End-to-end equalization with convolutional neural networks

Marco A. Martínez Ramírez, Joshua D. Reiss

Published in: *International Conference on Digital Audio Effects (DAFx) 2018*

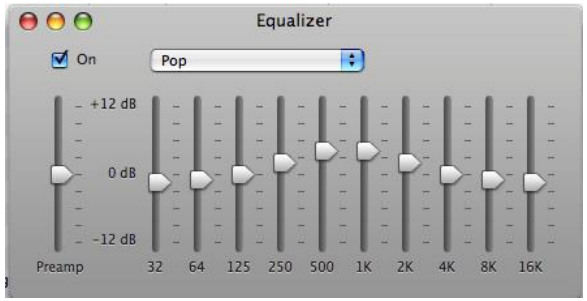
Centre for Intelligent Sensing
Queen Mary University of London

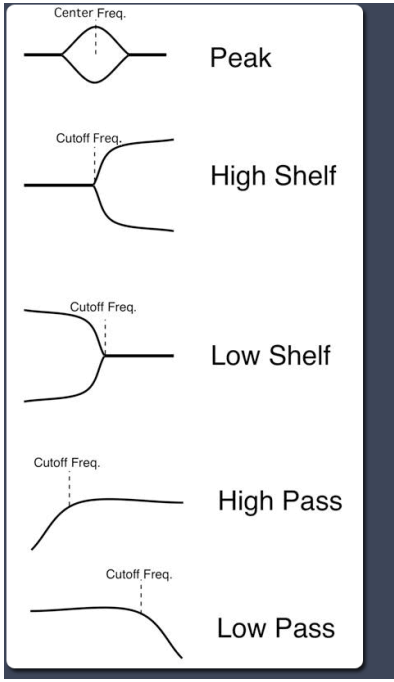
Introduction

- Equalization (EQ)
  - Audio effect widely used in the production and consumption of music.
  - Modification of frequency content through positive or negative gains which change the characteristics of the audio.
  - Recording, Mixing, Mastering, Home Audio Systems.

Introduction

- Equalization (EQ)
  - Audio effect widely used in the production and consumption of music.
  - Modification of frequency content through positive or negative gains which change the characteristics of the audio.
  - Recording, Mixing, Mastering, Home Audio Systems.





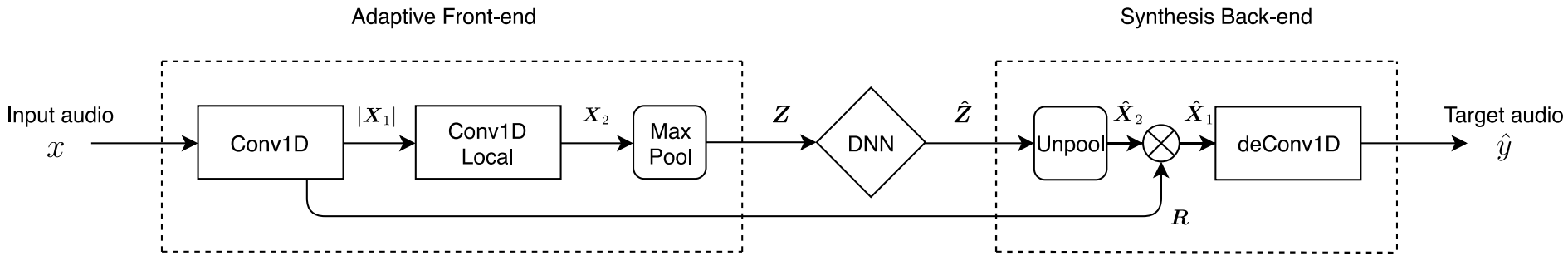
Task

- Given an **arbitrary** EQ configuration, our task is to train a deep neural network to learn and apply the specific transformation.

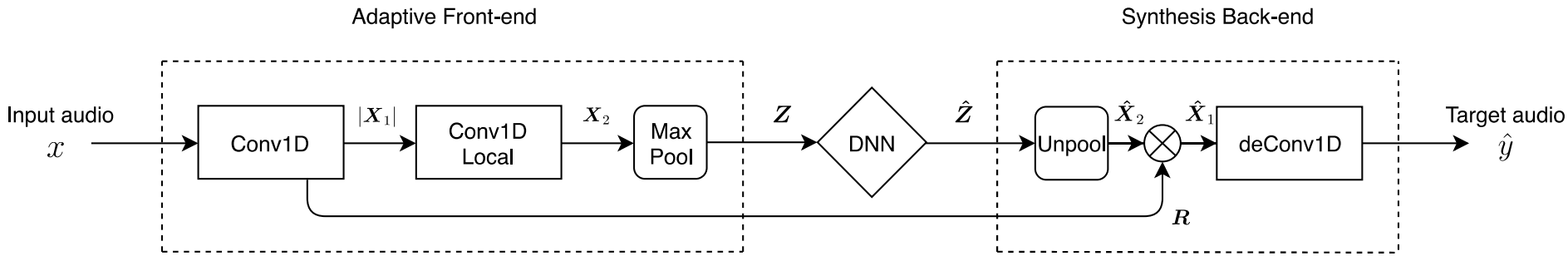
- Using an **end-to-end architecture**, where raw audio is both the input and the output of the system.

- We attempt to find a general deep learning architecture to perform audio processing in the context of matched equalization.

Model


Adaptive Front-end
Synthesis Back-end
Input audio
x
Conv1D
$|X_1|$
Conv1D Local
X_2
Max Pool
Z
DNN
$\hat{Z}$
Unpool
$\hat{X}_2$
$\hat{X}_1$
deConv1D
Target audio
$\hat{y}$
R

Model



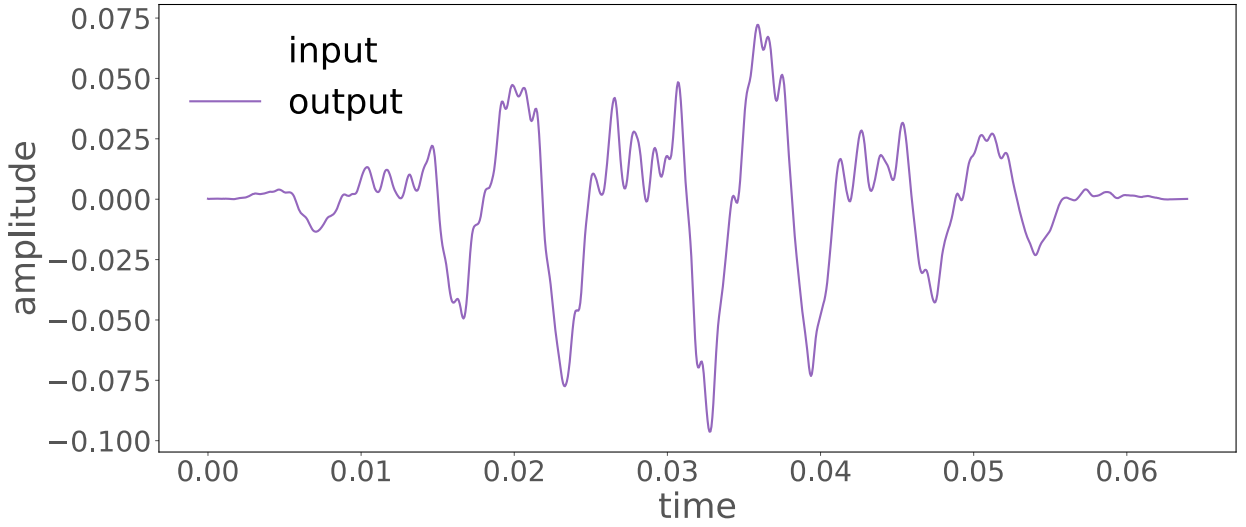
- Adaptive Front-end
  - 2 CNN layers, 1 pooling layer and 1 residual connection.
  - **Conv1D**: 128 filters of size 64, *absolute value* function.
  - **Conv1D Local**: 128 filters of size 128, softplus.
- Latent-space DNN
  - 2 layers of locally connected and fully connected dense layers
- Synthesis Back-end
  - 1 unpooling layer and 1 CNN layer.
  - **deConv1D**: deconvolution operation; implemented by transposing Conv1D filters.

