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. Those approaches can get categorized into different areas, as their workflow and principle differ in certain ways. As this field is related to the technique of image style transfer, a lot of works apply those methods to audio (spectrograms) and therefore call it explicitly "Audio Style Transfer". This is also because those solutions, are specifically defining a content and a style sound to combine, but more on that in section 2.2. Those methods who don't use this principle of content and style, can get categorized to the technique of "Neural Audio Synthesis" or simply just audio synthesis (see 2.1).

2.1 Neural Audio Synthesis

Neural Audio Synthesis is the field of creating/synthesizing novel sounds with the help of neural networks. The problem is similar and related to the field of Audio Style transfer. Like mentioned before, approaches in this domain differ in certain ways in those from style transfer. As a major difference, with Neural Audio Synthesis, no content or style sound is specified, which means, that for the creation of novel sounds, two sound sources are used equally. While Audio Style Transfer gets also a lot applied on whole audio samples or musical pieces, in synthesis the focus is more on the application for single notes. With a special look onto Autoencoder-Networks, Neural Audio Synthesis also includes the tasks of learning important sound features for compression and recreation of the input data. On how different approaches are designed, which (machine learning) techniques and which results could be obtained, will be shown in the following section.

Probably one of the most prominent solution, in the field of Neural Audio Synthesis, comes from Engel et al. [4]. With their work "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders" 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. Using this mechanisms they were able to create some new sounds which contain the characteristics of two different audio signals. The result can be explored via their online AI-Experiment called “Sound Maker”.[1] Furthermore can be mentioned at this point, that for their purpose they also created a huge dataset of audio samples (>306 000) that is open for public use.

The work by *Natsious et al.* does not explicitly mention the term Neural Audio Synthesis in its title, but deals with it throughout the article. [11] In their work they do a research on the reconstruction capacity of (stacked) convolutional audio encoders in terms of log-mel-spectrograms and carry out experiments on different configurations. In their experiments they evaluate the effectiveness of autoencoders in terms of neural audio synthesis whereas also possible improvements through additional techniques are measured. As they mention that with their work an exploration on musical timbre compression is made, the synthesis gets specifically refered to timbre synthesis. As audio spectrograms exist with different scales, this approach uses in contrast to others, the log-mel scale. They prove it beneficial, as it already captures the most significant properties with the effect of consuming less memory and computational power. For the training they also used the NSynth-Dataset proposed by *Engel et al.* [4], whereas just a sub-sample consisting of samples of different instruments of one single pitch was considered. The model(s) that where used throughout their experiments, followed the general structure of an (stacked) convolutional autoencoder network, which consists mainly of 2D convolutional layers. For experimental reasons, additional layers and techniques such as pooling layers, fully connected layers, dropout, kernel regularization got applied (added/removed). To measure the results of their experiments they where using error metrics such as root mean squared error (RMSE), structural similarity index (SSIM) but found out that those cannot accurately say something about the quality. Because of this reason, they also introduced a precision and recall score but also combined it in a F1_score. In order to generate from the spectrograms sounds, they were reusing the preserved phase information, unless there was no modification of the embedding. In the latter case the Griffin Lim phase estimation algorithm was applied, as no phase information is present.
Regarding the results that could be obtained by running these experiments by recon-

[1] “Sound Maker” https://magenta.tensorflow.org/nsynth-instrument

structing spectrograms (without modification in latent space), some interesting findings could be extracted. To their surprise, by reducing the size of the latent space, they found out that the smaller it is, more accurate spectrograms with a smoother distribution could be generated. Also in some cases where L1 or L2 regularization got applied, the spectrograms were more accurate, while with dropout layers no improvement could be achieved. The use of (max) pooling layers also resulted in a more accurate time-frequency resolution with less noise, then with just convolution layers. Finally removing the fully connected / dense layer showed that without it, the quality was significant better, as spatial information gets better preserved.

