Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Conditional Generation of Wavetables

Bachelor's Thesis

Gion Stegmann

stgion@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Luca Lanzendörfer, Florian Grötschla
Prof. Dr. Roger Wattenhofer

September 6, 2024

Acknowledgements

I would like to sincerely thank my supervisors, Luca Lanzendörfer and Florian Grötschla, for the invaluable opportunity they gave me to work on this project, as well as for their patience and guidance throughout the entire process. I am deeply grateful to my parents for their unwavering support during my studies. A special thank you goes to my friends from Team Zirma for always being there for me, and providing encouragement and friendship during the challenging moments.

Additionally, I acknowledge the use of generative AI tools, such as ChatGPT, for providing drafting suggestions, language refinement, and conceptual clarifications throughout the writing process.

Abstract

This thesis explores the application of Conditional Variational Autoencoders (CVAEs) for generating sonically meaningful waveforms in the context of wavetable synthesis. Wavetable synthesis allows for the creation of diverse and evolving sounds by using pre-recorded or algorithmically generated waveforms. However, designing custom waveforms is a complex and often inaccessible task for many music producers due to the specialized knowledge required. The goal of this research is to explore the facilitation of waveform generation by conditioning on perceptually relevant sonic parameters, thereby simplifying the sound design process. A CVAE model was developed and trained on a dataset of single-cycle waveforms labeled with spectral characteristics. The results demonstrate that the model can generate a variety of high-quality waveforms based on user-defined parameters, expanding creative possibilities in sound design. This approach not only democratizes the creation of custom wavetables but also offers an efficient and flexible tool for music producers seeking personalized and innovative sounds.

Contents

Acknowledgements i

Abstract ii

1 Introduction 1

1.1 Motivation and Background 1
1.2 Problem Statement 2
1.3 Objectives 2
1.4 Thesis Structure 3

2 Background and Related Work 4

2.1 Synthesizers: An Overview 4
2.2 Wavetable Synthesis 5
2.3 AI in Audio Synthesis 6
2.3.1 The Rise of AI in Audio Synthesis 6
2.3.2 Challenges and Opportunities 7
2.4 Sound and Timbre in Audio Synthesis 8
2.4.1 Timbre and the Fourier Transform 8

3 Methodology 11

3.1 Sonically Meaningful Parameters 11
3.1.1 Dataset Preparation 12
3.2 Model Architecture 13
3.2.1 Encoder 13
3.2.2 Reparameterization 13
3.2.3 Decoder 14
3.3 Model Training and Validation 14
3.3.1 Loss Function 14

3.3.2 Training Process and Hyperparameters 15
3.3.3 Normalization of Brightness Values 16
3.3.4 Validation 16

4 Results **18**
4.1 Model Performance 18
4.1.1 Training and Validation Loss 18

5 Discussion **23**
5.1 Analysis of Results 23
5.1.1 Consistency Across Configurations 23
5.1.2 Impact of Brightness Weight and KL Divergence 24
5.1.3 Evaluation of Overall Performance 24
5.2 Computational Complexity 25
5.3 Limitations 25
5.3.1 Lack of Sensitivity to Hyperparameters 25
5.3.2 Dataset and Generalization 26
5.3.3 Latent Space Exploration 26
5.4 Implications 26
5.4.1 Democratizing Sound Design 26
5.4.2 Expanding Creative Possibilities 26

6 Conclusion and Future Work **27**
6.1 Summary of Findings 27
6.2 Future Work 28
6.3 Conclusion 29

Bibliography **31**

Chapter 1

Introduction

1.1 Motivation and Background

Wavetable synthesis is esteemed for its ability to generate a wide range of complex and nuanced timbres, making it a highly valuable tool in the field of digital sound synthesis. This technique, which involves the rapid playback of single-cycle waveforms, enables the efficient creation of diverse sonic textures. Consequently, wavetable synthesizers have become a popular choice among both professional and amateur music producers.