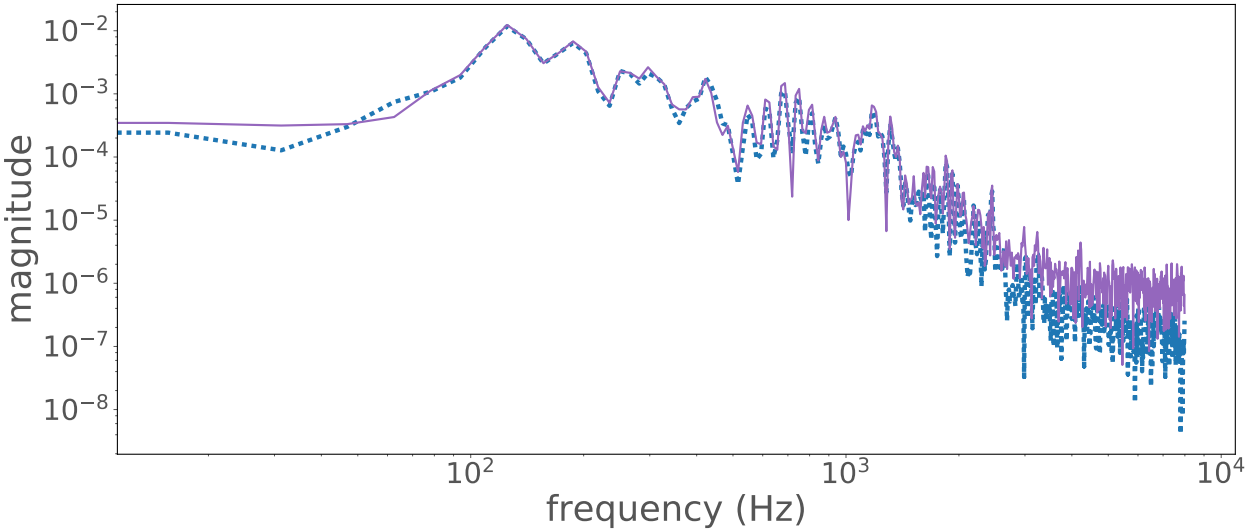
Training

- The training of the model is performed in two steps.
  - The first step is to train both the adaptive front-end and the synthesis back-end for an *unsupervised learning* task.
  - The second step consists of an *end-to-end supervised learning* task based on a given EQ target.

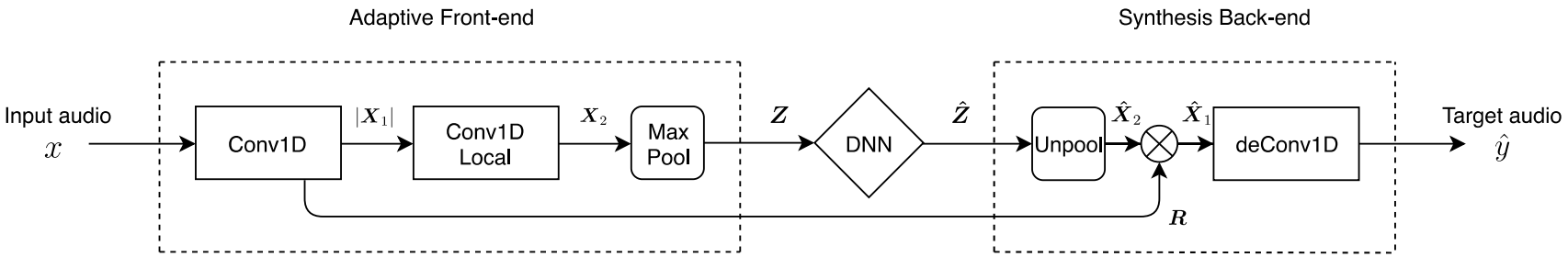
- The unsupervised and supervised learning steps were performed for each type of EQ target. Then, the models were tested with samples from the test dataset.

Results

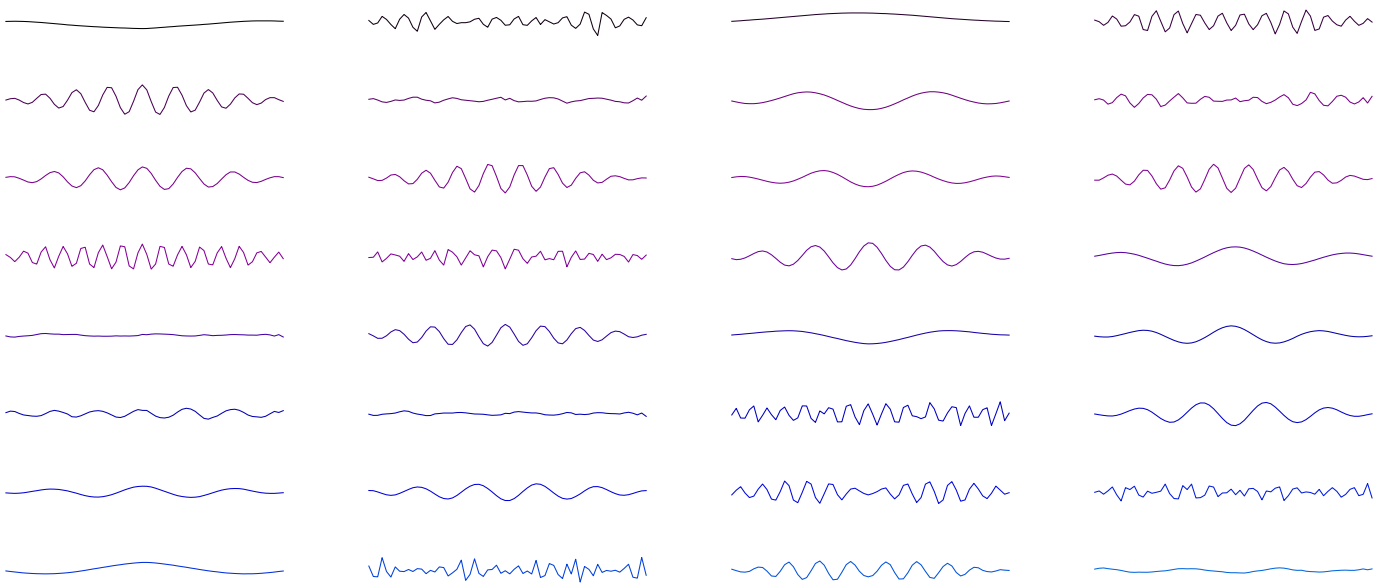

input
output
amplitude
time
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
0.00
0.01
0.02
0.03
0.04
0.05
0.06


magnitude
frequency (Hz)
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{2}
10^{3}
10^{4}

