

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

© Copyright 2023 Benedikt Langer, BSc

This work is published under the conditions of the *Creative Commons License Attribution-NonCommercial-NoDerivatives 4.0 International* (CC BY-NC-ND 4.0)—see <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

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 19, 2023

Benedikt Langer, BSc

Contents

Declaration	iv
Abstract	viii
Kurzfassung	ix
1 Introduction	1
1.1 Research questions	2
1.2 Outline of the thesis	2
2 Related Works	3
2.1 Neural Audio Synthesis	3
2.2 Audio Style Transfer	8
2.3 Image Style Transfer	12
3 Approach	13
3.1 Motivation	13
3.2 Overview	14
3.3 Pre-processing	15
3.3.1 Spectrograms and STFT	16
3.4 ML-Model	18
3.4.1 Neural Networks - Introduction	19
3.4.2 Convolutional Neural Networks	19
3.4.3 Autoencoder	20
3.4.4 Optimizer	23
3.5 Synthesis of novel sounds	24
3.5.1 Interpolation in latent space	24
3.6 Post Processing	25
3.7 Dataset	26
3.7.1 NSynth Dataset	26
4 Experiment	28
4.1 Implementation Environment	28
4.2 Training	29
4.2.1 Training configuration	29
4.2.2 Training Execution	30

Contents	vi
4.3 Initial Experiments	31
4.3.1 Whole Spectrograms as Input	31
4.3.2 Spectrograms of signalframes as Input	32
4.4 Experiments single frequency vectors	34
4.4.1 Experiments for Synthesis	36
4.5 Experiments with slices of spectrograms	37
4.5.1 Audio synthesis	39
4.5.2 Reconstruction and post-processing	39
4.6 Experiments with mel scale	40
4.6.1 Pre-processing	41
5 Results	43
5.1 Results regarding model with single frequency vectors	43
5.1.1 Experiments of single reconstruction	44
5.1.2 Experiments with interpolation in embedding	46
5.2 Results of experiments with spectrogram frames	48
5.2.1 Experiments of single reconstruction	49
5.2.2 Experiments with interpolation in embedding	51
5.3 Results with mel-spectrograms	53
5.3.1 Experiments of single reconstruction	55
5.3.2 Experiments with interpolation in embedding	57
6 Evaluation	62
6.1 Impact of model configurations	62
6.1.1 Impact on score	62
6.1.2 Impact on single reconstruction	63
6.1.3 Impact on interpolation	66
6.2 Impact of pre-processing	68
6.2.1 Configuration of STFT	68
6.2.2 Use of single frequency vectors vs. spectrogram frames	68
6.2.3 Magnitude vs. Mel	69
6.2.4 Spectrograms vs. Time-domain	72
6.3 Impact of post-processing	72
6.3.1 Energy correction	72
6.4 Differences between instruments	73
6.5 Comparison to other approaches	73
6.6 Limitations and difficulties	74
7 Conclusion and Future Work	75
7.1 Future Work	76
A Additional Graphics	77
B Supplementary Materials	83
B.1 PDF Files	83
B.2 Media Files	83

References	84
Literature	84
Software	86
Online sources	86

Abstract

In todays world, machine learning technologies are more and more present in everyday life. Besides the usage for simplifying common tasks those technologies are growing in their significance for entertainment including the audio domain, for example to generate sounds or even music with them.

Due to the high interest around this field, this thesis focuses on the exploration of applying machine learning techniques such as neural networks to synthesize novel sounds. As there already exist some approaches in this domain, this work emphasizes on the usage of convolutional neural networks shaped as autoencoders. Throughout this work experiments are conducted, that examine the suitability of convolutional autoencoders, that are differently parameterized regarding the convolution type (1D or 2D) but also the stride within the network. By performing those experiments on log-magnitude spectrograms, some further ones are also conducted on log-mel spectrograms for comparison. Those autoencoder networks are trained to reconstruct audio spectrograms. By introducing a step to interpolate the values generated of two audio samples in the smallest layer, the network is forced to construct a novel audio spectrogram. After converting this spectrogram back into time domain, this yields a novel sound containing characteristics of the two original spectrograms. While mainly the ability of convolutional autoencoders to synthesize audio is proven, those experiments also show which configuration but also kind of spectrogram, works out best. This happens in terms of model performance but more importantly on the output quality.

Throughout the experiments in this thesis it could be observed, that based on the results, different configurations of the network certainly influenced the model performance and output quality. Applying 2D convolutional networks with less strides performed best regarding the quality of the audible output. Otherwise while more strides or 1D convolutions were applied, the audible quality decreased significantly. This also happened when applying log-mel spectrograms in contrast to log-magnitude. Those findings could be confirmed either by reconstructing single spectrograms but also when interpolation in the smallest layer was applied. As a result of this thesis, it nevertheless got proven that by using convolutional autoencoders trained on spectrograms, novel sounds can be generated on the basis of two instruments characteristics.

Kurzfassung

In der heutigen Welt gewinnen Machine Learning Technologien immer mehr an Präsenz im Alltag. Während diese Technologien für die Erleichterung alltäglicher Aufgaben eingesetzt werden, gewinnen sie immer mehr an Bedeutung für die Nutzung in der Unterhaltung inklusive dem Bereich zur Klangerzeugung, nicht zuletzt um mit ihnen Klänge oder auch Musik zu generieren.

Da dieses Themengebiet von hohem Interesse zeugt, beschäftigt sich diese Masterarbeit mit der Erforschung der Anwendbarkeit von Machine Learning Techniken wie neuronaler Netzwerke um neuartige Klänge zu erzeugen. Da in diesem Bereich bereits Ansätze existieren, fokussiert sich diese Arbeit auf die Benützung von Convolutional Autoencoder Netzwerken. Im Zuge dieser Arbeit werden Experimente durchgeführt welche die Eignung von Convolutional Autoencodern untersuchen, die sich hinsichtlich der angewandten Art von Convolution (1D oder 2D) bzw. auch dem angewandten Stride unterscheiden. Da diese Experimente mit log-magnitude Spektrogrammen durchgeführt werden, gibt es zum Vergleich noch zusätzliche Experimente mit log-mel Spektrogrammen. Diese Autoencoder Netzwerke sind darauf trainiert, Spektrogramme zu rekonstruieren. Durch das Einbringen eines Interpolierungsschrittes der Werte von zwei Audio-Samples in der kleinsten Netzwerkschicht, konstruiert das Netzwerk ein neuartiges Spektrogram. Durch die Transformation dieses Spektrogramms in den Zeitbereich, resultiert daraus ein neuartiger Sound welcher die Charakteristiken von zwei Audio-Samples beinhaltet. Während hauptsächlich die Eignung von Convolutional Autoencodern zur Klangerzeugung erforscht wird, zeigen diese Experimente auch, welche Netzwerkkonfiguration bzw. Art von Spektrogrammen am besten geeignet sind. Dies erfolgt in Hinsicht auf die Model-Performance aber vor allem auch die Qualität des Outputs.

Durch die durchgeführten Experimente und deren Ergebnisse in dieser Masterarbeit konnte beobachtet werden, dass unterschiedliche Konfigurationen der Netzwerke die Performance des Models sowie auch die Qualität vom Ergebnis wesentlich beeinflussen. Erwähnenswert sind hier vor allem die Netzwerke mit 2D Convolutions und niedrigem Stride, da diese am besten punkto auditiver Qualität abgeschnitten haben. Andererseits durch zu hohem Stride bzw. Verwendung von 1D Convolutions verringerte sich die auditive Qualität signifikant. Dies konnte auch bei der Verwendung von log-mel Spektrogrammen im Kontrast zu log-magnitude Spektrogrammen beobachtet werden. Diese Erkenntnisse konnten einerseits durch das Rekonstruieren von einzelnen Spektrogrammen aber auch mit angewandter Interpolation in der kleinsten Netzwerkschicht, bestätigt werden. Als Ergebnis dieser Arbeit wurde nichtdestotrotz bewiesen, dass man mithilfe von Convolutional Autoencodern neuartige Sounds generieren kann, die auf den Charakteristiken von zwei vorgegebenen Audio-Samples basieren.

Chapter 1

Introduction

In todays world, more and more technologies based on Machine Learning are present within everyday life. Those grow in their applicability and already span numerous fields. Knowingly ML technologies are also used in the image domain, where they show the ability to alter but also generate pictures. Prominently image style transfer, is one of those, applying neural networks to extract and combine characteristics of two pictures to synthesize a new one. This leads to the idea to also apply ML technologies such as neural networks in the audio domain to synthesize novel sounds. The basic idea behind this topic is to take two instruments as a source, and generate one sound based on the characteristics of both instruments. For example to use a guitar signal and combine it with the characteristics of e.g. a synthesizer to form a novel output sound. By using technologies such as convolutional neural networks, combined with the knowledge of the image domain, this should enable the generation of interesting sounds in a new way. Therefore this topic is of high interest, as through it new ways of generating music can be explored.

This thesis therefore aims to explore the capabilities of convolutional neural networks to be used for the task of audio synthesis. As the emphasis in this work lies on using convolutional neural networks, the input audio data is provided as spectrograms. Like in some works that are discussed in chapter 2, through all experiments, this work applies a neural network shaped as an autoencoder, which has the feature of projecting the input data on a lower dimensional space, in the first part of the network. From this representation, which is also called embedding, the second part also called decoder, tries to reconstruct the input data from it. By introducing a step that interpolates the compressed data of two different input samples, a new embedding gets generated. The decoder part in addition generates a spectrogram containing characteristics of both input samples, which then gets converted back into time-domain resulting in a new synthesized sound. Finally through this technique it should be possible to generate interesting novel sounds based on different combinations of two audio samples.

1.1 Research questions

For this work some research questions have been defined that are answered by the evaluation of the conducted experiments, described throughout this thesis. The main goal of this thesis is to prove whether it is possible to create novel sounds, based on the characteristics of two instruments by using ML technologies such as convolutional neural networks. Furthermore it should be evaluated based on different neural network configurations, how those influence the quality of the output but also the models performance. As the pre-processing takes an important part in a ML-toolchain it will be shown how this also influences the model performance itself but also the quality of the synthesized sounds. Because the neural network extracts features of the input that are used for the synthesis, this work also gives an insight on what information the neural network learns. Finally as spectrograms are used as the input source for the models, a short evaluation shows if those are suited best for the task of neural audio synthesis.

1.2 Outline of the thesis

This section gives a short overview on the different chapters as well as a short explanation on what to expect in each. While this first chapter gives an introduction to this work, the following chapter 2 discusses already existing approaches around generating audio with neural networks. Specifically works concerning neural audio synthesis but also audio style transfer, get discussed as those approaches have a high relevance for this thesis. Chapter 3 explains the general methodology and applied technologies, to get a detailed look on how this work achieves the task of neural audio synthesis. Based on this methodology, in chapter 4 the conducted experiments will be described as those serve to derive the answers to the previously stated research questions. The succeeding chapter 5 depicts the results that were obtained by carrying out the stated experiments. These results incorporate visual representations of the output such as spectrograms and embeddings but also numbers showing the models error scores. Chapter 6 discusses the results in combination with an auditory evaluation of the output generated by the neural networks, and delivers the findings to answer the research questions. In the last chapter 7 the research questions will be answered by the findings obtained in the discussion. There also a future outlook is contained, that provides possible further research points to be carried out in the future.

Chapter 2

Related Works

There already exist a few good approaches around the topic of neural style transfer or generating audio using neural networks, that present rather good solutions. Some of these studies have proven, that with neural networks it is possible to generate synthesized audio up to a certain quality. Those approaches can be categorized into different areas, as their principle and methodology 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, and respective audio spectrograms and thus, call it explicitly audio style transfer. Secondly because those solutions are defining a combination specifically of content and style. This topic will be discussed in more detail in section 2.2. All methods which do not include this principle of content and style can be categorized to the technique of neural audio synthesis or simply just audio synthesis (see 2.1). Those methods incorporate mostly autoencoder networks.

2.1 Neural Audio Synthesis

Neural audio synthesis is the field of creating or synthesizing novel sounds with the help of neural networks. The approach is similar and related to the field of audio style transfer. As described before, approaches in this domain differ in certain ways to neural style transfer. As a major difference, with respect to 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 also gets frequently applied on whole audio samples or musical pieces, in audio synthesis the focus is more directed on the application for single notes. Especially concerning autoencoder networks, neural audio synthesis also includes the tasks of learning important sound features for compression and recreation of the input data. How different approaches are designed and which machine learning techniques and which results could be obtained, will be described in the following sections.

Probably one of the most prominent solutions in the field of neural audio synthesis comes from *Engel et al.* [6]. 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. In addition, there also exists a publicly available dataset called

“NSynth” [6] that contains a large amount of high quality audio samples. The dataset has been applied for the training of the model over the course of this specific project. In their work regarding the synthesis, *Engel et al.* developed and compared two different approaches with two different kind of networks. Basically they have a similar structure, as they are both designed as autoencoders. However they do accept different formats of audio data and thus have different components. While the one kind of network operates on time-continuous data the other one is trained on the spectral representation of audio samples. Throughout the work, the second technology using spectrograms is referenced as Baseline Model. This one consists of convolutional layers, with leaky ReLU activation, except the last layers in encoder and decoder. Furthermore batch normalization is applied as well. This spectral autoencoder is compared to the so called “WaveNet-style Autoencoders”, that are trained on continuous time signal. The autoencoder structure enables learning efficient encodings of the music data, which are representing essential features from the original audio.

In order to create new sounds they take the encoded data from the embedding space of two instruments and interpolate them linearly. In addition they applied the decoder part to reconstruct audio data. As a result they were able to create new sounds with the characteristics of two different audio signals. Comparing the performance of the two different networks used, they found the WaveNet-style autoencoder to be advantageous. This was proven both by the error scores for reconstructing the audios or auditory quality and by quantitative comparison with a pitch and quality classifier model. Nevertheless it was also concluded that the spectral baseline model has a strong performance.

The results regarding the WaveNet-style Autoencoder can be explored via Engel’s online AI-Experiment called “Sound Maker”¹.

In further publications and approaches, *Engel* continued the research on neural audio synthesis by applying other network structures for this purpose. With respect to generative adversarial networks (GANs) and recurrent neural networks (RNNs), two more works have been published in the field of neural audio synthesis [5, 10]. Similar to their work concerning WaveNet-style and convolutional autoencoder, they conducted experiments in (re)synthesizing audios, as an example interpolation of extracted features has been done. Eventually the discussed works show a suitability for audio synthesis and also highlight a major speedup in the computation of synthesized audio samples.

The work by *Natsiou et al.* does not explicitly mention the term neural audio synthesis in its title, but deals with it throughout the article [21]. In their work they research the reconstruction capacity of (stacked) convolutional autoencoders 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 feasible improvements through additional techniques are measured. Additionally, they mention that their work explores musical timbre compression. Here the synthesis gets specifically referenced to timbre synthesis. As audio spectrograms exist with different scales, this approach uses, in contrast to others, the log-mel scale. This scale proves to be beneficial, as it already captures the most significant properties with the characteristics of consuming less memory and computational power. For the training the authors used the NSynth Dataset proposed by *Engel et al.* [6], whereas just

¹“Sound Maker” <https://magenta.tensorflow.org/nsynth-instrument>

a sub-sample, containing samples of different instruments of one single pitch, was considered. The model(s) that were used throughout their experiments followed the general structure of a (stacked) convolutional autoencoder network, which is composed 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).

In order to measure the results of their experiments they applied error metrics such as root mean squared error (RMSE) and structural similarity index (SSIM). As a result these metrics cannot accurately express anything about the quality. Because of this reason, they also introduced a precision and recall score and combined it in a F1 score. In order to generate sounds from the spectrograms, they were utilizing 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 reconstructing 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 kernel regularization got applied, the spectrograms were more accurate, while with dropout layers no improvement could be achieved. The use of (max) pooling also resulted in a more accurate time-frequency resolution with less noise than with just convolution layers. Finally, after the removal of the fully connected layers it showed that the quality was significantly better, as spatial information got preserved better.

Regarding audio synthesis, *Colonel et al.* published a few works where they investigated the suitability of autoencoder networks in connection with audio synthesis [2, 3, 17]. Starting in 2017 they proposed an autoencoder based audio synthesis through compression and reconstruction of audio spectrograms [17]. In contrast to the previously mentioned approaches, this one is based on fully connected layers without convolutions. Also a self-made dataset was created, with their own synthesizer. In contrast to e.g. the NSynth dataset, this one 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 different optimizers. The mean squared error (MSE) was used as the error metric. Comparing these scores regarding networks of one or two hidden layers on each side show, the network using the Adam optimizer worked out best in contrast to using Momentum as the optimizer. These networks, based on sigmoid activation functions worked best when less compression is applied. The concept of 4 hidden layers showed having a mix of ReLU and sigmoid activation functions worked out best. Additionally by applying regularization methods such as dropout and L2 penalty, the latter proved to be better, as the results were of better auditory quality. Further results showed that sigmoid activations led to fuller sound than with ReLU. Furthermore by using bias terms, it could be observed that noise was present in the results. As a consequence, despite of the better convergence, they chose to let them out. The authors concluded that using a 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 a direct activation of the latent space. For an improved model convergence, the autoencoder was designed asymmetrically, via input augmentation. This means that they padded the input magnitude data with different permutations, being first or second order difference or mel-frequency cepstral coefficients (MFCC). Whereas in the previous work only MSE was contemplated as the error metric, here, a comparison of several cost functions was used. Within the cost functions, the mean absolute error (MAE), as well as the spectral convergence cost function (SC) with L2 penalty, was considered. The penalization of the total spectral power proves to be advantageous as the power in the output is more accurate compared to others. In comparison to their work from 2017, they also left out additional bias terms, but decided to just use ReLU activations instead of a mixture with sigmoid.

Eventually improvements to their previous results could be achieved regarding the additional methods that were applied. Concerning the augmentation of the input data, a significant improvement regarding score could be reached, whereas augmenting with first order difference outperformed all other approaches. Concerning the generated sound, it could be observed that by padding with the MFCCs a different sound palette was present. In another comparison to their baseline, *Colonel et al.* introduced the possibility to omit the encoder part of the network. This directly enables the activation of the innermost 8 neuron layers while the decoder can generate novel sounds. As no phase information was present, estimations were done via a method called real-time phase gradient heap integration, which enables the generation of a playable sound. In addition to this work, the authors implemented a small program including a GUI, where it is possible to directly interact and activate the innermost neurons (eight control values in latent space) to generate new sounds.

In a more recent work, *Colonel et al.* implemented and compared autoencoder networks with different topologies regarding their performance for musical timbre generation [3]. This work already utilizes findings and methodologies from previous works. Based on a study from 2018, they implemented a mechanism to directly activate and control the latent space of a trained autoencoder with a graphical tool. They found out that this technique proved to be difficult in terms of controlling the latent space. To overcome this issue and improve the work, they added chroma-based input augmentation to improve the reconstruction performance in this approach. The chroma values are based on the 12 note (western) scale to represent the dominant note present in an audio sample. Besides this type of input augmentation, they also implemented a so-called skip connection, where the latent space is conditioned by the chroma value. In this work the chroma values are represented via a one-hot encoded representation for each training sample, where the maximum value is set to one while all others are set to zero. Consequently these one-hot encoded chroma representations tell the note played in a single-note audio. With this technique the authors could shape the timbre around a specific note class. For the networks topologies, they varied the size of the bottleneck-layer (8, 3 or 2 neurons), the activation functions, the input augmentation, the use of the chroma skip connection, as well as different datasets. It should be mentioned that the authors trained and experimented with the self-generated dataset from

their previous works, containing five octaves of notes, a one octave subset of it and a separate violin note dataset. As a result, the network with an eight neuron bottleneck, with the chroma-based input augmentation worked out best. Thus, for the rest of the experiments *Colonel et al.* were using this technique. In case of the two neuron bottleneck network, the sigmoid activation functions without skip connection worked out best for the one octave dataset. The skip connection turned out to work best for the violin dataset (sigmoid and two neurons). Finally in the case of three neurons also the variant with the skip connection worked out best for both datasets. By analyzing the latent spaces some interesting observations could be made for the sake of audio synthesis. The authors applied a clustering method to see the distribution of the values in the latent space concerning their note and timbre. Using sigmoid activations turned out to limit the values to the range of (0,1) as well as distributing the values in a more uniform manner. Also, the skip connections led to a denser representation. By taking this as an advantage and moving forward with just sigmoid activations, sampling of the latent space (with a mesh grid e.g. 350x350 for two neurons bottleneck) was done to generate a new timbre. In combination with setting the additional chroma conditioning vector to a given note class, the decoder generates the timbre that matches the chroma vector and thus, the desired note is present in the output sample.

A comparative work on autoencoders, in terms of music sound modeling, has been published by *Roche et al.* in 2019 [26]. In this work they implemented four different types of autoencoder networks, that have been compared in terms of audio synthesis. Similar to the other techniques described earlier, this one also orientates itself on the principle of an autoencoder, to project the input data to a low-dimensional space, from which input can be (re)synthesized. In the described experiments, the proposed autoencoder networks consist of (shallow) autoencoders (AEs), deep autoencoders (DAEs), recurrent autoencoders (LSTM-AEs) and variational autoencoders (VAEs) which all got compared to principal component analysis (PCA) as baseline. As sound data for training and experimenting, they used a subsample of 10,000 different randomly selected notes from the publicly available NSynth dataset. The networks that where implemented got trained on the normalized log-magnitude spectra of those samples. Regarding the structure or the depth of the different networks, the researchers used two and three layers for the DAE on each side. In the case of the VAE just one version with two layers, and one version of the LSTM-AE with one layer on each side was applied. Regarding the size of the output from the encoder (latent space), they experimented with different values in a range from 4 to 100. The conducted experiments consist of a resynthesis-analysis where the reconstruction error (RMSE in dB) of the different methods got compared. Additionally to the RMSE so called PEMO-Q scores were introduced to calculate the objective measures of perceptual audio quality.

The results showed to their surprise, that PCA outperformed the shallow autoencoder network. Continuing with DAEs, the reconstruction performed almost 20% better than the shallow AE, having an encoding size of 12 and 16. Also the error decreased faster when the dimension of the latent space was reduced. Even better results with over 23% improvement compared to PCA could be achieved by using LSTM-AEs which brought them to the conclusion, that it is feasible to use more complex architectures. The fact that more compression and thus a small latent space can be generated, is even more important for sound synthesis. In comparison the reconstruction error from the VAE

was lying between the one of the DAE and shallow AE/PCA. As the size of the latent space influences the reconstruction error, it can be stated, that the bigger the size, the lower the error. Interestingly PCA outperforms all models having an encoding size of 100. In addition to the RMSE score, the perceptual audio quality got measured with the PEMO-Q score. The results are comparable to those with RMSE, with just the LSTM-AE having a slightly lower score as compared to RMSE. In this context it was also investigated how the latent space values can be used to be controlled by musicians, and thus, the correlation between those values has been calculated. Averaged over all samples per model, it was shown that the values from LSTM have the most correlation while VAE has the worst. Having less correlation makes VAE the better candidate in terms of using the latent values as control values for synthesis (less redundancy and clear perceptual meaning). In terms of audio synthesis, including the latent space variables, *Roche et al.* also demonstrated how it could be applied for sound interpolation like in the work of *Engel et al.*. For this task they selected the latent space vectors of two sounds with different characteristics and linearly interpolated each value. By decoding and in addition, applying the inverse STFT and Griffin-Lim, new interesting sounds could be generated.

2.2 Audio Style Transfer

The works discussed in this section all orientate themselves on the techniques of image style transfer. As those techniques have a significant impact on the development of audio style transfer algorithms, two important works are discussed in section 2.3. 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) are combined, to form a new stylized image [7]. This means that in the output image, the content image looks 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). In the image domain it is difficult to distinguish content from style, whereas it is a bigger question regarding audio, that appears in the different approaches. Most authors define the style as a musical instruments’ timbre or even a musical genre. Alongside, the content might be defined as global music structure with rhythmic constraints [9]. Those scientific questions also might be influenced if whole audio samples/musical pieces might be taken to get stylized or just some single notes from an instrument. Furthermore if speech is considered as audio data, style and content is differently defined as well. Here, style could be e.g. the emotion of the voice or the speakers identity and content of 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 [25]. In their study they developed a neural network that is constructed as an convolutional autoencoder. Here they officially described their system as an audio style transfer algorithm. In this case the process of generating audio containing characteris-

tics of two audio signals is slightly different as in the work of *Engel et al.* They worked with two networks, namely a transformation network and a loss network. This architecture and methodology is especially inspired by the neural style transfer algorithm by *Johnson et al.* 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 a spectrogram similar to the input of the network from the encoded data. For the training of the transformation network, the pre-trained weights of the loss network are used which speeds up the learning process. This means just optimization towards low level features/style has to be achieved. The trained transformation network is then able to transform an input spectrogram into a stylized 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 is minimized by backpropagation to the transformation network. Through this procedure it is possible to pass a single spectrogram through the transformation network. Subsequently the network outputs a new spectrogram containing the characteristics of itself (content) but also of a differently styled audio signal. Due to its architecture it also performed really quickly and could be used for real-time applications.

Verma et al. presented a new machine learning technique for the purpose of generating novel sounds, in their 2018 paper *Neural style transfer for audio spectrograms* [28]. In this approach they tried to apply the method for artistic image style transfer to audio and they specifically mention the approach proposed by *Gatys et al.* [7] (see section 2.3). Unlike *Gatys'* image style transfer approach, they adapted and trained an AlexNet architecture on the classification of audio samples. This kind of network is a so called convolutional neural network, where the audio gets converted into spectrograms, which can be seen as grey-scale images. An important aspect is, in this work they used the log-magnitude data of the STFT output. It also should be mentioned that they adapted the network towards a receptive size (kernel) of 3x3 instead of the larger ones in the original network, as it retains the resolution of the audio. Similarly to the image domain, the stylized output image gets initialized with random noise. Thus an input spectrogram consisting of a gaussian noise signal is utilized. This one gets iteratively optimized by minimizing the content, but also style loss via backpropagation. In the end this process creates a spectrogram combining the content of one audio with the style of a different audio sample. Additionally they 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 was published by *Liu et al.* exploring the application of technologies given from the image domain for “mixing audio” [20]. This means, that this approach focuses on using audio as spectrograms. As the previous study solely focused on the technique by *Gatys et al.* this experiment explores two more approaches. While one is inspired by *Johnson et al.*, a convolutional autoencoder coupled with a VGG classification network, the other one uses an approach with (cycle)GAN (Generative Adversarial Network).

In this work they call Gatys' approach specifically “slow transfer”, as the iterative computation from gaussian noise was proven to be really slow. Different to the previous work by *Verma et al.*, the authors adapted a VGG network (1 input channel in first layer instead of 3) for the “slow transfer” method. This approach has also been used in *Gatys'* image style transfer. The transfer process is also similar to the previous work, as *Liu* uses a spectrogram initialized as gaussian noise in order to iteratively minimize the content loss in the higher layers, and the style loss in the lower layers. Setting this transfer process as a baseline model, they also adapted a faster style transfer method by coupling the VGG network with a convolutional autoencoder network. The purpose of this network is to take the content spectrogram as input and then outputting a spectrogram containing also the style features of a style spectrogram. Comparing the described work to other approaches, this proves to be 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 serves as the initial spectrogram on which the content and style loss is computed in the VGG network. The gradient descent then gets applied to the autoencoder network, resulting after few iterations, in a stylized spectrogram. The researchers have proven that their approach is faster than the one with gaussian noise. As mentioned before, for the third experiment they adapted a CycleGAN, which accepts 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. Secondly the application of this method generates two images, as transfer is done in both directions, which means in case of audio that two new sounds are calculated.