Regarding audio synthesis, *Colonel et al.* proposed over the year a few works, where they investigated the suitability of autoencoder networks regarding this task. [2, 3, 9] Starting in 2017 they proposed an autoencoder based audio synthesis through compression and reconstruction of audio spectrograms. In contrast to the before mentioned approaches, this one uses an autoencoder based on fully connected layers without convolutions. Also a different dataset was used, as they generated it themselves using an own synthesizer. A difference to e.G. the NSynth dataset used in other approaches, is that it also contains polyphonic notes and thus more complex harmonies. During the experiment they trained different parameterized networks, where they vary the depth and width of the network and its layers as well as the activation functions and used different optimizers. As error metric in this work the mean squared error (MSE) was used. Comparing these scores regarding networks of one or two hidden layers on each side show, that using the Adam optimizer worked out best in contrast to using Momentum as optimizer. These networks, just using sigmoid activation functions worked best when less compression is applied. Having 4 hidden layers, they found out, having a mix of ReLU and sigmoid activation functions worked out best. To mention here also, by applying regularization methods such as dropout and l2 penalty, that the latter was proven better as the results where of better auditive quality. Some more interesting results that could get obtained, where that having sigmoid activations led to fuller sound then with ReLU. Furthermore by using bias terms introduced noise in the results, whereas despite of the better convergence, they chose to let them out. In the end they came to the result, that using the network with 4 hidden layers and a composition of sigmoid and ReLU worked out best also in terms of auditory quality.

Another work by *Colonel et al.* was proposed in 2018, which actually states an improvement of the method, described in their previous work from 2017. [2] Those improvements contain the use of a phase reconstruction method not used before, which allows in this method to directly activate the latent space. Furthermore to improve the models convergence, the autoencoder was designed asymmetrically, via input augmentation. As in the previous work only MSE was contemplated as error metric, this one made use and comparison of several cost functions.

Work by [2, 3, 7, 9, 13]

2.2 Audio Style Transfer

The works in this section have in common that they all entitle their work, as “Audio Style Transfer”. In their methodology they all orientate themselves at technique of image style transfer. Applying the method of image style transfer to audio also means, as audio is a time-continuous signal, that it has to be brought into a similar shape, which will be done mostly by generating spectrograms out of signals. As for image style transfer, a content and a style picture is needed, this principle also gets applied to audio style transfer. In image style transfer, the style (e.g. brush strokes, colors) and content of an other image (e.g. contours, scenery) get combined, to form a new stylised image. [5] This means that in the output image, the content image looks like painted with a certain “style”. Mapping this principle to the audio domain, this means, that there has to be a specific content sound (sample) that gets stylized with a certain style of a sound (e.g. style of a specific instrument). As in the image domain distinguishing content from style is already difficult, it is also a big or even bigger question that appears in the different approaches. Most of the time when defining the style of a certain audio, the authors define it as a musical instruments’ timbre or even a musical genre. Alongside this a content might get defined as global music structure containing rhythmic constraints. [6] Those questions also might be influenced if whole audio samples/musical pieces might be taken to get stylised or just some single notes from an instruments. Furthermore if as audio data, speech is considered, style and content also is different defined. Here style could be e.g. the emotion of the voice or the speakers identity and content the spoken words in an sample. The following works show different solutions specific to the problem of Audio Style transfer in which they also get compared and assessed.

One approach that applies this principle, is the solution proposed by Ramani et al. in 2018. [12] In their approach they developed a Neural Network that is constructed as an (convolutional) Autoencoder Architecture (like in the work from Engel et al.). 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. Both networks have the same structure and composition of layers. The loss network is trained to compress input spectrograms to lower dimensions which means that the encoder part learns to preserve the high level features of the input. In addition the decoder learns to reconstruct from the compressed data a spectrogram similar to the input of the network. For the training of the transformation network, the pre-trained weights of the loss network are used which speeds up training (just optimization towards low level features/style). Having the trained transformation network, it then is able to transform an input spectrogram into a stylised spectrogram. The loss network 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. presented in their paper in 2018 a new machine learning technique for the purpose of generating novel sounds [14]. In this approach they tried to apply the method for artistic image style transfer to audio where they specifically mentioned the approach proposed by Gatys et al.[5] (see section 2.3). Unlike to Gatys, they adapted and trained an AlexNet architecture on the classification of audio-samples. This kind of network is a so called convolutional neural network, whereas the audio therefore gets converted into spectrograms, as those can be seen as grey-scale images. An important note here is that in this work they used the log-magnitude data of the STFT output. Also to mention, they adapted the network to use a smaller receptive size (kernel) of 3x3 instead of the larger ones in the original network, as they claim that it retains the resolution of the audio. As in the image domain the stylised output image gets initialized with random noise, they also use here an input spectrogram consisting of a gaussian noise signal. The random noise spectrogramm afterwards gets iteratively optimized by minimizing the content- but also style loss via back-propagation. In the end this process creates a spectrogram containing the content of one audio with the style of one other audio sample. They also found out that including additional loss terms for temporal and frequency energy envelopes, helped to improve the quality, as otherwise temporal dynamics would not get incorporated. For their experiments they imposed the style of a tuning fork onto a harp sound and also transferred the style of a violin sound onto a sample of a singing voice. In this way they developed a novel method for achieving cross-synthesis by using image style transfer methods.

