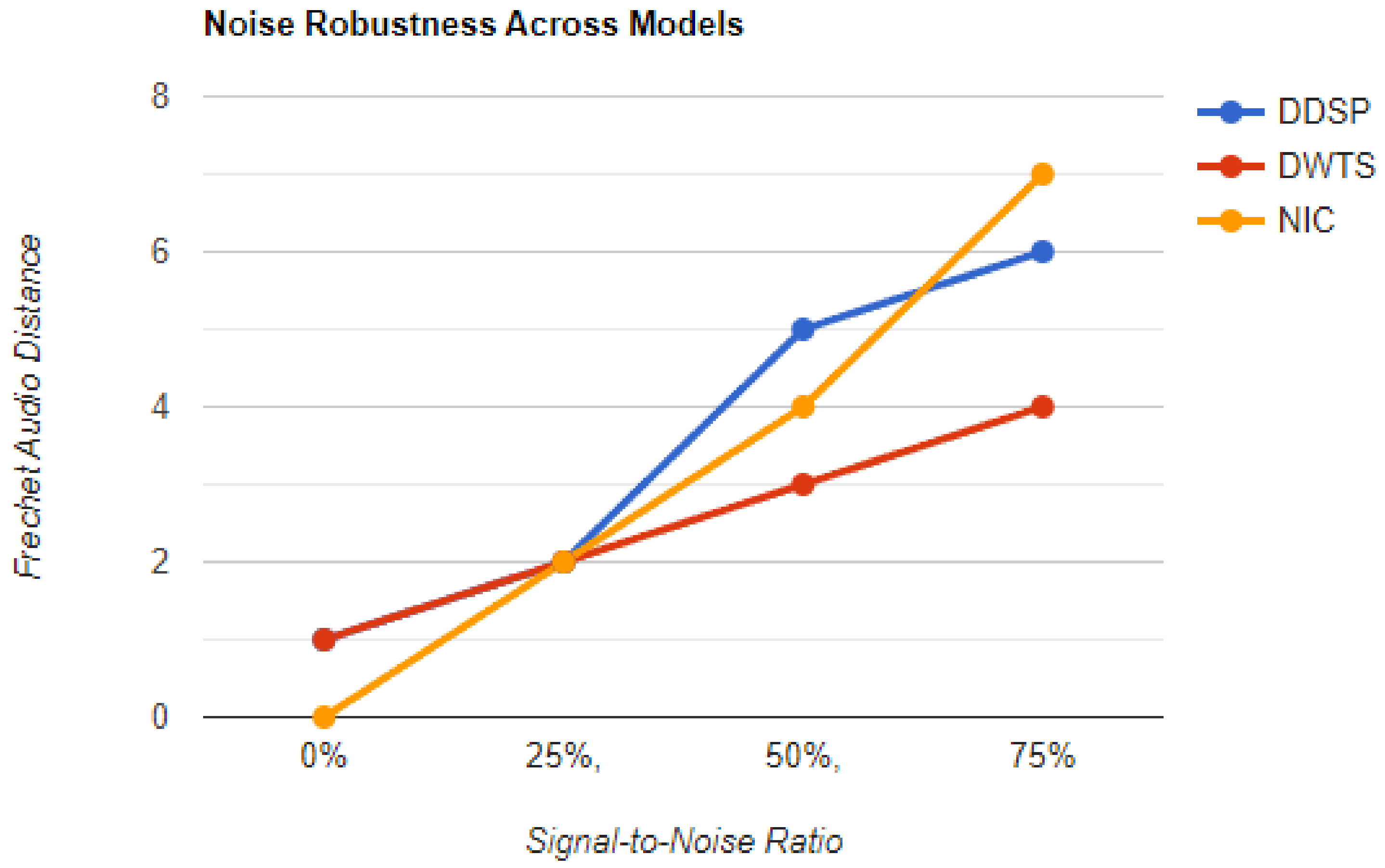
Subquestion #1

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?