As another result, this work showed that their approach generates the results in a shorter amount of time. For the purpose of evaluation, they listen to the outcome, but also apply objective mechanisms like visual assessment of spectrograms, consistency tests with classification and examination of signal clusters. Eventually the authors state that with the baseline approach features like the harmonic is not clear and high frequencies get discarded and that 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 online².

As the above mentioned approaches are working on single notes, the experiment of *Grinstein et al.* has been implemented for whole audio samples [9]. Within their work they were adapting several other approaches, with neural networks from the image domain. Besides neural networks, they also implemented a handcrafted sound texture model which was compared to the neural approaches. The latter one is composed of three sound processing steps, that in combination emulate the human auditory system. Taking a closer look at their work, especially on the neural networks, it can be said that they differ from existing ones in several ways. On the one hand, they do not use a random noise spectrogram, but use the content spectrogram which then gets stylized through their methods. On the other hand many audio style transfer approaches, are explicitly computing the result with a combined loss function, that incorporates both

²<https://www.xuehaoliu.com/audio-show>

style and content loss. *Grinstein et al.* do not make use of this concept, as they already initialize the future stylized spectrogram with the content spectrogram, like mentioned previously. On this target spectrogram, 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 is preserved.

In detail the authors 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 respective spectrogram. This style loss is equally computed as in *Gatys'* image style transfer approach. It is calculated by minimizing the error to a “style sound or spectrogram”, at specific layers in the network that extract the style. Via backpropagation the loss is minimized again at each layer, which results in a stylized content sound or spectrogram, after a few iterations. This workflow was applied to all three different network types and compared in the end. As their first network they used a VGG-19 network like *Gatys*, where the input spectrogram was replicated three times in order to match the input shape (RGB-like). By averaging all three channels in the end, they were able to obtain the final stylized spectrogram. The second network is called SoundNet, which is a convolutional network trained on unlabeled videos including sounds. This type of network operates on the raw waveform where no generation of spectrograms has to be done in advance. Finally a wide-shallow-random network was investigated with audio spectrograms consisting of just one layer CNN (like in the work of *Ulyanov and Lebedev* [30]). As the fourth and last method they made trials with a handcrafted sound texture model that emulates the human auditory system. Even if it is not a neural network, it consists of three layers doing cochlear filtering, envelope extraction with compressive non-linearity and modulation filtering.

Eventually the authors came to the following experimental conclusions: While using the VGG network no meaningful results could be obtained due to the noisiness. In comparison the SoundNet yielded more relevant results despite also containing some noise. Surprisingly the shallow random network performed best together with the sound texture model. For a better understanding, the results were provided online³.

The work of *Ulyanov and Lebedev* has to be mentioned here, as they are often referred to be one of the first that explored transfer algorithms for audio [30]. In fact, they took the architecture used in image style transfer by *Gatys et al.* and adapted it for usage on audio spectrograms. Rather than seeing the spectrogram as a picture, with the dimensions of frequency by time, they took the frequency values as channels for the CNN. The network itself was designed as a shallow network (1 layer) using 1D-Convolution with random weights. To obtain a final spectrogram containing content and style, an optimization is made on random noise, to minimize the loss values to a style and a content spectrogram. Instead of applying it on single notes, this approach also uses longer samples or music snippets.

³<https://egrinstein.github.io/2017/10/25/ast.html>

2.3 Image Style Transfer

In the previous sections different works about neural audio synthesis as well as neural audio style transfer were explored and discussed. A lot of these works especially those proposing solutions for neural audio style transfer, took their inspirations from the image domain. For this reason this section provides a short excursion on relevant image style transfer works of *Gatys et al.* as well as *Johnson et al.* [7, 15]. *Gatys et al.* were the first to implement a system of neural style transfer applied on visual data. By using a convolutional network trained on object recognition and localization (VGG-19) with images, they were able to extract the texture as style but also the content of an image. They found that especially from higher layers in the network, high-level features of the images, can be reconstructed. This includes objects and their arrangements in the scene, without the exact pixel information, which will be used as content representation further on. Using a special feature space for texture synthesis, the style of a content image can be extracted by using the feature responses at certain layers in the network. By combining these two principles, respective style losses and content loss can be computed which will be used for the style transfer. The generation of the target image starts by initialization of a random noise image, on which those losses get minimized by using gradient descent. *Johnson et al.* developed an improved image style mechanism on the basis of the former methodology, that particularly shows improvements regarding computational speed. For the computation of content and style losses they use a VGG network with 16 layers pre-trained on image classification tasks. Here they added a special transformation network, that is designed as an autoencoder. This one takes a target image as input (content image) and synthesizes an image on which the style and content loss is calculated in the VGG network instead of random noise. Backpropagation was performed just in the transformation network, while the VGG network stayed fixed. As a consequence the transformation network produced a stylized image after training. By comparing Johnson’s work to the method by *Gatys et al.*, it shows a significant improvement regarding computational speed but also yields promising results.

Previously described image domain methods significantly inspired the development of audio style transfer algorithms, which are presented in section 2.2. Finally, it should be noted that works by [25] and [20] adapted and applied the method of *Johnson et al.* while all others in section 2.2 mainly used the methodology proposed by *Gatys et al.*

Chapter 3

Approach

In the last chapter some approaches have been outlined and discussed, that have successfully implemented methods regarding the creation of audio signals using a neural network approach. As previously written, these have been mainly categorized in neural audio synthesis and neural audio style transfer. This work is mainly influenced from the area of neural audio synthesis, and can be categorized as such, as the methodology and workflow is strongly related to those works. Nevertheless regarding certain components, it is also influenced by the style transfer methods, despite not defining a specific content or style audio respective loss functions.

This chapter will therefore dive into the methodology and exact workflow of this works' solution, to the problem that also will help to derive the answers to the defined research questions. First a motivation should provide the reader with the intended idea and an overview of the applied methods, to get a general understanding of the idea (see section 3.1). Later on the separate steps and components that are needed to reach the desired functionalities, are described in detail, starting with the pre processing. Further on the ML-model (neural network) will be described, as well as the step that is done to synthesize new sounds. Further on, the required steps for (re)synthesizing a listenable audio as well as a description of the used dataset for training and also all experiments conducted later on are given (see chapter 4).

3.1 Motivation

Like mentioned in the beginning of this thesis, this work aims to explore the possibilities of machine learning techniques such as neural networks for applications in the audio domain for sound generation. This idea is mainly inspired by the idea of taking two distinct audio sources and mixing their characteristics in order to generate a new sound. As seen in the previous chapter, this idea is strongly related to the image domain, where the "synthesis" of a new picture based on two source images, is commonly known as image style transfer (section 2.3). This technique, having a content image to be stylised with a certain style from another image, would mean for the application in the audio domain, to have a sound style to be transferred onto a content or target sound. Such methods are specifically known as audio style transfer and can either be applied to single notes or also whole audio samples or songs. Having the principle of content and style

this would mean, that of one sound the global structure and rhythmical components get preserved while imposing style (e.g. the timbre) on it to generate audios. The details to these approaches have already been outlined in the previous chapter, when describing some existing work around this topic.

Neural audio synthesis is another method for neural sound generation, which does not apply the principles of style and content audio. In the previous chapter, some insights could be gained, how neural audio synthesis can look like, as well as how it can be achieved using different methods and neural networks. Most of those methods were showing promising results, either concerning the auditory quality but also the possibilities that arise in experimenting and designing sounds. Those methods were applying so called autoencoder networks most of the time, that can be used for dimensionality reduction of input data, as they have a so called bottleneck in the middle [12]. Because of this structure, the compressed data in this bottleneck, can be seen as a representation for essential features that either can be combined/interpolated or directly synthesized. To generate synthesized audio, the solutions described in section 2.1 took advantage of the “decompression part” of the network in order to generate audio data. The exact workflow and methodologies for sound creation have already been mentioned in the chapter related works (see chapter 2, sectionn 2.1).

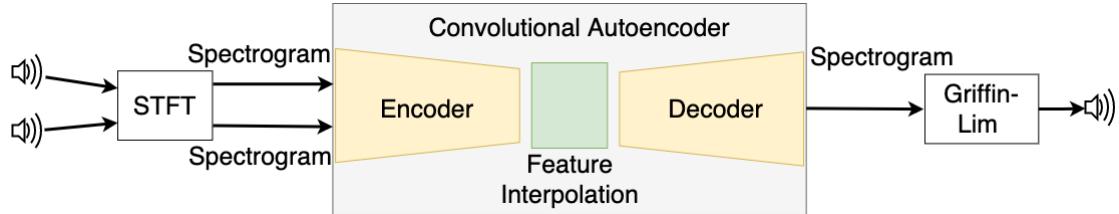
Out of those methodologies, when having the idea of using two instruments’ characteristics, to generate audio, the approach of *Engel et al.* [6] using convolutional and WaveNet-style autoencoders yielded the most promising and interesting results. This can be said especially in terms of output quality but also concerning its implementation and reproducability. With a provided interactive web application, the results of these solutions can be explored, whereas different sounds can be mixed based on a certain ratio. The results in the web application are based on the WaveNet-style autoencoder but according to the scientific article, the convolutional model (baseline) also provides strong results. Implementing an approach with a WaveNet-style network would also go beyond the scope, not at least as the computational costs would be too high. As also some audio style transfer methods, especially the approach by *Ramani et al.* [25], are using convolutional autoencoders, this kind of network was chosen to be preferable, to be applied in this work. How (convolutional) neural networks work, especially concerning autoencoders, will be described later on with special focus, on how those functionalities help to carry out neural audio synthesis, but also to gain general knowledge and a better understanding.

3.2 Overview

Based on the motivation and existing approaches, this work aims to propose a system that uses a convolutional autoencoder network, for the task of neural audio synthesis. This systems’ goal is to take two distinct audio samples as input, whereas the significant features of those are extracted and interpolated, to (re)generate a novel sound in the end. In figure 3.1 the general workflow of the toolchain is depicted in order to get an understanding of how this system is built up.

Starting on the very left, two audios are taken and have to be brought into a suitable representation for this type of network. As audio is in its raw form a time-continuous signal and the input for convolutional networks are of a different shape (e.g. images) some

Figure 3.1: Overview of the proposed solution.



pre-processing has to be done. In this case the short-term Fourier transform (STFT) is applied in order to generate a spectrogram, that shows the frequency spectra over time. The frequency spectra contain on the one hand the magnitude (power) of the frequencies but also the phase information. For this purpose, only the magnitude data is used, as recent publications stated that it contains the most descriptive data of an audio (spectrogram). The autoencoder model then takes the magnitude data as input, from which a compressed representation with the essential features gets generated by the lefthand (encoder) side. Having those features of two different audio samples, those get linearly interpolated, to generate one feature vector representing the “mixed” features of two instruments. This new vector gets passed through the righthand (decoder) side of the network, which regenerates again spectral magnitude data of the same dimension as the input. In order to obtain a “playable” audio sound, it gets transformed back into time domain with the Griffin-Lim algorithm [8] or the inverse short-term Fourier transform (ISTFT). The latter will be applied if there was no interpolation in the embedded space, as the phase information can be reused. Corresponding terminologies as well as a detailed insight into each step and its functionalities are given down below in the following points.

3.3 Pre-processing

Pre-processing is the task of preparing raw data for a specific purpose. Moreover it is an important component of machine learning techniques, with respect to training neural networks. Deciding which pre-processing technique(s) to use on the one hand depends heavily on the type of ML problem that has to be solved or even the training method that is chosen, but of course also on the type of data itself. As written before, for this work a neural network consisting of convolutional layers has been chosen to be applied to the problem of neural audio synthesis. As convolutional neural networks are known for image processing tasks, they also can be applied for audio data, which has already been outlined in recent works in this field (see chapter 2). In contrast to image data which most of the time has a 3D shape (width x length x RGB-colors), raw audio has a different structure in its data representation, as it is a time-continuous signal (1D-shape). In order to bring the audio data in a similar shape, it has to undergo some pre-processing steps. Some representations of audio that have an image-like shape and that got proven beneficial regarding neural audio synthesis, include e.g. log-magnitude spectrograms or Mel-spectrograms but also chromagrams and Constant-Q Transform as stated by [1]. Taking recent works into account, this work chooses to use the first

representation as this one also got used more frequently and had promising results. As for comparison in the experimental part of this work, also the use of Mel-spectrograms will be assessed and discussed concerning the synthesis task and the performance of the neural network.

As the practical part of the project to this thesis is implemented in Python, a special library was used for the pre-processing part. For this part the library *librosa* [29] was used, as it provides practical functions for audio processing, that were considered useful for this work. The functionalities of calculating spectrograms (STFT) but also transforming spectral data back into time domain to generate playable audio data (ISTFT, Griffin-Lim) are of special interest here.

3.3.1 Spectrograms and STFT

Spectrograms represent a 2D-representation of an time-continuous signal, which essentially shows the presence and change of frequencies over time. Like previously said, there exist different forms of spectrograms e.g. log-magnitude and log-mel spectrograms. Especially speaking of the log-magnitude spectrogram whose calculation is based on the short-time Fourier transform (STFT) and thus on the Fourier transform. The Fourier transform takes a frame of N values of an (audio) signal and transforms it from the time domain into the frequency domain. Generally said, that the bigger the frame, the better is the frequency resolution. What this means in terms of calculating the spectrogram, will get outlined shortly. What's also important to mention at this point is, that the result of the Fourier transform consists of an array of N complex numbers, which are mirrored around the middle. Every complex number in this array stands for a so called frequency bin in the signal. The real part of these numbers would represent the power/magnitude of this “bin” and the imaginary part gives information about the phase. Coming back to the frequency resolution, this for example means that when taking a one-second signal with a sampling rate SR and performing the Fourier transform with length $N = SR$ on it, this would yield an array of N values. The first value in the result depicts the signal's offset whereas all values from 1 to $N/2$ are the frequency bins with a resolution of 1 Hz per bin. This means that each of these bins shows the magnitude and also phase of each frequency from 1 to $N/2$ Hz. The ongoing values in the result show the same values except they are mirrored, as they depict the negative frequencies. Because of this behaviour, the second part can be omitted for further use. Now these values just show the frequency spectra of one time frame and do not incorporate more information about the change. To overcome this shortcoming the Fourier transform can be applied to a series of frames of the signal in order to obtain multiple frequency spectra over time that are depicted as a spectrogram.

The calculation of multiple frequency spectra over time is done via the so called short-time Fourier transform. This form of calculation is widely used for pre-processing of audio data for ML-tasks (see chapter 2). When applying this transform, a few parameters have to be considered, as those influence the result but also the quality for the later workflow. As *librosa* is used for the sound processing steps, the mentioned parameters are specifically concerning this library. One of the the most important parameters is `n_fft` as it specifies the actual length of the signal frame, on which the FFT (fast Fourier transform) gets applied. This parameter therefore influences the frequency- but

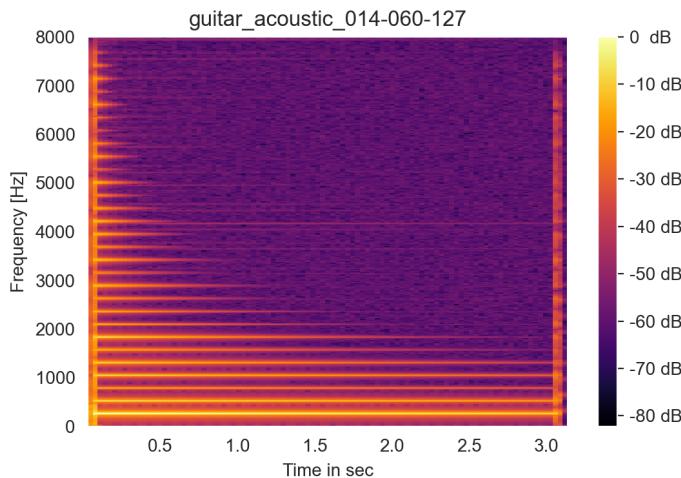
also time-resolution in the final spectrogram. Regarding the official Librosa documentation, this should be a value of a power of two, as it speeds up the computation of the FFT. Another important parameter would be the `hop_length`, which defines how many audio values are between the beginning of the first and the following frame. This means that when defining this parameter to `n_fft/2` this would yield a 50% overlap of the following frame. Modifying this parameter would mean to either increase or decrease the overlap and also the amount of time columns in the result, as more overlapping frames occur. The overlap of the frames is also coherent with the chosen window function for the STFT. As every time when the FFT gets applied to a frame, this one gets multiplied with a so called “window”. Multiplying a signal frame with a window has to be done, as the FFT assumes, that the transformed signal is periodic (repeating itself infinitely) [11]. This becomes problematic when the input signal does contain frequencies that may not directly fall into a frequency bin, due to the FFT’s frequency resolution. Due to the assumed cyclic continuation the Fourier transform will “think” that there is a discontinuity and will spread therefore the power over the spectrum. There are various window functions such as “Hann”, “Hamming”, “Blackman”, etc. which start at (almost) zero, rise to a maximum in the middle but fall again to (almost) zero at the end (symmetric). Multiplying the signal frame with such a window function helps to overcome this issue as it removes the discontinuity. Which window to chose depends on the use-case of the application, whereas throughout this work a “Hann” window was chosen. It got chosen, because it is good suitable for most sound processing tasks. Moreover it reaches zero at both ends and thus eliminates the discontinuity problem. Coming back to the relation with the window overlap, if no overlap would be used a lot of information of the signal would get lost. This is because when multiplying the signal frames with the window functions, this would result in very small or zero values at the beginning and end of the frame [11]. When having overlapping frames this issue would be corrected. Important here is again the amount of overlap as this is dependent on the window and its wideness. Using the “Hann” window a common value for the overlap would be 50% which was also considered throughout this work. This value is also beneficial later on when performing the inverse transformation, back into time domain, but more on that in section 3.6. Beside of these parameters some more exist, for example for specifying the padding of the signal whereas in this work a constant padding on both sides of the signal has been used, which is also the default setting.

Now having knowledge about the STFT and its parameters, it can be applied to a signal to generate a spectrogram. For example by using an audio sample with a sample rate of 16 kHz and applying the STFT with a `n_fft` value of 1024 and a `hop-length` of 512 this would result in a spectrogram with a frequency resolution of 15,625 Hz and time resolution of 64 ms. As explained before, the values of the result consist of complex numbers which contain the magnitude but also the phase at each frequency bin. By setting this result absolute, or calling the function `librosa.magphase(spectrogram)` the real magnitude data can be obtained, whereas the latter also retrieves phase information in a separate vector. The magnitude here displays the energy values of the spectrogram, whereas for further processing and also to be better displayable those get converted into a dB-scale¹. This function also takes a reference value that is set to 0 dB, which in

¹normally the magnitude would need to get squared to obtain the power, but in this case magnitude without squaring was taken

this case will be the maximum value of the magnitude spectrum. As for post-processing when converting the dB-scaled data back into energy, also a reference value is needed, this one gets preserved, in order to get the same scaling as in the original input. Finally when having the log-mag spectrograms in dB, those were considered for the training of the neural network afterwards. An example of a log-mag spectrogram can be seen in figure 3.2. This spectrogram shows the frequency spectra of a guitar sample over time. It can be seen, that at the beginning and at the end there are broadband spectra, which represent the guitar stroke (transient) at the beginning and the noise of damping the strings at the end. As also the phase information was obtained when calculating the magnitude data, this one was also preserved next to the energy reference value for the recreation of signals, as it is needed there (this will mainly affect the recreation of single samples, without interpolation in embedded space, but more on that later on).

Figure 3.2: STFT log-magnitude spectrogram of a guitar note.



This section describes the general workflow of the pre-processing from taking a signal and converting it into a spectral representation. This workflow is a basis on which different experiments with different parameterization (size of n_fft, etc.) of the calculation of the spectrograms but also with additional steps (log-mel scale, additional framing, etc.) are being made. Those steps will get mentioned later on in chapter 4 when describing the experimental part of the thesis.

3.4 ML-Model

The main or core component of every machine learning project is of course the model itself, as it achieves the main task of prediction or inference to a given problem. Those models exist as different technologies that perform regression tasks or even classification tasks. Dependent on the use case, but also the kind of data that is present, different models are better suited or not. To enumerate technologies, there exist the KNN algorithm, Decision-Trees, Random Forests, Support-Vector-Machines (SVM) but also neural networks which can be applied in a variety of use cases. Especially the latter,

the neural networks, are able to achieve a variety of different tasks, as they are highly adaptive regarding their topology, used layers, but also their size and shape. This variety of different tasks spreads across different domains, including images and audio.

3.4.1 Neural Networks - Introduction

Generally said a neural network can be seen as a graph of connected nodes with numeric values that can achieve transformations between patterns using message-passing algorithms [16]. Those nodes are commonly structured in layers, where there especially exist certain nodes or even layers that are seen as input nodes/layers and some as output nodes/layers. Between the input and the output there can also exist so called hidden layers, expanding the depth of the network. The links between the nodes, that are also called neurons, are connected via links, that are parameterized with weights, that get optimized using learning algorithms. Each neuron receives its weighted input (activities) of its connected predecessors, which get converted into a single output that gets broadcast to all its connected successors [13]. The latter involves a so called input-output function which is also commonly known as activation function (e.g. ReLU, Sigmoid, Softmax, etc.). Important to know is that the weights on the connections define how much this value influences the input of the connected node. When a neural network is trained to a specific problem (e.g. classifying certain images), using predetermined training data, the output of the neural network is compared with the desired one, resulting in a certain error (metric). To minimize this error, the weights in the network are adapted by back-propagating the error through all the layers, to the beginning. On this way it changes the influence of certain connections and therefore the overall outcome. This procedure is repeated on all training data over several iterations, until the error becomes low to produce the desired output. The initialization of the network and its weights is often random, which also means that every training run starts and progresses differently.

Depending on the problem at hand, neural networks can be trained using labeled data (supervised) but also just by minimizing a cost function (unsupervised) [22]. More details on the learning will get mentioned later on when explaining the model itself.

3.4.2 Convolutional Neural Networks

In this approach, due to the promising usage in existing solutions, a convolutional neural network has been chosen as the model. Convolutional neural networks are a type of network, that get primarily used for tasks in the image domain for example to recognize patterns in pictures or classification, but also like seen in chapter 2 for image style transfer. They have the advantage over traditional neural networks, that they can deal with the dimensionality of pictures (width by height by colors/depth). The layers containing convolutional nodes have kernels as learnable parameters [22]. Those kernels, if taking a 2D-convolutional layer, are normally small in width and height (e.g. 3x3) but span the whole depth (channels) of the input (in the case of RGB pictures depth of three). Those kernels calculate the scalar product for each value contained in the kernel and the input map. This yields, having e.g. a 3x3 kernel operating on a 3x3 field, in a single value. This value is the weighted sum of the kernel's values and from those of the input vector. This operation will be applied to each 3x3 field along the spatial dimension of the input, resulting in a smaller activation map. One can also apply padding around

the input to preserve the dimensionality. Furthermore a stride can be defined, which defines how much these convolved fields overlap, as using a bigger stride would result in a smaller overlap. Having also a smaller overlap would result in a much smaller activation map. For example taking a 7x7 input, by applying a 3x3 kernel with no strides and also no padding, this would yield an output field of 5x5. With a padding of 1x1 the output would be of the same size. Finally if the stride would be 2, the output field would be of 3x3. Through training of the network and back propagating the error, those values in the kernel get adapted in order to learn certain important features or patterns on which a classification or pattern recognition can be made (easier). As it got described for 2D convolutions, depending on the input dimensionality it can also be applied as 1D convolutions or even 3D convolution.

With the knowledge that such convolutions are successful on image data, those can be also applied on audio provided in the shape of a spectrogram. As described in the previous section (see 3.3) spectrograms can be described in a “picture-like” shape having the dimensions of frequency by time with a depth of one. Speaking of that, the spectrogram can be seen as a grey-scale image. The energy in different frequency bands over time with its variations, can be seen as “recognizable” patterns. Taking a 2D convolution, the kernel (e.g. 3x3) takes a 3x3 frame of frequency by time which results in one value. Summed up, this results in an activation map smaller or equal (if zero padding is used) than the input. After training on different samples and iterating several times, this activation map would contain the most significant features or characteristics of the spectrogram for example when training on classification. Depending on the chosen hyperparameter for the amount of output channels (depth), the resulting activation map can be of depth one or even deeper.

As mentioned before, each neuron in a neural network has an activation function, which is also the case for convolutional neural networks. As there exist different kinds of activation functions, for this approach mostly ReLU (Rectified Linenar Unit) activation functions got used but also LeakyReLU. Additional to this activation function, Batch-Normalization gets applied after each convolutional and ReLU non-linearity component. The choice is mainly based on already existing approaches like from *Ramani et al.*[25] and *Engel et al.*[6], as it was proven to yield promising results in combination. According to *Ioffe and Szegedy* [14] applying Batch Normalization also improves the training speed (number of iterations/epochs) and enables the use of much higher learning rates. Furthermore it acts as a regularizer, so that overfitting-reducing technologies, such as Dropout, can be omitted.

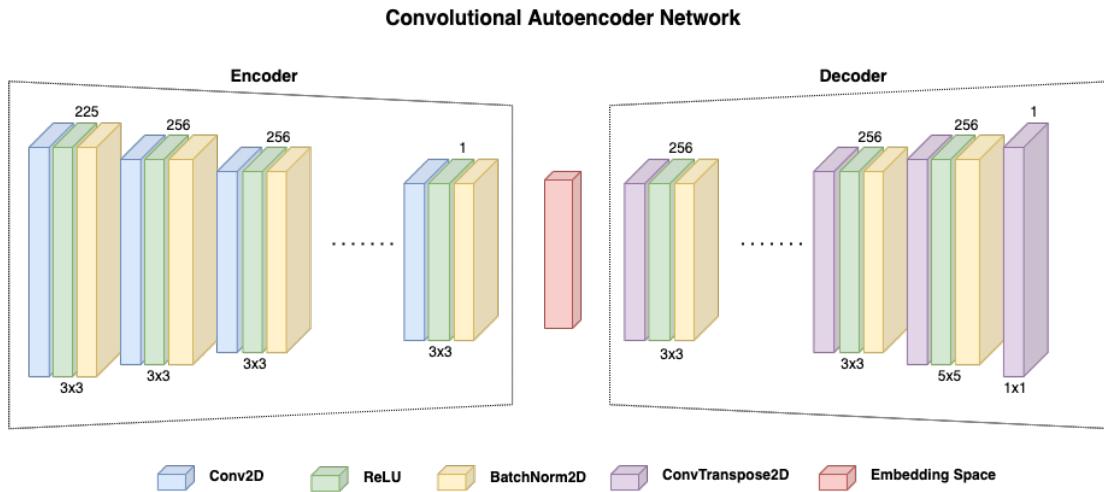
3.4.3 Autoencoder

Neural networks exist in different shapes and compositions, depending on the desired work it should fulfill. Whereas some networks e.g. for classification of a given input, might reduce the width of the layers towards the end, some are designed to have a so called bottleneck. This means that this kind of network contains a smaller central layer then the input, but at the end again a bigger layer (eventually same size of input layer) [12]. Those networks are called autoencoders, in which the first part of the network (until the smaller central layer) gets called “encoder”. The second part beginning at the small central layer, that is getting bigger towards the end, gets called “decoder”. These

parts are called this way, as for the encoder part, it “encodes” the high-dimensional input data to a low-dimensional representation (output of small central layer). The counterpart is therefore called “decoder” as it decodes the low-dimensional data, to bring it again to a higher-dimensional representation (mostly same as input). The lower dimensional output of the small central layer, can be named differently, as for example “code”, what *Hinton et al.* does when describing the concept of autoencoders in his publication [12]. Some other common names would be e.g. latent space, embedding space, encoding or embedded data. This principle of dimensionality reduction therefore means, that the encoder part extracts the most important or characteristic data, from which the decoder is able to reconstruct the input data.