Despite its advantages, wavetable synthesis presents a significant challenge in the realm of sound design. The process of designing waveforms that accurately reflect a producer's creative vision is inherently complex and specialized. For many producers, the task of translating abstract auditory concepts into concrete waveforms is difficult, particularly because this requires a deep understanding of both the technical aspects of waveform construction and the perceptual qualities of sound. This complexity means that the creation of waveforms is often relegated to expert sound designers who possess the necessary skills to craft these intricate audio elements.

These experts typically create and commercialize waveform packs or provide them to synthesizer manufacturers. While this practice ensures the availability of high-quality waveforms, it also introduces a notable limitation: producers are constrained to using pre-designed waveforms, which can restrict their creative potential. The reliance on pre-made waveforms inherently limits the range of sonic possibilities, as producers are confined to the specific sounds provided, reducing the opportunity for more personalized and innovative sound design.

This thesis aims to address these limitations by exploring the application of Artificial Intelligence, specifically Conditional Variational Autoencoders (CVAEs), in facilitating the process of waveform creation. By enabling the generation of waveforms conditioned on user-selected, sonically meaningful parameters, this approach seeks to democratize the process of sound design in wavetable synthesis. The ultimate goal is to empower music producers to transcend the constraints of

pre-made waveforms, thereby unlocking greater creative potential and enabling the realization of a more personalized sonic vision.

1.2 Problem Statement

The process of sound design in wavetable synthesis, while powerful, remains complex and largely inaccessible to the average music producer. The creation of custom waveforms that accurately reflect a producer's auditory intentions typically requires specialized knowledge and skills, which most producers lack. Consequently, the task of waveform creation is often left to professional sound designers, whose products—whether in the form of waveform packs or as part of commercial synthesizers—inevitably limit the creative flexibility of the user. This reliance on pre-designed waveforms constrains the producer's ability to fully explore and realize their creative potential in sound design. There is a clear need for a solution that can democratize the creation of sonically meaningful waveforms, allowing producers of all skill levels to translate their creative ideas into tangible audio forms without being confined to existing presets or packs.

1.3 Objectives

The primary objective of this thesis is to develop a method for generating sonically pleasant and contextually relevant wavetables using Conditional Variational Autoencoders (CVAEs). Specifically, this research aims to:

- **Explore the use of Autoencoders for audio synthesis:** Investigate the suitability of Autoencoders, specifically CVAEs for generating single-cycle waveforms in the context of wavetable synthesis.
- **Define and utilize sonically meaningful parameters:** Identify key auditory parameters (e.g., brightness, warmth) that can be used to condition the generation process, ensuring that the output waveforms align with the intended sonic characteristics.
- **Automate the process of waveform creation:** Develop and train a CVAE model that can generate waveforms based on user-specified parameters, thereby reducing the complexity and expertise required for custom sound design.
- **Evaluate the generated waveforms:** Assess the quality and usability of the generated waveforms through both quantitative measures and qualitative listening tests, comparing them to traditional, manually designed waveforms.

1.4 Thesis Structure

This thesis is structured as follows:

- **Introduction:** Introduces the topic, presents the motivation for the study, outlines the problem statement, and specifies the objectives of the research.
- **Background and Related Work:** Provides a detailed overview of wavetable synthesis, discusses the challenges in waveform creation, and reviews existing approaches in AI-driven audio synthesis, with a focus on related work in the field.
- **Methodology:** Describes the architecture of the Conditional Variational Autoencoder (CVAE) used in this study, details the selection of sonically meaningful parameters, and outlines the data preparation, model training, and validation processes.
- **Results:** Presents the findings of the study, including the performance of the CVAE model, the characteristics of the generated waveforms, and their evaluation in comparison to traditional methods.
- **Discussion:** Analyzes the results, highlighting the strengths and limitations of the approach, and discusses the implications of these findings for the field of sound synthesis.
- **Conclusion and Future Work:** Summarizes the key contributions of the thesis, reflects on the research outcomes, and suggests directions for future work to further enhance the capabilities of AI-driven waveform generation.