ACSML

Example Visualizations

Noise Robustness



Frequency Masks

FAD Across Mask Bandwidth (K=50)

Mask Bandwidth	DDSP	DWTS	NIC
None	0.01	0.20	0.12
Narrow (200 Hz - 500 Hz)	0.20	0.01	0.24
Medium (1000 Hz - 2000 Hz)	0.46	0.69	0.40
Wide (4000 Hz - 8000 Hz)	0.78	0.96	0.85

Fragmented Audio

FAD by Fragmentation Level

Fragmentation Width	DDSP	DWTS	NIC
None	0.01	0.20	0.12
Narrow (10 - 20 ms)	0.16	0.21	0.16
Medium (300 - 500 ms)	0.39	0.40	0.51
Wide (1000 - 2000 ms)	0.87	0.78	0.91

*These are all fake data for the purpose of visualizations

Research Questions

💡 **Subquestion #2** 💡

How does the synthesized output generated by the models compare to the original audio input in terms of perceptual quality?

ACSML

Example Visualizations

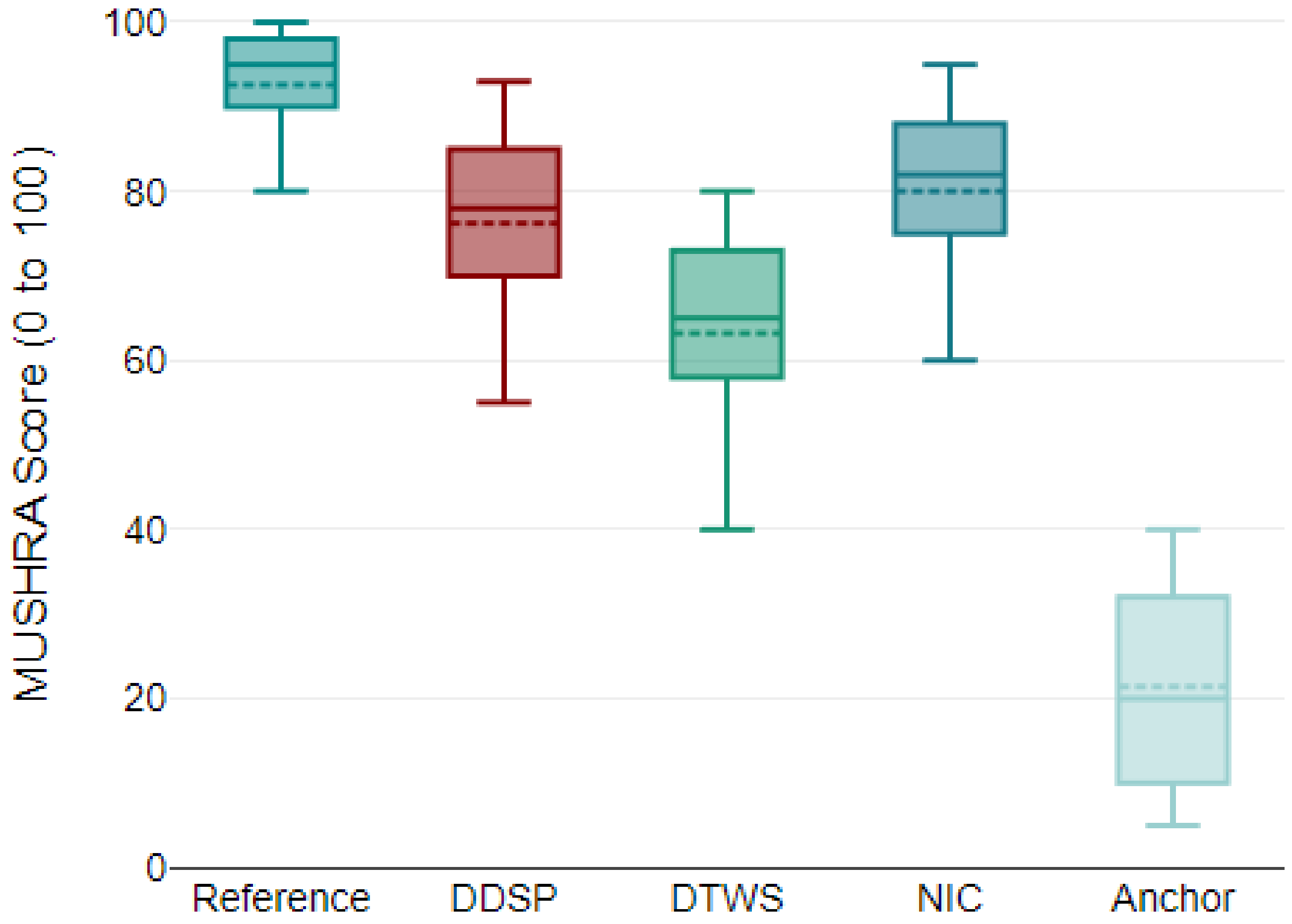
MOS

Mean Opinion Scores for Each Stimuli for Each Model

Stimulus	DDSP	DWTS	NIC	Original
Violin	3.8	3.2	4.3	4.9
Trumpet	3.3	3.9	3.6	4.6
Piano	4.1	3.0	4.0	4.8