In the related works chapter 2 a few works have made use of this principle to extract “characteristic” features for audio data and synthesize audio from it, by altering them. This knowledge encouraged this work to also make use of this principle to synthesize audio by using autoencoders. Especially when having the extracted audio features as encodings, those can be modified (easier), to in order synthesize novel sounds from them using the decoder part of the network. Combined with the knowledge to apply convolutional networks to audio spectrograms and the advantages of autoencoders to extract features from the input, the model in this work is designed as a convolutional autoencoder. Also the works of *Ramani et al.* or *Engel et al.* applied these kind of network to generate novel sounds. Figure 3.3 shows the basic autoencoder network structure which is used throughout this thesis. To be mentioned, the whole implementation of the neural network for this work has been achieved with the deep learning library PyTorch [23].

Figure 3.3: Basic autoencoder structure used throughout this thesis.



In figure 3.3, the depicted autoencoder, gives an insight of which components and layers it is composed of. The amount of layers as well as the parameters shown in this sketch, are just an example, as throughout the experimental part, those get varied. Therefore, the dotted lines in this figure act as a placeholder for possibly more layers on each side of the network. The numbers above the individual layers represent the amount of output channels (depth) and those on the bottom represent the kernel size. It has

been mentioned that in this approach the convolutional layer gets equipped with ReLU activation functions and batch normalization. This also gets visualized in this figure, where each layer is shown as a combination of sub-layers that incorporate those three components. Speaking of the composition of the layers, it can be seen in the decoder part, that instead of a convolutional layer, a convolutional transpose layer is used. That is because of the nature of convolutions, as their output is always smaller or equally sized, which has been mentioned under point 3.4.2. As the decoder part generates output that is bigger than its input, the layers have to perform upsampling. In convolutional nets this is typically done via the convolutional transpose layer.

This transposed convolution is not the reverse operation of a convolution, but it is more of a operation to recover the shape from the convolutions input. [4] Taking the convolution example from above, by taking as input a 5x5 field and applying the transposed convolution with a kernel of 3x3, this would result in a field of 7x7. The equivalent of this operation would be a convolution with the same kernel, on an input of 5x5, with 2x2 padding (padded input = 9x9), as this would result in a 7x7 output too. When in the convolution, striding was applied, this also works for the transposed convolution. Again when having the 3x3 input, by applying a 3x3 kernel with a stride of 2, this again results in a 7x7 output field. To be mentioned, the striding parameter for the transposed convolutions defines how much zeros are added between the values of the input. This means, when taking the previous example, that the 3x3 input gets one column respective one row of zeros inserted after each value to result in a 5x5 field.

With this knowledge, convolutional transpose layers are best suited to be in the decoder part of convolutional autoencoder networks. Coming back to the architecture of the convolutional autoencoder in figure 3.3, those convolutional transpose sublayers are also coupled with ReLU activations and batch normalization, with an exception to the last layer.

Having this kind of autoencoder for this approach and the idea to extract features of spectrograms using convolutions, this models task is to encode and decode spectral audio data. The network is therefore configured to produce an output of the same dimensionality as the input of the encoder. Like mentioned above, the output should be a reconstruction of the input, that gets inferred by decoding the extracted features in the embedded space. To achieve this goal of reconstructing the input, the network has to be trained through minimizing a specific error. In some cases also maximization is desired, but is dependent on the error score and the desired outcome. In the case of an autoencoder, this works by comparing the output of the decoder with the input of the encoder, by calculating the difference. Generally said, depending on the outcome and the goal of a training, there exist different metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE) but also more specific formulas. Throughout the literature, those error functions are also often called cost (functions) or loss (functions). For the experimental part of this work, the choice has been made to use MSE as the error metric, as it has been used successfully by some existing works.

To minimize this error an optimization of the network has to be made, which will be done through backpropagating the error score. Through this step, the parameters (weights, bias, convolutional kernels, etc.) get adapted according to the error, which in the best case improves the score and thus the output of the network. In this work

the output of the network are reconstructions of spectral data. As having the encoder-decoder structure, this output data is reconstructed through the decoder, which takes the encoder output as input. As mentioned before, this generated data of the encoder is a compressed vector, consisting of the most significant information extracted from the input. Depending on the configuration of the network, especially the encoder part, this vector can be of different sizes. This is because when using more convolutional layers with strides, the output of each layer gets more downsampled, which results in a smaller encoding. As a consequence, the decoder part has to learn to regenerate spectral data from this encoded vector. This vector therefore is a compressed representation of the models input, which contains the most significant characteristics of this specific spectrogram and further on, of the sound. All in all regarding the learning, it can be said, that the encoder learns to extract essential features, from which the decoder learns to regenerate the input as well as possible.

Altering those encodings has therefore the consequence, that the output of the decoder is different, and thus will be a novel, synthesized sound. More on how it should get altered is discussed later on. Choosing the right compression is also an important part that influences the outcome and thus the quality of the desired model output. Too much compression may lead to the fact that the decoder has too little information to infer the desired spectral output, which in order results in poorer quality of the resulting audio. On the other hand having less compression results in embeddings being not significantly smaller than the input, containing less important data too (e.g. noise). It also may become more difficult to alter those encoded vectors. Throughout this work, different amounts of compression have been applied in the experimental part, which get discussed later, including the impacts and observations made on that. Knowing those properties and behaviour of the autoencoder model strengthens the idea of applying autoencoders for the use of neural audio synthesis.

3.4.4 Optimizer

Coming back to the training process, where the network gets optimized, in order to minimize a certain error function. For this optimization, different strategies exist, where hyperparameters such as learning-rate or weight-decay play an essential role. Those optimizations are on a large scale, stochastic gradient-based techniques, to which algorithms such as stochastic gradient descent (SGD) or Adam can be counted. Those algorithms influence and improve the convergence which means to find a minimal error. Throughout this implementation the Adam optimizer [19] has been chosen, as it is used widely in recent publications where promising results could be achieved. Also regarding the training process in this work, Adam optimizer has proven to be advantageous, in contrast to SGD. The parameters that have been found to have the most impact on the optimization process during training, are the previous mentioned learning-rate and weight-decay. To be mentioned, the learning-rate specifies how “fast” the model actually learns while weight-decay works as a penalty for the weights optimization to prevent overfitting. If the learning rate is chosen rather large, then the network learns faster, but because of its big steps or jumps, it could miss the optimal solution in the solution space. Also it could happen that when a local optimum is found, that it “jumps” out again. In this case a smaller learning rate would be desirable, as it makes smaller steps.

Choosing it too low would end up in a slow training where also large areas of the solution space are not visited. The latter leads to a training process stuck in a local minimum. Therefore it is important to choose the right size of learning rate, but this issue depends also on the problem size and type of network. In this work different learning rates have been applied throughout the experimental part where also different findings could be made, but this will be shown and discussed later on in this thesis.

With this knowledge, it can be stated, that a high learning rate could be advantageous at the beginning of the training process, in order to rather find a global optimum and explore the solution space. In order to prevent jumping out of a minimum, a smaller learning rate would be desirable later on in the training. For this case, there exist some mechanisms to decrease the learning rate later on in the training, especially when detecting oscillations of the error due to a too large learning rate.

In this work's implementation, for the start of the training a specific starting learning rate has been set, while throughout the training it gets adapted, when no more optimization and eventual oscillation gets detected.

3.5 Synthesis of novel sounds

Having now covered the important properties of the pre-processing but also of the applied machine learning model, this section explains the methodology to synthesize novel sounds. In the chapter 2 where related works got discussed, an insight could be gained, on how different works tried to synthesize audio with their neural networks. For example *Colonel et al.*[2, 3, 17] suggested in their works, to synthesize novel sounds through directly activating the innermost layer. This means after training, for the sound creation process, the encoder part gets omitted. For example in the case where they had a network with 8 neurons at the encoders bottleneck, 8 different values could get defined, within a certain range. The decoder part then created, based on its training on recreating spectrograms, spectral data that got converted back to time domain to form a synthesized playable sound.

3.5.1 Interpolation in latent space

Having this as one possibility, to use in particular autoencoders as a tool for audio synthesis, there also exist some more interesting approaches. One of those got applied by *Engel et al.* [6] but also *Roche et al.* [26] where they also made use of the latent space encodings. Contrary to omitting the encoder part, those works aimed to utilize the whole network, as no direct activation of the innermost layer is considered. Instead the authors proposed to take the encoded values of two different audio samples (possibly two distinct instruments) and combine them via linear interpolation. The interpolation process yields a vector of interpolated values. This new vector can be seen as a modified encoding, which gets processed by the decoder part, resulting in a spectrogram-like vector. The result of this process is then a synthesized spectrogram, that aims to contain features of the two input sounds.

As the latter methodology corresponds the most with the initial idea, to generate audio based on the characteristics of two instruments, this method was chosen to be implemented to carry out experiments on the creation of novel sounds.

To explain this method in more detail, the encoded vectors of two audio samples serve as the basis. Then each of this vector is taken, to interpolate a value that lies on a linear line between the value at a given index in one vector with the corresponding value of the same index in the second vector. To mention at this point, those vectors are of the same length. The result then is a new array containing the interpolated values of those two vectors. This procedure gets repeated for each encoded output of one sample with the corresponding output of the second sample. Knowing the fact that those encoded values represent the most significant features and probably the characteristics of the note, the result can be seen as a combination of those characteristics. Decoding those will then end up in having a spectrogram containing the characteristics of both instruments. How this procedure and its results look like, with the actual experiments, is shown later on in this work.

3.6 Post Processing

It has been discussed, that regarding neural audio synthesis, the audio data can appear in different shapes. As there exist approaches, that focus on time domain signal like the WaveNet-style autoencoder from *Engel et al.*, there also exist those who operate on spectrograms using convolutional networks. As mentioned before, this work emphasizes the use of a convolutional autoencoder, that takes spectrograms or spectral data as input. In the case of this autoencoder, the output is of the same shape as the input and therefore also a spectrogram. To generate again a playable or listenable audio, this one has to get converted back into time domain. There exist many different methods, to achieve this, while this also depends heavily on the data that is present. As stated in section 3.3, when spectrograms are calculated via the STFT, the output vectors are complex valued. To be mentioned, that without modification or further utilization, via the inverse STFT (ISTFT), this result vector can get converted back into time domain without loss. For achieving this task, it is necessary that the magnitude but also the phase information has to be present. As this autoencoder just operates on the magnitude data of spectrograms, there is no phase information present in the output of the network. At this point, multiple ideas can be applied, depending of what is done throughout the process. This means that when there is no modification in latent space, i.e. value interpolation, the original phase information can be reused to apply the ISTFT. As mentioned previously the phase information gets preserved for exactly this case.

In the second case, where modification steps are performed, like here the interpolation of two sounds' embeddings, no phase information is present that can be used. To overcome this issue, there exist techniques that can approximate the phase information. One of those techniques, which is also probably one of the most prominent ones, is the Griffin-Lim [8] algorithm that tries to estimate the time domain signal based on just the magnitude data. With this algorithm the phase gets randomly initialized, and with alternating forward- and inverse STFT, estimated. For the calculations of the audio signal, again the python library *librosa* [29] has been used, as it provides the inverse STFT but also implements the Griffin-Lim algorithm. According to the documentation of *librosa*, a so called "fast" Griffin-Lim algorithm is applied, which got developed by *Perraudin et al.* [24] which utilizes an additional momentum parameter which helps to accelerate the convergence of the estimation.

Some further note here, as when the pre-processing steps (see section 3.3) have been examined, that some certain parameters have to be taken care of. These are especially the `n_fft`, `hop_length` but also the `window`. It has been mentioned, that e.g. the `hop_length` was chosen to be half of the `n_fft`, to ensure a 50% overlap of the STFT frames. Furthermore the “Hann”-window was chosen, to be multiplied with the signal frames, to avoid discontinuities. Again, those also appear in the inverse STFT as well as Griffin-Lim. In order to obtain the best result, it is important to apply the same values with those parameters. As the inverse Fourier transform gets applied to each vector in the spectrogram, this results in single time domain frames that have the length of `n_fft` (original length). With the knowledge of the window and the overlap, those single frames, get “overlap-added”, which results again in the full length audio sample.

Before applying the the inverse calculations on the autoencoders output, it has to be considered that the autoencoder works on the db-scaled magnitude. As a consequence the output therefore is also db-scaled. To apply the inverse calculations, the output becomes energy again. For this case a reference value has been preserved from the pre-processing stage, in order to obtain the (almost) same scaling again in the output signal. For the experimental part some more steps also got applied, like scaling the energy according to the average energy present in the original spectrograms. This step should also correct and improve the resulting sounds. More on that in chapter 4, but also later on when examining and discussing the results.

3.7 Dataset

As machine learning models including neural networks, have to get trained, in order to deliver accurate results, data is needed on which it should get trained on. To get compelling results, it is not only important to have an appropriate model configuration or pre-processing chain. The choice of an appropriate dataset therefore is also of high significance. As seen in related works, datasets can either be self-generated or taken from a publicly available data source. As in the case of this work, the model operates on audio data, a dataset of musical notes is desirable. Generating sufficient data, by oneself, is a task that would take a significant amount of time. Not only as this dataset preferably should contain a large amount of audio samples, those also should be highly diverse such as different instrument sources or different pitches. Not only is it an advantage for the training of the neural network to have lot of samples and diversity, as it helps to improve the learning process and generalization within the neural networks. Moreover it is also advantageous for this kind of work, as when having many different instrument sources and available notes, more interesting combinations with regard to audio synthesis can be made.

3.7.1 NSynth Dataset

Exactly for this kind of approach a large dataset consisting of instrument samples called “NSynth”, has been made publicly available by *Engel et al.* [6]. This dataset consists of a total of 306.043 musical notes that have a unique pitch, timbre but also envelope and has been created for the idea of neural audio synthesis. This amount of musical notes

incorporate monophonic audio snippets, sampled at a rate of 16 kHz, of 1.006 different instruments. Every note is of a specific pitch, ranging over every note of a standard midi piano (21-108). This results in having 88 different pitched notes, in the best case, as not every instrument is capable of producing all different pitches. The average amount of pitches is therefore, according to the scientific publication, 65.4 per instrument. With more detail, each audio sample belongs to a certain instrument family which could for example be a keyboard, guitar, organ, bass, brass and so on. Further on they can be distinguished by their source, being either produced acoustically, electronically or synthetically. All these specifications make this dataset highly attractive for this kind of work, besides being publicly available² for free and already successfully used for neural audio synthesis. Regarding the use for machine learning tasks, on their website, they provide the dataset already split up into a training, validation and test set which do not overlap at all. To specify, the training set consists of 289.205, the validation set of 12.678 and the test set of 4.096 examples. For this purpose, it has been chosen, to take these splits as they are, for the training, validation and finally testing stage.

²<https://magenta.tensorflow.org/datasets/nsynth>

Chapter 4

Experiment

This chapter describes the experiments conducted in order to be able to answer the defined research questions. In the previous chapter 3 the general methodology and used technologies were described, to gain an understanding of the implementation and its components. Those can be parameterized in order to influence the outcome of this work. Furthermore some additional steps can be introduced that have a significant impact on the result. Based on this knowledge and implementation, some experiments were conducted. Those experiments should deliver some interesting insights on how different configurations influence the workflow but also the final result being synthesized audio. Those experiments span almost every stage from the pre-processing to post-processing. This section therefore describes how the proposed methods were utilized in order to answer the questions.

4.1 Implementation Environment

In the previous chapter it has already been mentioned, that this approach was developed in Python using specific libraries. For pre- and post-processing the Python audio library *librosa* [29] has been chosen, as it provides all necessary functionalities that are needed for the approach and experiment. For all steps regarding the neural network model such as configuration, training, inference etc., *PyTorch* [23] has been utilized. The project though has been implemented and applied on two different machines, depending on the task that had to be done. Generally speaking, the toolchain comprised of all stages, has been developed on a local machine running Python, except the training itself.

Speaking of that, the training has been mainly performed on a remote Jupyter Notebook that has access to one of eight NVIDIA GeForce RTX 3090 high-performance GPUs. Not at least, as in the case of training a convolutional neural network, this is a rather time and computational power-consuming task. Of course, this not only depends on the kind and complexity of the network, but also on the amount of data that is used for the training. Furthermore using GPU-acceleration means to significantly have more computation power and speed. Using the local machine, just the CPU could be utilized for training, which would mean that training is done sequentially and thus significantly slower while also the local machine has to be running constantly. For the training on the remote instance, the pre-processing also was done there as the data is directly

loaded there. As just the training was performed remotely, all other steps, including the evaluation towards audio (re)synthesis having the trained model, were done locally. Not only because no time consuming tasks had to be performed, but also as it was more convenient, as the remote service is not always accessible.

4.2 Training

The training, as mentioned previously, was performed on a remote Jupyter Notebook service with access to a GPU. As outlined in the previous chapter, for the whole experiments, the NSynth data set proposed by *Engel et al.*[6], was used. This dataset is already split into a training, validation and test part. For the training on the remote notebook, the training and validation data set was used therefore, which in order also was pre-processed there. As a side note, in the beginning of the project, the training was just done locally, with a small subset of the already small test set (mostly of one instrument). This was just done to make a low-level proof that the autoencoder model can produce meaningful results.

4.2.1 Training configuration

To take a closer look into the training process, this one consists of several important stages and components. First of all the PyTorch model, defined as a class, was initialized. As a metric is needed, to measure the error of the output, the mean squared error (MSE) was utilized. This error metric calculates the difference between all values of the desired and actual output, squares them, and takes the average over all. Furthermore to optimize the network, as explained before, the Adam optimizer got applied, in which the (starting) learning rate, but also the weight decay was defined. The right learning rate depends heavily on the amount of training data but also complexity of the model. In the case of this work, this means it is in the range between $1e - 5$ to $1e - 7$. In chapter 3 it was also mentioned, that a technique to reduce the learning rate during the training process was utilized. Generally speaking, this function called `torch.optim.lr_scheduler.ReduceLROnPlateau(...)` reduces the learning rate, by a given factor, if within a certain patience period (epochs) no optimization was detected. This method improves the training process as further convergence can be achieved. Of course for the training process, the pre-processed training and validation data set had to be loaded. To mention, the pre-processed data was calculated and stored on disk, in advance, to not always have to run through it. For the training it therefore was loaded and brought into the desired shape for the corresponding model. This shape was varied throughout the experiments, to observe its impact on the training process but also quality of the output. As this also depends on the chosen network, this aspect is described later on in this chapter. For the training process, this was done with the training dataset, but also on the validation dataset. The latter is an important step during the training, to validate the models performance on never seen data. The data in the right shape had to be converted to a tensor and in further steps, to a (custom) dataset object. Finally a so called “DataLoader” had to be initialized, either for the training but also for the validation. In this DataLoader the batch size can be specified, which is an important parameter regarding the training. Throughout a few batch sizes have

been tried out, whereas in the end, the preferred batch size was 32. This means that the input data was portioned in equally sized chunks of 32 tensors, that were fed into the network at once. Further on it can be set that after each training epoch, the dataset is shuffled. Setting this to true, the samples in the batches are in a different order and constellation. Otherwise, the batches always consist of the same data in the same order. Throughout the experiments, this setting has proven to be advantageous regarding the convergence of the model, as the error could be minimized further (see chapter 5).

4.2.2 Training Execution

Having the configured dataloaders, which are also iterables, it is possible to iterate over the batches of the dataset. Before the training begins, a number of epochs has to be set, which defines how often the training should be performed on the whole dataset. In advance it cannot be said, how many epochs are needed, therefore a preferably big number was chosen e.g. 1000. To clarify, the training never ran until the final epoch, as it got stopped at a point, where the result is sufficient (more on that shortly). In each iteration, the corresponding batch of data got put into the network with a forward pass which calculated an output. This output was compared to the desired value, which in this case was the same as the input, and further on the error was calculated. Further on the gradients got computed and optimization via parameter updates was done. The loss value was added up, and after all batches are done, the average loss got computed, to show the progress.

After all training batches have ran through and optimization was done, the model was validated using the held out validation set. This one was equally batched and ran through the same process, except there was no optimization. Here the error was just calculated to see how the model performs on data that was not used for training and therefore never was seen before by the network. This technique helps to prevent overfitting the training data, which would be signalized through an increasing validation error despite training error becoming smaller. Having the validation error after each epoch, this value was used for the learning rate scheduler to decrease the learning rate if the error does not decrease in a certain period.

To ensure to have a sufficiently trained network, the error scores got observed periodically. If the validation score did not improve more, i.e. the network convergence stagnated, and was sufficient, the training was stopped. Important to know that after each training iteration, which is also called epoch, the state of the model got saved, for further use. As also the scores for each model state were known, it was therefore possible to determine the best model, having the lowest MSE-score on the validation data. This one then was further tested and analyzed towards the applicability for audio (re)synthesis. Those further steps, which do not involve training, as well as the pre-processing regarding the evaluation data, were performed locally.

4.3 Initial Experiments

In the beginning of this project, initial experiments were done in order to produce a very basic proof of concept implementation. Like mentioned before, those implementations where entirely held on the local machine, because at the time there was no access to the previously mentioned high-performance GPU. This was also possible due to just taking a small subset of the test dataset. Of this test dataset, samples of one instrument (*keyboard_synthetic*) were taken out, and considered for test-wise training.

4.3.1 Whole Spectrograms as Input

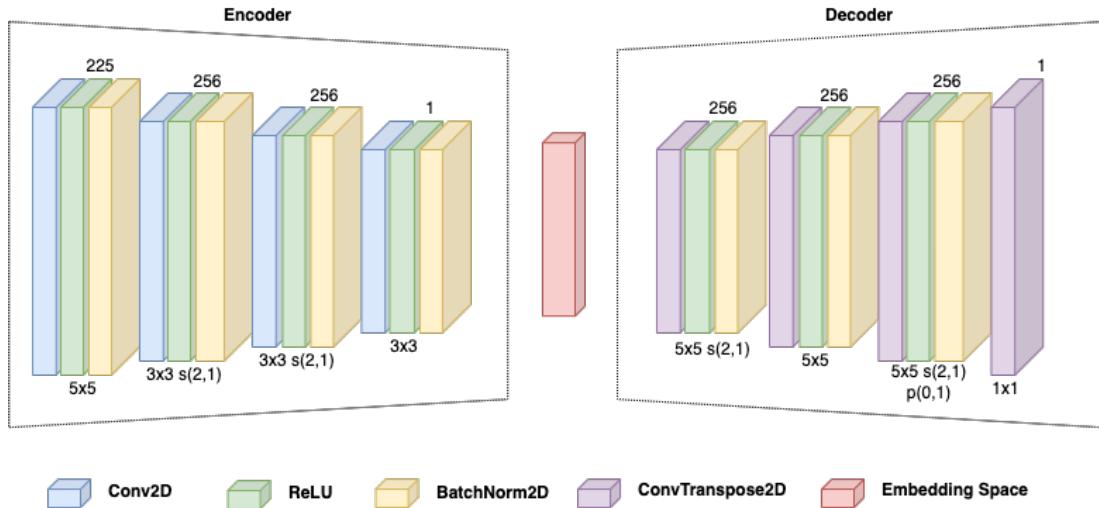
Pre-processing

For the first experiments, all the samples of the subset have been converted to log-magnitude spectrograms as a whole. As already known, those samples all have the same length of 4 seconds and are sampled at a rate of 16 kHz. Some samples are padded with zeros, as those do not contain audio data over the full length. As parameters for the STFT a `n_fft` of 512 and `hop_length` of 256 were chosen. In combination with the sampling frequency of 16 kHz this resulted in spectrograms having 257 frequency bins with a resolution of 31,25 Hz. Regarding the time-resolution it could be said, that each frequency vector represented 16 milliseconds of the original signal with a 50% overlap. This procedure resulted in spectrograms having a dimension of 257x250.

Model and training

As the spectrograms can be seen as grey-scale images, and thus 2D convolutions were applied, a third dimension was added resulting in 257x250x1 spectrograms. Finally to form a dataset ready for training, all the spectrograms of the mentioned subset were concatenated to a 4D array, which was converted into a tensor and subsequently into a dataset. The following training was performed on 80% of this subset, whereas the other 20% were used for evaluation. Regarding the duration of the training, this was held for a short time (~20 epochs).

To also mention the model configuration, this consisted of 4 convolutional layers in the encoder including ReLU activation and batch normalization. The decoder part had therefore also 4 layers, which contain respective convolutional-transpose layers with also ReLU and batch normalization except the very last layer. Another important detail was, because of the shape of the input, that the number of input channels had to be 1. Throughout the network, this parameter got varied, whereas in the first layer an expansion was set to 225, subsequently in the second to 256 output channels. To the innermost layer it was again reduced to 100. In the decoder part, again an expansion was made, whereas at the end it was reduced to 1 channel again, in order to match the input size. The next graphic (fig 4.1) shows the autoencoder that was used for the initial experiments. To be mentioned the kernel-size (e.g. 5x5), strides (s) and padding (p) here is already the configuration for the succeeding experiment. The channels, the general structure as well as amount of layers was the same for all initial experiments.

Figure 4.1: Initial 2D convolutional autoencoder.

Evaluation

Regarding the evaluation of this first model, it had been tested out on the remaining 20% of the small subset. These first experiments consisted just of evaluating the outcome of the decoder, by passing single spectrograms through the network. By this, the ability of the network to recreate audio spectrograms was proven, which further on serves as a base for the next experiments. Regarding the final outcoming sounds, the original preserved phase information was reused to recreate audios.

4.3.2 Spectrograms of signalframes as Input

In the previous initial experiment, the samples from the NSynth dataset have been taken as a whole for the experiment. It has been mentioned that all samples are 4 seconds long, but some contain padding in order to reach a length of 4 seconds. As those zero-paddings are then also part of the training data, this could affect the behaviour but also the outcome of the model. If those zero-paddings would be left out, this would lead to unequal samples, which results in training data unusable for the model. This is because the network has a fixed size of neurons at the input, which implies to have a fixed input shape. Furthermore this also means that having a fixed size of 4 seconds, the input always has to be of 4 seconds, which is not desirable. Not at least, if the system should be used in real-time applications for audio synthesis, one cannot wait to have 4 seconds of a signal, to perform audio synthesis with it. Therefore it would be desirable to perform audio synthesis on smaller “frames or chunks” of an audio signal, respective spectrogram.

Pre-processing

This leads to the idea to take chunks or frames of audio data that get transformed into distinct spectrograms of same size. For a start the length of those frames was set to 500 ms. Furthermore it was chosen that those frames are not consecutive, but have an overlap of 50%. Additionally those frames were multiplied with a window function, like it is used in the STFT. Similar to the window function used in the STFT, here a Hann-window is used. Again as those trimmed signals were all differently long, they had to be padded to a multiple of the frame size respective hop-length. This ensures to have equally long chunks of the signal. As of the windowed frames, the first and last frame didn't have overlapping parts at the beginning respective end. To overcome this issue, additional zeros were padded to form one frame on each side, to also get an overlap. In combination, having the framed and windowed signal chunks, by overlapping each again with 50% and adding the values, this would yield the original signal again. This would be especially helpful when reconstructing the final signal in the end.