Chapter 2

Background and Related Work

2.1 Synthesizers: An Overview

Synthesizers have been a cornerstone of modern music production, revolutionizing the way sound is created and manipulated. Since their inception, synthesizers have evolved through various technological advancements, each bringing new capabilities and expanding the sonic palette available to musicians.

The earliest synthesizers, such as the Moog Modular Synthesizer introduced in the 1960s, were analog devices that used voltage-controlled oscillators (VCOs), filters, and amplifiers to generate and shape sound [1]. These early analog synthesizers were primarily based on subtractive synthesis, a method where complex waveforms (rich in harmonics) are generated and then shaped using filters to remove (or subtract) certain frequencies, creating the desired timbre. Subtractive synthesis became the foundation of many classic synthesizers, such as the Minimoog and the Roland Jupiter series, which remain iconic in music production [2].

The 1980s marked a significant shift with the introduction of frequency modulation (FM) synthesis, popularized by the Yamaha DX7. FM synthesis generates complex sounds by modulating the frequency of one waveform with another, creating intricate harmonic structures that were difficult to achieve with subtractive methods [3]. This era also saw the rise of digital synthesizers, which allowed for more precise control and the ability to produce entirely new types of sounds through methods such as additive synthesis (building sounds by adding together simple waveforms) and sample-based synthesis (using recorded audio samples as the source material) [4].

As digital technology advanced, synthesizers became more compact, affordable, and accessible, leading to widespread adoption across various genres of music. The 1990s and 2000s saw the emergence of virtual analog synthesizers, which emulated the sound and behavior of classic analog synths using digital signal processing (DSP). These developments paved the way for modern software synthesizers (soft synths), which offer immense flexibility and integration within

digital audio workstations (DAWs). [5]

Each type of synthesis method—subtractive, FM, additive, sample-based, and virtual analog—has contributed to the vast and diverse landscape of sound design in music production. However, as technology continues to evolve, newer methods, such as wavetable synthesis, have gained prominence due to their ability to produce highly detailed and dynamic sounds with greater ease and flexibility.

2.2 Wavetable Synthesis

Wavetable synthesis represents a significant advancement in the evolution of synthesizers, providing a powerful tool for generating complex and evolving sounds. Unlike traditional synthesis methods, which often rely on manipulating simple waveforms or samples, wavetable synthesis uses a series of pre-recorded or algorithmically generated waveforms stored in a table (hence the name). These waveforms, each representing a different harmonic content or timbre, can be interpolated, modulated, or sequenced in real-time to produce sounds that can evolve dynamically over time. [6]

The concept of wavetable synthesis was first introduced in the late 1970s and early 1980s, with the PPG Wave synthesizer being one of the earliest and most influential examples. The PPG Wave, designed by Wolfgang Palm, utilized digital wavetables in combination with analog filters, creating a hybrid instrument capable of producing a wide range of unique sounds. [7]

One of the key advantages of wavetable synthesis is its ability to offer a vast sonic palette within a single instrument. By storing multiple waveforms within a wavetable, a synthesizer can play different timbres, creating complex sounds that would be difficult to achieve using traditional methods.

Moreover, wavetable synthesis allows for detailed control over the harmonic content of a sound. Producers can design custom wavetables by either selecting or creating the waveforms that best match their creative vision. This level of control is particularly valuable in modern music production, where there is often a need to craft highly specific and unique sounds.

However, the complexity of wavetable synthesis also presents challenges, particularly in the area of sound design. Creating effective and sonically pleasing wavetables requires not only a deep understanding of sound synthesis but also an ability to translate abstract auditory ideas into concrete waveforms. This challenge is compounded by the sheer number of possibilities offered by wavetable synthesis, which can make the process of designing custom wavetables both time-consuming and technically demanding.