MUSHRA

MUSHRA Scores for Each Model



*These are all fake data for the purpose of visualizations

ACSML

Research Questions

Subquestion #3

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

ACSML

Example Visualizations

Different Instrumental Domains

Average NDB/k for each Model on Each Instrumental Domain

Domain	Instrument Inputs	DDSP (NDB/k)	DWTS (NDB/k)	NIC (NDB/k)
Training Set (Orchestral)	Violin, Oboe, French Horn, Piano	0.03	0.10	0.14
Unseen Set 1 (Non-Western Folk)	Sitar, Erhu, Koto	0.35	0.45	0.28
Unseen Set 2 (Electronic)	Synth Bass, FM Pads	0.43	0.86	0.91

Scope and Limitations

Scope

- Limited to implementations of DDSP, DWTS, and NIC as presented in their original papers
- Audio input: any melodic recording of an instrument (i.e. violin, marimba, etc.)
- Training data: NSynth dataset
- Metrics: FAD, NDB/*k*, MOS, MUSHRA

Limitations

- None of the metrics are individually sufficient
- Models have only been trained on monophonic audio (i.e. NSynth)

ACSML

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Significance of the Study

Significance of the Study

- This study lies at the intersection of computer science and computational musicology
- By evaluating three of the most recent generative models in the field, we could see which holds the greatest potential for further advancements
    - These models could be used for developing new tools in the realms of **music production, audio restoration, processing or enhancement, and sound design**

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Review of Related Literature

WaveNet and GANs (Generative Adversarial Networks)

WaveNet

- Groundbreaking work by Engel et al. that paved the way for computational musicology by using conditional autoencoders on raw audio waveforms.
- Birth of the NSynth dataset, composed of 300,000+ notes from various instruments.
- Found links between instantaneous frequencies and vibrato, enhancing harmonic blend.
- **Limitation:** Sample-based audio recreation without full global context

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

Generative Adversarial Networks

- Also by Engel et al., uses Progressive GAN architecture and musical pitch conditioning.
- Employs Short Term Fourier Transform (STFT) and Instantaneous Frequency on the Mel scale (IF-Mel) for 50,000x faster sample generation than WaveNet.
- Latent features and pitch information yield smooth timbral interpolation and consistency.
- Represents a breakthrough in GAN audio generation with superior quality to WaveNet.

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

ACSML

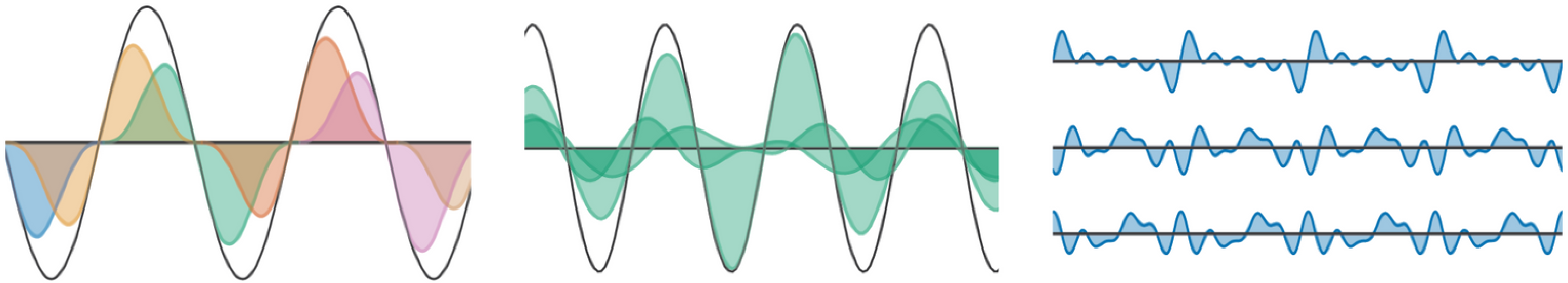
ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Review of Related Literature