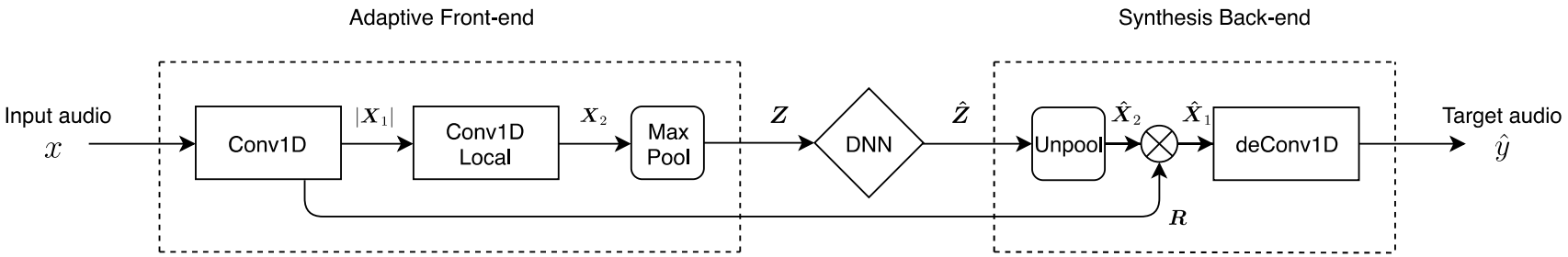
Results



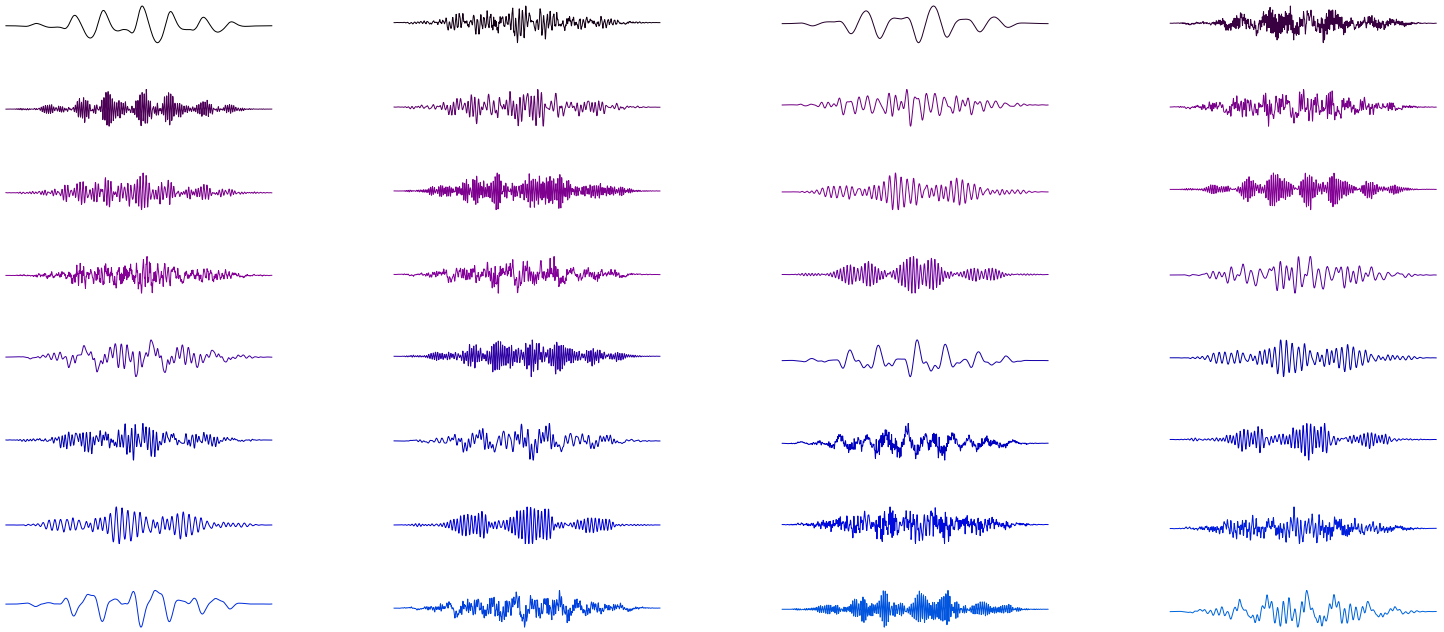
Conv1D filters



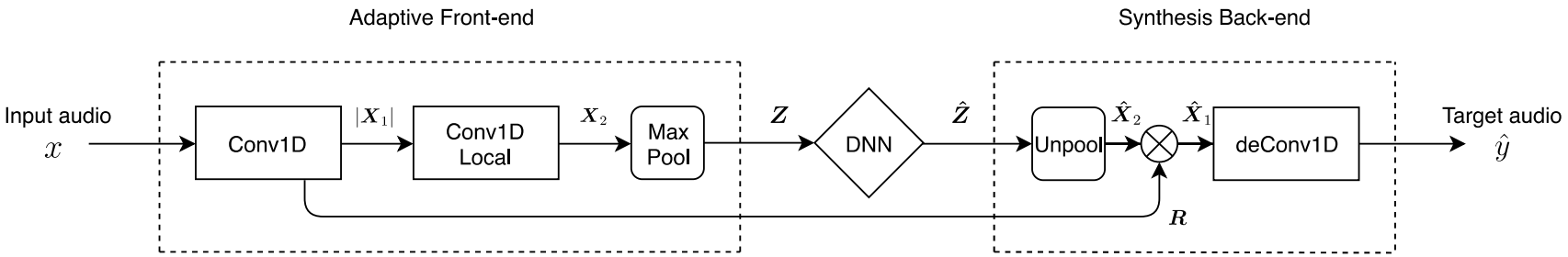
Results


Adaptive Front-end
Synthesis Back-end
Input audio
x
Conv1D
$|X_1|$
Conv1D Local
X_2
Max Pool
Z
DNN
$\hat{Z}$
Unpool
$\hat{X}_2$
$\hat{X}_1$
deConv1D
Target audio
$\hat{y}$
R