More work in that field is coming from *Liu et al.* [10] which also explored the application of technologies given from the image domain for “mixing audio”. This also means, that this approches focuses on using audio as spectrograms. As the previous work solely investigated on the one technique by Gatys et al. this one explores two more approaches in addition to compare the results. While one of those two additional is inspired by Johnson et al. which is a convolutional autoencoder coupled with a VGG classification network the other one uses an approach with GAN (Generative Adversarial Network). In their work they called Gatys’ approach specifically slow transfer, as the iterative computation from gaussian noise was proven really slow. In contrast to the previous work by Verma et al., they used for the “slow transfer” method an adapted VGG network (1 input channel in first layer instead of 3) which has also been used in Gatys’ image style transfer. The transfer process is also similar to the previous work, as they use a spectrogram initialized as gaussian noise to iteratively minimize the content loss (in the higher layers) and the style loss (lower layers). Using this one as baseline model, they also adapted a faster style transfer method as coupling the VGG network with a convolutional autencoder network. The purpose of this network is to take as input the content spectrogram and outputting a spectrogram containing also the style features of a style spectrogram. Comparing it to other approaches this is very similar to the one of Ramani et al. having a transformation network. The only difference is the second network as here they are using a VGG classification network and no second autoencoder. Having the output of the autoencoder network (also called generative network) this one is the initial spectrogram on which the content and style loss gets computed in the VGG network (just like previously with gaussian noise). The gradient descent then will get applied to the autoencoder network, resulting after few iterations, in a stylised spectrogram. They have proven that this approach is way faster than the one with gaussian noise. As already mentioned before, for the third experiment they adapted a cycleGAN to accept audio spectrograms instead of images. In the image domain this kind of network is able to apply style transfer to only a portion of the input images. Also when using this method, two new sounds are calculated (in both directions). They also mentioned, that this approach generates the results in a shorter amount of time. For comparison, they listen to the outcome but also apply objective mechanisms like visual assessment of spectrograms, consistency tests with classification and examination of signal clusters. With the baseline approach e.g. the harmonic is not clear and high frequencies are discarded, also the faster transfer emphasizes on lower frequencies but is missing out on beginnings of the notes. With cycleGAN also the lower frequencies get emphasized while higher ones get discarded. The listenable results of each approach are provided online.[2]