Differentiable Digital Signal Processing (DDSP)

Differentiable Digital Signal Processing

- Addresses limitations of other audio generation methods
  - Strided Convolutions (“sliding window”): **phase alignment**
  - Fourier Representaton: **spectral leakage**
  - Autoregressive: **waveform != perception**



Strided Convolution
Phase Alignment
(WaveGAN, SING)

Fourier Representation
Spectral Leakage
(Tacotron, GANSynth)

Autoregressive
Waveform != Perception
(WaveNet, SampleRNN)

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

Differentiable Digital Signal Processing

- Introduces **components**
  - Spectral Modeling Synthesis
    - Harmonic plus Noise model
  - Harmonic Oscillator / Additive Synthesizer
  - Envelopes
  - Linear Filter Design
    - Time-varying FIR filter applied to non-overlapping audio frames

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

Differentiable Digital Signal Processing

- Introduces **components**
    - Subtractive Synthesizer
    - Reverb Module

$$x(n) = \sum_{k=1}^{K} A_k(n) \sin(\phi_k(n)),$$

$$\phi_k(n) = 2\pi \sum_{m=0}^{n} f_k(m) + \phi_{0,k},$$

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

Differentiable Digital Signal Processing

- DDSP Autoencoder
    - Encoders
        - f-encoder that outputs fundamental frequency **f(t)**
            - CREPE
        - l-encoder that outputs loudness **l(t)**
            - A-weighted log of power spectrum
        - z-encoder that outputs residual vector **z(t)**
            - MFCC from log-mel-spectrogram to normalization layer

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

ACSML

Differentiable Digital Signal Processing

- DDSP Autoencoder
    - Decoder
        - Input: (f(t), l(t), z(t)) (250 timesteps)
        - Output: Parameters for synthesizers

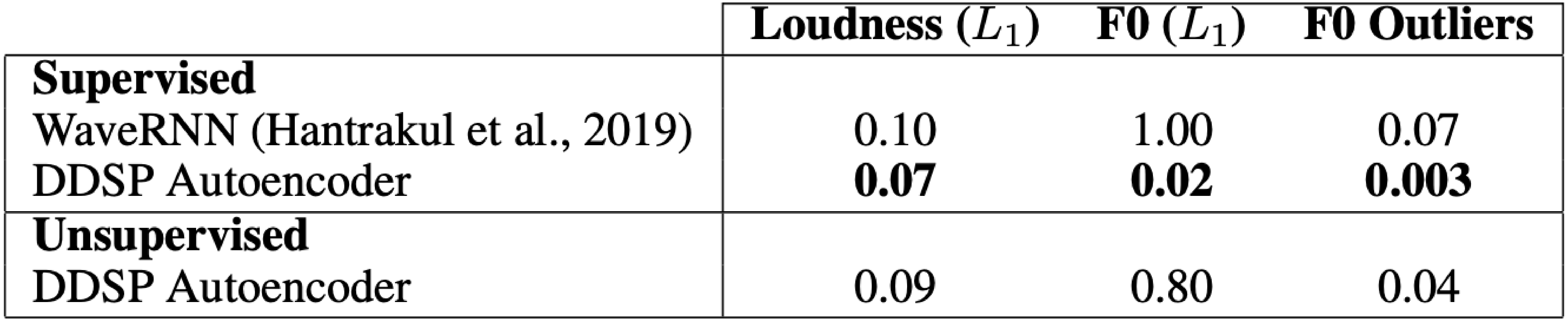
Engel, J., Hantrakul, L., Gu, C., and Roberts, A. 2020. DDSP: Differentiable Digital Signal Processing. Retrieved April 1, 2024 from http://arxiv.org/abs/2001.04643

Differentiable Digital Signal Processing

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.