Having those framed and windowed signal chunks, the STFT gets applied on those, having the same configuration as in the first experiment. This then leads to have multiple spectrograms for the length of 500 ms with again a frequency resolution of 31,25 Hz. Again for the training and testing, additional to the log-mag data, the phase information, reference value and name of the sample including a number to identify the frame got preserved for later use.

Model and training

The configuration of the model is rather similar to the one used in the first setting. It has the same amount of layers on each side, despite different strides, but also different kernel-sizes got applied (see figure 4.1). Again this one has been trained on 80% of the *keyboard_synthetic* test dataset samples. Of course the significant difference here is, that now the input data are not whole spectrograms but overlapping frames. This also means that the amount of data has increased. When collecting the spectrograms for the dataset, the names get shuffled, in order to not have the same order. In contrast, the single frames don't get shuffled during training. Again the training has been performed over 20 epochs on the local machine.

Testing/Evaluation

For the purpose of evaluating the test score this has been done with the remaining 20% of the samples. To evaluate the ability of the autoencoder to reconstruct spectrograms, the whole samples were used including those from the training. Here the spectrograms were provided in the order as they appear. Having all reconstructed spectrograms, the inverse STFT was applied with the preserved phase information. Resulting in the frames corresponding to every single input note, those got overlapped and added (in their right order). By this procedure the signal with the original length could be obtained and further on evaluated auditorily. For results see chapter 5.

A step for interpolating two different sources has not been investigated here, up to these experiments. Also the embedded space also has not been evaluated so far but will be subject of further experiments.

4.4 Experiments single frequency vectors

The above mentioned experiments, were a proof regarding the ability of convolutional autoencoders to recreate audio spectrograms. From this point on the experiments were done using the whole training dataset and also include audio synthesis. In the previous experiment the model was trained on frames of audio data which were 500 ms long. Having the idea to synthesise audio on real-time input, this would mean that always a signal frame of 500 ms has to be present, in order to have an input for the model. Therefore it is desirable to have an input that is as short as possible. An important note at this point is therefore, that spectrograms consist of frequency by time data. Each vector of the spectrogram along the time axis represents a short frame of time.

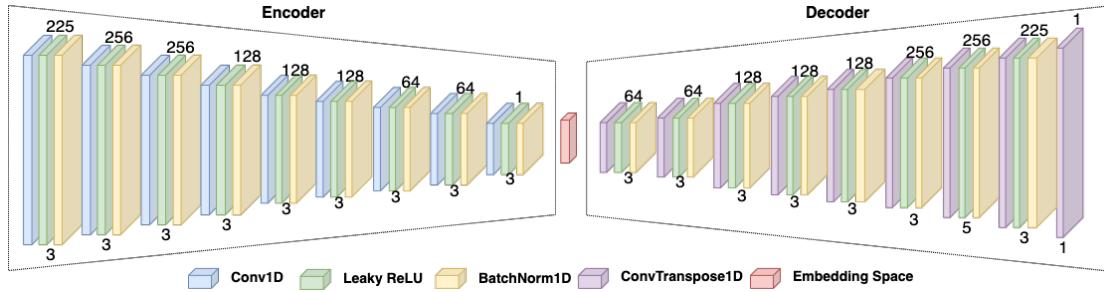
As mentioned before, when pre-processing was done on 16 kHz sampled data with an `n_fft` of 512, respective `hop-length` of 256, a time resolution of 16 ms was therefore present. Therefore the idea was to take those single frequency vectors as input for the model. The shape therefore is frequency by channels which in the case of the previous parameters is 256x1. Having this idea, also the silence at the end of the signals could be omitted as just the frequency domain was used for the models input. This therefore enabled taking samples of different length in time. Throughout the experiment the value for the `n_fft` was increased to 1024, as an increase of the models performance could be achieved. Increasing this parameter therefore meant to decrease the time resolution, but increased the amount of frequency bins and thus having a better frequency resolution. Resulting in a number of 513 frequency bins (15,625 Hz/bin) and time resolution of 32ms per vector.

Neural network

As a consequence a different model had to be used, as 2D convolutions were no more suited. Therefore a model was designed that uses 1D convolutions. 1D convolutions are performing the same calculations, with the difference of having just a 1-dimensional kernel. This one dimensional kernel just operates on the frequency axis and tries to extract important features. Equally to the 2D convolutions they also apply the principle of channels. Regarding the channels, those also get expanded but then subsequently until the innermost layer reduced to 1 in order to form a single dimensional vector. Until the end again those channels get expanded but reduced again to 1 to have the same shape as the input vector. In figure 4.2, the structure and configuration of this autoencoder is depicted.

Training

The training process for this experiment differs in some major points from the previous ones. The main difference lies in the shape of the data. This can be concluded as here just the single frequency vectors were used for this network and therefore the network itself has 1D convolutions as discussed before. Furthermore, this experiment was originally trained locally with the small subset like before, but later on, access to a GPU-accelerated instance was made available. This GPU instance, as discussed at the beginning of this chapter, enabled training complex ML models with a large amount of data, over a long time. With this possibility, the full training dataset could be utilized

Figure 4.2: Deep 1D convolutional autoencoder.

as it consists of several thousands of audio samples and would be too big to use locally. This does not mean that the whole dataset was used all the time for training on this instance. With the progress of the project, several trainings with different amounts of data were made, that impacted the training process.

As here single frequency vectors were used, the process of creating a tensor and subsequently the dataset for the dataloader, was different. Here all spectrograms of the pre-defined and pre-processed dataset were utilized, and the single frequency vectors get concatenated to form one big 3D array in the shape of amount by channels by frequency. To note the amount of channels here is also initially one. While experimenting, regarding the training, also different strategies regarding shuffling were used. First on just the samples were shuffled resulting in the instruments being mixed, but the frequency vectors were in the same order as in the spectrogram. This stayed the same as after each epoch the data didn't get shuffled. One more successful strategy was to also shuffle the whole generated array, resulting in the frequency vectors of all instruments being mixed. Furthermore shuffling was also done after each epoch. The same strategies were also applied to the validation dataset, as from this point on also the validation dataset was considered, partly due to more computational power was present from this point on. A batch-size of 32 was considered.

Introducing the validation step from this experiment, also the scheduler to adapt the learning rate during the training was applied. Taking into account that the network is rather complex and a huge amount of data has been used, the learning rate also had to be chosen adequately. Throughout these experiments, a small learning rate of $1e - 7$ was chosen, as this led to a more stable training and better convergence.

Testing

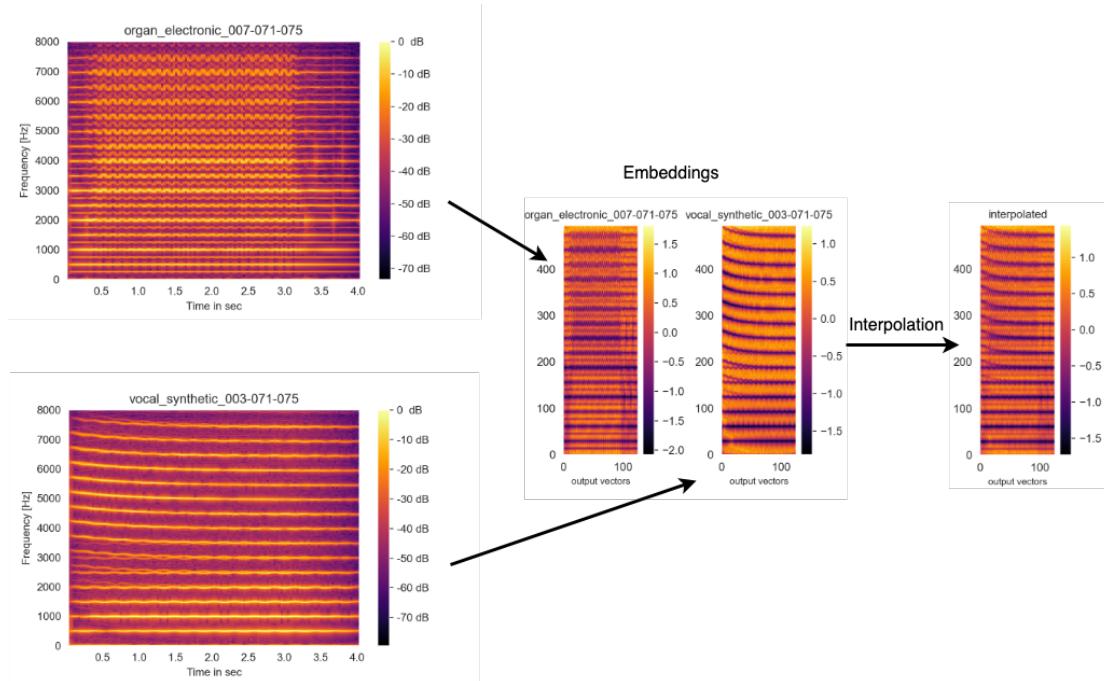
As described above, after each training, the current model state was saved, those can be used for the testing and evaluation regarding synthesis. As having numerous states that got saved, the one with the lowest error score on the validation set got considered for further steps. For the steps around testing and evaluation, the held-out test dataset was used. This stage was configured to either take the whole test dataset, a specific pitch, or a specific instrument source. By this technique its made possible to get the scores and therefore performance for certain instruments or pitch.

4.4.1 Experiments for Synthesis

Having this kind of network that was trained on the training dataset, to reconstruct single frequency vectors, some experiments have been conducted for generating audio. As during the training the network learned how to reconstruct frequency vectors, one experiment is to examine the quality of the output for a single non-modified audio. Those are similar to the ones conducted in the first experiments above. Also by reconstructing just single notes, the preserved phase information could be reused, which enabled application of the ISTFT. For a comparison also the Griffin-Lim algorithm was utilized.

As the main objective of this work was to examine the capability of creating novel sounds, from this point on in the project the interpolation step was introduced. With the interpolation step the encoded features in the embedded space of two instruments were taken and value-wise interpolated. As it is known at this point, the encoder part of this network takes as input a vector of 513x1 and creates a lower-dimensional representation of it with the size of 495x1. Those representations can be seen as the essential features, and were considered for the interpolation task. For this task, the frequency vectors of two instruments of probably the same pitch were passed through the network. Here the data did not get shuffled, as it was important to produce the values in the same order as they come from the spectrograms. The output of the encoder for each instrument was concatenated to a 2D array 495xN where N is the number of encoded frequency vectors. To get a better idea of this concept, figure 4.3 shows two spectrograms with the corresponding accumulated output of the encoder.

Figure 4.3: Input spectrograms with embeddings and interpolated embedding.



Having those two representations, interpolation was performed. Like discussed in chapter 3, along the x-axis each output vector of one sample was interpolated with the corresponding output vector of the second sample. This process is also shown in figure 4.3 where the result of the interpolation process is shown. This process got applied equally for each subsequent experiments as the encoder output is always of the same shape, but more on that later on. This interpolated vector was subsequently passed through the decoder part which forms again frequency vectors of 513x1 that were accumulated to a spectrogram. This spectrogram then depicts the result of audio synthesis combining features of two different instruments.

Further on this spectrogram was converted back to audio domain. As in this case no phase information is present, the Griffin-Lim algorithm [8] for phase estimation was applied. By listening to the final sound, it should be possible to hear the characteristics of both instruments combined in this sound.

4.5 Experiments with slices of spectrograms

Based on the results and insights that could be gained with the previous experiments (see chapter 5) some more experiments were made. Not at least to examine how different representations of the input data and as a consequence different model configurations influence the task of neural audio synthesis. For this case the following experiments in contrast worked with 2D representations as input. 2D data has already been used in the initial experiments with the whole spectrograms, but also spectrograms based on frames of the audio signal. As a difference for the following experiments, the spectrograms have been taken, but sliced into overlapping chunks. The spectrograms taken in this experiment were generated with the same configurations as in the previous experiment (`n_fft=1024, hop_length=512`). The idea to use chunks of spectrograms as input arose to improve the models performance with regard to synthesis and recreation of spectrograms. As by using single frequency vectors, just the frequency information is present and no information about its change over time. As a theory to also incorporate the temporal axis as input, important information about the temporal frequency changes could be captured. Those frequency changes could deliver more important characteristics of the samples that could be extracted. As not the whole spectrograms should be used as input, chunks of them are favorable, to keep the input as small as possible. Therefore the choice has been made to take three consecutive frequency vectors to form a frame with the shape of (513x3). Throughout the experiments it has been tried out to either take the frames subsequently or overlapping, leading to the preference in taking the latter. The overlap has chosen to be always 2 vectors from the preceding frame.

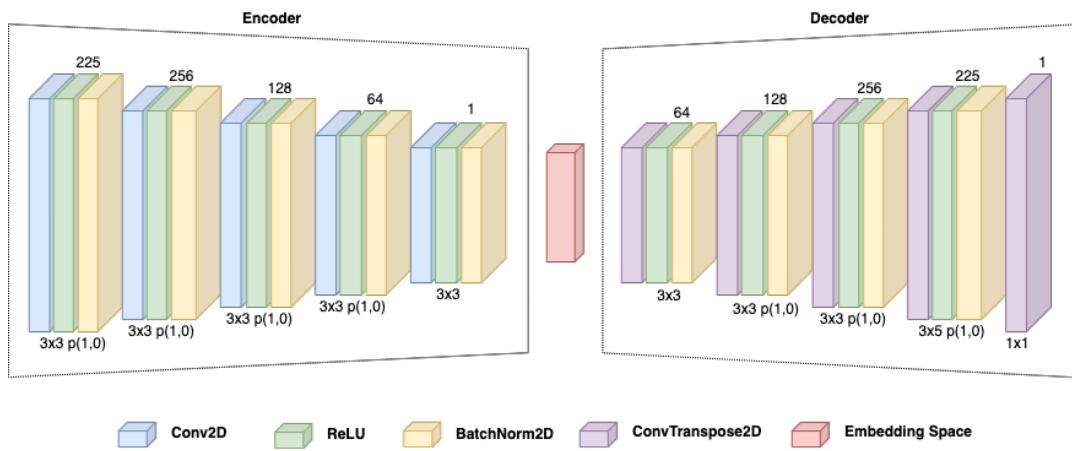
For demonstration, let there be a spectrogram $spec = [f_0, f_1, f_2, f_3, f_4, f_5, \dots, f_n]$ of length n where f_i is the frequency vector at index i . This results in an array of $specs_slices = [[f_1, f_2, f_3], [f_2, f_3, f_4], [f_3, f_4, f_5], \dots, [f_{n-3}, f_{n-2}, f_{n-1}], [f_{n-2}, f_{n-1}, f_n]]$. The idea behind this overlap was to capture every change or time-pattern in the source spectrogram. This technique also had the advantage that all the silence of the source spectrograms could be cut away, allowing to use differently long spectrograms as input. A positive side-effect here was also to gain additional input data, on which the model could be trained on. As the models output are also overlapping frames, to reconstruct the final spectrogram and audio, this also had to be considered, but more on that later

on. Again when creating the tensor and dataset, the spectrograms were taken in a random order. At first the resulting frames of the training set didn't get shuffled. Later on during the training, the frames in the batches, which again are of size 32, were shuffled like in previous examples. The same happened to the validation dataset, while it also was shuffled initially and after each epoch. An important note is that for the training and validation the whole dataset were used all the time.

Neural Network

As again input data in a 2D shape was used for these experiments, the model again had to be adapted. Instead of 1D convolutions and 1D batch normalization again, 2D convolutions and 2D batch normalization has been used. Also in contrast to the previous network, a normal ReLU activation has been used. These experiments incorporate different model configurations whereas a special focus has been given to the amount of striding and thus input compression. Here it should be evaluated, how more compression and thus a smaller latent space influences the quality of the decoder output and further on of the generated audio. Both in terms of single note reconstruction and interpolation based synthesis. The following figure 4.4 shows, the basic network structure for these kinds of experiments.

Figure 4.4: 2D convolutional autoencoder.



This network is not as deep as the one used in the previous experiment because it consists of 5 layers on each side (10 in total). As the input is of size 513x3, the convolutions had to use a padding along the time axis. If no padding would be applied, just one time a kernel of 3x3 could be applied. Thus 5 times a 3x3 kernel could be applied with the last layer having no padding. This results in a single 1D vector in the embedded space. In the decoder part, padding also had to be applied to regain the same dimensionality in the end. As previously said, these experiments should give an insight on how the striding and thus the size of the embedding influences the performance and quality of the output. Choosing more strides and a smaller embedding, the network has to learn to extract more efficient encodings, from which it can reconstruct the input data. Thus by choosing a small embedding size, the network should learn to extract the

most important features of the input. Therefore the size is also of significance regarding the audio synthesis task by interpolating the embeddings. In the following table 4.1, the different configurations regarding the striding are shown.

	Encoder	Embedding-size	Decoder
Single-Stride	e2	250	d4
Double-Stride	e2, e4	124	d2, d4
Triple-Stride	e2, e3, e5	62	d1, d3, d4

Table 4.1: Setting of strides in network.

To explain, each row represents a network configuration, that either uses one, two or three times a striding of (1,2) on each side of the network. As of the shape of the input, striding can just be applied on the frequency axis. The columns with the names “Encoder” and “Decoder” show the respective layers where the stride was applied. For example e2 is the second layer in the Encoder and d4 the fourth layer in the decoder. The center column called “Embedding-size” explains the size of the embedded space vector. By having these model configurations, interesting findings and results could be obtained.

4.5.1 Audio synthesis

In the previous experiment with the 1D convolutional network, the interpolation step has been introduced. This step was also applied with this network. As the embedded space vector also has the shape of a 1D vector, the interpolation procedure here was exactly the same.

Analysis of the embedded space

As with neural audio synthesis novel interesting sounds should get generated, it is also of interest to find interesting combinations for the interpolation. Combined with the fact that the embeddings contain the extracted features/characteristics of a sound, those can be utilized for this task. It therefore has been implemented to take the output of the encoder of several notes (e.g. from the same pitch) and compute the correlation coefficients between those. The result got depicted in a correlation matrix to see which samples have the lowest correlation coefficients. Low correlation coefficients between two data samples mean that those have little similarities and thus different characteristics. By taking those embeddings with the lowest correlation coefficients and interpolating them, interesting novel sounds can be generated. More on that in the next chapters.

4.5.2 Reconstruction and post-processing

As the shape of the output of the network is also the same as of the input of the network, this had to be considered when recreating the spectrogram. The output therefore were also again frames of 3 that had to be overlapped. During the experiments different strategies have been tried out, whereas it was preferred to average the overlapping parts of the output frames, in order to form the final spectrogram.

When having the final spectrograms, some further experiments regarding the improvement of the sound quality have been carried out. In this case this includes to correct energy in the frequency bands of the output. With this technique important properties of the sound like the transient or impulse (e.g. guitar stroke) should be preserved. For this task in advance the energy-values of the frequencies were summed up for each frequency vector of the original audio spectrograms. This was also done for the output spectrogram and its frequency vectors, after converting from dB to energy. By comparing those sums with the corresponding values of the input spectrogram, a factor could be calculated. This factor then was multiplied with the corresponding output frequency vector in order to have the same amount of energy present in the output spectrogram. If the output spectrogram was generated by an interpolated embedding, the energy sums of the two input samples were taken and averaged. These averaged values then were taken as a reference to correct the energy. By performing the inverse STFT or Griffin-Lim algorithm, the resulting audio sample should therefore also have a corrected amplitude and thus improved sound. Regarding the results, those will be discussed as well later on in the thesis.

4.6 Experiments with mel scale

Up to this point, the experiments have all been conducted on log-magnitude spectrograms. Those spectrograms depict the frequency energies on a linear scale with a certain resolution (e.g. [0 Hz, 15,625 Hz, 31,25 Hz ... 8000 Hz]). Those spectrograms have been computed of a dataset consisting of musical notes. Those notes can be grouped into musical intervals describing the distance of a note to another, whereas the interval of an octave would be a doubling in frequency. As an example the note a' has a frequency of 440 Hz (if perfect sine), whereas its octave on top would be an a'' with 880 Hz. If going an octave down it would result in 220 Hz having an a (to be continued A 110 Hz, A' 55 Hz, ...). This also means that the higher the note gets, the larger the distance in Hz and the lower it gets, the less distance in Hz between the notes. Having this principle one can come to the conclusion that having a linear scale like in log-magnitude spectrograms, the resolution with lower notes is worse than with higher notes. With a look onto machine learning and neural networks, this also means, that there is less data available to train on for low notes. Keeping this in mind, the choice has been made to perform comparative experiments based on a different scale, which is called mel scale [27]. This scale can be seen as a compressed form of a spectrogram. Furthermore it is an empirical scale that is based on the human auditory perception, as they perceive frequency logarithmically rather than linearly. Mapped on the mel scale, in contrast to magnitude, low-frequency contents have a better resolution in contrast to the high-frequency range [18]. This means that with this scale lower notes can be better differentiated and thus are more emphasized than with the linear scale. Even though that this scale is based on observations it was proven to be beneficial regarding machine learning tasks. Not at least as several approaches, mentioned in related work, applied this scale for their audio synthesis task. Furthermore this scale could also be seen as a compressed form of a spectrogram containing the most significant properties.

4.6.1 Pre-processing

The pre-processing step does not really differ from the ones performed in previous experiments. A difference here is in order to obtain spectrograms with the mel scale to call the *librosa* function `librosa.feature.melspectrogram()` which takes as input a pre-computed STFT-spectrogram. Throughout this experiment the spectrograms created with an `n_fft` of 1024 and `hop_length` of 512 have been taken to be converted to mel-spectrograms. The resulting mel-spectrogram has a size of $128 \times t$ and thus is smaller than the mag-spectrogram ($513 \times t$). Having the mel-spectrograms the same steps as with log-magnitude were performed (dB conversion, preserving the power reference, ...). Also for the model training, frames of 3 vectors of the mel-spectrogram were taken as input for the model. In the next figure 4.5 a mel-spectrogram with its corresponding log-magnitude spectrogram is displayed, to see the difference.

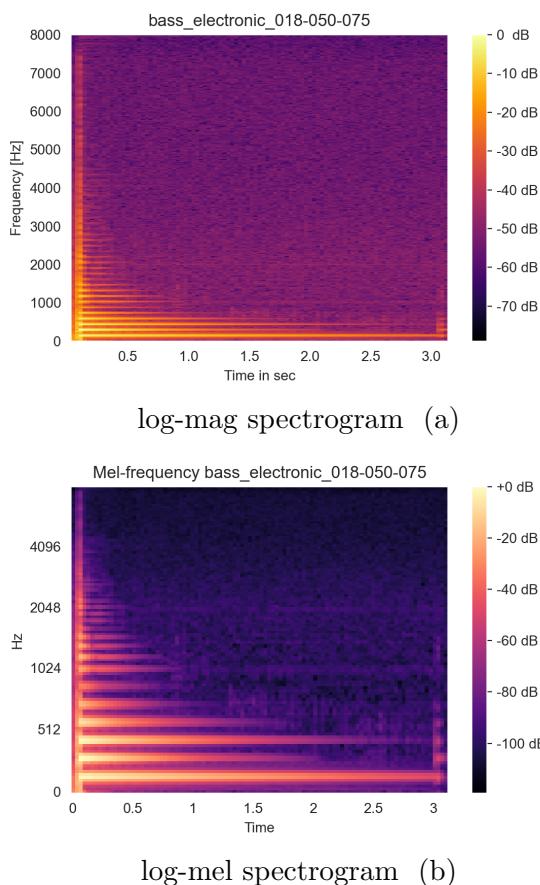


Figure 4.5: Difference between log-magnitude and log-mel spectrograms.

By comparing those, it can be seen, that the scale is logarithmic and the distance between the low frequencies greater than between the high-frequencies. As of this known characteristics and its already compressed form, the experiments using these kind of spectrograms thus can be expected to bring different but interesting results. Here also the same experiments with and without interpolation got conducted to get a comparison.

Neural Network

Like in the previous experiment, also three different networks with single-, double- or triple-strides on each side were applied and assessed further on. The basic network structure and configuration does not differ of the one used in the previous experimental setting as having the same amount of layers and same configuration of the channels. Nevertheless some differences arise, because the last layer applies a kernel of 1×2 to regain the same input shape. Also in the networks using three or two times strides on each side, padding in the layer e_4 and d_4 is set to 1,1. Regarding the network with three strides on each side, those have been set onto layers e_2 , e_3 , e_4 and d_2 , d_3 , d_4 . As the input of the network is already of smaller size, the corresponding embeddings are therefore significantly smaller like in the previous experiments (56, 28, 14). Those were also altered like in the previous experiments by interpolation to synthesize novel sounds.

Reconstruction and post-processing

The steps around the reconstruction and post-processing are similar to the previous experiments. Because as input also frames of 3 were used, the output consists of 3 frames too. Those also were overlapping and averaged to form a final log-mel spectrogram. Again also the energy got corrected to have the same amount of energy present like in the input. Having the final log-mel spectrograms, those could be either directly converted to an audio using the function `librosa.feature.inverse.mel_to_audio` or first to an magnitude spectrogram. The latter is used when just single audio samples were reconstructed (without interpolation) as here the phase information could be reused, and thus the ISTFT can be applied to obtain the final sound. Otherwise the mentioned function is performing the conversion to magnitude and Griffin-Lim at once to obtain an estimated signal.

Chapter 5

Results

This chapter shows the results that could be obtained through the experiments, mentioned in chapter 4. Those results contain either numbers, for assessing the model performance with a certain error performance, but also of graphics, displaying the outputs of the models. The outputs of the models, being spectrograms, are compared with their corresponding input spectrograms. As the output of the encoder part (embedding) plays a crucial role regarding audio synthesis, those also get displayed and assessed further on. Some additional graphics containing additional signal plots can be found in the appendix A. Regarding the final listenable sounds (that are provided in supplementary material), those get discussed in the next chapter.

5.1 Results regarding model with single frequency vectors

As described in chapter 4 experiments have been conducted by using the single frequency vectors of the spectrograms. A 18 layer deep neural network with 1D convolutions has been trained on the reconstruction of single frequency vectors. Throughout the development different settings have been tried out in order to find a model with an optimal training process and performance. First trainings have been made on `keyboard_synthetic` mixed with `guitar_acoustic`. As a note, those two terms are labels from the dataset representing the instrument family and source. Here it could be observed that no sufficient convergence could be reached ($MSE > 200$). Training a distinct model per instrument source showed that the instrument source has a significant influence on the convergence. For example a model trained solely on `keyboard_synthetic` converged better than a model with `guitar_acoustic`. Comparing the performance of a network trained on `guitar_acoustic` with one trained on e.g. `guitar_electronic`, showed that with the latter a better score could be reached (~17). Therefore the decision was made to combine `keyboard_synthetic` and `guitar_electronic` instead of acoustic, for the training. Compared to the one mixed with `guitar_acoustic` a score of ~78 could be reached. Subsequently more instrument sources were added, until the whole training dataset was utilized. This model showed the best performance regarding its error scores and was chosen to be considered for the ongoing experiments. Table 5.1 shows the error scores of this model. As a short note, each trained model was chosen based on the best validation error score.

	MSE-Score
Training	4,237
Validation	4,638
Test	1,190

Table 5.1: MSE-Scores 1D convolutional autoencoder.

Here it can be seen that the error scores for training and validation are rather close, while the score on the held out test set was significantly smaller. This score could be reached with a training over 8 epochs and was considered as the best score, as after those 8 epochs the validation score increased again (despite decreasing training score). Table 5.2 is also interesting as it shows the error scores on the test set, regarding different pitches. As displaying all pitches would go beyond the scope, pitch classes from 030 to 100 have been chosen. In this case it can be seen that all error scores have roughly the same value and do not differ significantly. Nevertheless there is a small trend visible if the score increases depending the pitch.