2.3 AI in Audio Synthesis

Artificial Intelligence has increasingly become a transformative force in various fields, and audio synthesis is no exception. The integration of AI into audio synthesis has opened up new possibilities for sound design, composition, and production, allowing for the creation of more complex, nuanced, and innovative sounds than ever before. This section will explore the role of AI in audio synthesis, with a particular focus on key developments such as the NSynth project [8] and the emergence of AI-generated music.

2.3.1 The Rise of AI in Audio Synthesis

The application of AI in audio synthesis is rooted in the broader field of machine learning, where algorithms learn patterns from data and use these patterns to generate new content. In audio synthesis, this typically involves training neural networks on large datasets of sounds, enabling the models to learn the characteristics of different audio signals and generate new sounds based on this learned knowledge [9].

Historically, much of the research in AI and audio has focused on tasks that are either upstream or downstream of synthesis. On the upstream side, significant advancements have been made in voice generation and text-to-speech (TTS) systems [10], where AI models generate human-like speech from text inputs. These systems are now widely used in virtual assistants, automated customer service, and accessibility tools. On the downstream side, AI has excelled in speech recognition and dictation, where models convert spoken language into text, as well as in tasks like audio classification and music recommendation [11], which involve categorizing and analyzing audio content. While these applications are critical and have seen widespread adoption, they differ fundamentally from the creative and generative challenges posed by musical audio synthesis [12].

One of the pioneering efforts in AI-driven audio synthesis is Google's NSynth (Neural Synthesizer), developed as part of the Magenta project [8]. NSynth uses a deep neural network to analyze and interpolate between sounds from a large database of musical notes played by various instruments. Unlike traditional synthesis methods that rely on deterministic mathematical algorithms to generate sounds, NSynth leverages the learned latent space of the neural network to create new, unique sounds by blending the characteristics of different source sounds. This allows for the generation of hybrid timbres that are not constrained by the limitations of traditional synthesis techniques, offering a new frontier in sound design.

NSynth operates by encoding audio samples into a latent space where similar sounds are positioned closer together. By manipulating this space, NSynth can generate novel sounds that have characteristics of multiple instruments or timbres

[8]. For example, it can create a sound that blends the qualities of a flute and a violin, or a guitar and a trumpet, producing new sonic textures that are both familiar and entirely original. This approach not only expands the palette of available sounds but also introduces new creative possibilities for musicians and sound designers.

Beyond NSynth, AI has also been applied to more comprehensive music generation tasks. AI systems such as OpenAI's Jukebox [13] and AIVA (Artificial Intelligence Virtual Artist) [14] have demonstrated the capability to compose entire musical pieces autonomously. These systems use deep learning models trained on large datasets of music to understand and replicate the structure, style, and emotional content of different genres. AIVA, for instance, has been used to compose classical music pieces that are indistinguishable from human-composed works in terms of complexity and emotional depth [14]. This level of sophistication in AI-generated music represents a significant milestone in the integration of AI into the creative process.

AI-driven tools have also made their way into more practical applications within music production. For instance, tools that use AI to automate tasks such as mastering and audio restoration are becoming increasingly common. These tools can analyze audio tracks and apply adjustments that would traditionally require expert knowledge, thereby making high-quality music production more accessible to a broader range of users [15].

2.3.2 Challenges and Opportunities

While the integration of AI in audio synthesis offers exciting opportunities, it also presents several challenges. One of the primary challenges is ensuring that AI-generated sounds are not only novel but also musically useful and emotionally engaging. The subjective nature of music and sound means that evaluating the quality of AI-generated content is inherently difficult, as it must resonate with human listeners on both a technical and emotional level. [16]

Moreover, there is the issue of control and interpretability. In traditional synthesis methods, sound designers have a clear understanding of how their input parameters will affect the output. With AI, particularly deep learning models, the relationship between input and output can be more opaque, making it harder for users to predict or fine-tune the results [17]. This can limit the usability of AI tools, especially for professionals who require precise control over their sound design.