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

Differentiable Digital Signal Processing



	Loudness (L_1)	**F0 (L_1)**	**F0 Outliers**
Supervised			
WaveRNN (Hantrakul et al., 2019)	0.10	1.00	0.07
DDSP Autoencoder	**0.07**	**0.02**	**0.003**
Unsupervised			
DDSP Autoencoder	0.09	0.80	0.04

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

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Review of Related Literature

Differentiable Wavetable Synthesis (DWTS)

Differentiable Wavetable Synthesis (DWTS)

- Shan et al. extend DDSP concepts for dynamic audio synthesis.
- Uses a **dictionary of wavetables** extracted from one-shot audio samples (e.g. NSynth dataset).
- Wavetable morphing enables **changing timbres over time**, leading to expressive output.
- Employs a **mathematical formula to combine wavetables, controlling amplitude, attention, and fractional indexing for precise audio generation.**

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

ACSML

Differentiable Wavetable Synthesis (DWTS)

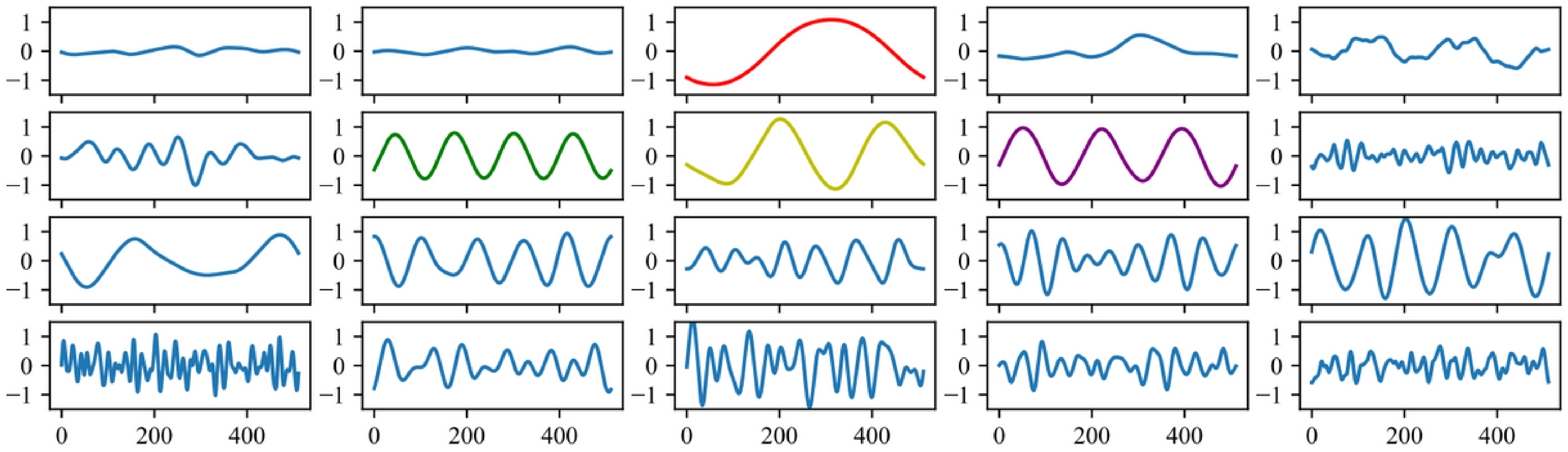


Fig. 1. Learned wavetables ordered with highest average attention weights appearing first (normal English reading order). Wavetables of key harmonics are highlighted: f_0 (red), f_1 (yellow), f_2 (purple) and f_3 (green). The remaining wavetables are data-driven combinations of higher harmonics. The first two wavetables appear to be silence.

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

Differentiable Wavetable Synthesis (DWTS)

- Results (Reconstruction Error):
    - DDSP (Normal): 0.5834
    - DDSP (DWTS): 0.5712

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

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Review of Related Literature