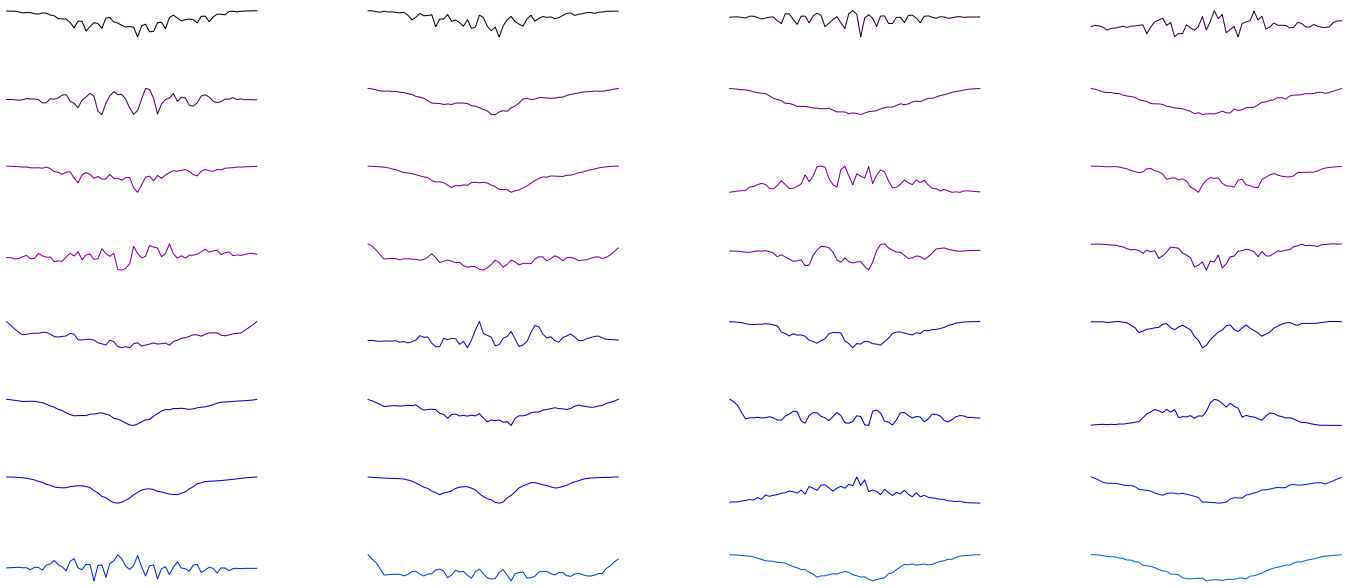
X_1



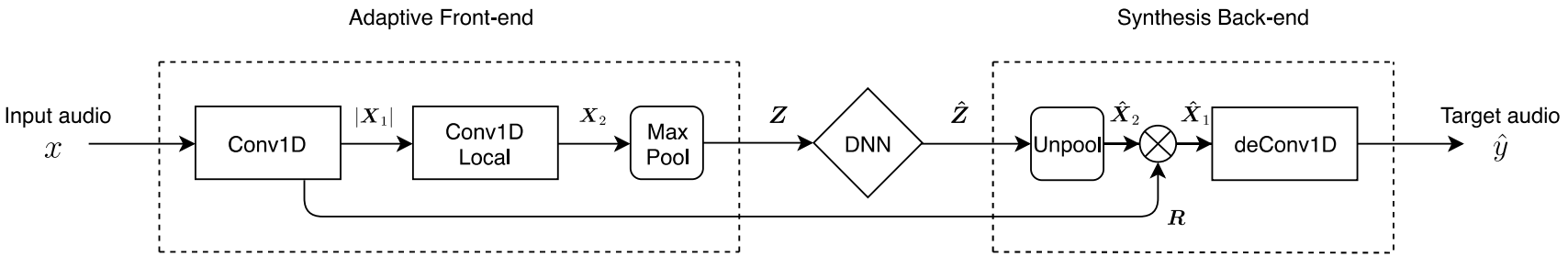
Results


Adaptive Front-end
Synthesis Back-end
Input audio
x
Conv1D
$|X_1|$
Conv1D Local
X_2
Max Pool
Z
DNN
$\hat{Z}$
Unpool
$\hat{X}_2$
$\hat{X}_1$
deConv1D
Target audio
$\hat{y}$
R