Despite these challenges, the potential for AI to revolutionize audio synthesis is significant. By enabling the creation of entirely new types of sounds and automating complex sound design tasks, AI has the potential to democratize music production, making it more accessible and expanding the creative possibilities available to musicians and producers. As AI technology continues to evolve, it

is likely that we will see even more sophisticated applications in audio synthesis, pushing the boundaries of what is possible in sound design and music creation.

2.4 Sound and Timbre in Audio Synthesis

Timbre is one of the most fundamental aspects of how we perceive sound. The Acoustical Society of America (ASA) defines timbre as "that attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar," adding that "timbre depends primarily upon the frequency spectrum, although it also depends upon the sound pressure and the temporal characteristics of the sound" [18]. This characteristic allows us to distinguish between different sounds, even when they share the same pitch and loudness. For example, a piano and a guitar playing the same note at the same volume will still sound distinctly different due to their unique timbral characteristics. Timbre is often described as the "color" or "quality" of sound and is influenced by the harmonic content, dynamic behavior, and spectral distribution of the waveform.

2.4.1 Timbre and the Fourier Transform

Timbre can also be understood visually through a **Fourier transform** representation of a sound, which converts a time-domain signal (i.e., a waveform) into the frequency domain [**?**]. The Fourier transform reveals the harmonic content of a waveform by showing the individual frequencies that make up the sound and their relative amplitudes.

For example, in the Fourier transform of a waveform, a single fundamental frequency will appear as the largest peak, while its overtones, or harmonics, will appear as smaller peaks at integer multiples of the fundamental frequency. The distribution and amplitude of these harmonics are what primarily define the timbre of the sound.

Consider three waveforms—a pure sine wave, a square wave, and a triangle wave—all with the same fundamental frequency:

- A **sine wave** is the simplest waveform, containing only the fundamental frequency with no overtones. In the Fourier transform, it appears as a single spike at the fundamental frequency. Its sound is perceived as very pure and smooth.
- A **square wave**, on the other hand, has a richer harmonic content. Its Fourier transform shows not only the fundamental frequency but also odd harmonics (i.e., $3f_0$, $5f_0$, etc.). The higher the amplitude of these harmonics, the "brighter" or "harsher" the sound becomes.

- A **triangle wave** has a harmonic structure similar to the square wave, but the harmonics fall off more quickly. Like the square wave, it contains only odd harmonics, but their amplitudes decrease at a faster rate (proportional to $1/n^2$, where n is the harmonic number). This results in a softer, more "mellow" sound compared to the square wave.

The plots below show the time-domain representation (left) and the corresponding Fourier transform (right) for a sine wave, square wave, and triangle wave, respectively:

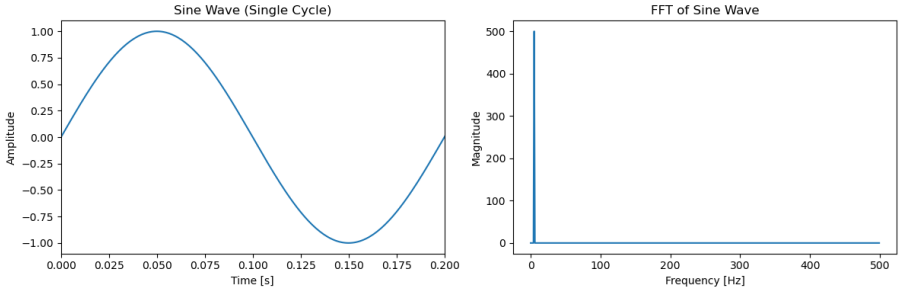

Figure 2.1: Sine wave (left) and its Fourier Transform (right).