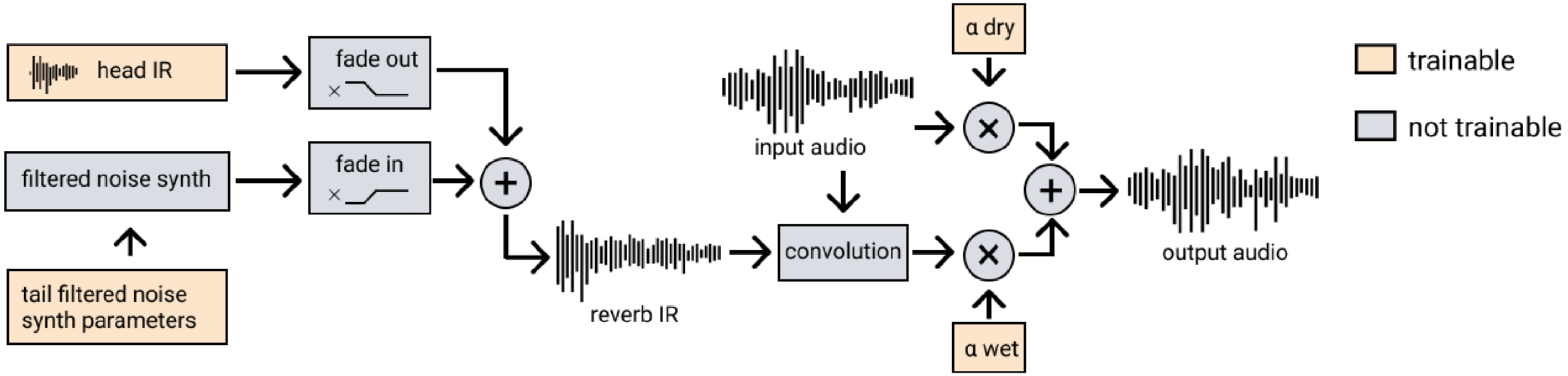
Neural Instrument Cloning (NIC)

Neural Instrument Cloning (NIC)

- Reduced required audio data from **10 minutes to (4 to 256) seconds** by drawing inspiration from speech voice cloning.
- Incorporated ***F0* confidence** (*pitch estimation data*) and novel **two-part reverb design** into the model architecture.
- While adding a natural quality, the synthesized audio sometimes exhibited a **distant or echo-like character**.
- While pre-training did not directly improve quality, it **significantly accelerated instrument cloning processes**.

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

Neural Instrument Cloning (NIC)



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

Neural Instrument Cloning (NIC)

- Results:
    - F0-confidence lessened artifacts around note onsets and offsets
    - Two-part reverb design: diminishing returns

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

Review of Related Literature

DDSP (Other)

Other Applications of DDSP

- **Neural Waveshaping: using NEWTs** (Hayes, 2021)
  - Neural Waveshaping Synthesis that employs neural waveshaping units that learn from shaping functions from unlabeled audio.
- **Implementations: Synthesizer Sound Matching DDSP** (Masuta and Saito, 2021)
  - Implemented a more precise DDSP with controllable parameters.
- **Hybrid Models: DSPGAN** (Song et al., 2022)
  - Combines GAN and DSP-inspired methods for high-fidelity speech synthesis
- **Accessibility: Tone Transfer** (Carney et al., 2021)
  - Modified DDSP for efficient web use

ACSML

Review of Related Literature

Audio Synthesis Evaluation Metrics

Objective Evaluation

- **Inception Score (IS):** Popular but has limitations, now outperformed by FID.
- **Fréchet Inception Distance (FID):** Measures quality and diversity, more reliable than IS. Requires replacing Inception component with audio-specific features.
- **Fréchet Audio Distance (FAD):** (Based on FID) Compares generated audio against a reference set, measures audio quality and distortion (Kilgour et al.).
- **Number of Statistically Different Bins (NDB/*k*):** Evaluates diversity of generated audio versus the original dataset (Richardson and Weiss).

ACSML

Subjective Evaluation

- **Mean Opinion Score (MOS):**
    - Participants rate audio on a Likert scale across quality-related questions (clarity, naturalness, etc.).
- **Multiple Stimuli with Hidden Reference and Anchor (MUSHRA):**
    - Builds upon MOS with hidden benchmark audio and low-quality anchors for a more nuanced comparison.

