Creation of Audio Effects Using Deep Learning

Author: Ivan Kazakov
Supervisor: Petr Grinberg

Faculty of Computer Science, HSE

April 16, 2025

Introduction

Audio effects can be described as the controlled transformation of a sound based on some control parameters. They are:

- Used to shape acoustics, tone, timbre, and other characteristics to change the perception of transformed sound
- Widely utilized in music production and sound design industries
- Implemented via analog devices or plug-ins in digital audio workstations (DAWs)

(a) Analog devices

(b) Plugins in contemporary DAW

Motivation

The deep learning approach helps overcome the following limitations of standard audio effect modeling methods:

- Modeling audio effects often requires knowledge of the analog circuitry, but detailed analysis of the target device is not always possible
- Allows to create a model that is able to generalize from one effect to another, thus, streamlining the development

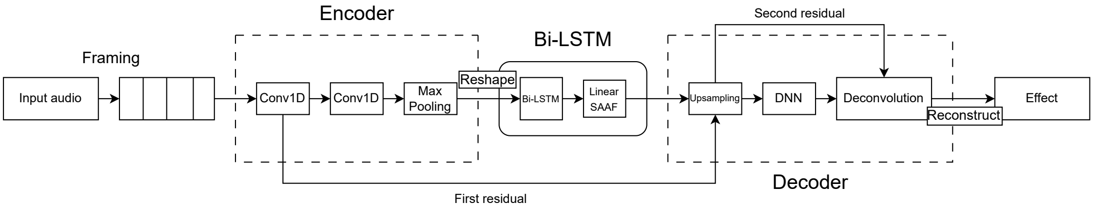
Objective

The objective of this coursework is to employ the deep learning approach to emulate audio effects.

Methodology: baseline

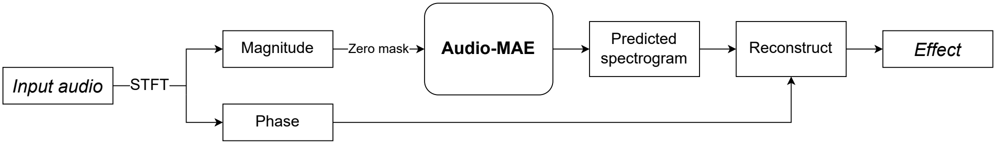
Our first approach is based on the existing autoencoder-based solution proposed by Martinez et al. We choose this model because it is:

- Capable of emulating a broad range of time-varying and nonlinear audio effects
- Lightweight: easy to train and run



Methodology: Audio-MAE

We extend the ideas from the first approach and consider the more novel Transformer-based autoencoder:



Experiments and results: comparison with the reference

We compared our implementation with the reference using the energy normalized mean absolute error metric (MAE) $\downarrow$ that was proposed in the original paper:

Audio Effect	Reference	Lightweight	Audio-MAE
	MAE	MAE	MAE
Chorus	0.0190	0.0615	0.0564
Distortion	Not provided	0.0906	0.2101
Tremolo	0.0100	0.0341	0.0139

The performance of the implemented model was found to be comparable to that of the reference, but there may be a slight mismatch between our data and those in the reference, because the authors of the original paper did not specify the sample partition index.

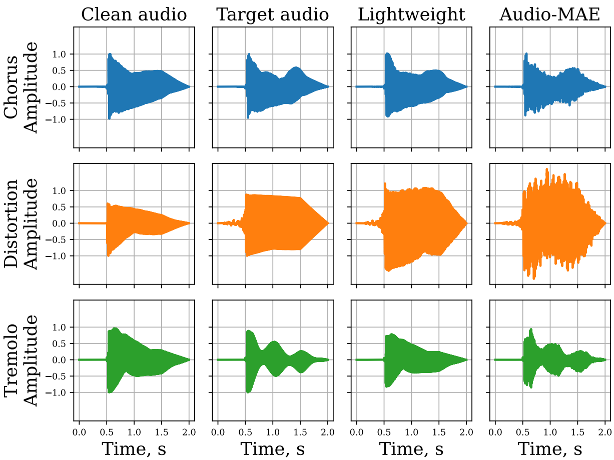
Experiments and results: quantitative analysis

We employed additional metrics (all $\downarrow$) to further explore the quality of the implemented models:

Audio Effect	Lightweight					Audio-MAE				
	CF	STFT	SCE	RMS	MSE	CF	STFT	SCE	RMS	MSE
Chorus	**0.9158**	**0.0638**	**0.0007**	**0.8503**	0.2709	2.0421	0.0761	0.0027	1.6041	**0.1536**
Distortion	**6.8552**	**0.0623**	**0.0068**	**1.5204**	0.5557	16.0442	0.1232	0.0145	2.7963	**0.3778**
Tremolo	3.6102	0.0939	0.0024	2.3386	0.4134	**1.4498**	**0.0548**	**0.0009**	**1.1328**	**0.0998**

Experiments and results: qualitative analysis

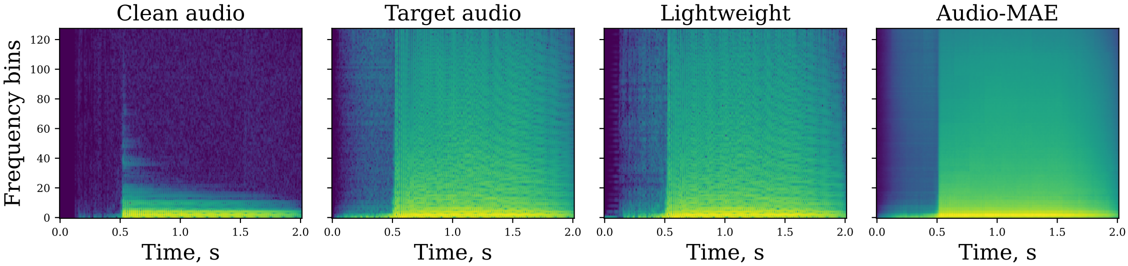
We conducted qualitative analysis for each effect:



Demo samples are available at `https://ylxsbn.github.io/demos.html`.

Experiments and results: distortion

We found that the Audio-MAE model struggles to extract features from the distortion spectrogram, so the time-domain approach can be advantageous in this case:



Conclusion

In this coursework:

- The lightweight model from the article was implemented
- The pipeline of a more novel Transformer-based approach was reworked for the purpose of creating audio effects
- Chorus, tremolo, and distortion audio effects were emulated using both models

It was seen that:

- Novel approach showed better results for chorus and tremolo, however, the distortion effect was better for the lightweight approach
- Time domain can be better for some tasks

Future work:

- Larger and more diverse datasets
- Other loss functions
- Diffusion models