As the already mentioned approaches are working on single notes/sounds, the work of *Grinstein et al.* has been implemented for whole audio samples [6]. Within their work they were adapting several other approaches with neural networks from the image domain for his idea. Besides of neural networks, they also implemented a handcrafted sound texture model which got compared to the neural approaches. The latter one is composed of three sound processing steps, that in combination emulates the human auditory system. Taking a closer look on their approach, especially with the neural networks, it can be said that it differs in several ways. On the one hand they do not use a random noise spectrogram, moreover they already use the content spectrogram which then gets stylised through their methods. Most/many approaches that deal explicitly with Audio Style Transfer, are computing the result with a combined loss (function), that incorporates a style loss but also a content loss. *Grinstein et al.* do not make use of this concept, as they already initialize the future stylised spectrogam with the content spectrogram, like mentioned previously. On this one, just the style loss gets optimized, as the content is already present. To mention here, they proved this method to have compelling results, as the global structure of the content sound also is preserved.
With the neural network-based approach they investigated the use of three different network architectures for the purpose of audio style transfer. Concerning all three network types, they minimized the style loss on the content sound/spectrogram. The style loss is equally computed as in Gatys' image style transfer approach, to a "style sound/spectrogram", at specific layers in the network that extract the style. Via back-propagation the loss gets minimized again at each layer, which results after a few iterations in a stylised content sound/spectrogram. This workflow was applied to all three different network types and compared. As the first network they used a VGG-19 network like Gatys, whereas the input spectrogram was replicated three times in order to mach the input shape (RGB-like). By averaging all three channels in the end they obtained the final stylised spectrogram. The second network they used SoundNet which is Convolutional network learned on unlabeld videos including sounds. This type of network operates on the raw waveform wheras no generation of spectrograms has to be done in advance. As

[2] https://www.xuehaoliu.com/audio-show

final network a wide-shallow-random network was used with audio spectrograms consisting of just one-layer CNN (like in the work of Ulyanov and Lebedev [15]). As the fourth and last method they used a handcrafted sound texture model that emulates the human auditory system. Even if its no neural network, it consists of three layers doing cochlear filtering, envelop extraction with compressive non-linearity and modulation filtering.
Having the results of their experiments using those approaches, a comparison could be made. While using the VGG network no meaningful results could be obtained (extremely noisy), the SoundNet yielded more relevant results despite also containing some noise. To their surprise the shallow random network performed best together with the sound texture model. for a bether understanding, they provided their results online.[3]

<write something about [1]

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 [5, 8]>

[3] https://egrinstein.github.io/2017/10/25/ast.html

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] Jiyou Chen et al. “Audio style transfer using shallow convolutional networks and random filters”. *Multimedia Tools and Applications* 79 (2020), pp. 15043–15057 (cit. on p. 8).

[2] Joseph Colonel, Christopher Curro, and Sam Keene. *Autoencoding Neural Networks as Musical Audio Synthesizers*. 2018. eprint: 2004.13172 (eess.AS) (cit. on p. 4).

[3] Joseph T Colonel and Sam Keene. “Conditioning Autoencoder Latent Spaces for Real-Time Timbre Interpolation and Synthesis”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9207666 (cit. on p. 4).

[4] 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 pp. 2, 3).

[5] 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 pp. 5, 6, 8).

[6] Eric Grinstein et al. “Audio Style Transfer”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 586–590. DOI: 10.1109/ICASSP.2018.8461711 (cit. on pp. 5, 7).

[7] Lamtharn Hantrakul et al. “Fast and Flexible Neural Audio Synthesis.” In: *ISMIR*. 2019, pp. 524–530 (cit. on p. 4).

[8] 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. 8).

[9] joseph colonel joseph, christopher curro christopher, and sam keene sam. “improving neural net auto encoders for music synthesis”. *journal of the audio engineering society* (Oct. 2017) (cit. on p. 4).

[10] 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. 6).

[11] Anastasia Natsiou, Luca Longo, and Sean O'Leary. *An investigation of the reconstruction capacity of stacked convolutional autoencoders for log-mel-spectrograms.* 2023. DOI: 10.48550/ARXIV.2301.07665 (cit. on p. 3).

[12] 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. 5).

[13] Fanny Roche et al. *Autoencoders for music sound modeling: a comparison of linear, shallow, deep, recurrent and variational models.* 2019. arXiv: 1806.04096 [eess.AS] (cit. on p. 4).

[14] Prateek Verma and Julius O Smith. "Neural style transfer for audio spectograms". *arXiv preprint arXiv:1801.01589* (2018) (cit. on p. 6).

Online sources

[15] Dmitry Ulyanov and Vadim Lebedev. *Audio texture synthesis and style transfer.* 2016. URL: https://dmitryulyanov.%20github.%20io/audio-texture-synthesis-and-style-transfer (visited on 03/14/2023) (cit. on p. 8).

Check Final Print Size

— Check final print size! —

width = 100mm
height = 50mm

— Remove this page after printing! —