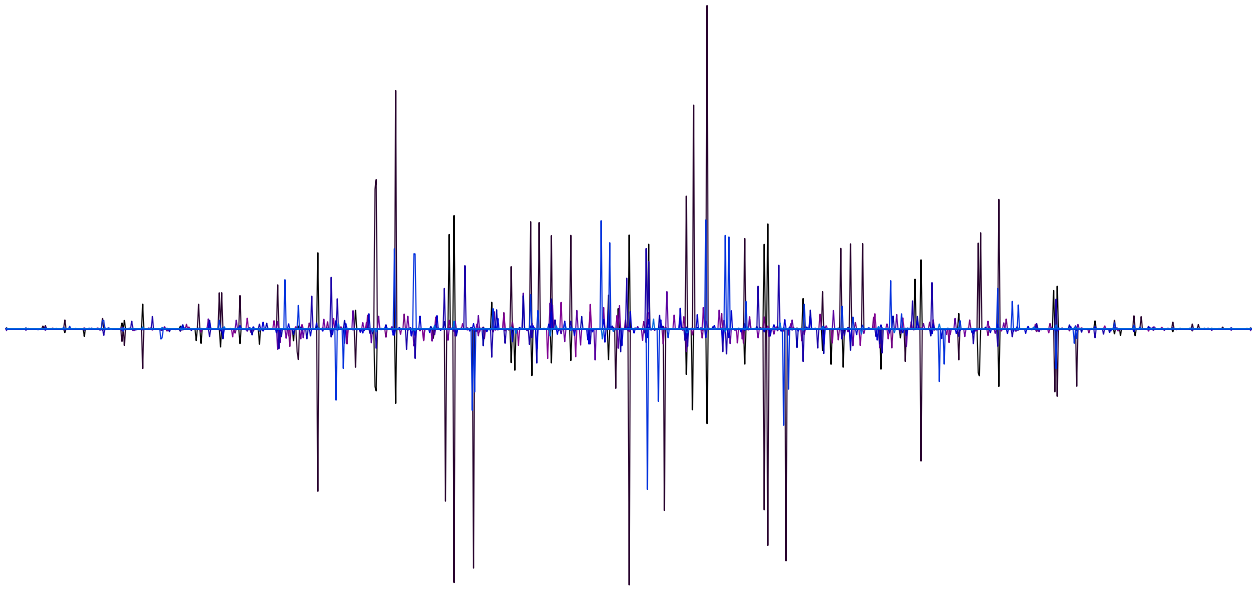
Z



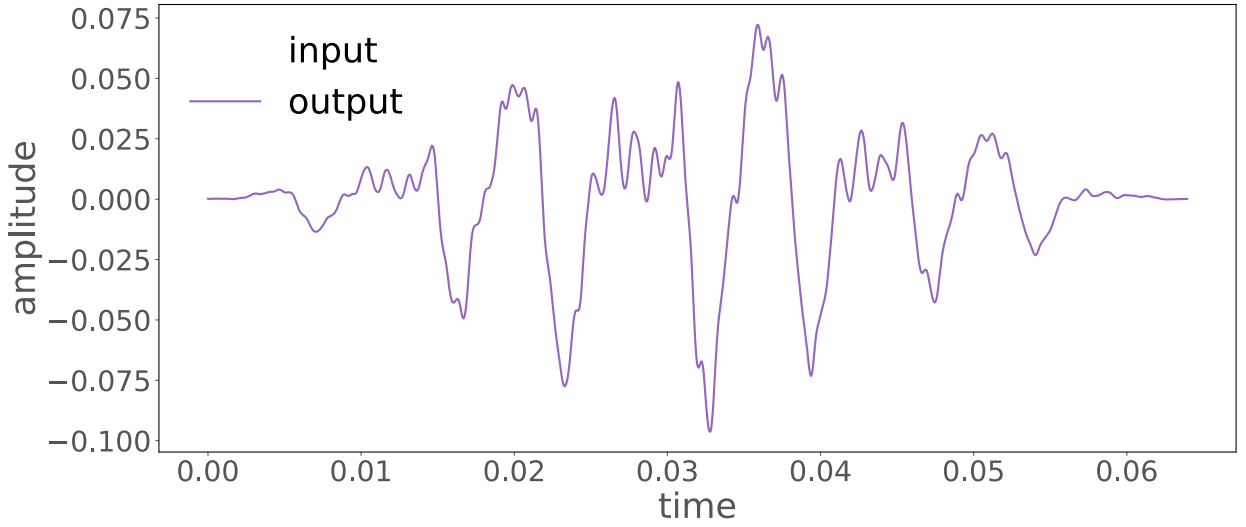
Results

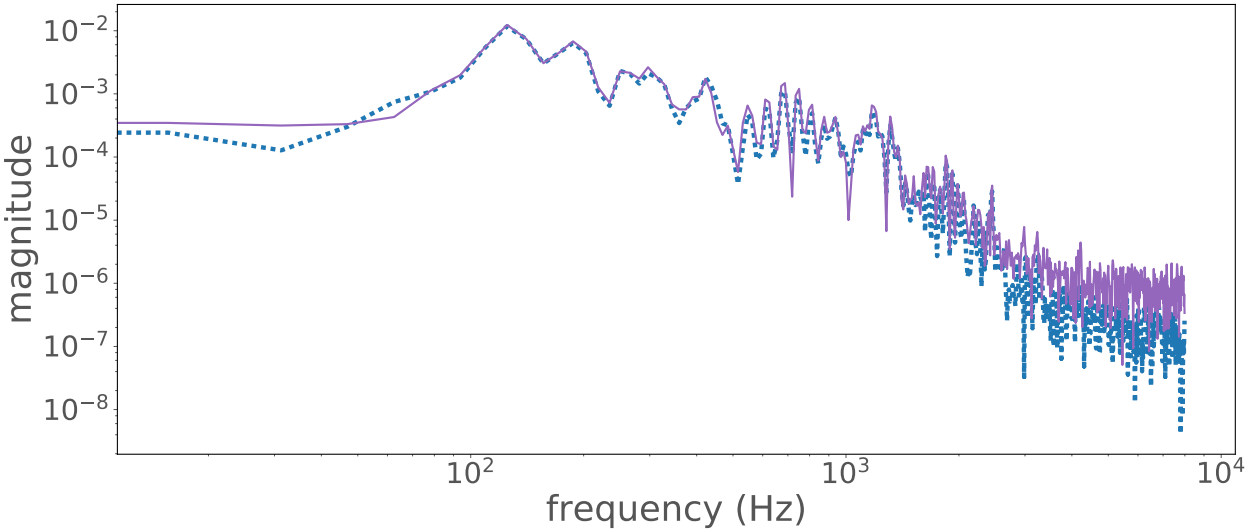


$\widetilde{X_1}$



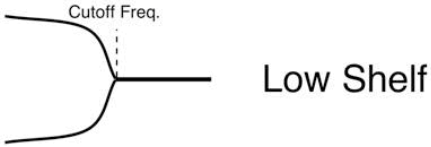
Results

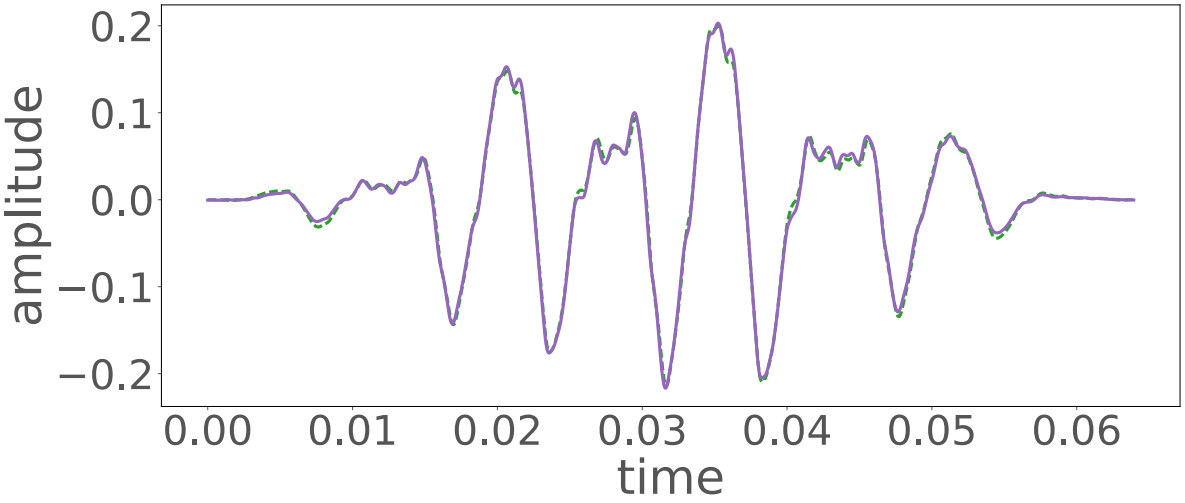

input
output
amplitude
time
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
0.00
0.01
0.02
0.03
0.04
0.05
0.06

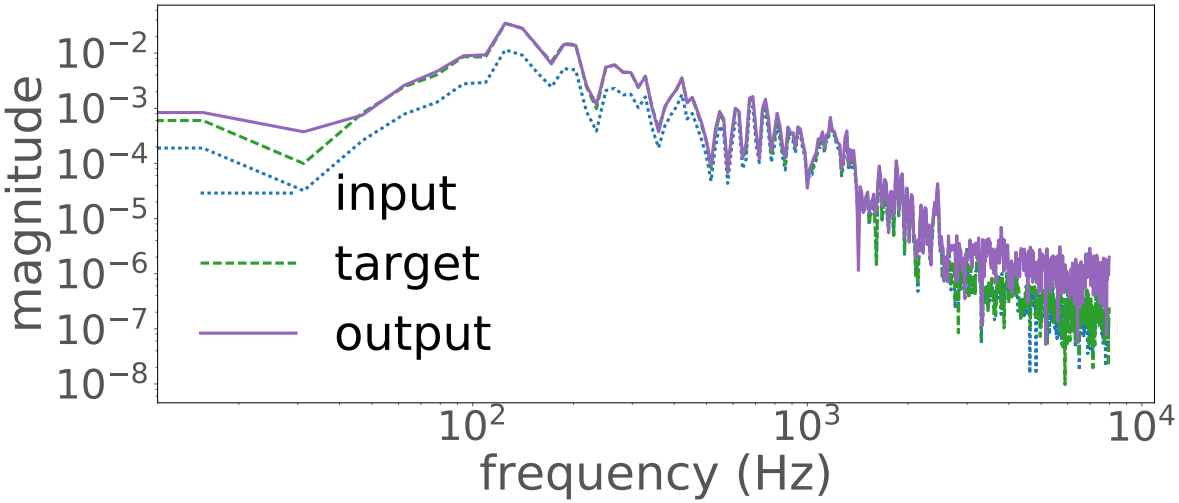

magnitude
frequency (Hz)
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{2}
10^{3}
10^{4}

