ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Thesis!!!!

Hehe

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Research Topic (FINAL!!!)

Research Topic

Replicating Instrumental Tones by creating a model that replicates and synthesizes tones from audio input for music composition and sound design

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Research Questions

Research Questions

How do methods in audio synthesis and replication (DDSP, Tone Transfer, and GANs) compare in their accuracy and efficiency in replicating instrumental tones?

Sub-question #1:

How does the quality and the choice of encoding and representation for the input audio impact the efficiency and accuracy of sound replication?

Research Questions

How do methods in audio synthesis and replication (DDSP, Tone Transfer, and GANs) compare in their accuracy and efficiency in replicating instrumental tones?

Sub-question #2:

How stable are different methods in audio synthesis and replication in handling noise in the audio input and adversarial attacks?

Research Questions

How do methods in audio synthesis and replication (DDSP, Tone Transfer, and GANs) compare in their accuracy and efficiency in replicating instrumental tones?

Sub-question #3:

What role does the choice of audio input's hyperparameters play in the performance of the model in terms of replicating various instrumental tones?

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Technology

Technologies we will use/look into

- Generative Adversarial Networks (GANs)
- Differentiable Digital Signal Processing (DDSP)
- Neural Networks, particularly:
  - Convolutional Neural Networks (CNNs)
  - Recurrent Neural Networks (RNNs)
- ... and more that we will mention later
- ..... and more that we may find over the course of getting more RRLs (we have so many RRLs)

ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

Methodology

Very rough methodology

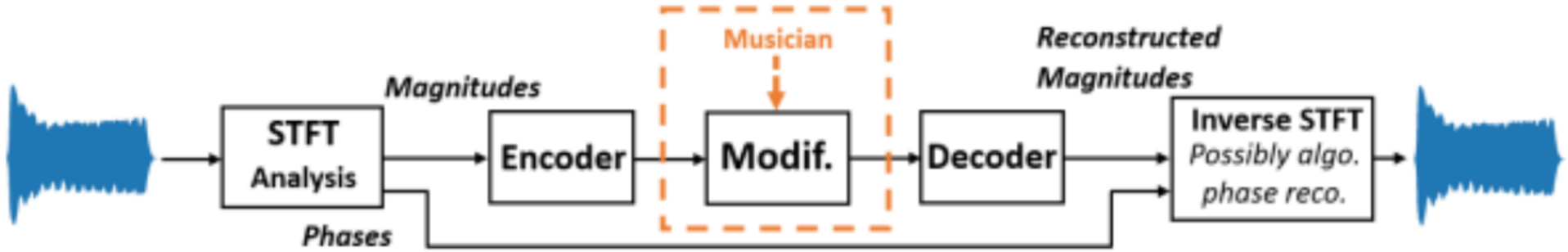


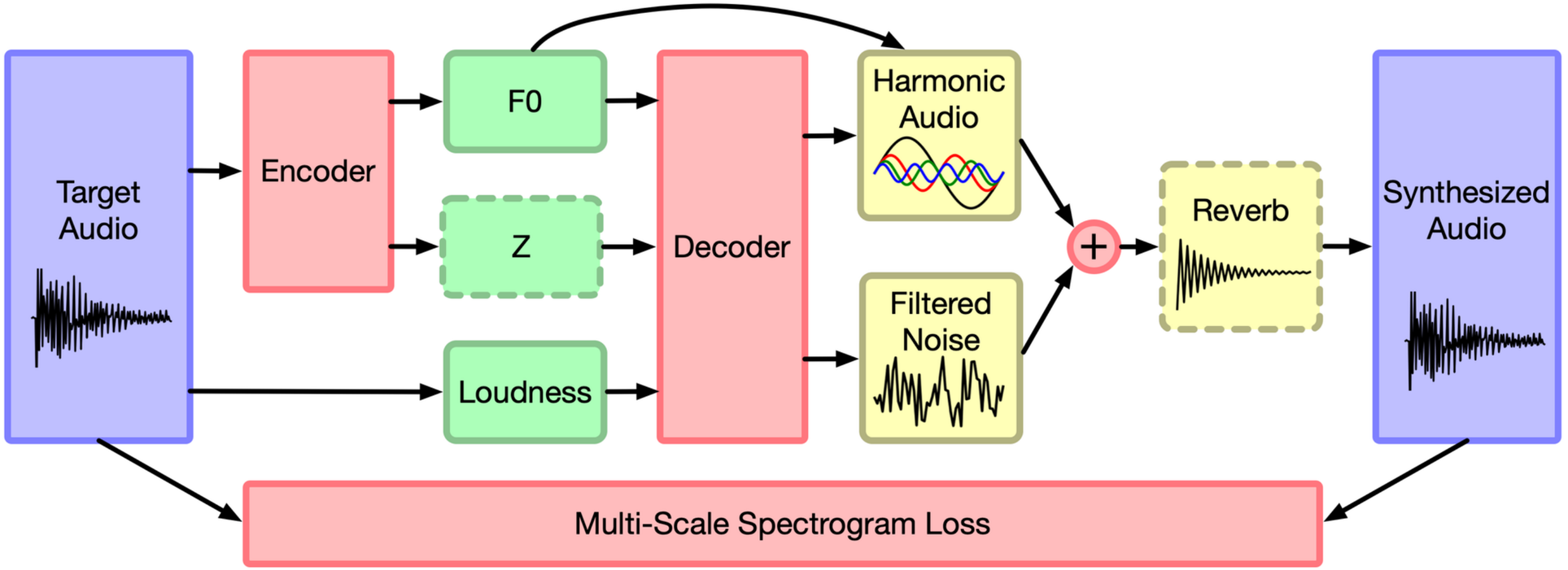
Figure 3.1: Global diagram of the analysis-transformation-synthesis process.

Very rough methodology

- Predicting waveforms (SING, MCNN, WaveGAN)
- Predicting Fourier coefficients (Tacotron, GANSynth)
- Autoregressive waveforms (WaveNet, SampleRNN, WaveRNN)
- Directly generating audio w/ oscillators (DDSP, very new)
- DDSP led to “FM Tone Transfer with Envelope Learning” (most recent paper)

Differentiable Digital Signal Processing (DDSP)

The model has three encoders: f-encoder that outputs fundamental frequency $f(t)$, l-encoder that outputs loudness $l(t)$, and a z-encoder that outputs residual vector $z(t)$.



DDSP Components

- Harmonic Additive Synthesizer + Subtractive Noise Synthesizer (very mathy)
- Filter Design -> Neural Network that:
  - Output vector (Inverse Discrete Fourier Transform is used on the vector components)
  - Audio is divided into non-overlapping frames
- Training uses **ADAM Optimizer**

To apply the time-varying FIR filter to the input, we divide the audio into non-overlapping frames $\boldsymbol{x}_l$ to match the impulse responses $\boldsymbol{h}_l$. We then perform frame-wise convolution via multiplication of frames in the Fourier domain: $\boldsymbol{Y}_l = \boldsymbol{H}_l \boldsymbol{X}_l$ where $\boldsymbol{X}_l = \mathrm{DFT}(\boldsymbol{x}_l)$ and $\boldsymbol{Y}_l = \mathrm{DFT}(\boldsymbol{y}_l)$ is the output. We recover the frame-wise filtered audio, $\boldsymbol{y}_l = \mathrm{IDFT}(\boldsymbol{Y}_l)$, and then overlap-add the resulting frames with the same hop size and rectangular window used to originally divide the input audio.