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Methodology

Comparative Analysis

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

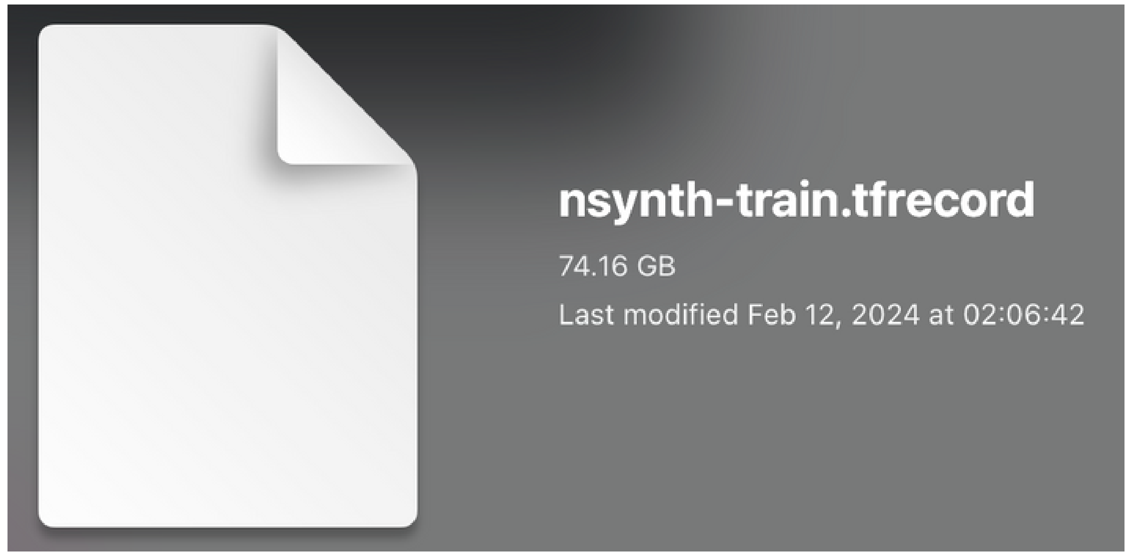
1. **Differentiable Digital Signal Processing (DDSP)** by Engel et al.
2. **Differentiable Wavetable Synthesis (DWTS)** by Shan et al.
3. **Neural Instrument Cloning From Very Few Samples (NIC)** by Jonason and Sturm

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.

Dataset

NSynth Dataset

- Annotated already!
- Used in DDSP
- ...is an audio dataset containing **305,979 musical notes, each with a unique pitch, timbre, and envelope.**
- Created by Engel et al.



nsynth-train.tfrecord

74.16 GB

Last modified Feb 12, 2024 at 02:06:42

https://magenta.tensorflow.org/datasets/nsynth

Dataset Preparation

Why? → **monophonic and generated in isolation**

What? →

- F0 (*fundamental frequency*) ← CREPE
- Loudness
  - Root-mean-squared / A-weighted log power spectrum

Dataset Use

Smaller Subset

- “70,379 examples composed mostly of strings, brass, woodwinds and mallets with pitch labels within MIDI pitch range 24-84”
- Considering training size and availability of resources