Results

Shelving

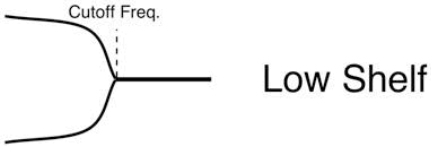






Results

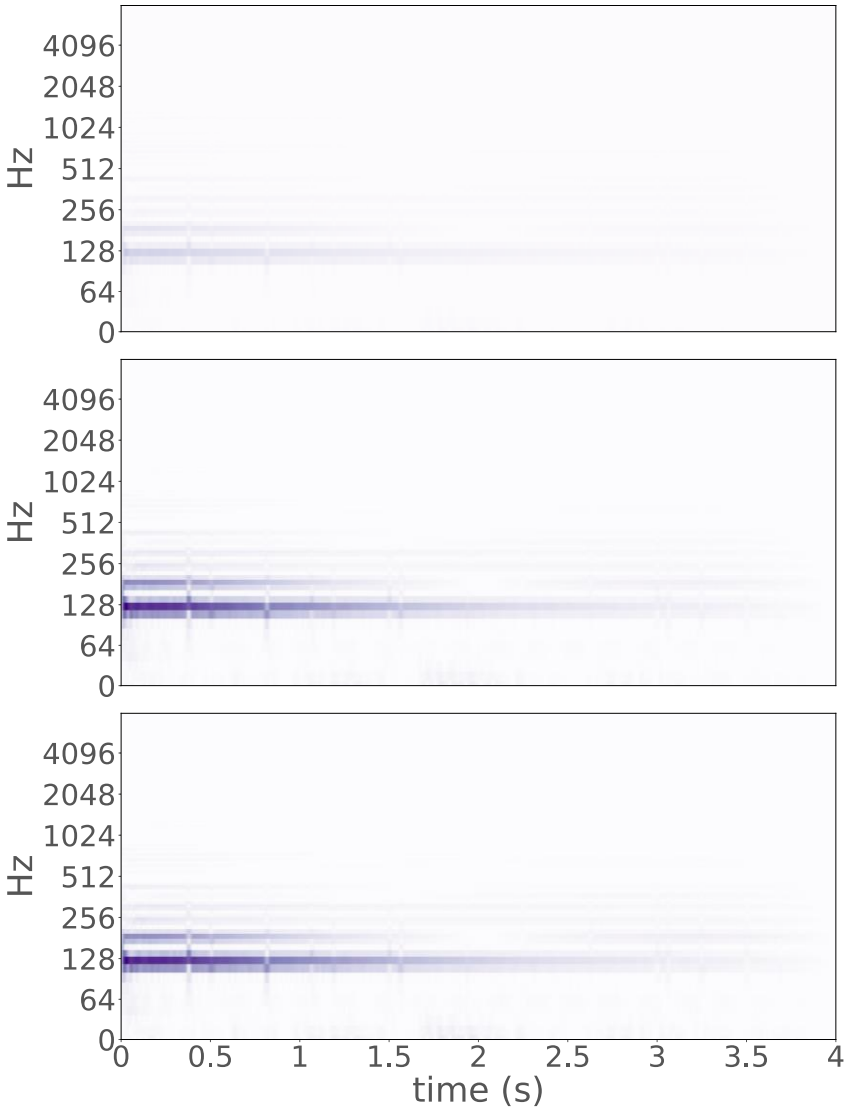
Shelving



Input

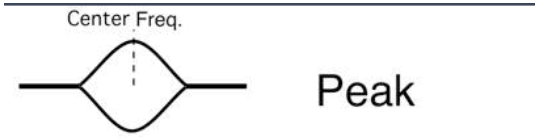
Target

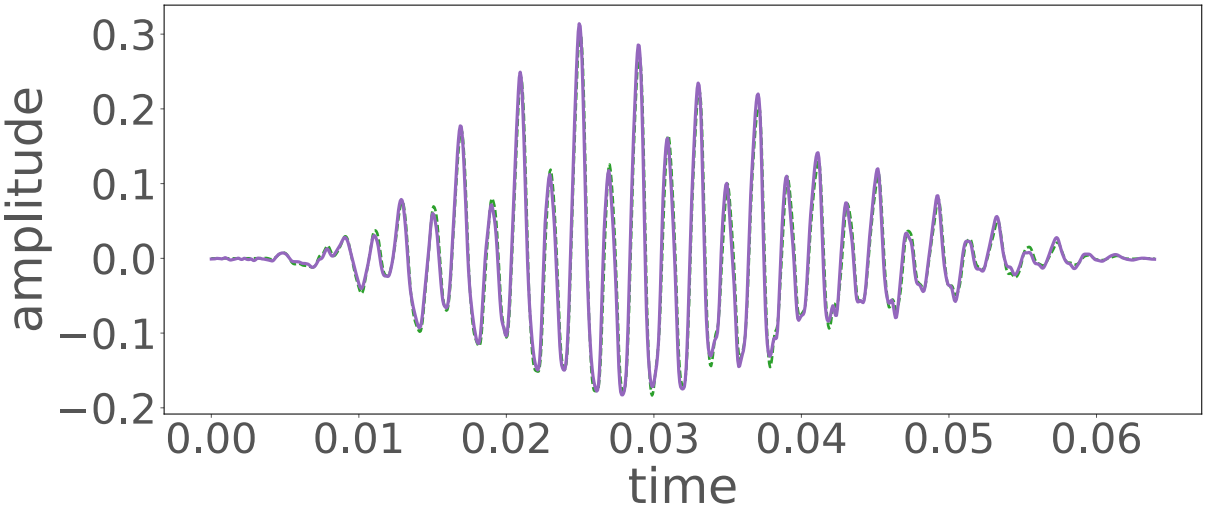
Output

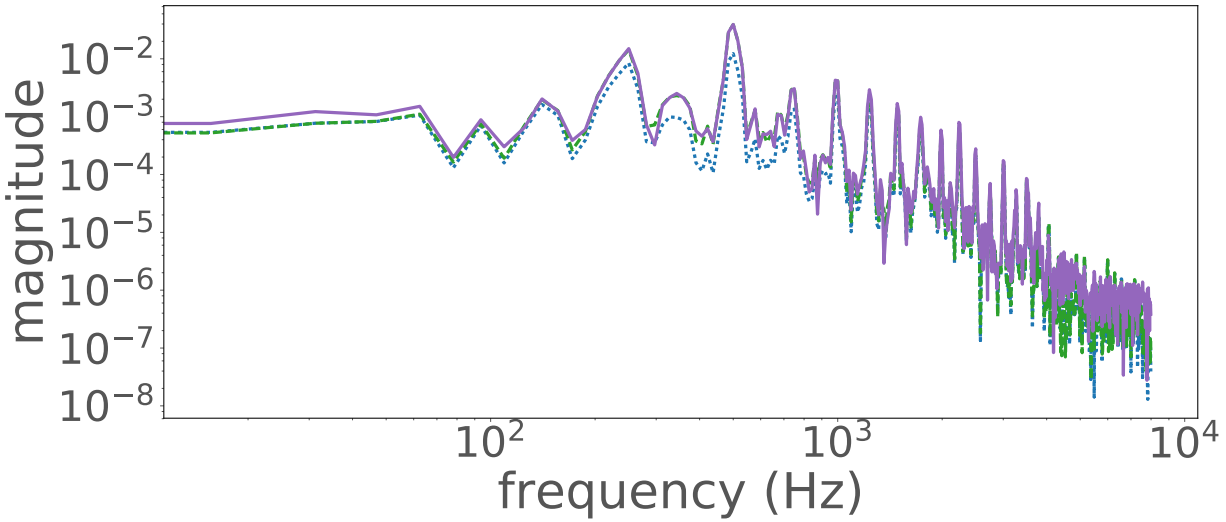


Results

Peaking

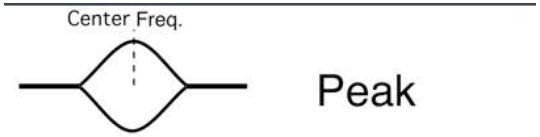






Results

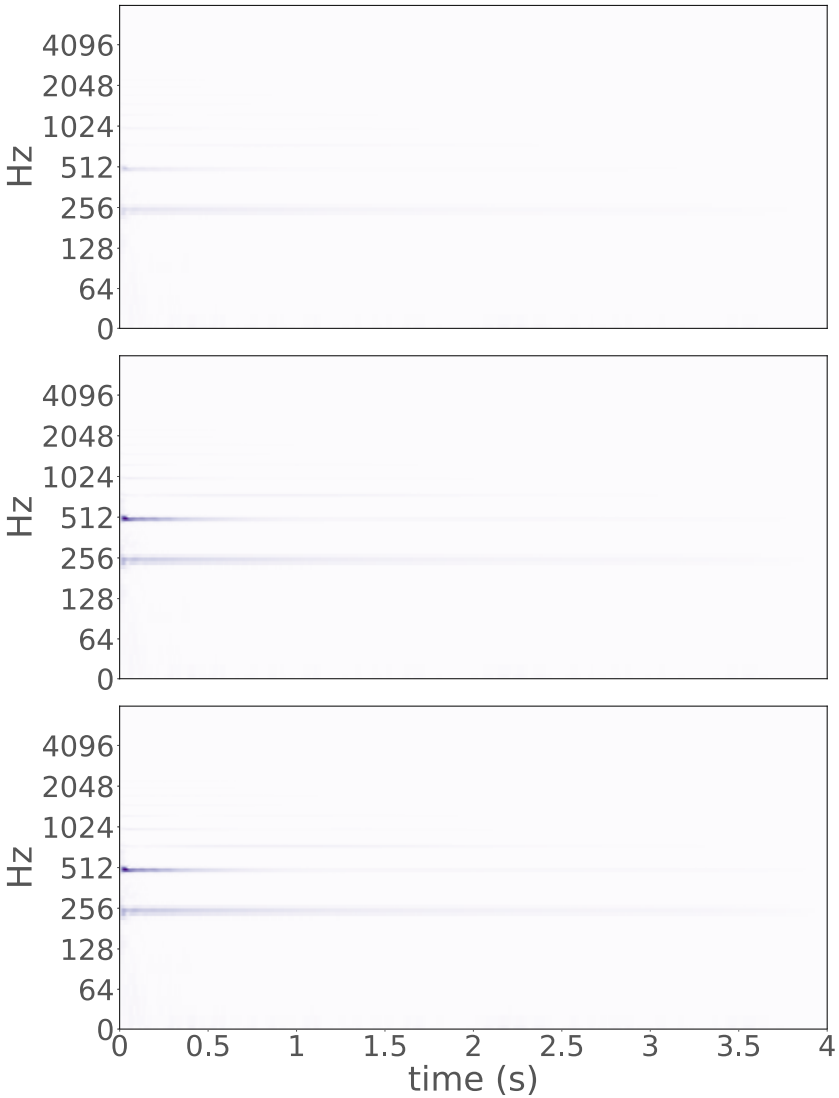