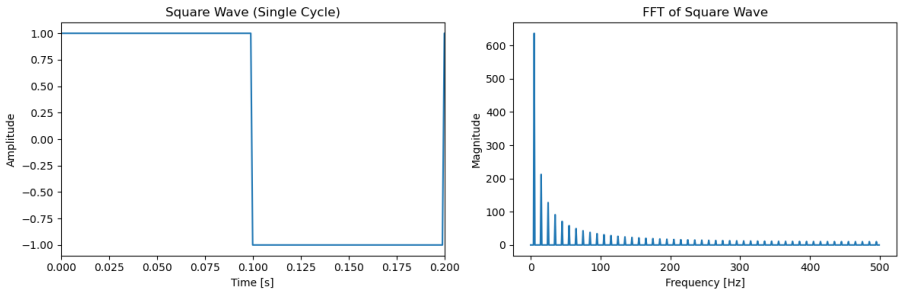

Figure 2.2: Square wave (left) and its Fourier Transform (right).

In wavetable synthesis, controlling and shaping timbre is essential, as each single-cycle waveform has a distinct harmonic structure that defines its timbral characteristics. Traditionally, sound designers manipulate timbre by adjusting the harmonic content of a waveform through its Fourier representation—shaping the frequency components to craft the desired sound. This involves modifying the amplitude and distribution of the waveform's harmonics to achieve specific qualities such as brightness, warmth, or harshness.

In this thesis, however, rather than directly manipulating the Fourier space, we aim to shape the sound in the time domain, allowing the waveform's harmonic structure to emerge naturally from the generated time-domain signal. The Conditional Variational Autoencoder (CVAE) is designed to generate wavetables

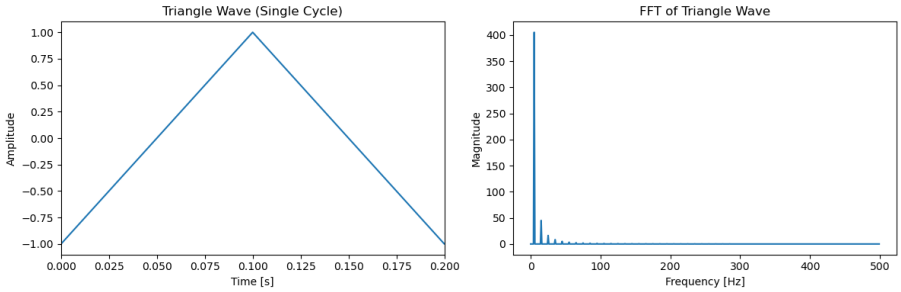

Figure 2.3: Triangle wave (left) and its Fourier Transform (right).

conditioned on perceptually meaningful parameters like brightness and warmth, which are closely tied to the harmonic content of the waveform. By controlling these parameters in the time domain, we implicitly shape the Fourier representation, just as a sound designer would when sculpting harmonics.

Chapter 3

Methodology

3.1 Sonically Meaningful Parameters

In order to guide the waveform generation process, the CVAE is conditioned on a sonically meaningful parameter that corresponds to perceptual characteristics of sound. This parameter is chosen to ensure that the generated waveforms align with the creative goals of music producers and sound designers. The parameter used in this research is defined as follows:

- **Brightness**: Brightness refers to the balance of high-frequency content in a sound. Sounds with more energy in higher frequency ranges are perceived as "brighter," while sounds with more emphasis on lower frequencies are perceived as "darker" or "duller." To quantify brightness, the spectral centroid of the waveform is calculated, which serves as an indicator of where the majority of the sound's frequency energy is concentrated. [19] [20]

  The calculation of brightness begins by applying a real fast Fourier transform (FFT) to the waveform, which converts the time-domain signal into its frequency-domain representation. The spectral centroid is then computed using the following formula:

  $$\text{spectral\_centroid} = \frac{\sum f \cdot |X(f)|}{\sum |X(f)| + 10^{-8}}$$

  where $X(f)$ represents the magnitude of the waveform's frequency component at frequency f, and the sum is taken over all frequencies in the spectrum. This ensures that the spectral centroid reflects the "center of mass" of the frequency distribution, providing a direct measure of brightness.

  A higher spectral centroid indicates that more energy is concentrated in the higher frequencies, resulting in a brighter sound. Conversely, a lower spectral centroid corresponds to a darker sound. This calculated brightness

is used as an input to the CVAE, allowing the model to generate waveforms with varying degrees of brightness, based on user-defined preferences.