Pitch	MSE-Score
030	1,049
035	1,393
040	0,935
045	1,096
050	1,019
055	1,108
060	1,086
065	1,111
070	1,164
075	1,041
080	1,177
085	1,753
090	1,100
095	1,796
100	2,043

Table 5.2: MSE-Scores for specific pitch classes using 1D convolutional autoencoder.

5.1.1 Experiments of single reconstruction

Having the trained network, this one got evaluated towards the ability of reconstructing audio spectrograms and further on recreating the sounds. The next graphics 5.1, 5.2 and 5.3 show the result of taking a guitar sample as input and reconstructing it.

With a special look onto the reconstructability of spectrograms, figure 5.1 shows the input spectrogram and the generated output spectrograms from the network. Here it can be seen that the reconstructed spectrogram differs in a few points from the input

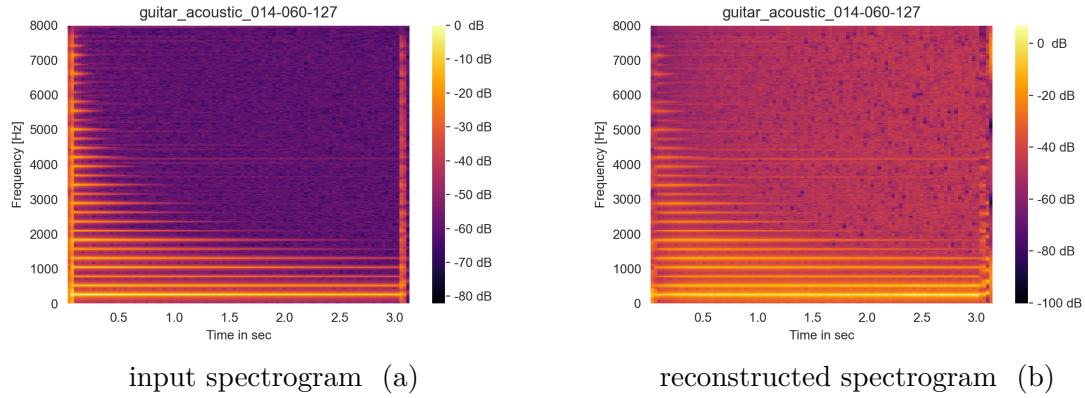


Figure 5.1: Difference between input and reconstructed spectrogram.

spectrogram. First of all it can be seen, that across the frequency areas that contain little energy in the input spectrogram, in the output more energy is present. This also means that between those areas and the sound-characteristic high energy areas, less difference is present. Furthermore regarding the original broad spectra at the beginning and at the end, those are hardly present in the output spectrogram. Knowing that these represent the stroke and the damp of the string, it can be said that these are not present in the output. This is also confirmed through looking at the time-domain signal.

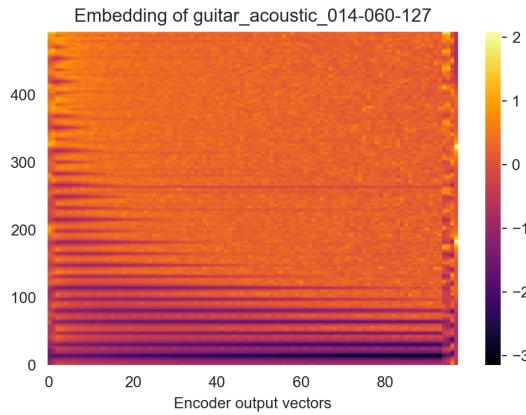


Figure 5.2: Embedding of guitar acoustic.

Looking at the graphic that depicts the embedded space, there it already can be seen that those broad spectra are not preserved. Despite of that the high energy areas (harmonics) in the input spectrogram are preserved as negative values in the embedding while the original low energy areas get represented by positive values. Also as there are no strides applied in the network, there is almost no compression in the embedding.

A final look onto the time domain plots in figure 5.3 also reveal, that there's no impulse at the beginning of the signal. Furthermore it also can be said that the amplitude in general differs in its course but also in how strong it is.

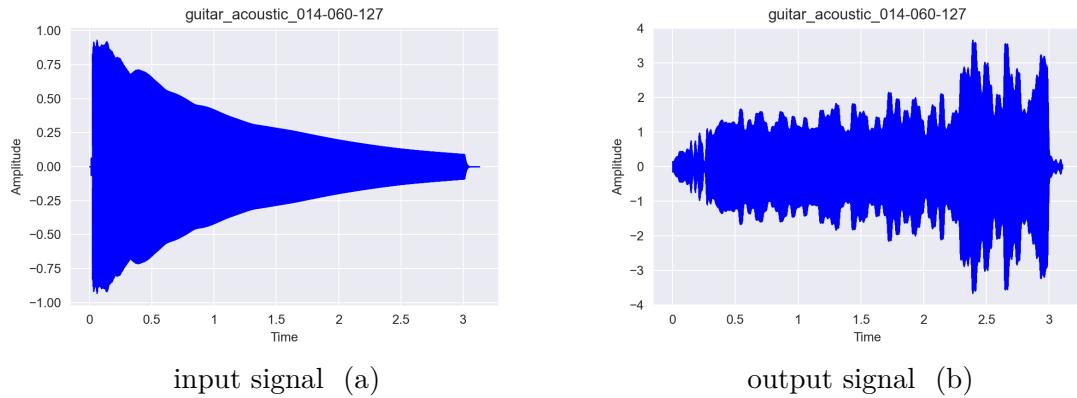


Figure 5.3: Differences between output and reconstructed time signal.

5.1.2 Experiments with interpolation in embedding

With this kind of 1D convolutional autoencoder first experiments were made, that use interpolation in embedded space to synthesize novel sounds. In chapter 3 and 4 the concept has already been described in detail. Having this concept, instrument sources were chosen, which in order got encoded and interpolated together in order to form a new embedding. Figure 5.4 shows the spectrograms of the two input instruments that were chosen to create a novel sound. For this representation an acoustic guitar and acoustic brass sample were taken and encoded. By comparing those two spectrograms, it can be seen that they differ significantly in their structure. While it can be seen that the guitar loses energy in its higher frequency ranges, the brass sample keeps its frequency energy over the whole time. Furthermore as the harmonics stay constant in the guitar sample, those alternate in the brass. Those properties make it interesting to use them for the interpolation and thus synthesise a novel sound. For comparative reasons, in further experiments also those two instruments have been used for the synthesis task which can be seen later on.

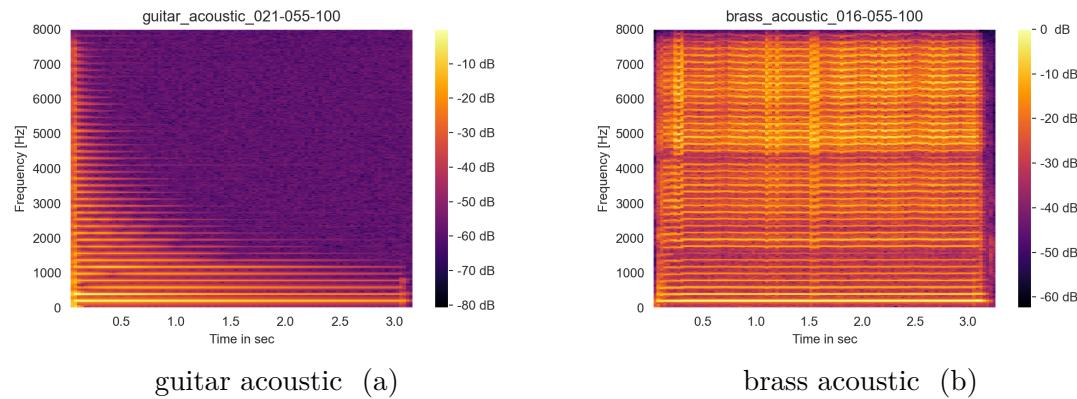


Figure 5.4: Input spectrograms.

After encoding, the outputs were taken and interpolated as can be seen in figure 5.5a. As it can be seen, those embeddings are a compressed form of the input spectrograms. When looking onto the generated interpolated embedding, it can be said that this one incorporates both instruments encoded features. Having this new vector, this one was fed into the decoder network in order to generate an output spectrogram that can be seen in figure 5.5b.

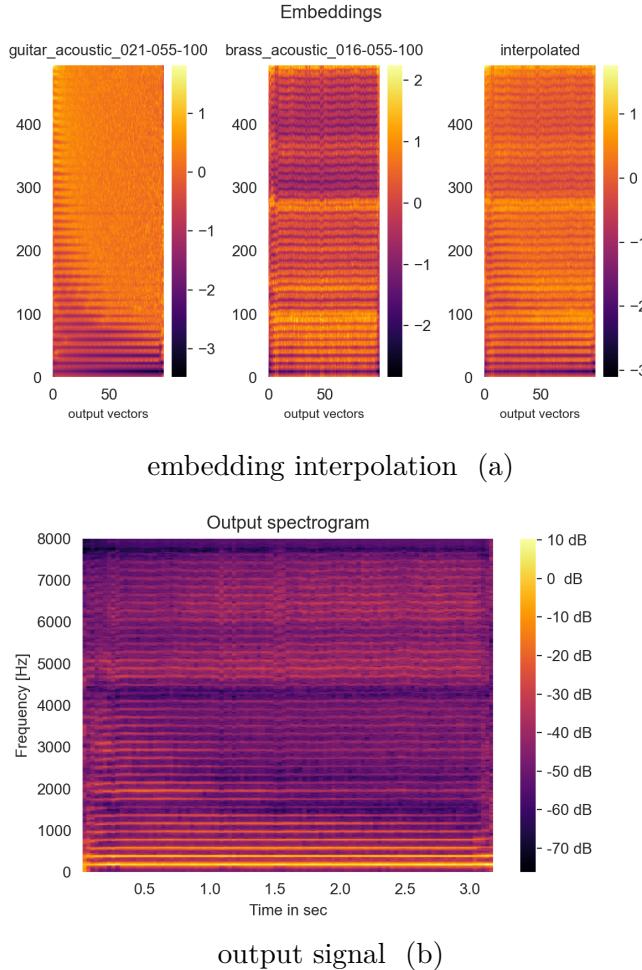


Figure 5.5: Embeddings of input samples with interpolated embedding and its corresponding output spectrogram.

Having the final spectrogram here as output, it can be seen that spectral data of both input spectrograms are basically contained. Similar to the output spectrogram in figure 5.1b this one also does not contain the impulse (guitar stroke) at the beginning. To finally obtain a listenable sound, the output spectrogram has to be converted back to time-domain, which in this case is done with the Griffin-Lim algorithm [8] as no phase information is present. To improve the quality of the output but also to examine the performance of other types of networks some further experiments have been made using 2D convolutions but also additional post-processing steps.

5.2 Results of experiments with spectrogram frames

The here shown results correspond to the described experiments in section 4.5. Here again a 2D convolutional autoencoder has been utilized to (re)synthesize audio. As input data for the neural networks, overlapping spectrograms snippets with the length of 3 along the time axis have been considered. In contrast to the previous network, this was supposed to bring better results, regarding the output but especially to preserve transients. As different networks with different stridings (compression) were trained and used, the results should also show the differences in the output but also in their quality regarding audio (re)synthesis. Additionally the encoded data gets examined by calculating the correlation coefficients between samples (mostly of the same pitch). By that it can be seen how much the encodings of different instruments differ. Additionally it also can be examined how good the different networks extract the essential features of the input data. Throughout the project these correlation coefficients got depicted in a so-called correlation matrix. Displaying one here would not work out sufficiently, as because of its many values, those are hardly readable without zooming in. Because of this, just the findings are discussed later on in chapter 6. Finally also additional post-processing steps get applied, which help to correct the energy and thus the quality of the output. The results shown here are without energy correction, but the effect will also be discussed later on in the discussion.

The following table 5.3 shows the MSE error scores of the different 2D convolutional networks. As mentioned in chapter 4, three networks were trained, that differ in their configuration regarding striding and thus have different embedding sizes. These networks therefore are called single, double- and triple stride networks, regarding the amount of striding on each network side. By this they can be distinguished better in this work. In table 5.3 it can be seen that by using a different amount of strides, this has an impact on the score. The network with just one stride on each side has the lowest error whereas the networks with two or three strides have a significantly higher error. Regarding the difference between training, validation and testing error all three networks show the same behaviour as the validation error is higher than the training error with the test score being the best.

	Single-Stride	Double-Stride	Triple-Stride
Training	9,779	13,717	18,292
Validation	10,094	14,056	19,056
Test	7,831	10,711	16,655

Table 5.3: MSE-Scores 2D convolutional autoencoder - single-stride, double-stride, triple stride.

Compared to the error score of the 1D convolutional network, it can be said that all error scores are significantly higher. In case of this work, the score does not mean that the quality of the output is worse, especially regarding the final output sound. The latter will be discussed later on in the discussion of this work.

As during the previous experiment also the scores regarding different pitches were calculated, the next table 5.4 shows the different scores regarding pitches ranging from

30 to 100. Similar to the 1D convolutional network, the error scores regarding the pitch do not show a specific trend. Interestingly the scores of the triple-strided network have some rather strong outliers regarding samples around the classes 045-050. Having the same error scores this also does not mean that all pitches have the same audible quality, but more on that in chapter 6.

Pitch	Single-Stride	Double-Stride	Triple-Stride
030	7,084	9,175	12,241
035	7,912	12,294	15,426
040	7,292	11,205	18,639
045	7,251	12,434	22,151
050	6,840	9,903	20,287
055	7,535	10,434	18,102
060	7,444	10,128	17,670
065	8,313	10,286	16,497
070	7,410	9,868	14,894
075	7,849	9,949	14,273
080	8,624	10,529	15,486
085	9,299	11,090	16,598
090	8,798	11,435	16,415
095	9,948	12,032	17,049
100	12,346	13,179	19,812

Table 5.4: MSE-Scores for specific pitch classes using 2D convolutional autoencoder.

5.2.1 Experiments of single reconstruction

Having mentioned the error scores regarding reconstructing spectral audio data, those cannot be taken solely to assess the performance of the network. Therefore experiments, like with the previous network, were conducted in recreating single spectrograms. With those the ability towards reconstructing audio spectrograms becomes assessed visually and auditorily for the purpose of audio (re)synthesis. For comparative reasons, the same spectrogram source has been chosen like with the previous network. As the original input spectrogram has already been shown in figure 5.1a, here just the output of the encoder part (embedding) but also the total output gets depicted (see figures 5.6, 5.7 and 5.8). The spectrograms in each of these graphics were generated with single-, double- and triple-stride networks and do not contain the energy correcting post-processing mechanism. Despite of this fact, it can be seen when looking onto all the output spectrograms (5.6a, 5.7a and 5.8a), that all preserve the broad spectra at the beginning and ending of the spectrograms, especially by looking onto the first one. When comparing again the low energy areas of the input spectrogram to the output spectrograms, it can be said that they also contain similarly little energy. Again it can also be seen that the embedding looks like a spectrogram but in a compressed form of the input, as it also contains similar structures. Contrary to the embedding in figure 5.2 the high energy areas have positive numbers while original low energy area have negative numbers.

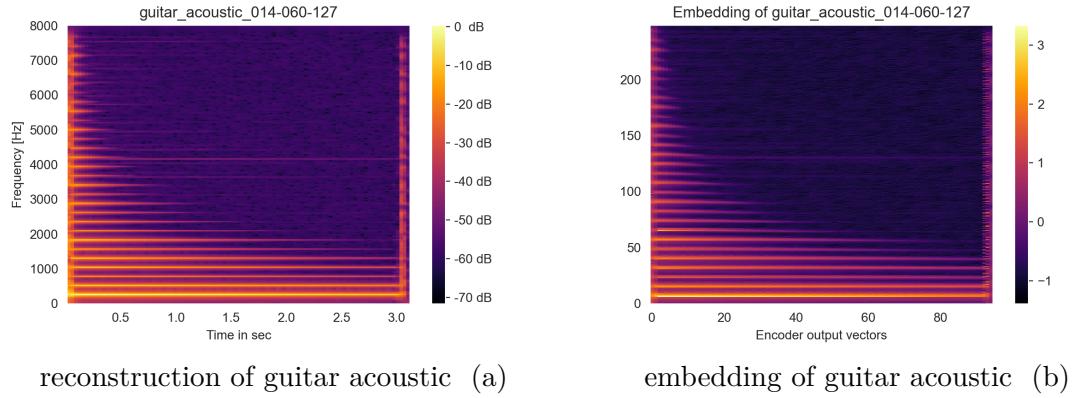


Figure 5.6: Reconstructed spectrogram and corresponding embedding - single-stride network.

With a look on the output of the double-stride network (figure 5.7a), it also can be said that the broad spectra are preserved but with less energy. This can be also noticed when looking on the output spectrogram of the triple stride network (figure 5.8). The latter also shows less energy in the high energy areas and less “precise harmonics” (washed out).

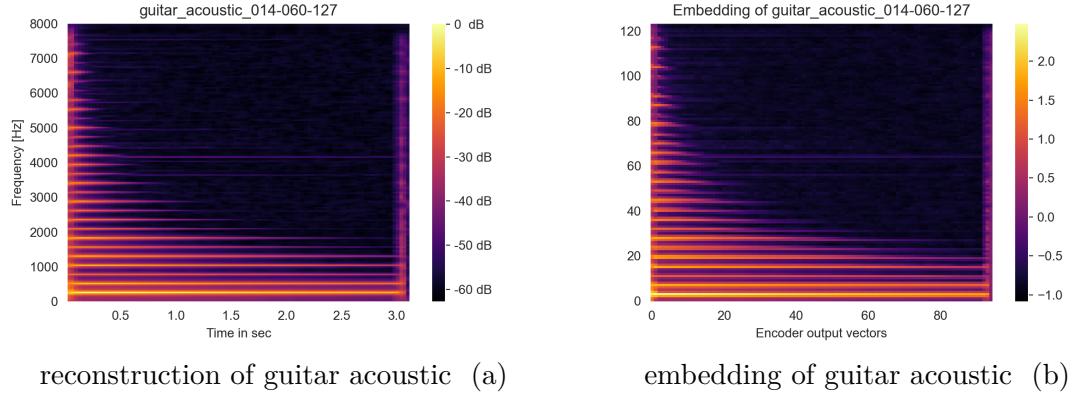


Figure 5.7: Reconstructed spectrogram and corresponding embedding - double-stride network.

Looking onto the embeddings of the two networks it can be said, that those contain significantly less values across the y-axis. Comparing it to the input but also the output, despite of the significant compression, the significant harmonic features and general structures are present as positive numbers (regarding the colorscale). As concerning the double-strided network, the embedding still has a fine granularity contrary to the one obtained by the triple strided network.

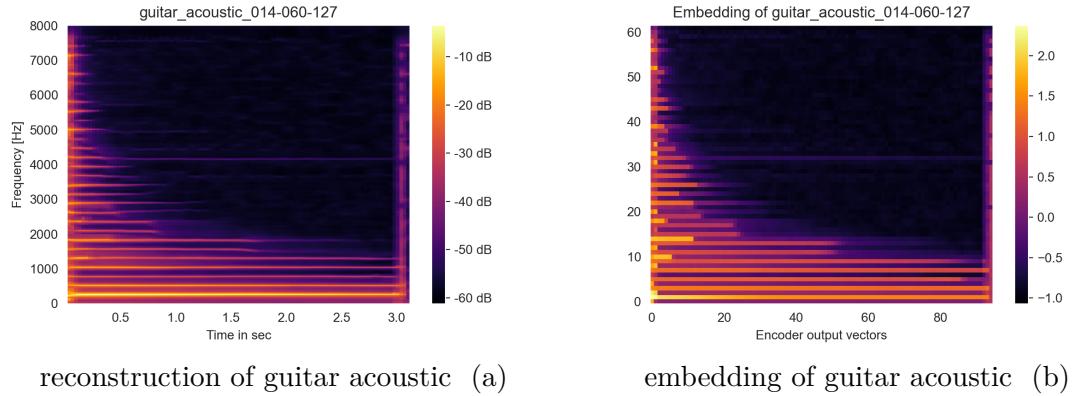


Figure 5.8: Reconstructed spectrogram and corresponding embedding - triple-stride network.

5.2.2 Experiments with interpolation in embedding

The results in this section show the resulting output spectrograms that got generated by interpolation of the embedding space vectors. This has already been done with the 1D convolutional network where novel sounds could be generated. As with the previous experiments with reconstructing single spectrograms, these experiments yield interesting results too. Not at least as the embedding got more compressed, this also effects synthesizing new audio, as this is done by interpolating the embeddings. The following graphics (5.9, 5.10 and 5.11) show again the interpolation between the same two instrumental sources, for comparative reasons. Again also the resulting output spectrograms are displayed to see the final result. The auditory quality again gets assessed and discussed in the next chapter. When looking at the embeddings in the three graphics, that after interpolating, the features of both instruments are visible in the result. Here it can be seen that the “harmonic” features of the brass sample do not fade contrary to the guitar samples. Nevertheless because of the interpolation the features are present but having lower values as it interpolates mostly between negative and positive numbers. The areas where both samples have common features (lower harmonics), rather stay equally valued as those have close values in both embeddings.

By looking again on the decoded output it can be seen that again both instruments are present. Compared to the result in the first experiments, the guitar sample is more present with these kind of networks. Contrary to the experiment with single sample reconstruction, here the difference between the different strided networks can be noticed even more. With special notice to the higher harmonics of the original brass samples. Those harmonics appear more precise with the single-strided network (see figure 5.9b). In the output of the double-stride network in figure 5.10b the “contours” coming from the brass sample are less sharp than with the single-stride.

Taking the output of the triple strided network (figure 5.11) into account, the difference to the other networks can be seen clearly. The encoded features again contain the same structure as the input spectrograms, but due to the three-fold striding, just the most significant information is present. Comparing the interpolated embeddings it

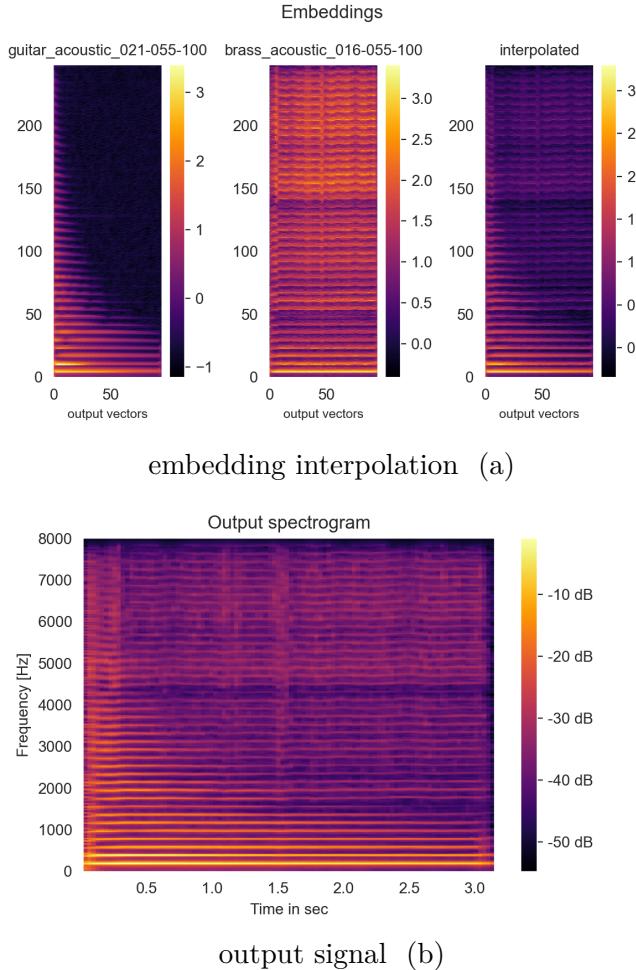


Figure 5.9: Embeddings of input samples with interpolated embedding and its corresponding output spectrogram - single-stride network.

can also be said that in the latter, the guitar features can be recognized better. Looking at the final output spectrogram in 5.11b it can be seen that the original fine harmonics aren't present anymore. The energy values of the different frequencies over time don't remain constant and show "washed out" contours. Also the fine changes over time are not as precisely present as with previous network configurations. Nevertheless both instrumental features can be recognized in the final output. Again in the discussion part, the auditive quality of all outputs presented here, with and without energy correction, is assessed and brought into relation with the displayed spectrograms here.

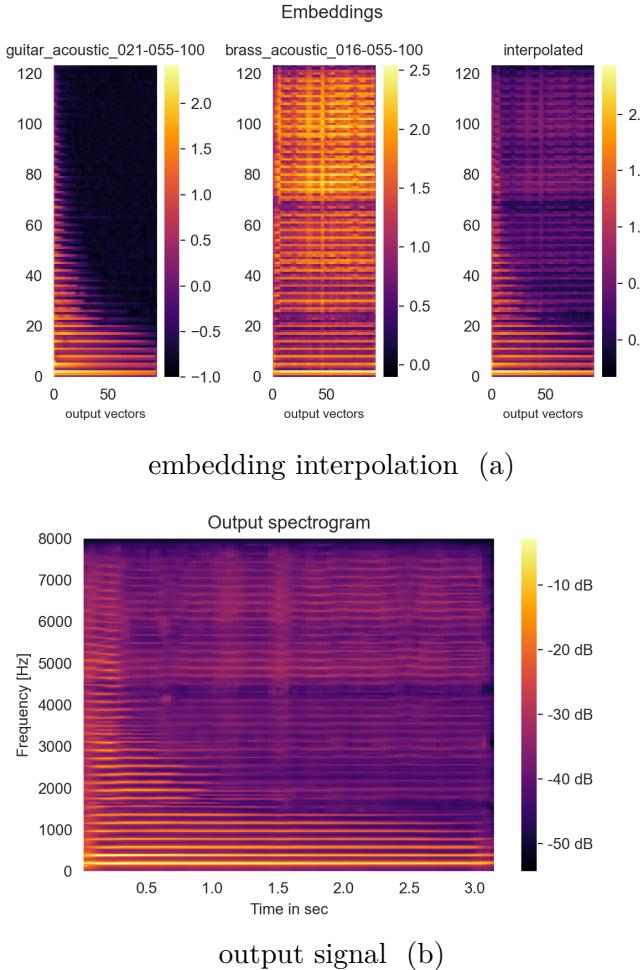


Figure 5.10: Embeddings of input samples with interpolated embedding and its corresponding output spectrogram - double-stride network.

5.3 Results with mel-spectrograms

The final experiments are again done with a 2D convolutional network, but using mel scale spectrograms instead of log-magnitude. In chapter 4 under section 4.6 an introduction to the mel-scale has been given. As those log-mel spectrograms are a compressed version of the log-magnitude spectrograms and emphasize the lower frequencies, these experiments are expected to bring different but interesting results. This is again meant regarding the model performance but also towards the quality of reconstructing spectrograms with or without modification of the embeddings. The results again get depicted in tables filled with the MSE scores of training, validation and testing but also the scores regarding specific pitch classes of the test set. Later on again the graphics contain the output spectrograms and embeddings with and without interpolation. Again three different networks with single, double and triple strides were used. Those are similar in their structure to the ones used with log-magnitude spectrograms.