Peaking



Input

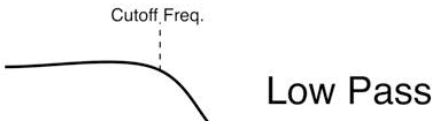
Target

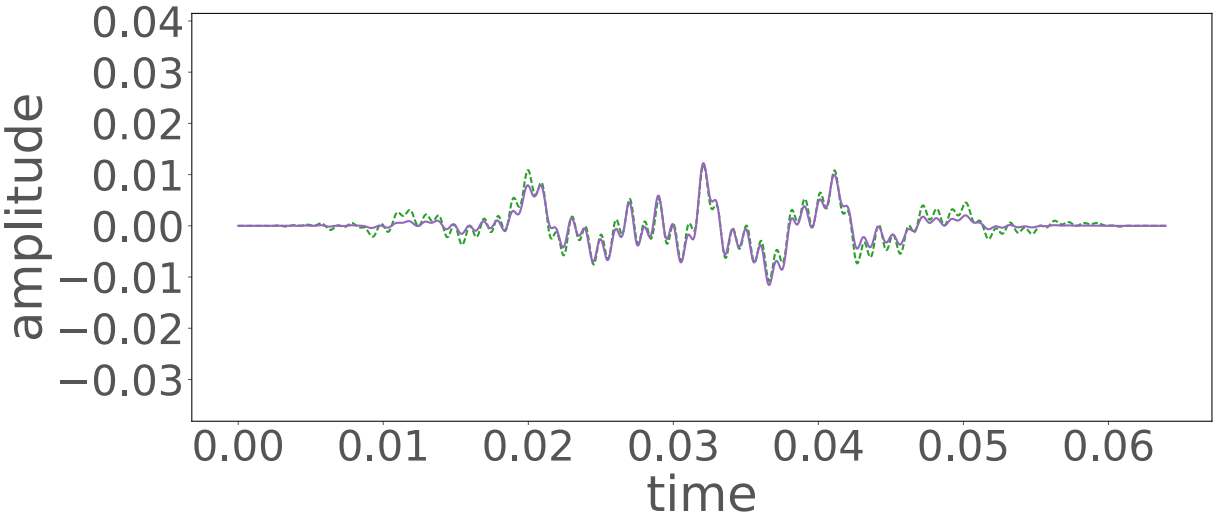
Output

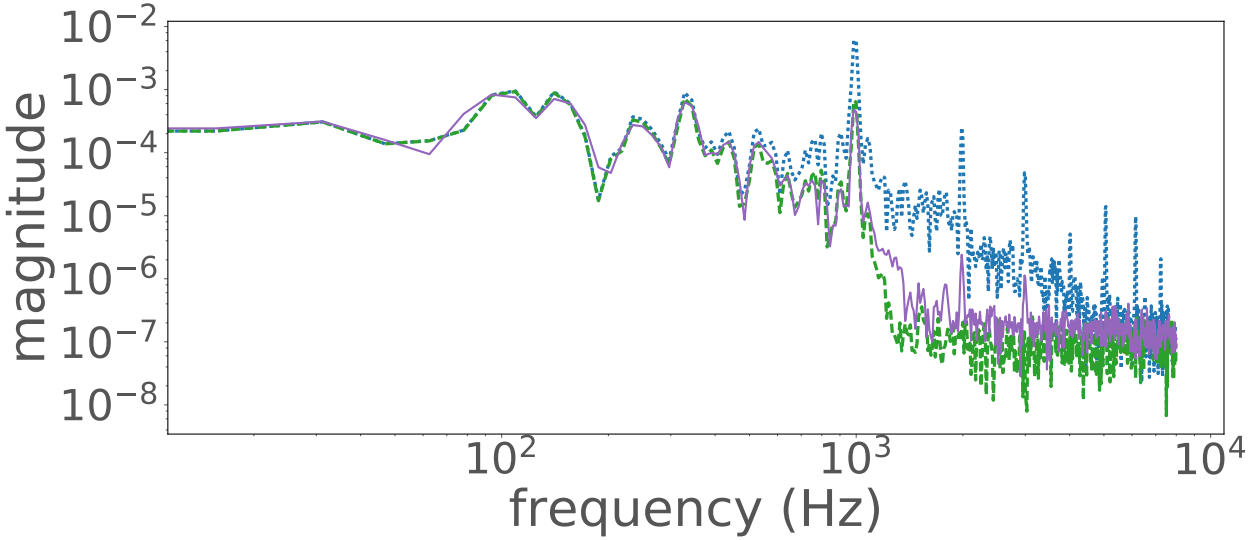


Results

Lowpass

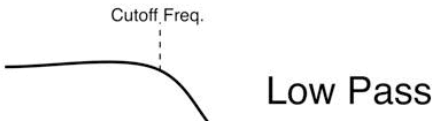






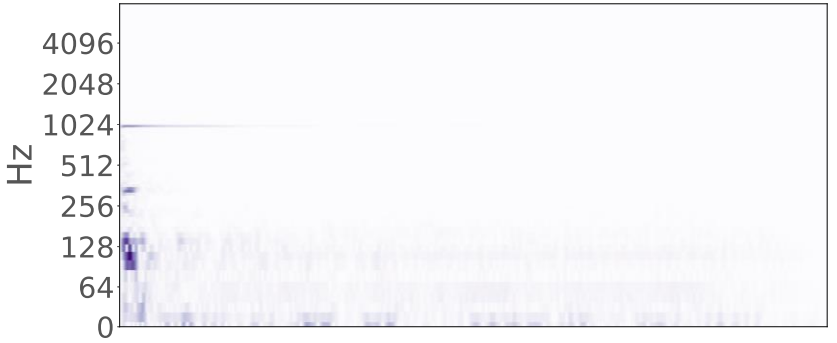
Results

Lowpass

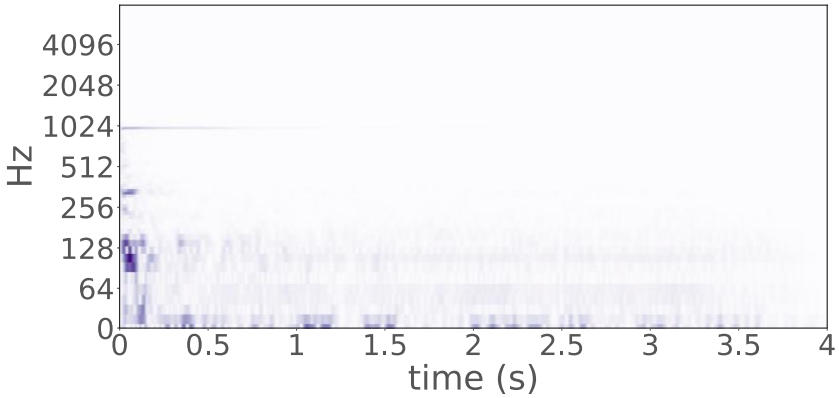


Input

Target

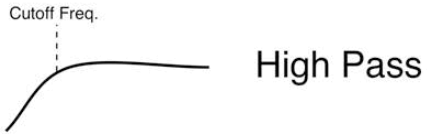