The brightness parameter is numerically encoded and passed as part of the conditioning input c, allowing the model to learn how variations in this parameter affect the resulting waveform. This approach enables intuitive control over the synthesis process and facilitates the generation of a wide range of sounds by simply adjusting the brightness value.

3.1.1 Dataset Preparation

To train the Conditional Variational Autoencoder (CVAE) model, a dataset of single-cycle waveforms was prepared using the AKWF dataset. The AKWF dataset is a collection of approximately 4000 single-cycle waveforms, provided under the Creative Commons CC0 1.0 license.[1] These waveforms, sampled at 600 samples per cycle, are widely used in synthesizers and samplers for wavetable synthesis. This makes the dataset ideal for research in automated waveform generation.

The preparation of the dataset involved several key steps:

- **Waveform Normalization**: Each waveform from the AKWF dataset was processed to ensure consistency in amplitude by normalizing it to the range $[-1, 1]$. This step was crucial to ensure that all waveforms had comparable levels and could be used uniformly in the training process.
- **Brightness Calculation**: The perceptual quality of brightness was computed for each waveform using its spectral centroid, which measures the distribution of energy across the frequency spectrum. The spectral centroid is a widely used metric for quantifying how much of the waveform's energy is concentrated in higher frequencies. The brightness values were normalized to a range of 0 to 1, ensuring that they were comparable across all waveforms in the dataset.
- **Data Labeling and Storage**: After computing the brightness values, each waveform was labeled with its corresponding brightness value. These labels were stored alongside the waveforms, allowing the CVAE model to condition its waveform generation on these perceptual characteristics during training.
- **Dataset Structuring**: The dataset was organized for efficient loading and training. The waveforms and their corresponding brightness labels were structured to be used in batches, allowing for scalable and effective training of the model.

[1] https://www.adventurekid.se/akrt/waveforms/adventure-kid-waveforms/

The use of the AKWF dataset ensured that the waveforms provided a wide range of timbral characteristics, making it a robust resource for training the CVAE. The careful normalization of both waveform amplitude and brightness values facilitated the generation of diverse and perceptually meaningful waveforms, aligned with the creative goals of sound designers and music producers.

3.2 Model Architecture

The Conditional Variational Autoencoder (CVAE) used in this research is designed to generate single-cycle waveforms conditioned on a perceptual feature—brightness. The CVAE consists of two primary components: an encoder and a decoder, with a reparameterization step in between to facilitate sampling from the learned latent space. This section describes the architecture of the CVAE, as well as the loss function used during training.

3.2.1 Encoder

The encoder takes as input both the waveform and the brightness condition. The waveform is represented as a vector of 600 time-domain samples, and the brightness is a scalar value representing the perceptual quality of the sound. These two inputs are concatenated and passed through a series of fully connected layers to extract high-level features. The final output of the encoder is two vectors: one representing the mean (μ) and the other representing the logarithm of the variance ($\log \sigma^2$) of the latent distribution. These values are used to define the Gaussian distribution from which latent variables are sampled.

Mathematically, given the input waveform x and the conditioning variable c (brightness), the encoder computes:

$$\mu = f_{\text{enc},\mu}(x, c), \quad \log \sigma^2 = f_{\text{enc},}(x, c)$$

where $f_{\text{enc},\mu}$ and $f_{\text{enc},}$ represent the neural network transformations that output the mean and log-variance, respectively.

3.2.2 Reparameterization

The reparameterization step is crucial for enabling backpropagation through the stochastic latent space. To sample from the latent space, the mean and log-variance produced by the encoder are used to generate the latent variable z. This is done by sampling from a Gaussian distribution with the learned mean and variance:

$$z = \mu + \epsilon \cdot \sigma, \quad \text{where } \epsilon \sim \mathcal{N}(0, I)$$

Here, $\sigma = \exp(0.5 \cdot \log \sigma^2)$ represents the standard deviation, and ϵ is a random noise sampled from a standard normal distribution. This process, known as the reparameterization trick, allows gradients to be passed through the sampling process during training.

