Interpolation Based Neural Audio Synthesis using Convolutional Autoencoders

Benedikt Langer, BSc

MASTERARBEIT

eingereicht am

Fachhochschul-Masterstudiengang

Mobile Computing

in Hagenberg

im Juni 2023

Advisor:

FH-Prof. DI Stephan Selinger
Alexander Palmanshofer, BSc MSc

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere. This printed copy is identical to the submitted electronic version.

Hagenberg, June 27, 2023

Benedikt Langer, BSc

Contents

Preface

Abstract

This should be a 1-page (maximum) summary of your work in English.

Kurzfassung

An dieser Stelle steht eine Zusammenfassung der Arbeit, Umfang max. 1 Seite. ...

Chapter 1

Introduction

Chapter 2

Related Works

Despite this technology is not that well explored and popular as in the image domain, there exist a few proposed approaches that have developed a rather good solution. Some of these approaches have proven, that with Neural Networks it is possible to generate synthesized audio up to a certain quality.

2.1 Audio Synthesis

Probably one of the most prominent solution is from Engel et al. [1]. With their work they have proposed a system that is capable of synthesizing audio as well as interpolating/morphing encoded audio data of two instruments to create new audio. Not only they have proposed a system, but also a public available Dataset called “NSynth” that contains a large scale of high quality musical notes. The latter has been used for training of this specific project. In their work regarding the synthesis, Engel et. al. developed and compared two different approaches with two differ- ent kind of networks. Nevertheless they have a similar structure, as they are both designed as Autoencoders but accept different kinds of data and thus have different components. While the one kind of network operates on time domain data the other one is trained on the spectral representation of audio samples. Throughout their work the second technology using spectrograms is referenced/used as Baseline Model as they focus on the use of so called “WaveNet Autoencoders” that are trained on continuous time signal. With using the Autoencoder-Structure they make use of its ability in learning efficient encodings of the music data. These encodings are representing essential features from the original audio. To create new sounds they take the encoded data from the embeddding space of two instruments and interpolate them linearly. In addition they used the decoder part to reconstruct it back to audio data. With this mechanisms they were able to create some new sounds which contain the characteristics of two different audio signals. The re- sult can be explored via their online AI-Experiment called “Sound Maker”.

2.2 Audio Style Transfer

Another approach is proposed in a paper by Ramani et al. [5]. In their approach they developed a Neural Network that is also, like the previous paper, based on an Autoencoder Architecture. As also the title says, they speak officially about their system as “Audio Style Transfer Algorithm”. The process of generating an audio containing characteristics of two audio signals is here slightly different as in the work of Engel et al.. as they use in order two networks, namely a transformation network and a loss network. At first the transformation network transforms an input spectrogram into a stylised spectrogram whereas it is using the pre trained weights from the loss network. The latter is subsequently used to calculate the style-loss but also content-loss between the respective spectrograms and the output from the transformation network. This loss gets minimized by back- propagation to the transformation network. By this procedure it is possible to pass a single spectrogram through the transformation network which in order outputs a new spectrogram containing the characteristics of itself (content) but also of one other style audio. To be also mentioned due to its architecture it also performs really fast and could be used for real-time use.

Verma et al. also
Another source is coming from [4]

2.3 Image Style Transfer

<In this section briefly mention the approaches of Gatys and Johnson for image style transfer and it’s relation to audio style transfer [2, 3]>

Chapter 3

Approach

Chapter 4

Experiment

Chapter 5

Results

Chapter 6

Discussion/Evaluation

Chapter 7

Conclusion

Chapter 8

Future Work

Appendix A

Technical Details

Appendix B

Supplementary Materials

List of supplementary data submitted to the degree-granting institution for archival storage (in ZIP format).

B.1 PDF Files

Path: /

thesis.pdf	Master/Bachelor thesis (complete document)

B.2 Media Files

Path: /media

*.ai, *.pdf	Adobe Illustrator files
*.jpg, *.png	raster images
*.mp3	audio files
*.mp4	video files

B.3 Online Sources (PDF Captures)

Path: /online-sources

Reliquienschrein-Wikipedia.pdf	**WikiReliquienschrein2022**

Appendix C

Questionnaire

Appendix D

LaTeX Source Code

References

Literature

[1] Jesse H. Engel et al. "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders". *CoRR* abs/1704.01279 (2017). arXiv: 1704.01279. URL: http://arxiv.org/abs/1704.01279 (cit. on p. 2).

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. "Image Style Transfer Using Convolutional Neural Networks". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2414–2423. DOI: 10.1109/CVPR.2016.265 (cit. on p. 3).

[3] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European conference on computer vision*. Springer. 2016, pp. 694–711 (cit. on p. 3).

[4] Xuehao Liu, Sarah Delany, and Susan Mckeever. "Sound Transformation: Applying Image Neural Style Transfer Networks to Audio Spectograms". In: Aug. 2019, pp. 330–341. DOI: 10.1007/978-3-030-29891-3_29 (cit. on p. 3).

[5] Dhruv Ramani et al. "Autoencoder Based Architecture For Fast & Real Time Audio Style Transfer". *CoRR* abs/1812.07159 (2018). arXiv: 1812.07159. URL: http://arxiv.org/abs/1812.07159 (cit. on p. 2).

Check Final Print Size

— Check final print size! —

width = 100mm
height = 50mm

— Remove this page after printing! —