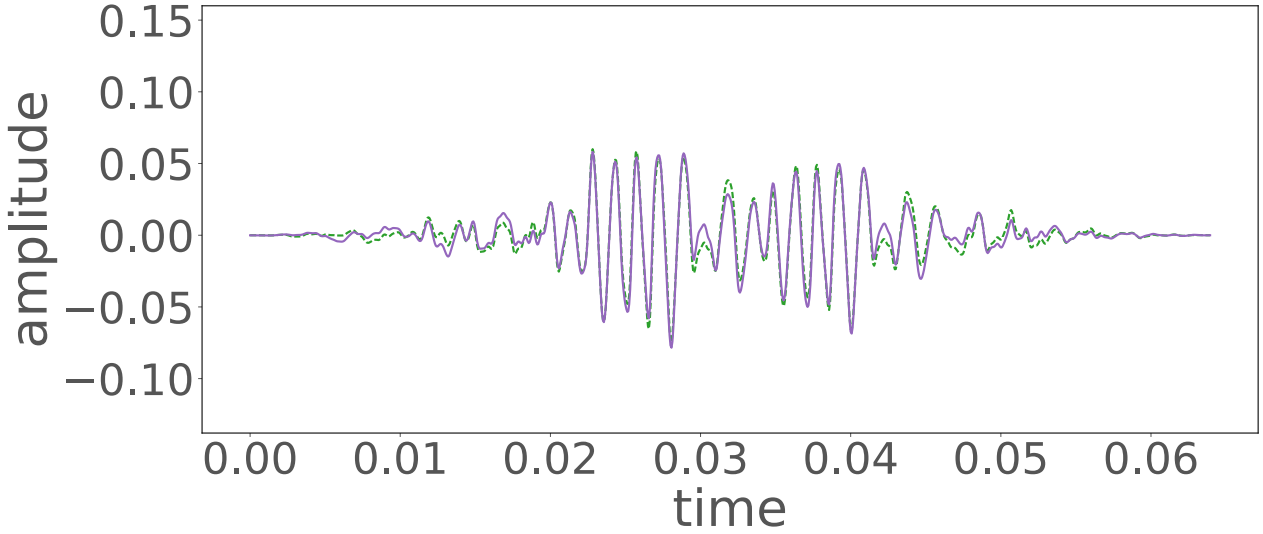
Output

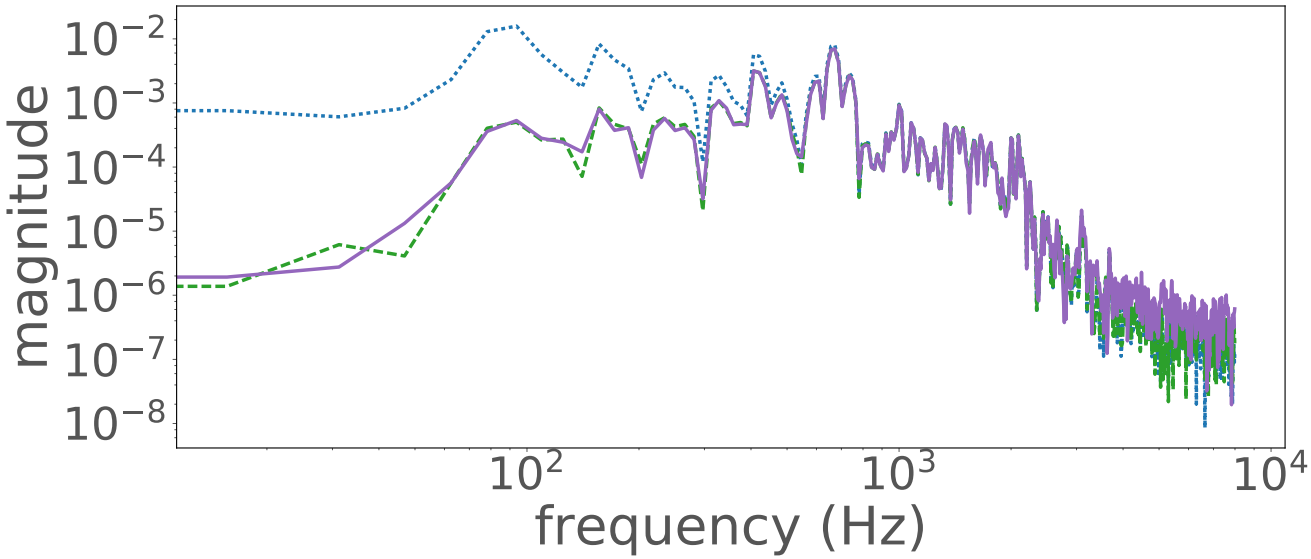


Results

Highpass

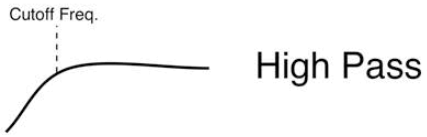






Results

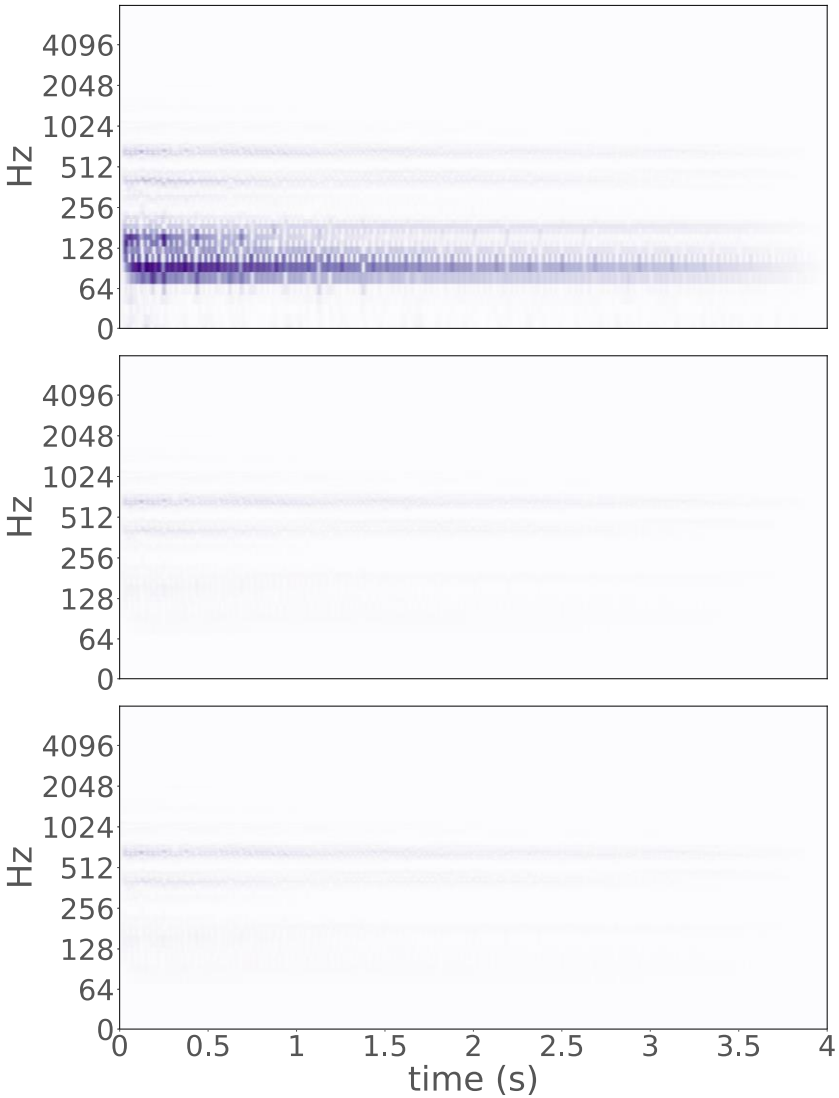
Highpass



Input

Target

Output



Future work

- Style-learning of a specific sound engineer.
- Model much more complex audio effects.
- Generalization capability.