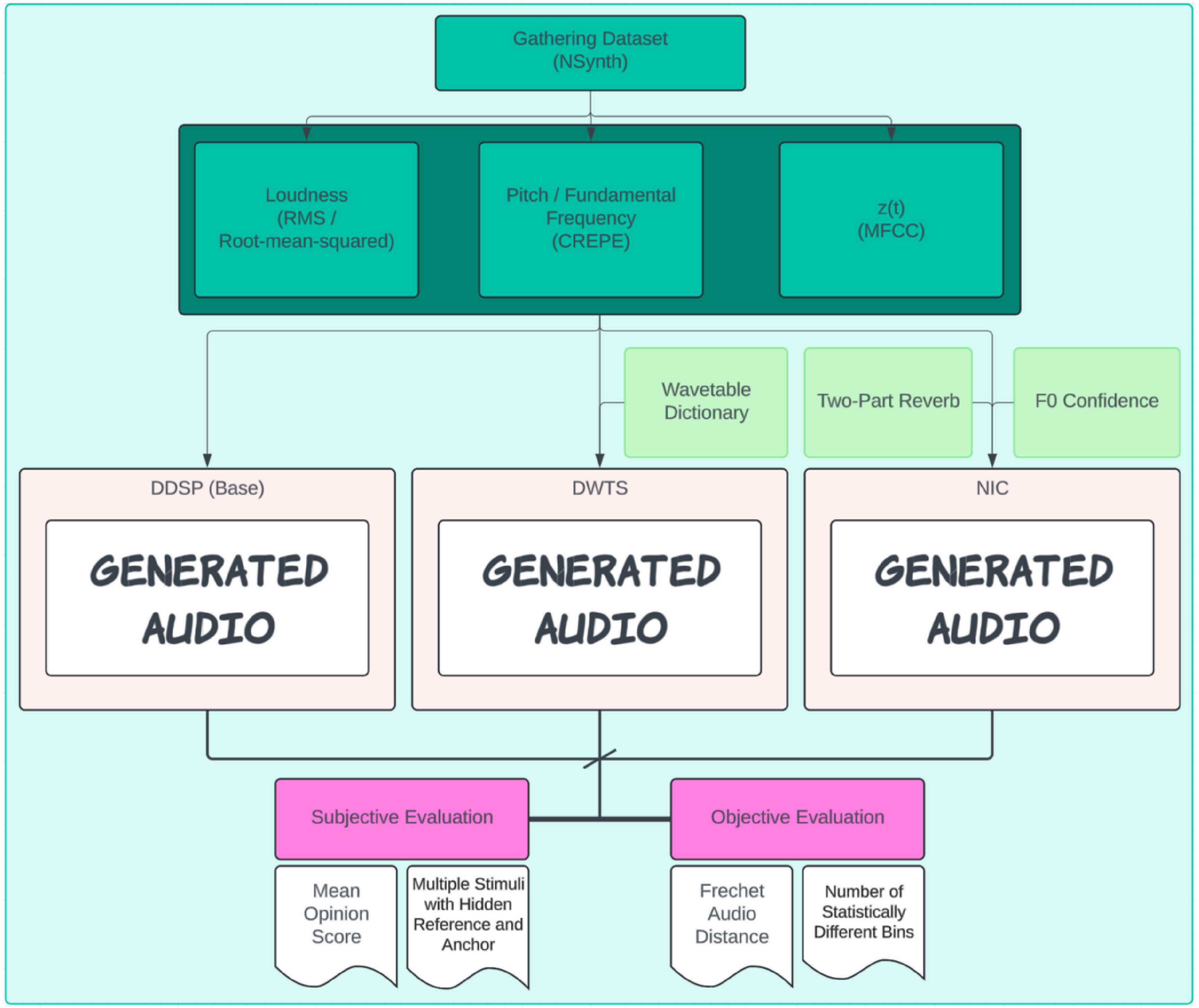
Model Training

Multi-Scale Spectral Loss

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

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).

General Flowchart



ACSML

Evaluation - Objective Metrics

- **Fréchet Audio Distance (FAD)**: Measures how close the generated audio is to the distribution of real, clean music. Lower FAD indicates better audio quality.
- **Number of Statistically Different Bins (NDB/k)**: Measures the diversity of generated audio samples. Higher NDB/k indicates a wider range of unique timbres and characteristics.

Evaluation - Subjective Metrics

- **Mean Opinion Score (MOS)**: Participants will rate the audio on a Likert scale.
- **Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)**: Participants rate the perceived quality of generated audio in relation to hidden reference and anchor samples.

ACSML

Audio Input Set-ups

- 1 base set of audio inputs will be procured
- 4 set-ups for audio input configuration
    - **Control set-up**: base set
    - **Noise set-up**: base set with white noise added of various ratios
    - **Mask set-up**: base set passed through notch filter
    - **Fragmented set-up**: base set with silence interspersed
- Another set will be procured with **instruments not found in NSynth**
- All set-ups will be used as input and synthesized audio will undergo metrics

ACSML

Timeline - CSCI 199.1

April 1 - 6	**Midterm Defense**
April 7 - 13	**Post-Midterm Revisions**
April 14 - 27	**Refining of RRL, Methodology** RRL: More research into methods Methodology: Refining of subjective testing (i.e. logistics)
April 27 - May 12	**Preliminary Model Creation**
May 13 - 18	**Final Defense**

Timeline - CSCI 199.2 and 199.3

CSCI 199.2, First Half	**Training of Models** **Experimentations (Audio Synthesis)**
CSCI 199.2, Second Half	**Data Gathering (Metric Testing)** **Progress Report**
CSCI 199.3, First Half	**Data Gathering, cont.** **Data Analysis**
CSCI 199.3, Second Half	**Data Analysis, cont.** **Revisions** **Final Defense**

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

THANK YOU!