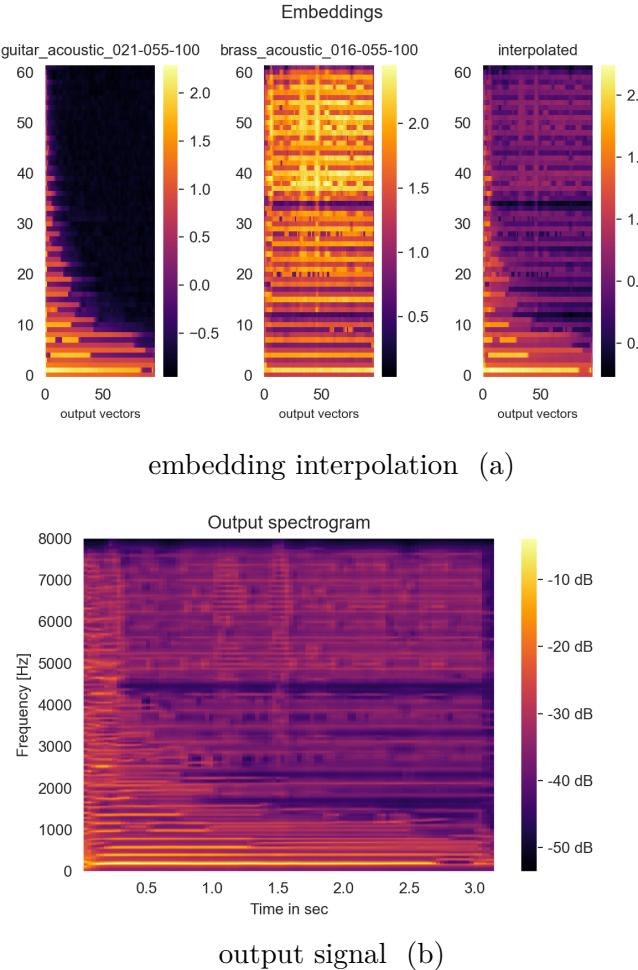


Figure 5.11: Embeddings of input samples with interpolated embedding and its corresponding output spectrogram - triple stride model.

The following table 5.5 shows the training, validation and testing scores of all three different networks. First of all the MSE scores of all networks are significantly, higher than the ones of the log-magnitude networks. An exception is the testing score of the single-stride network as it is equal to the training score of the triple-strided log-magnitude network. The scores between the networks also increase with the amount of compression, making the single-stride network again the best performing network regarding its score. Interestingly the validation score of the single-stride network is the only one that is higher as its corresponding training score. The double- and triple-stride networks therefore show that they have a lower score on reconstructing unseen data. Again the auditory quality will be assessed in the next chapter 6.

Like in the previous experiments, the network performance has been evaluated towards certain pitch classes. Table 5.6 shows the MSE scores of different pitch classes. For comparative reasons, again they got categorized on a scale from 30 to 100 in steps of 5. Contrary to those scores in the previous experiments, here the values show that

	Single-Stride	Double-Stride	Triple-Stride
Training	30,579	49,685	62,833
Validation	31,084	48,439	60,471
Test	18,322	33,689	43,300

Table 5.5: MSE-Scores 2D convolutional autoencoder with log-mel spectrograms - single-stride, double-stride, triple stride.

the lower pitched samples have the best score. With some exceptions/outliers, the error scores increase with a clear trend, the higher pitched the used samples were. Interestingly the difference in the score between the highest and lowest used samples increases the more compression is used.

Pitch	Single-Stride	Double-Stride	Triple-Stride
030	17,178	24,567	30,302
035	18,380	26,142	36,509
040	19,382	34,125	44,062
045	22,036	55,770	51,097
050	22,398	41,101	49,810
055	22,159	43,215	56,664
060	21,472	40,493	54,690
065	22,303	40,140	51,926
070	23,875	41,950	50,904
075	24,571	37,424	49,297
080	29,053	43,572	60,926
085	29,395	46,822	62,750
090	33,316	51,408	72,206
095	31,385	51,593	70,607
100	32,129	51,426	70,890

Table 5.6: MSE-Scores for specific pitch classes using 2D convolutional autoencoder using mel-scale.

5.3.1 Experiments of single reconstruction

Similar to the other experiments to assess the ability of reconstructing spectrograms and therefore recreate audio samples, in this section the embeddings and corresponding output spectrograms are depicted. There again the properties and respective differences to the other experiments get assessed. Regarding the auditory quality the results are discussed in chapter 6. The here shown results again correspond just to the reconstruction of single samples without interpolation. As in these last experiments the mel scale as different scale was used, the corresponding input spectrograms are shown in advance. Figure 5.12 shows the input spectrogram of the single reconstruction experiments.

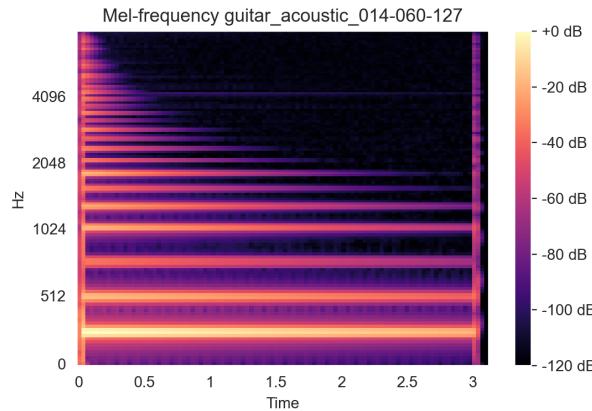


Figure 5.12: Original mel-spectrogram of guitar acoustic.

The next three figures (5.13, 5.14 and 5.15) show the output spectrograms but also the embeddings of the guitar sample depicted in figure 5.12. As its known at this point, the mel-spectrograms are already a compressed version log-magnitude spectrograms, as they contain 128 values instead 513. Thus it can be seen on the y-axis of the embeddings, that those embeddings are even more compressed than in the last experiment (size of 56, 28 and 14). Furthermore it can be noticed regarding the embeddings of the first two networks (figures 5.13b and 5.14b), that contrary to the previous experiments with log-magnitude spectrograms, the low energy areas are positively valued. The high-energy areas and harmonics are therefore negatively valued. Concerning the scale it can be also said that those embeddings have a narrower value range then those in the previous experiments.

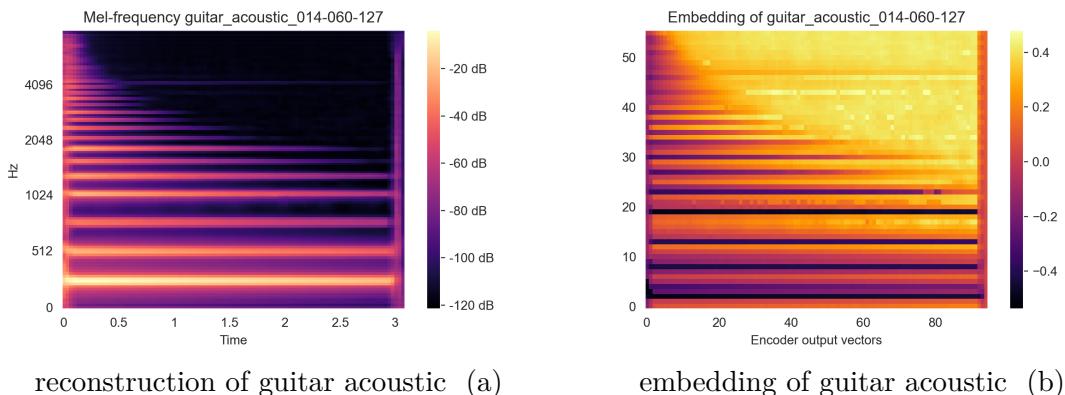


Figure 5.13: Reconstructed log-mel spectrogram and corresponding embedding - single-stride network.

In figure 5.13a the output of the single-strided network can be seen. In comparison to the input spectrogram, its reconstruction is the best out of those three networks, as it contains the most details. Despite of this, it can be seen that the broadband

spectra at the beginning and end are not as intense as with the input. Further on it can be seen that changes over time in the frequencies between the lower harmonics are not present in the output. The embeddings in figures 5.13b and 5.14b show that the features representing the lower harmonics are recognizable as negative values. The latter already shows a lack of features in the higher harmonics. This is caused by the input log-mel spectrograms having a shorter distance between the higher harmonics than in the lower ones. Regarding the output of the double-stride network it can therefore be said that the higher harmonics can be less distinguished from each other than with the single-stride network (see area over >2048 Hz).

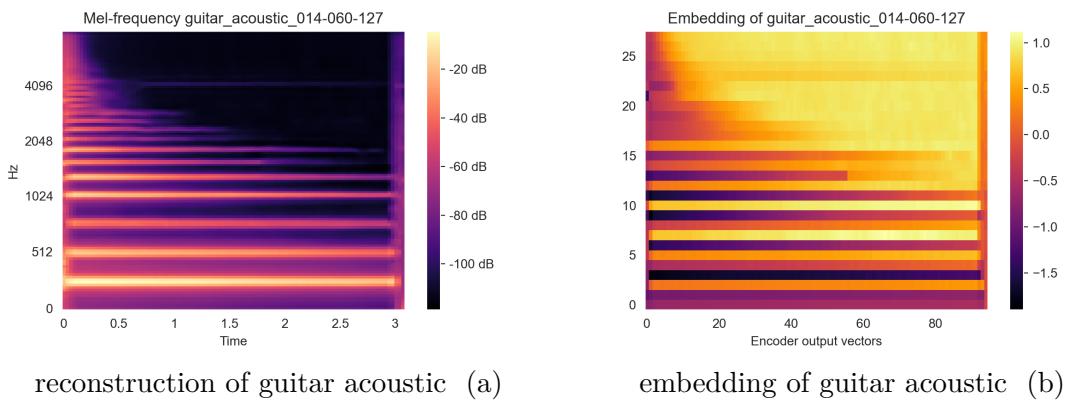


Figure 5.14: Reconstructed log-mel spectrogram and corresponding embedding - double-stride network.

Comparing the result of the double-strided network with the one of the triple strided network, significant differences can be discovered. Especially when looking onto the embedding, it can be discovered that due to the small amount of values on the y-axis the features are less distinguishable. Thus it is also more difficult to describe them. Nevertheless it can be said for this input spectrogram, that the colored areas still represent the most significant features that describe the input. Interestingly when looking onto the output spectrogram in 5.15a, regarding the lower harmonics, no differences can be seen to the previous output spectrograms. The part around 4096 Hz also shows some distinct spots that contain more energy, which are not present in the input spectrogram (artefacts).

5.3.2 Experiments with interpolation in embedding

The final experiments in this work also deal with the introduction of the interpolation step to synthesize novel sounds. With the experiments of reconstructing single audios, it could be seen that despite the small compressed embeddings, the input spectrograms could be reconstructed adequately. Having those smaller embeddings, less values are available to reconstruct spectrograms which in conclusion means that altering those also brings significant changes. This fact makes these experiments even more interesting, as here again, the embeddings of two input samples are interpolated, to generate novel sounds. As an example, the same guitar and brass samples that were used in the previous

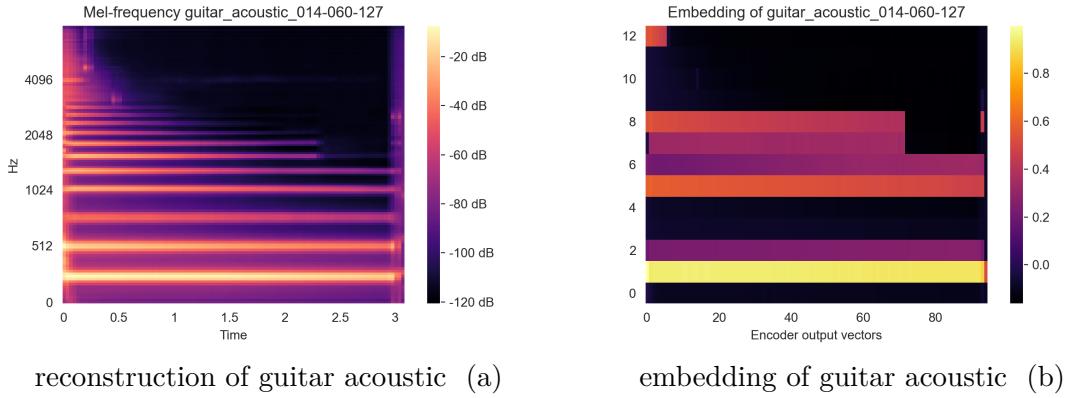


Figure 5.15: Reconstructed log-mel spectrogram and corresponding embedding - triple-stride network.

experiments are utilized to synthesize audio. The following graphics 5.16 show the input log-mel spectrograms of the two source signals. To mention again, the higher harmonics are closer in the scale, while the lower harmonics have a greater distance. Especially when looking at the brass sample this can be noticed as the energy does not fade.

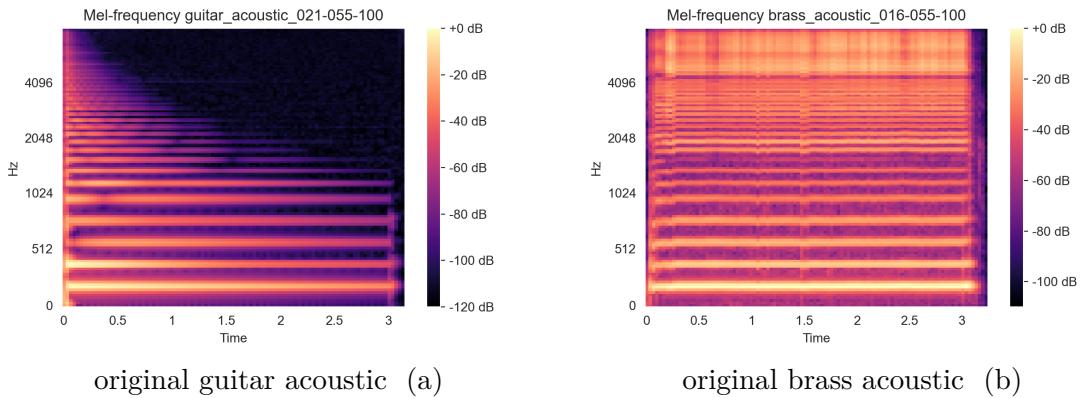


Figure 5.16: Input log-mel spectrograms.

Knowing the source log-mel spectrograms, the following graphics show the embeddings of those spectrograms, the interpolated embedding as well as the resulting output spectrogram using again the three differently strided networks (single-, double- and triple-stride). By comparing all the output spectrograms to the two input spectrograms, it again can be said that the output of the single-stride network in figure 5.17b is the most precise and contains the most characteristics of both instruments. The upper harmonics of the brass sample can still be recognized in the output spectrogram despite the “fading” influence of the guitar sample. Fine changes in the harmonics over time of the trumpet, like at 1,5 seconds, cannot be recognized anymore. Furthermore the guitar sample can also be recognized in the output spectrogram. With a look onto the embed-

dings those again show the high energy areas with low or negative values. The original low energy areas in turn have positive values. Furthermore due to the little distance between the higher frequency ranges in the input log-mel spectrograms, the embeddings in this area just depict the high-energy areas. This leads the output spectrogram to not have a high precision in the higher harmonics. Nevertheless through interpolation the resulting embedding and the output spectrogram contains features of both instruments.

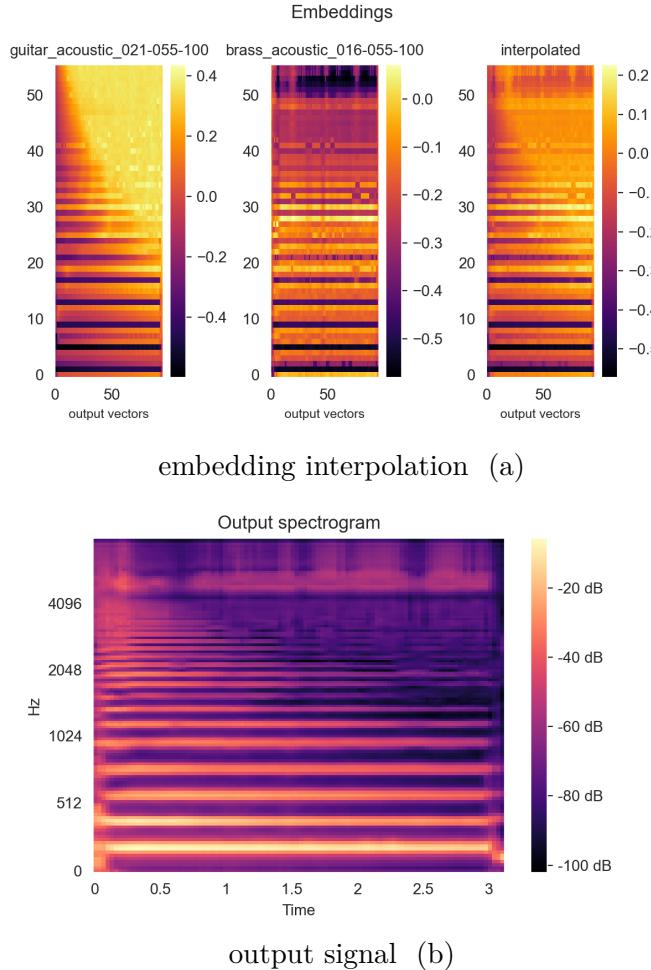


Figure 5.17: Embeddings of input samples with interpolated embedding and its corresponding output log-mel spectrogram - single-stride network.

Figure 5.18 shows the output of the double-stride network where already significant differences to the output of the single-strided network can be seen. First of all regarding the embeddings, the features representing the higher harmonics of both instruments (range $>=15$) are less intense while the features of the lower harmonics are still distinguishable. Having the interpolated embedding the features of both embeddings can still be recognized, despite the lower precision on the y-axis above 15. As a result in the output spectrogram, the higher harmonics of both instruments are rather “washed out” and appear blurry in the spectrogram.

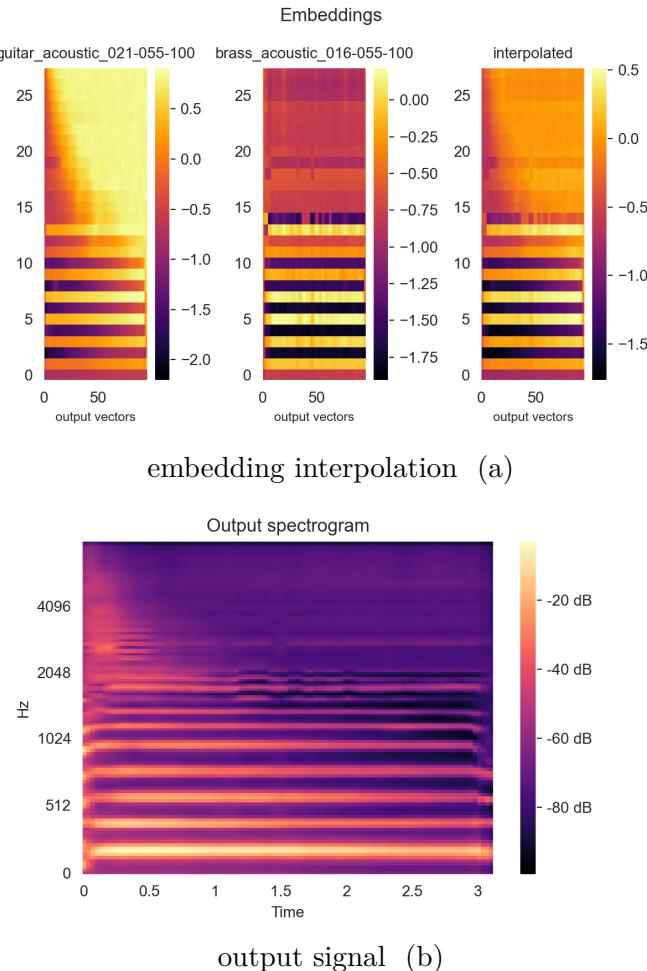
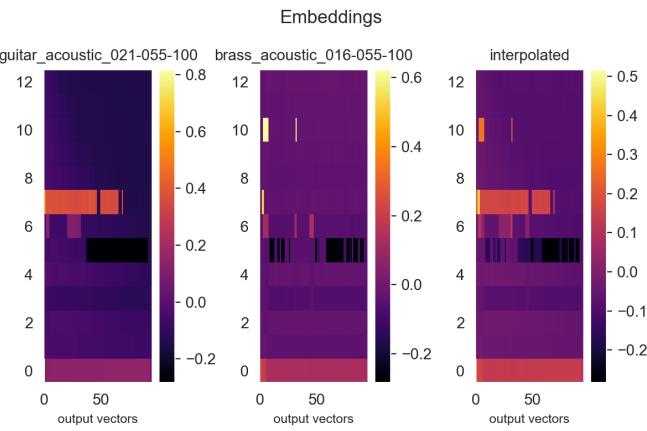
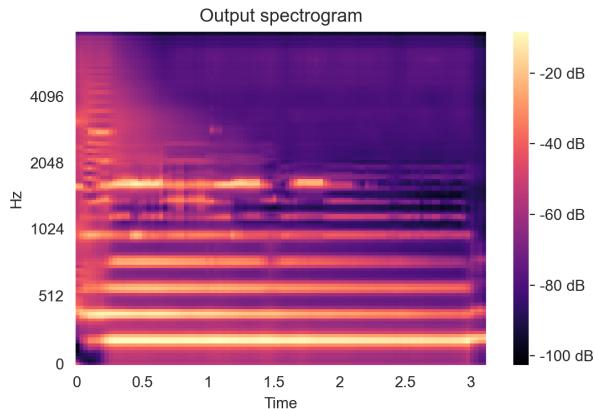


Figure 5.18: Embeddings of input samples with interpolated embedding and its corresponding output log-mel spectrogram - double-stride network.

Coming to the output of the triple-stride network in figure 5.19, here it can be seen that the features contained in the embeddings have no common structure with the input spectrograms. The values therefore cannot be described sufficiently. Also those appear different to the one depicted in figure 5.15b. When looking at the output spectrogram generated of the interpolated embedding, it can be seen that in the area around 1024 to 2048 significant energy changes regarding certain frequencies are present which cannot be recognized in the input. Similar to the output of the previous network in figure 5.18b fine structures of the input are also not present anymore.



embedding interpolation (a)



output signal (b)

Figure 5.19: Embeddings of input samples with interpolated embedding and its corresponding output log-mel spectrogram - triple-stride network.

Chapter 6

Evaluation

In this chapter the results (see chapter 5) that could be obtained by the experiments shown in chapter 4 are discussed. As the results of the experiments cannot solely be described through graphics and numbers shown in the previous chapter, the evaluation also takes the auditory results into account. Based on the combination of the graphical and auditory results, some interesting findings and observations can be derived that get discussed in the following sections.

6.1 Impact of model configurations

Throughout the experiments different models have been trained and utilized for reconstruction or synthesis of audio spectrograms. With these experiments, the goal was to find a model that is suited for the task of (re)synthesizing audio data. Furthermore by exploring different models with different configurations, insight could be gained, on how these affect the performance but also the output of the model. In the case of this work, the results of the experiments can be presented in different ways. In the previous chapter 5, the results have been depicted as error scores, but also the embedded space data and the resulting spectrograms. With those alone, the approaches cannot be fully evaluated, as it is also important to consider and discuss the auditory quality, as it dealt with audio data. The following section therefore takes the numerical and graphical but also auditory results into account to evaluate the models with regard to their configuration. In total the evaluation takes the results of 7 different models into account whereas 3 of them get discussed later on in section 6.2.

6.1.1 Impact on score

To train a neural network, but also other machine learning models, an error score is needed. With the help of this score, the network gets optimized to minimize the error and thus improve the networks performance and quality of output. In chapter 5 the error scores regarding each network configuration are displayed separately regarding the network type (1D or 2D convolutional) and input type (log-mel vs. log-magnitude spectrograms). By comparing the error scores regarding the network type it can be said, that the 1D convolutional network performed best, as it shows the lowest scores on training and validation, but especially testing. All networks that use 2D convolutions

and operate on snippets coming either from log-magnitude or log-mel spectrograms have a worse score. To mention those scores just show the amount of error between the reconstructed and the input data. Furthermore when also considering the amount of strides being used an interesting observation can be found regarding the score. Here it can be said, that the more striding is applied in the network, the higher gets the error. This was observed using the 2D convolutional networks, as with the 1D convolutional networks no strides have been applied. Mentioning also the scores concerning specific pitch classes, it can be said that the 1D convolutional network has the best scores in all pitch classes. Moving on to the 2D convolutional networks, the scores of the single pitch classes are also higher than the ones from the 1D convolutional network. Also they increase with the amount of striding. Nevertheless it can be seen that concerning the 2D convolutional networks with log-magnitude spectrograms and 1D convolutional networks, they all show constant scores alongside the pitch classes. An exception are the three networks trained on log-mel spectrograms as those increase with the pitch class. As this is mainly influenced by the type of input, this gets discussed later on in section 6.2.

In total it can be said that 1D convolutional networks yield a better score than 2D convolutional networks having also the best scores among specific pitch classes. Furthermore regarding the 2D convolutional networks it can be said that the amount of striding certainly influences the models score, as the more striding is applied, the higher the error gets in the network. Nevertheless just of this score, it cannot be concluded that the output sound using a 1D convolutional network is also of better quality. This can also be said concerning the stridings within the network, as the output has to be evaluated further on by listing to the final sounds. Furthermore as the error scores for the specific pitch classes do not show a strong trend (models with log-magnitude spectrograms), it also cannot be said, that the output of all pitch classes is equally good in its auditory quality. Therefore in the next section 6.1.2 the outputs being spectrograms, embeddings but also the generated audible signals, get discussed in order to evaluate the impact of the models configuration properly.

6.1.2 Impact on single reconstruction

As in the previous subsection 6.1.1 the impact of model configuration on the score has been discussed, here the generated experimental artefacts such as spectrograms, embeddings and audios get included into the evaluation. Like in the chapter 5 the discussion here gets divided into two parts, whereas in this part the impact on the single reconstruction and in the following part (subsection 6.1.3) the impact on the interpolated (synthesized) outcome gets evaluated. This subsection also mainly deals with the outcome of the networks using log-magnitude data. The discussion regarding the log-mel spectrograms will follow later on in section 6.2.

Impact on generated spectrograms and audio

In this section the artefacts like spectrograms and audios gained by reconstructing single audios without altering the embeddings get assessed. This should bring insights into how the model configuration affects the ability to reconstruct audio spectrograms but also the quality of the output. As an instrument which represents the results, the guitar sample shown in the result chapter is chosen. This guitar sample contains characteristics such as the guitar stroke but also the damping of the strings at the end of the sample. Also it has the characteristic that it fades over time. To take the spectrograms and embeddings shown in chapter 5 some first observations can be made. First on by comparing the output spectrograms of the 1D convolutional network, and those of the 2D convolutional network, it can be seen that the original areas containing little energy (dark) get better preserved in the 2D convolutional networks. The output of the 1D convolutional network appears to have more energy distributed in areas of little energy. Further on it also can be observed that especially the single-stride 2D convolutional network is able to preserve the broadband spectra at the beginning and the end of the spectrograms. In comparison the 1D convolutional network has difficulties in preserving those broadband spectra. This can be also justified when listening to the final audio. As a note, unless it is mentioned, the evaluation on the audible output is done on samples generated with the original phase information. Concerning the output of the 1D convolutional network the stroke and damping is not present but it is in the ones coming from the 2D convolutional networks. Also the sample of the 1D convolutional network has a worse listening quality as it gets louder from the beginning on and becomes noisy. The latter especially gets proved by the fact that the spectrogram contains more energy across all frequencies that have usually little energy contained. Further on by listening to the outputs of the 2D convolutional network, those are overall better in their audible quality but also when looking onto the spectrograms. Especially concerning the single-stride network its output has the best quality, as it preserves the important characteristics such as stroke and the damp best, but also fades like the original sample. By comparing especially this one, hardly no difference can be noticed, except a small lack in the intensity of the stroke. When listening to the outputs of the double- and triple-strided network, no significant differences can be noticed, as also characteristics, like stroke, fade and damping are still contained. Just the output of the triple-stride network is slightly distorted, being not as clear as the output of the others, when listening closely. Taking also its corresponding spectrogram into account, those distortions can be justified by the less precise harmonics. Similar observations can be also made by using a sample of a different instrument (e.g. brass). Despite this one not fading over time and does not have characteristics like a guitar (stroke, damping), the output is also of similiar quality like with the guitar. Also the same conclusions can be made, as the quality using the 2D convolutional network is better than with the 1D convolutional network, having equally good sounds using single- and double-striding.