FM Tone Transfer

1. Envelope Dataset Generation
    a. Making dataset w/ randomized **velocity, note, duration**
    b. Randomized ADSR with **Yamaha DX7 emulator**
2. Envelope Learning
    a. "...we train a Recurrent Neural Network model $g()$ to learn the correspondences between the features a, f and the controls ol reflected in the dataset."
3. FM Tone Transfer
    a. Feature Extraction
    b. Control Prediction

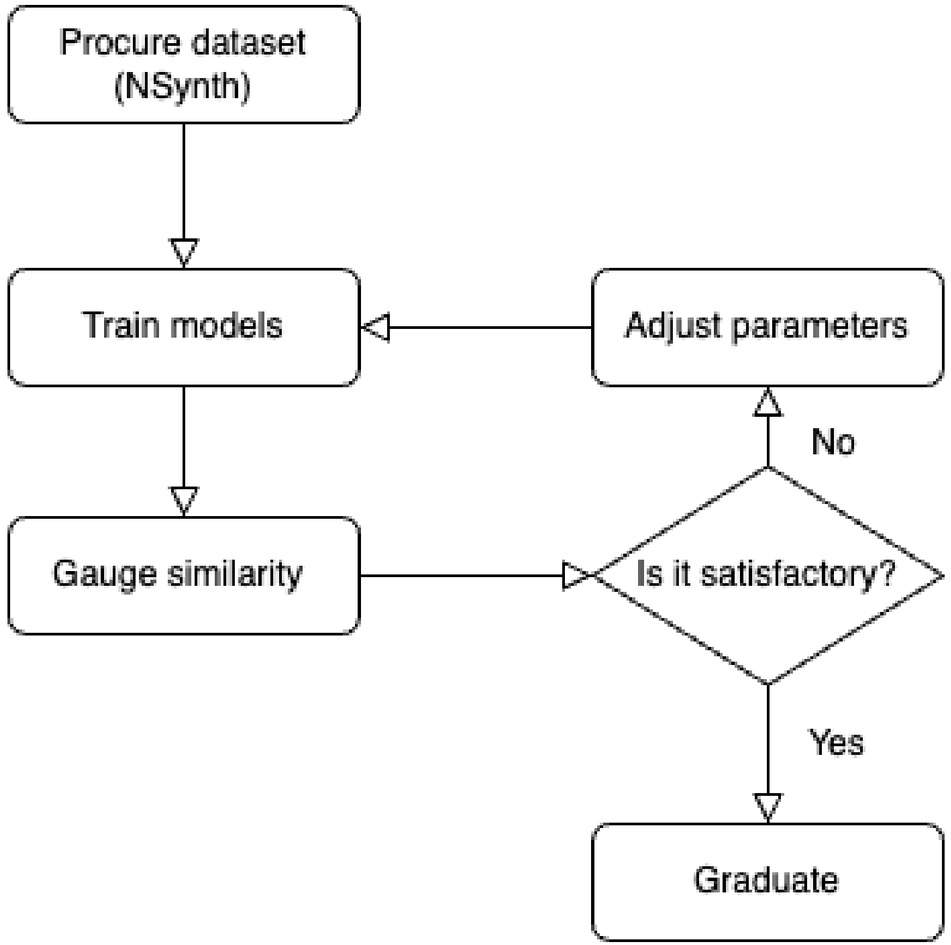
(2) Control Prediction, we use our neural network g_ϕ to infer a set of frame-wise FM synthesis controls, the oscillator output levels $\hat{ol}_k$, from the conditioning signals $\hat{a}_k$ and $\hat{f}_k$.

$$\hat{ol}_k = g_\phi(\hat{a}_k, \hat{f}_k) \quad (3)$$

Technology / Terms

- Fourier Transforms
- GRUs
- MLPs
- Autoencoders
- CREPE <-- pitch detector used that could be useful
- Multi-scale spectral loss
- As an ES major I do not know what the differences are between RNNs, GANs, CNN. Thank you

Very rough methodology



ATENEO COMPUTATIONAL
SOUND AND MUSIC
LABORATORY

THANK

you.