As those samples were not altered in their embedding, the original phase information could be reused for the inverse STFT to obtain the sounds. Those have been taken into account in the last paragraph. For the comparison but also as for synthesized spectrograms no phase information is present, samples have been generated using the Griffin-Lim algorithm [8] to estimate the signal. Using the same guitar samples that

have been used for discussion before but generated with the Griffin-Lim algorithm introduced some differences in the sound quality especially concerning the 2D convolutional network. Taking again the output of the 1D convolutional network into account, by listening to the signal, the quality does not change. Interestingly while listening to the samples generated by the 2D convolutional networks, drawbacks between the ones using Griffin-Lim and the others with the original phase information, can be observed. As the ones with the preserved phase fade clearly towards the end, the ones being estimated by Griffin-Lim, alternate slightly in their loudness.

Previously in subsection 6.1.1 the scores regarding their pitch class have been discussed. As those cannot solely be considered for grading the performance and output quality, the audible outputs have to be taken into account. As this cannot be done for all samples among all instruments, guitar acoustic samples were picked for evaluation. When listening to the generated samples of the 1D convolutional network, it can be noticed, that among the pitch classes (going from 030 to 100) all the samples have the characteristic in becoming louder towards the end. Furthermore also especially when the pitch gets higher, that the samples contain more distortion. If the final sound gets reconstructed with the Griffin-Lim algorithm, an interesting behaviour can be observed. With special regard to the lower pitched samples e.g. midi note 030, the pitch is not recognizable in the output. Later on when using samples of higher pitch classes, the pitch becomes recognizable. Coming to the samples generated by the 2D convolutional network, here the output samples of the single- and double-stride networks remain the same in its quality alongside the pitch (original phase). Also no difference between the two networks can be perceived. Regarding the triple-stride network, more distortion can be perceived in the lower pitched samples, while when ascending in the pitch, the distortion gets less and less perceptible. Again when taking the results obtained by applying the Griffin-Lim algorithm, the pitch cannot be recognized in the output samples using lower pitched samples. Especially when applying more striding like in the triple-stride network, this behaviour appears more significantly. Having the original phase information this would not be an issue, but in the case of synthesis where the output spectrogram gets reconstructed from an altered embedding and no phase information is present, it is.

Impact on generated embeddings

Finally after discussing the impact of the model configuration on the spectrograms but especially outputs, it is also of interest, how the embeddings differ from model to model. Especially as those represent the features from which the decoder part reconstructs a spectrogram. In chapter 5 the embeddings of a guitar sample have been depicted together with their corresponding recreated spectrograms. First of all it can be seen, that all embeddings in general are a projection of the input spectrograms on a lower dimensional space. Different to the embeddings generated by the 2D convolutional networks, the ones from the 1D convolutional network preserve the high energy areas with negative values and vice versa. As also in the output spectrograms and samples of the 2D network, the broadband spectra are preserved better, those can be also seen in the embeddings at the beginning and the end. This happens in contrast to the 1D convolutional network, as like in the output, the broadband spectra are not recognizable. With a final look

onto the spectrograms generated by the 2D convolutional networks, it can be seen, that the more striding is applied the less values are available to preserve the inputs features. With regard to the triple-stride network, the embedding contains the essential features of the input, but regarding its output it lacks to preserve the desired quality which was discussed before. Therefore in this case, the embedding contains little information to reconstruct a spectrogram with a desired quality. This also makes the single- and double-stride network favorable as they preserve enough data to reconstruct spectrograms with a sufficient quality.

6.1.3 Impact on interpolation

In subsection 6.1.2 the developed models got evaluated towards their ability to reconstruct single audio spectrograms. There already promising results and interesting findings could be obtained that are important for synthesizing audio. Again like previously in this section the 1D convolutional but also the single-, double- and triple-stride 2D convolutional networks using log-magnitude spectrograms get taken into account. The evaluation for synthesizing sounds with networks taking log-mel spectrograms as input will be also be done later on in section 6.2. Discussed in this section will be the artifacts like embeddings, output spectrograms, but also output sounds. Furthermore during the experiments correlation coefficients between the embeddings were calculated to obtain interesting combinations of embeddings that differ in their characteristics. This was not done with the embeddings of the 1D convolutional network. Nevertheless interesting combinations for the audio synthesis could be found. For the purpose of these experiments, two instruments (guitar acoustic and brass acoustic) were selected that have significant differences and should synthesize an interesting sound.

Observations regarding correlation of embeddings

Regarding their correlation coefficients those have been 0,16 (single-stride), 0,0052 (double-stride) and -0,17 (triple-stride). By comparing those values to other combinations, those signalize a low correlation and thus being highly different. As a note it could also be examined that some examples being of the same instrument family have a high correlation which means they are similar and contain similar characteristics. With this knowledge it can also be approved that the networks learn to capture features that describe the instrumental source. Some correlation matrices can be seen in the supplementary material to the thesis as they cannot be depicted sufficiently on paper. Comparing the correlation matrices over those three 2D convolutional networks no significant differences can be observed.

Impact on generated artefacts

When taking a look on the results in chapter 5 besides the embeddings and spectrograms concerning single reconstructions also the results of synthesized spectrograms are shown. There the generated embeddings of two instruments were shown in combination with the resulting interpolated embedding. Alongside those, the generated output spectrograms are shown, that represent the synthesized audio. Like mentioned before, as source samples a guitar and brass sample has been chosen to serve as base for the

synthesis task. First taking the results using the 1D convolutional network into account, there it could be seen that by interpolating the embeddings the resulting new embedding contains the features of both instruments. Nevertheless by comparing it to both input spectrograms it can be seen, that especially the guitar features cannot be recognized that easily. This behaviour can also be recognized by looking onto the resulting spectrograms and comparing it to the two input spectrograms. Further on by taking the results using the 2D convolutional networks into account, improvements can be detected. First of all when looking onto the interpolated embeddings, here the instruments features can be distinguished better, especially the guitar sample can be recognized better. This behaviour also varies having different amounts of striding in the networks. In particular when examining the interpolated embeddings of the double- and triple-stride network the features of both instruments can be visually distinguished the best.

When moving on onto the spectrograms, the best quality in terms of preserving the fine structures of the source instruments delivers the single-stride network. Also contrary to the output spectrogram of the 1D convolutional network both instruments can be recognized better. The output spectrogram of the double-stride network appears to preserve the distinct sounds even better, but lacks in preserving the fine structures and changes of the input spectrograms. With a final look onto the output of the triple-stride network, this one has hardly any fine structures preserved as the the harmonic features over time also appear “washed out” and being not recognizable precisely. This can be brought into relation with the contained features in the embedding, as less features are preserved.

To also bring the quality of the audible output into relation some further interesting findings can be obtained. Like also with the single reconstruction experiments, the output of the 1D convolutional network does not contain a guitar stroke, and also proves that the guitar is less audible then the brass sample. Furthermore it is also in its amplitude louder then the original samples and contains distortion which is certainly not desired. In comparison the outputs generated by the 2D convolutional networks have an overall better perceptual quality. Interestingly as the single- and double-stride networks’ output have differences in the spectrograms, those cannot be perceived in the final output sound making them sound equally. The guitar also can be perceived better with the sound characteristics of the brass sample. Furthermore also the stroke is preserved in the output. As a further advantage to the output of the 1D convolutional network also no distortion is contained, making it sound clearly. Finally listening to the output of the triple-stride network it can be said that the quality is again significantly worse then with the single- and double-stride networks. Nevertheless the instruments are both perceptible and distinguishable in the output, but it appears to have missing frequency parts, making it sound more “metallic”. The latter can also be brought into relation with the “washed out” harmonics in the spectrogram. Furthermore also the pitch is unstable as it rises slightly in the end of the sample.

6.2 Impact of pre-processing

Having discussed the impact of the model configuration and composition, in this section the impact the pre-processing on this kind of machine learning problem has, gets discussed. Like it has been described in chapter 3, section 3.3, the step of pre-processing is an important component of every machine-learning task. Especially as it also has a significant influence on the performance of the machine-learning model. This means also, that it has a significant influence on the output and its quality, with special regard to the outcome of this work. Therefore this section describes the influence of pre-processing that has been observed during this work.

For the task of neural audio synthesis, there exist different methods that differ mainly in the shape of their input data. While there exist methods that use the time-domain signal, a few methods including this one, use spectrograms for this task. To generate spectrograms from audio, the pre-processing stage incorporates this task.

6.2.1 Configuration of STFT

Like discussed earlier, to obtain spectrograms, the STFT was applied which could be parameterized. This parameterization includes the size of the signal window, on which the Fourier transform gets applied to. Throughout this work this parameter got altered, as it influences the frequency- but also time-resolution of the spectrogram. First on this parameter was chosen to be 512 which means a frequency resolution of 31,25 Hz and time 32 ms (with sampling rate of 16 kHz). Due to the unsatisfactory model performance, this value has been changed to 1024 to have a frequency resolution of 15,625 Hz and time-resolution of 64 ms. This also has the advantage of having a better frequency resolution and therefore also more features to learn from along the frequency axis. Therefore this value was chosen throughout all experiments. When wanting to develop a real-time capable system the time resolution has to be taken into account, as having a better frequency resolution the resulting time frames increase. This means that the duration of the audio that has to be available for the model increases. Therefore it has to be said that the right configuration needs to be found to obtain the right frequency-time trade-off that is best for the problem.

6.2.2 Use of single frequency vectors vs. spectrogram frames

As it could be seen throughout the experiments, having the spectrograms, those can also be differently prepared and used for the model. First on just the single frequency vectors have been chosen to be used as input for the model. Later on overlapping frames of spectrogram slices were utilized as model inputs. Using the first one, this has the advantage that just one frequency vector is needed as input to generate an output. Especially in terms of aiming to develop a system that performs real-time audio synthesis this would be the ideal choice. Nevertheless by comparing the results of those two approaches, the latter is proven to have a better output quality. The details to this discussion and results have already been discussed in section 6.1, as this also deals with the model configuration.

6.2.3 Magnitude vs. Mel

As mentioned before, there are also other methods on how to use audio data for machine learning tasks. Except from using the raw time-domain signal, there also exist different types of spectrograms that can be used. First on during this work log-magnitude spectrograms have been used as the source input data for the models. Further on in the final experiments log-mel spectrograms have been used, which are based on the mel scale. By performing the same experiments as with log-magnitude spectrograms, the results should bring an insight on the impact, the pre-processing or use of a different scale has. Again three differently configured networks have been implemented and considered for the experiments. Those again differ in the amount of striding (single, double, triple) on each network side. In the last chapter the error scores, but also the output spectrograms and embeddings of the models, using both types of spectrograms have been shown. With those results it already can be seen that it has a significant impact on how the audio data is pre-processed and provided for the neural network.

Impact on score

First of all regarding the error scores it can be said that, without taking the networks configuration into account, better scores could be reached, using the log-magnitude spectrograms. The scores between the single-, double- and triple-stride networks with log-mel spectrograms also rise (more significantly) the more striding is applied. Regarding the scores for single pitch classes, different behaviour could be observed. As the scores in the log-magnitude spectrogram trained networks remain equal among the pitch classes, the ones using log-mel spectrogram increase with the pitch. This means that using the log-mel spectrograms, the networks have a bigger error and thus more difficulties in reconstructing higher pitched samples than lower ones. On the other side the networks with log-magnitude spectrograms can reconstruct the input equally well, with no regard to the pitch (concerning the scores). As this is just the performance of the network, regarding the reconstruction error, this does not automatically mean, that the listening quality is better or worse. Thus again the artefacts like embeddings, spectrograms but also audio samples have to be taken into the evaluation part which gets done in the following subsection. The observations regarding log-magnitude spectral data has already been discussed before in subsection 6.1.2.

Impact on single reconstruction

Again like with log-magnitude spectrograms, experiments have been conducted in using log-mel spectrograms to be reconstructed. The results in chapter 5 show again the reconstructed spectrograms with their corresponding embeddings among the three different networks. For comparative reasons, again the same instrument source have been elected for these experiments. Therefore key differences between applying log-mel and log-magnitude should be discovered. Back in the previous chapter, it already has been said that the output spectrograms are lacking in preserving the higher harmonics. This problem also increases the more striding is applied. Furthermore with the triple-stride network, some artifacts in the upper harmonics could be perceived in the output spectrogram. Having less precise upper harmonics can be also brought into relation with

the nature of the mel scale, as lower frequencies are more emphasized due to the larger distance on the scale as the higher frequencies. Despite of that as a guitar sample gets examined, the broadband spectra at the beginning and end are also present. Compared to the log-magnitude embeddings, the here generated embeddings again value the original high-energy areas with negative numbers and vice versa. Nevertheless the lack of precision in the higher harmonics can also be seen in those embeddings, as with increasing the stride, less values can be captured for the original dense higher harmonics. Despite of this the lower harmonics are still visible as negatively valued areas.

By also listening to the corresponding output samples, some final findings can be formulated. First of all if not explicitly mentioned, the original phase information was reused concerning the discussed output samples. To the overall quality it can be said, that it is less good as with log-magnitude spectrograms, as they do not have a clear sound (alternating loudness). Nevertheless features like stroke and damping of the strings can be perceived, as well as the fading of the sound. The output of the double-stride network sounds less clear, having more alternations in loudness than the one coming from the single-stride network. The output of the triple-stride network again is of the lowest quality, as it contains audible artefacts at the beginning of the sound, which are also visible in the output spectrogram. Listening to the same sounds that were regenerated with the Griffin-Lim algorithm minor drawbacks in quality can be perceived. While the one coming from the single-stride network introduces more alternation in the loudness, the outputs of the double- and triple-stride networks are unstable in their pitch.

Bringing the pitch into account, the scores for the pitch classes alongside the three different strided networks, have been examined. Like with the log-magnitude networks also the audible output samples have to be considered. Again the acoustic guitar has been chosen, from which samples along the pitch classes were picked to evaluate the quality. As it has been discussed, the scores increased while ascending in the pitch class which should mean that higher pitched samples are of a lower reconstruction quality. By listening to samples of all three different strided networks, no difference can be observed in the quality regarding the pitch. Taking again the output obtained with Griffin-Lim algorithm into account, similar observations can be made like with log-magnitude spectrograms. Especially with the double- and triple-stride networks, the effect of having no recognizable pitch information present regarding lower pitch-classes. Ascending in the pitch class the pitch is present, but mostly remaining unstable or slightly alternating over time. The single-stride network again performs best. Despite of the lack of pitch in the lower classes, the overall quality remains equal.

Impact on interpolation

As in the previous section the results of reconstructing single audios using log-mel spectrograms, this point evaluates the capability of synthesizing audio based on interpolating embeddings. This has already been done using log-magnitude spectrograms in subsection 6.1.3. The evaluation for log-mel spectrograms will be done in the same manner, as the results depicted in chapter 5 will be taken into account in combination with assessing the audible quality. By this the impact of using log-mel spectrograms compared to log-magnitude gets discussed. Again the same two instrument sources were used (guitar acoustic and brass acoustic) to synthesize audio.

During the experiment, like in previous ones, correlation matrices were calculated to obtain an overview, of which samples are least similar and thus form an interesting combination. For this combination the correlation coefficients have been 0,31 (single-stride), 0,62 (double-stride) and 0,53 (triple-stride) which are significantly higher than those regarding the log-magnitude networks. Also when comparing the correlation matrices of the experiments in 6.1.3 with those generated with the embeddings of the log-mel networks it can be seen, that those values are significantly different. Also within the log-mel networks the correlations matrices and their values differ a lot. This means that while in one network two instruments have a low correlation, in another one those have a high correlation. For example as seen with the values mentioned before it can be said that the embeddings generated of log-magnitude spectrograms have more distinct features than those generated of log-mel spectrograms. Thus it can be said that those networks extract the features differently.

Coming to the results of the experiments that were obtained by synthesizing audio using interpolation in the embedded space. By comparing all output spectrograms that are shown in the previous chapter, the one generated by the single-stride network, preserved the input features the best. Especially because the features of both instruments can be recognized sufficiently in the output. With special regard to the characteristics of the brass samples, those are preserved best, as the fine harmonic structures still can be seen in the output. The output spectrogram of the double-stride network already lacks in preserving most of the harmonical features of the upper frequency range of the brass sample. Nevertheless the low frequency harmonics are still preserved in a good manner. This is also mostly because the features representing the energies of the lower frequencies, are similar regarding both instruments. Furthermore the features representing the original high-frequency areas, are also less contained in the embeddings and thus those cannot be sufficiently reconstructed in the output. With a final look onto the output of the triple-stride network, this one also does not contain the harmonics in the upper frequencies and also introduces features that are not contained in the original input spectrograms. By listening to the output samples, those findings can also be confirmed. There also the output of the single-stride network has the best quality, as it preserves both instruments and makes them perceivable. Just like in the output spectrogram the resulting sound of the double-stride network misses the higher harmonical features of the brass. Finally when listening to the output of the triple-stride network, this one contains audible artefacts like harmonics that are not present in the input samples but can be seen in the output spectrogram. Also the high harmonical features are not present anymore.

Making the fact, that in terms of using log-mel spectrograms, the single-stride network generates audios with the best quality while more striding introduces undesired artefacts but also misses features especially in higher harmonical ranges. Generally the quality compared to the approaches using log-magnitude spectrograms and 2D convolutional networks is lower.

6.2.4 Spectrograms vs. Time-domain

Knowing now that the pre-processing has a significant impact on the model performance but also quality of the audible output, it can be also discussed if spectrograms are suited best for this task. Another form of input except of spectrograms would be e.g. using time-domain signals. One has been discussed in the work of *Engel et al.* [6] where time-domain signal was used as input for a WaveNet-style autoencoder network. In their work they proved the WaveNet-style autoencoder approach to be advantageous, especially regarding the audible quality of the output. Having discussed the impacts the generation of spectrograms has on the task of audio synthesis, it can be said, that a few limitations showed up. As discussed before, the configuration of the STFT has an impact on the frequency resolution but also on the time resolution. For example for going towards real-time applications a high time resolution is desireable, as shorter time-frames can be taken as input for the model. This has the disadvantage of having a bad frequency-resolution which means to have less information present per time-frame. Furthermore as especially with this approach convolutional networks are applied, the phase information cannot be used. Leading to the fact of having magnitude data as output, to which than the phase has to be estimated at the end to regenerate the audible signal, as it is missing (regarding interpolated output). Like evaluated before this also affects the quality as the output regenerated with the original phase information is of higher quality than the estimated one. Knowing these problems, those would not be present, when using the time-domain signal as input. Also no calculations regarding transformations have to be made, which also speeds up the runtime. By taking the time-domain signal, this would request a different model type. Taking the by *Engel et al.* proposed WaveNet-style autoencoder, this one operates on raw time-domain signal but has a more complex structure then convolutional autoencoders.

6.3 Impact of post-processing

Having now discussed the findings and observations by applying different network configurations but also using different pre-processing settings, this section deals with the effect of additional post-processing steps. With regard to the already discussed results, those were not using any additional steps. As discussed, especially with the 1D convolutional network, certain features of like the stroke or damping of guitar strings do not get preserved in the output sufficiently. By introducing 2D convolutional networks that use overlapping log-magnitude and log-mel spectrograms, these issues could be improved significantly.

6.3.1 Energy correction

Nevertheless to further improve such features but also to better approximate the original amplitude, additional steps were introduced that have been mentioned in chapter 4 under subsection 4.5.2. To mention again briefly, the sum of energy values across each frequency vector of the input spectrogram has been computed in advance. The same has been done on the output spectrogram, to compare it to the original energy values and calculate a coefficient to adjust the energy in the output. This technique certainly improved the sound of the output, especially if single audios were reconstructed. With

special regards to the guitar samples, this improved the presence of the guitar stroke or the damping. But also when listening to other reconstructed samples, like brass, it can be observed, that also the loudness is better aligned with the input.

When applying this post-processing step with the interpolated outcome, some further interesting findings can be formulated. Especially as there are two input spectrograms the average energy between the two input spectrograms, respective their frequency vectors, is taken for the correction. For example taking the synthesized sounds that have been discussed before, those were created without additional post-processing. By applying the energy correction, it can be observed that especially the loudness is affected by the correction. Without it still contains the fading of the guitar, whereas by equalizing the energy, the samples rather keep the same loudness or even increase in loudness. Nevertheless the guitar stroke is better perceivable. This concerns the approaches using log-magnitude but also log-mel spectrograms.

6.4 Differences between instruments

No experiments were done that observed differences that could be observed regarding the instrumental source. Nevertheless, when reconstructing single samples, it could be observed that some instruments classes or even samples have better scores than others. Especially in the initial experiments, where originally just two instrument sources were considered for training. There it could be observed, that e.g. acoustic guitar samples led to a significantly worse score in training than synthetic keyboard samples. But also after training with the whole dataset, it could be observed, that some classes performed worse regarding their test score. As this has not been a main research question during the recent work, this would be subject of further studies.

6.5 Comparison to other approaches

In the chapter 2 a few approaches regarding neural audio synthesis have been described, which also served as a base for this research. Especially mentioning those that implemented autoencoder networks to synthesize audio. Mentioning those, some of the approaches used the output of the encoder parts to interpolate them in order to generate novel sounds. This approach was also applied in this work while some approaches like those coming from *Colonel et al.* synthesized audio by omitting the encoder part and directly activating the embedded space. Most of the related work, that are specifically referenced as approaches with neural audio synthesis, also implemented other approaches that use different types of neural networks but in the shape of autoencoders. This approach solely emphasized on convolutional neural networks designed as autoencoders, that work with audio data shaped as spectrograms. *Engel et al.* compared the use of a convolutional autoencoder to a WaveNet-style autoencoder, where the first one acted as a baseline model. Despite proving the WaveNet-style encoder as qualitatively better approach, the spectral autoencoder also yielded promising results. This can also be said for the outcome of this work, as this work also generated promising results and interesting findings.

Similar to other works, this work also implemented different networks to assess which type or which configuration can be applied best for the task of audio synthesis. These configurations differ mainly in the type of convolutions (1D or 2D), the amount of layers but also configuration of striding, while the basic composition of layers (1D/2D convolution - (leaky) ReLU - Batch Normalization) stayed the same. This happened contrary to the other approaches, as those mainly did research on different compositions not only incorporating convolutions. Furthermore this work solely focused on the work on pure log-magnitude and log-mel spectrograms as input data, without additional input augmentation. Also no additional conditioning of the embedding (e.g. for the pitch) has been done in this work. Nevertheless, compelling and interesting results could be obtained, that prove the ability of convolutional autoencoders with spectrograms to synthesize sounds based on interpolation in the embedded space. Furthermore this work gives a special insight on how different parameterized convolutional autoencoders influence the quality of the output with regard to neural audio synthesis.

6.6 Limitations and difficulties

During the course of this work, some difficulties but also limitations occurred that had an impact on the work in progress but also on the outcome. First of all one of the major difficulties, especially in the beginning, was the ability to properly train a (complex) neural network, as high computational power is needed, which was not present at the beginning. Nevertheless this difficulty could be overcome, as for the purpose of this project, access to a GPU-accelerated computing service could be provided. This one improved and sped up the training process which made it possible to perform trainings on a large scale.

Another difficulty was how to properly display and the outcome, as this work deals with audio data. As it could be seen this could be solved by displaying the results as spectrograms but also with error scores. Nevertheless a full evaluation based on those artefacts could not be made, as it doesn't have a meaningfulness about the sound quality. For this case the outcome also had to be evaluated by listening to the resulting sounds to make a grading about the sound quality. Furthermore in case of this work it is difficult to produce a general evaluation of the performance, as the neural networks have been trained on several instruments data, which also have different characteristics and thus result in different behaviour. For this case the evaluation has been made on an example on which the behaviour for e.g. acoustic guitar could be observed. This is also the case when evaluating the interpolation based synthesis, as here this was also done on an example combination of two instruments. Especially observing and evaluating the behaviour of all instruments and their combinations regarding synthesis, would be a relatively time-consuming task. Therefore the findings in this thesis respond solely to the observations being made on the experiments with the selected samples.

Chapter 7

Conclusion and Future Work

This thesis and the work underlying it dealt with the exploration of using machine learning methods such as neural networks for the generation of audio data. As the title of this thesis states, this work emphasized on the use of convolutional neural networks that are shaped as an autoencoder. Due to this shape, the first part (encoder) of these networks produces a projection of the input data to a lower dimensional space, from which the second part (decoder) is able to regenerate the original input, which in this case are audio spectrograms. By modifying this latent space through interpolation, interesting findings and results could be obtained which are relevant for the field of neural audio synthesis. Not at least as they give important findings regarding the use of spectrograms and convolutional networks to synthesize audio. Having the final evaluation, the research questions defined at the beginning of this work can be answered.

First of all as discussed in the previous chapters 6 and 5, it was proven that by the application of convolutional autoencoder networks, novel sounds can be generated, through interpolating the embeddings of two different samples. The networks, that used log-magnitude spectrograms and 2D convolutions with single- or double-strides, showed the most promising results. Promising in the way that the output spectrograms and moreover the output sounds contain the characteristics of both input samples best in a novel sound. Moreover those sounds also have a good listening quality without noise or distortion. By having evaluated different types of networks containing 1D or 2D convolutions or different amount of stridings it also can be answered, that the configuration and composition has a significant impact on the model performance but also the quality of the output. Especially to mention the 1D convolutional network and the triple-strided 2D convolutional networks as there too much information is lost and distortion is included. Further to mention the reconstruction or synthesis of lower notes in combination with Griffin-Lim phase estimation, leads to the loss of the pitch. Nevertheless as this is a general problem, which appears among all network types, it gets influenced by the amount of striding. Not only the model configurations influence the quality of the output, but also the pre-processing as seen in the evaluation. With special respect to applying log-mel spectrograms compared of log-magnitude, the overall model performance but also the final sound is of a lower quality. All these findings can be concluded for either reconstructing single samples but also with the introduced interpolation step. Regarding the learnt information within the neural networks, it can be concluded that the networks mainly learn to project the harmonical structures of an input spectrogram

onto a lower dimensional representation. Finally, regarding the question whether audio spectrograms are suited best to synthesize audio, it can be concluded, that despite generating already promising results, limitations and disadvantages occur. This is especially true when aiming to research towards a real-time capable system.

Despite initial difficulties regarding the ability to properly train a complex convolutional network, with a large-scale dataset, those could be overcome and resulted in a promising outcome. The evaluation of this outcome also turned out to be difficult, as audio data is the main output and thus is difficult to display on paper. Nevertheless by displaying the outcome as error scores, spectrograms, embeddings or signal plots and evaluating them in combination with listening relevant findings could be gained. While those findings originate of experiments based on a small subset of the dataset, these nevertheless give an important insight on how a convolutional neural network can be applied to synthesize audio and what has to be taken care of in terms of the output quality.

7.1 Future Work

Having this work with already promising results, those results still motivate further research. Furthermore to gain a more detailed knowledge on the behaviour and differences of distinct instruments, further experiments across more instruments should be conducted. Not at least as this would provide additional insights on how the performance of the models differ by using different instruments. Further on, like in some related works, experiments by applying different activation functions, error scores or introducing additional layer types like fully-connected should be conducted. Some works in the field of neural audio synthesis also introduced pitch conditioning techniques, which are also a possible subject to be introduced in future studies. By adding those studies more observations could be made regarding the model performance respective the output quality and if eventual improvements can be achieved.

As the field of neural audio synthesis is of high interest for the generation of audio respective music, an implementation that is capable for real-time usage would be desirable. Despite the fact that this implementation mainly focused on the general ability to synthesize audio, it would nevertheless serve as a solid base to construct such a system.

Appendix A

Additional Graphics

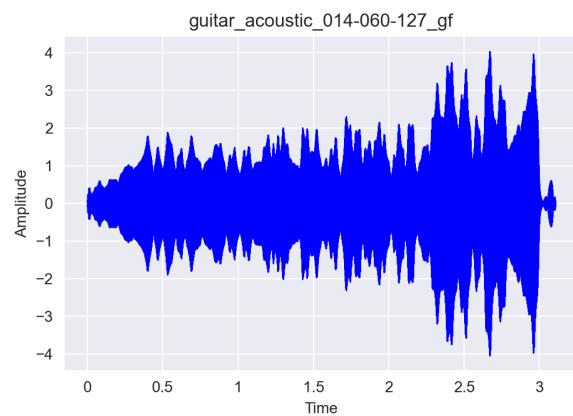


Figure A.1: Acoustic guitar output with Griffin-Lim (1D conv).

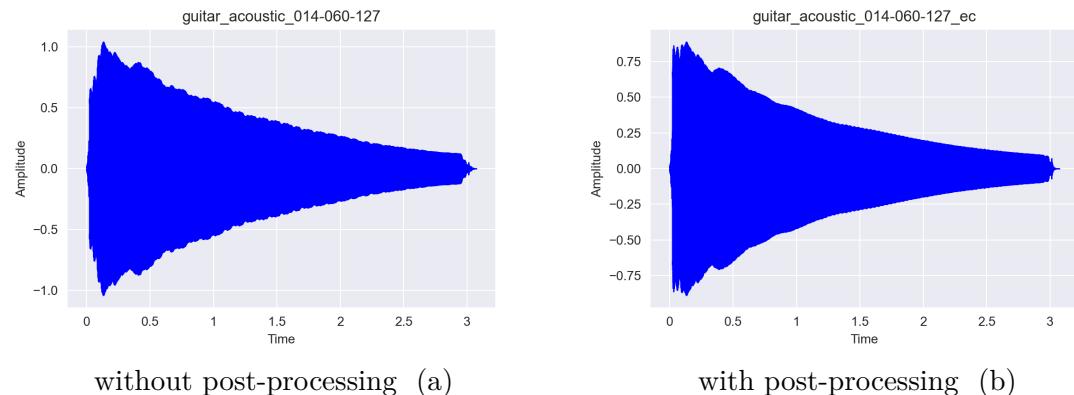


Figure A.2: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional single-stride network and original phase.

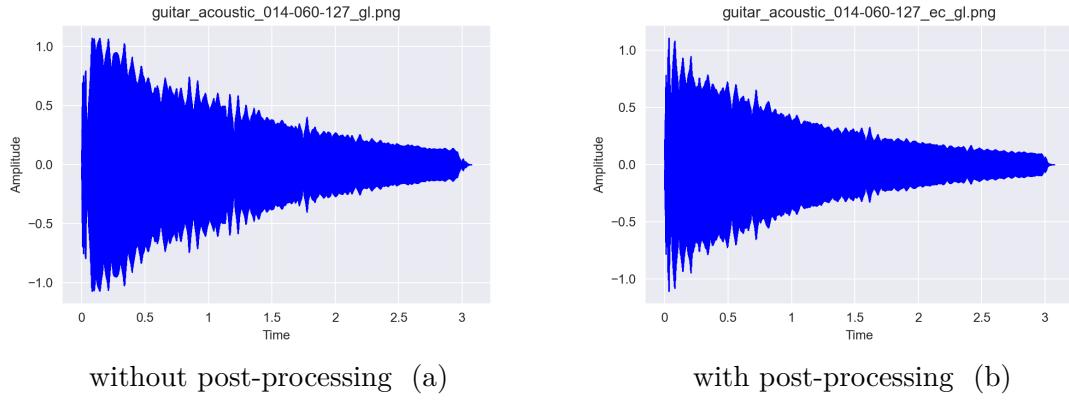


Figure A.3: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional single-stride network and Griffin-Lim.

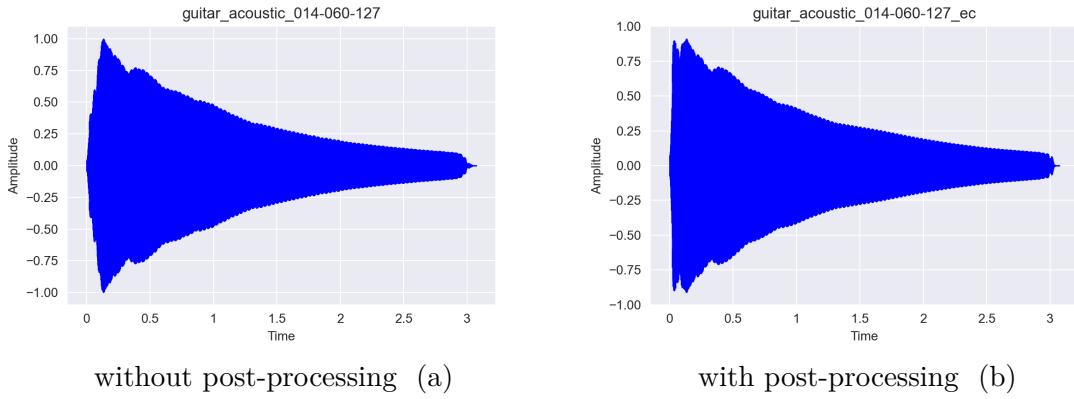


Figure A.4: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional double-stride network and original phase.

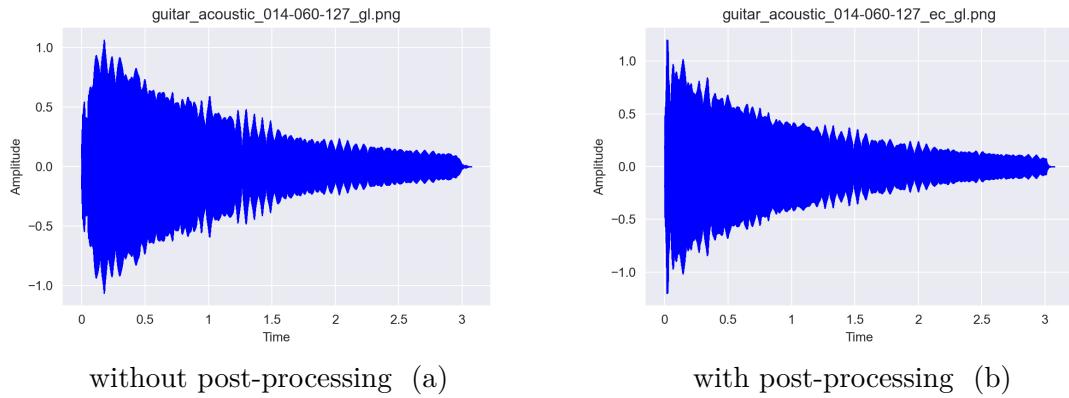


Figure A.5: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional double-stride network and Griffin-Lim.

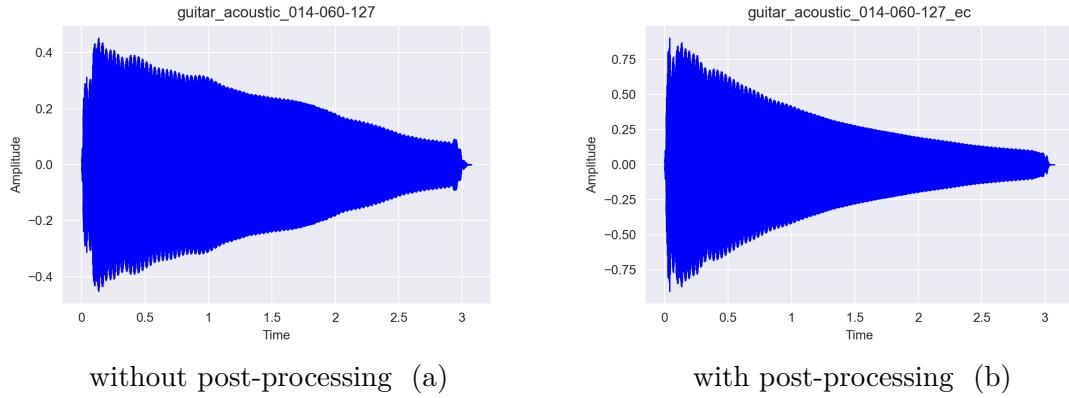


Figure A.6: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional triple-stride network and original phase.

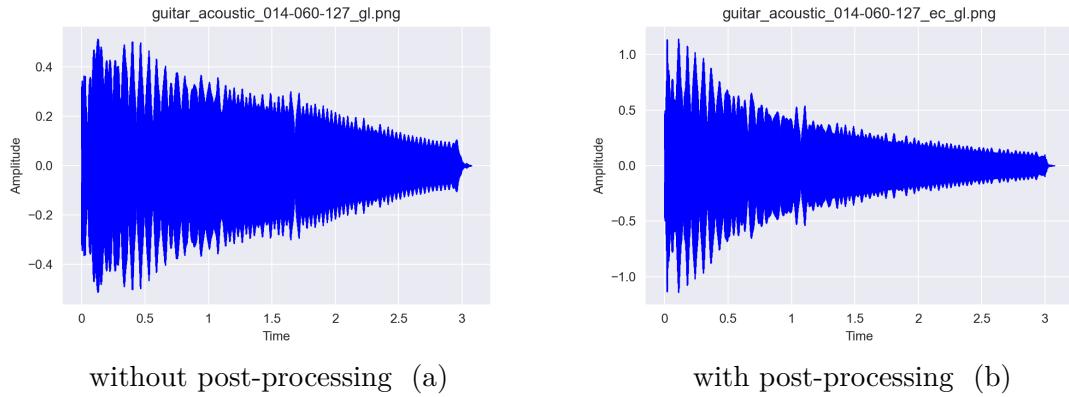


Figure A.7: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional triple-stride network and Griffin-Lim.

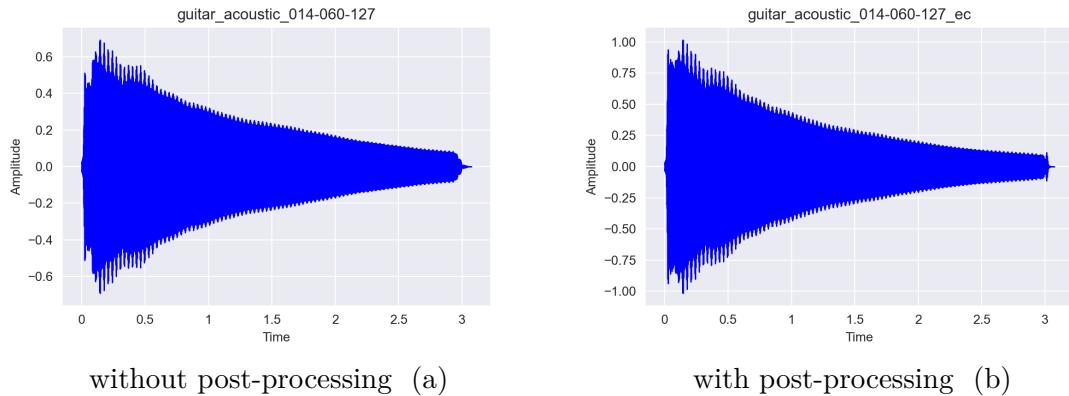


Figure A.8: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional single-stride network and original phase with log-mel spectrograms.

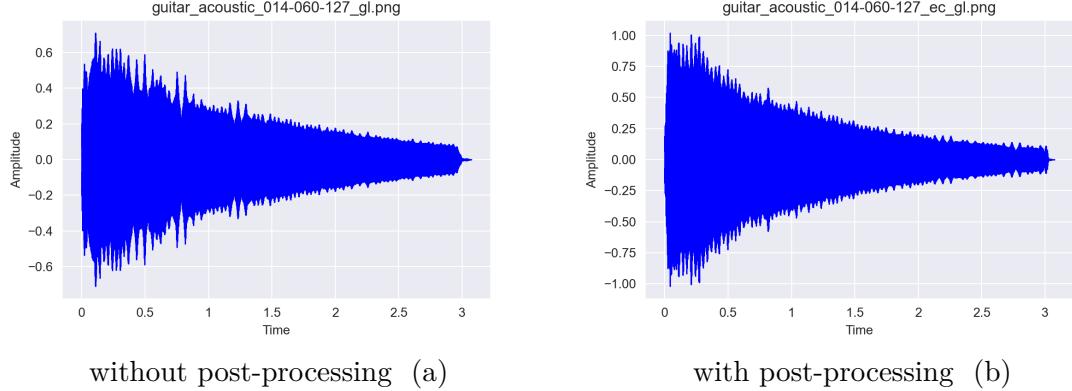


Figure A.9: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional single-stride network and Griffin-Lim with log-mel spectrograms.

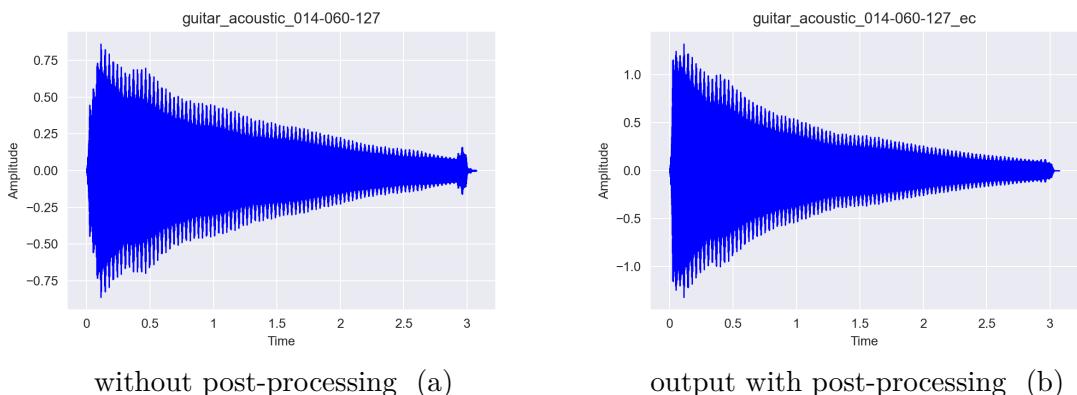


Figure A.10: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional double-stride network and original phase with log-mel spectrograms.

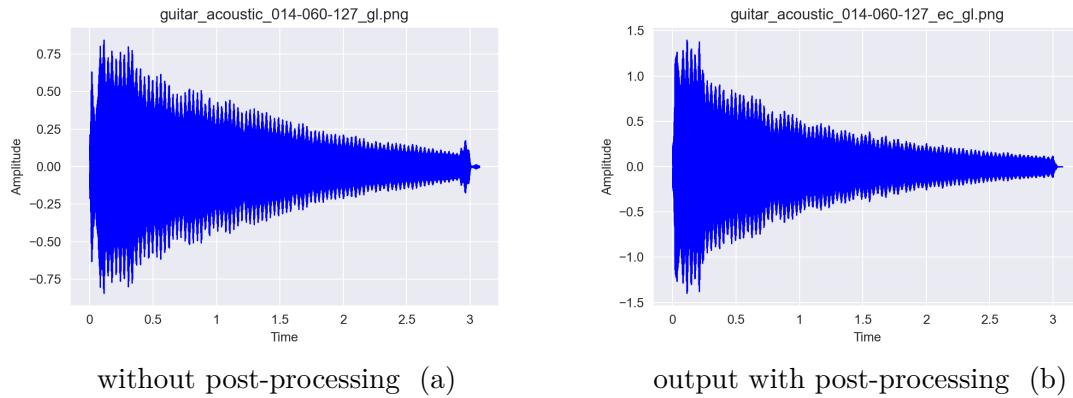


Figure A.11: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional double-stride network and Griffin-Lim with log-mel spectrograms.

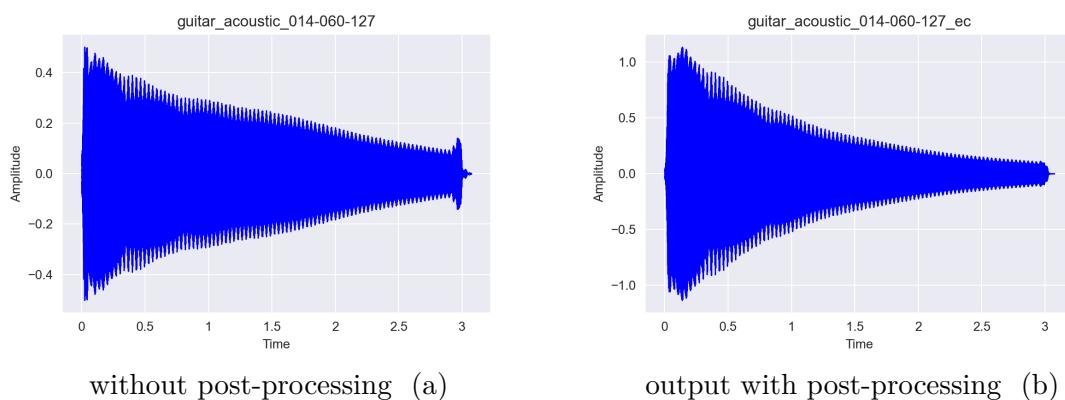


Figure A.12: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional triple-stride network and original phase with log-mel spectrograms.

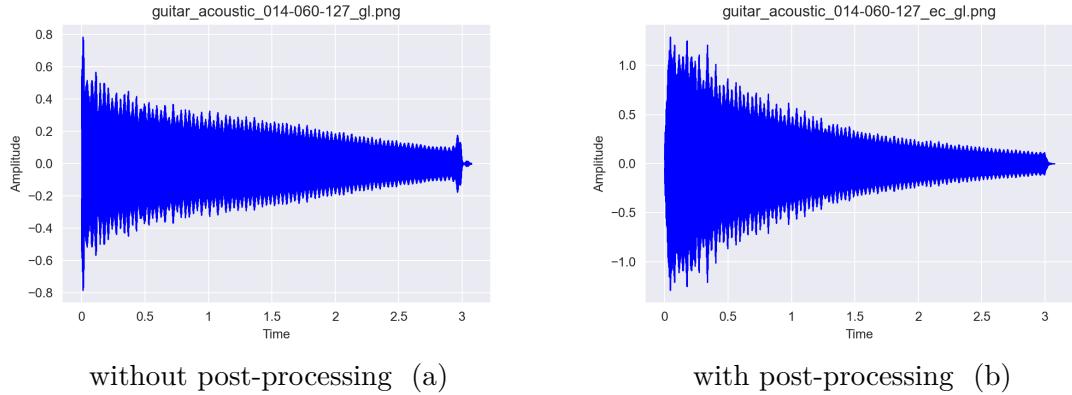


Figure A.13: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional triple-stride network and Griffin-Lim with log-mel spectrograms.

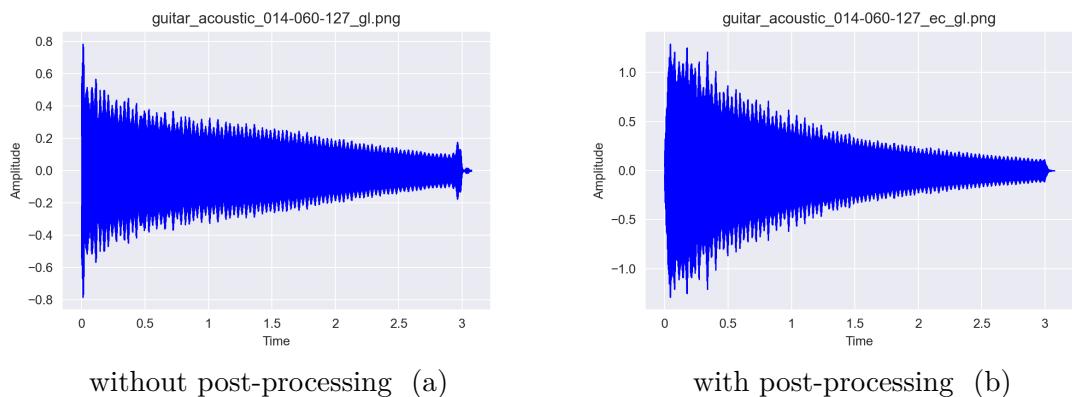


Figure A.14: Difference between acoustic guitar output signals without and with post-processing using 2D convolutional triple-stride network and Griffin-Lim with log-mel spectrograms.

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/

thesis.pdf Master thesis (complete document)

B.2 Media Files

This is a list where the synthesized audio files can be found. As a note “<stride>” is a placeholder for single-, double- or triple-stride. Furthermore the names of the output samples regarding single reconstruction contain endings like “ec” or “gl”, which mean specifically that those samples are post-processed (energy corrected) and/or generated with Griffin-Lim.

Path: /Thesis/media/correlation_matrices/

*.png correlation matrices

Path: /Thesis/media/wav/interpolated/

1D/*.wav output 1D convolutional network

2D_log_magnitude/*.wav output 2D log-magnitude

2D_log_mel/*.wav . . output 2D log-mel

Path: /Thesis/media/wav/single_reconstruction/

1D/*.wav output 1D convolutional network

2D_log_magnitude/<stride>/*.wav output 2D convolutional network
log-magnitude

2D_log_mel/<stride>/*.wav output 2D convolutional network log-mel

References

Literature

- [1] Keunwoo Choi et al. *A Tutorial on Deep Learning for Music Information Retrieval*. 2018. arXiv: 1709.04396 [cs.CV] (cit. on p. 15).
- [2] Joseph Colonel, Christopher Curro, and Sam Keene. *Autoencoding Neural Networks as Musical Audio Synthesizers*. 2018. eprint: 2004.13172 (eess.AS) (cit. on pp. 5, 6, 24).
- [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 pp. 5, 6, 24).
- [4] Vincent Dumoulin and Francesco Visin. *A guide to convolution arithmetic for deep learning*. 2018. arXiv: 1603.07285 [stat.ML] (cit. on p. 22).
- [5] Jesse H. Engel et al. “GANSynth: Adversarial Neural Audio Synthesis”. *CoRR* abs/1902.08710 (2019). arXiv: 1902.08710. URL: <http://arxiv.org/abs/1902.08710> (cit. on p. 4).
- [6] 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. 3, 4, 14, 20, 24, 26, 29, 72).
- [7] 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. 8, 9, 12).
- [8] D. Griffin and Jae Lim. “Signal estimation from modified short-time Fourier transform”. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), pp. 236–243. DOI: 10.1109/TASSP.1984.1164317 (cit. on pp. 15, 25, 37, 47, 64).
- [9] 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. 8, 10).
- [10] Lamtharn Hantrakul et al. “Fast and Flexible Neural Audio Synthesis.” In: *ISMIR*. 2019, pp. 524–530 (cit. on p. 4).

- [11] Gerhard Heinzel, Albrecht Rüdiger, and Roland Schilling. “Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new at-top windows” (2002) (cit. on p. 17).
- [12] G. E. Hinton and R. R. Salakhutdinov. “Reducing the Dimensionality of Data with Neural Networks”. *Science* 313.5786 (2006), pp. 504–507. DOI: 10.1126/science.1127647. eprint: <https://www.science.org/doi/pdf/10.1126/science.1127647> (cit. on pp. 14, 20, 21).
- [13] Geoffrey E Hinton. “How neural networks learn from experience”. *Scientific American* 267.3 (1992), pp. 144–151 (cit. on p. 19).
- [14] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG] (cit. on p. 20).
- [15] 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. 12).
- [16] Michael I. Jordan and Christopher M. Bishop. “Neural Networks”. *ACM Comput. Surv.* 28.1 (Mar. 1996), pp. 73–75. DOI: 10.1145/234313.234348 (cit. on p. 19).
- [17] 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 pp. 5, 24).
- [18] Hemant K. Kathania et al. “On the Role of Linear, Mel and Inverse-Mel Filterbank in the Context of Automatic Speech Recognition”. In: *2019 National Conference on Communications (NCC)*. 2019, pp. 1–5. DOI: 10.1109/NCC.2019.8732232 (cit. on p. 40).
- [19] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/ARXIV.1412.6980 (cit. on p. 23).
- [20] 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 pp. 9, 12).
- [21] 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. 4).
- [22] Keiron O’Shea and Ryan Nash. *An Introduction to Convolutional Neural Networks*. 2015. arXiv: 1511.08458 [cs.NE] (cit. on p. 19).
- [23] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: 1912.01703 [cs.LG] (cit. on pp. 21, 28).
- [24] Nathanaël Perraudin, Peter Balazs, and Peter L. Søndergaard. “A fast Griffin-Lim algorithm”. In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2013, pp. 1–4. DOI: 10.1109/WASPAA.2013.6701851 (cit. on p. 25).

- [25] Dhruv Ramani et al. “Autoencoder Based Architecture For Fast & Real Time Audio Style Transfer”. *CoRR* abs/1812.07159 (2018). arXiv: 1812.07159. URL: [ht tp://arxiv.org/abs/1812.07159](http://arxiv.org/abs/1812.07159) (cit. on pp. 8, 12, 14, 20).
- [26] 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 pp. 7, 24).
- [27] Stanley Smith Stevens, John Volkman, and Edwin Broomell Newman. “A scale for the measurement of the psychological magnitude pitch”. *The journal of the acoustical society of america* 8.3 (1937), pp. 185–190 (cit. on p. 40).
- [28] Prateek Verma and Julius O Smith. “Neural style transfer for audio spectrograms”. *arXiv preprint arXiv:1801.01589* (2018) (cit. on p. 9).

Software

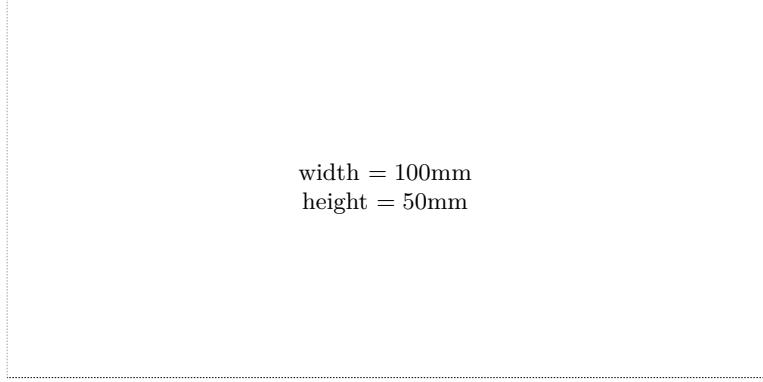
- [29] Brian McFee et al. *librosa/librosa: 0.9.1*. Version 0.9.1. Feb. 2022. doi: 10.5281/z enodo.6097378 (cit. on pp. 16, 25, 28).

Online sources

- [30] Dmitry Ulyanov and Vadim Lebedev. *Audio texture synthesis and style transfer*. 2016. URL: <https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer> (visited on 03/14/2023) (cit. on p. 11).

Check Final Print Size

— Check final print size! —



width = 100mm
height = 50mm

— Remove this page after printing! —