3.2.3 Decoder

The decoder takes the sampled latent variable z and the brightness condition c as inputs, concatenates them, and passes them through a series of fully connected layers. The decoder aims to reconstruct the original waveform from the latent representation. The output of the decoder is a vector of the same dimension as the input waveform, and a `tanh` activation function is applied to ensure that the reconstructed waveform has values in the range $[-1, 1]$.

Formally, the decoder reconstructs the waveform $\hat{x}$ as follows:

$$\hat{x} = f_{\text{dec}}(z, c)$$

where f_{dec} represents the neural network that decodes the latent variable z and the conditioning variable c back into a waveform.

3.3 Model Training and Validation

Training the Conditional Variational Autoencoder (CVAE) involved optimizing the model's ability to reconstruct input waveforms and ensuring the latent space supports smooth interpolation while adhering to the desired conditioning parameters. The following details the training process, including the loss function, optimization, and hyperparameters used.

3.3.1 Loss Function

The total loss function used during training is a combination of three key components:

- **Reconstruction Loss**: This is computed using the Mean Squared Error (MSE) between the input waveform x and the reconstructed waveform $\hat{x}$. The reconstruction loss encourages the model to accurately reproduce the

original waveform from the latent space and the conditioning parameters:

$$\text{MSE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

- **KL Divergence Loss**: This term encourages the latent space to follow a normal distribution, allowing the model to generate new waveforms by sampling from this distribution. The KL divergence between the learned latent distribution $q(z|x)$ and a standard normal distribution $p(z)$ is given by:

$$\text{KL}(q(z|x)||p(z)) = -0.5 \sum_{i=1}^{N} (1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2)$$

  where μ and σ represent the mean and variance of the latent space.

- **Brightness Loss**: This measures the error between the predicted brightness of the reconstructed waveform and the actual brightness of the input waveform. The brightness is computed as a scalar value for each waveform, and the loss is calculated as:

$$\text{MSE}(b_{\text{pred}}, b_{\text{true}}) = \frac{1}{N} \sum_{i=1}^{N} (b_{\text{pred}_i} - b_{\text{true}_i})^2$$

  where b_{pred} is the predicted brightness and b_{true} is the true brightness value.

The total loss is a weighted combination of these three components:

$$\mathcal{L} = \alpha \cdot \text{MSE}(x, \hat{x}) + \beta \cdot \text{KL} + \gamma \cdot \text{MSE}(b_{\text{pred}}, b_{\text{true}})$$

where α, β, and γ are the weights assigned to each loss term. In the final implementation, these weights were set as follows:

$$\alpha = 1.0, \quad \beta = 0.1, \quad \gamma = 1.0$$

3.3.2 Training Process and Hyperparameters

The Conditional Variational Autoencoder (CVAE) model was trained using a variety of configurations to explore how different hyperparameters and latent dimensions affected the model's performance. The following parameters were explored:

- **Waveform Input Dimension**: Each waveform was represented as a vector of 600 time-domain samples.

- **Condition Dimension**: The brightness parameter was used as the single conditioning input.
- **Latent Dimension**: Different values of latent dimension $(32, 64, 128)$ were tested to determine the optimal capacity for encoding the waveform data.
- **Hidden Dimension**: All configurations used a hidden dimension of 256 for the fully connected layers in both the encoder and decoder.
- **Loss Weights**: The relative weights for the reconstruction loss, KL divergence, and brightness loss were adjusted across experiments. The brightness loss was weighted more heavily in some configurations to improve perceptual accuracy in generated waveforms, with values ranging from 1.0 to 4.0.
- **Optimizer**: The Adam optimizer was used for training, with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.
- **Learning Rate**: A fixed learning rate of 1×10^{-3} was used in all experiments.
- **Batch Size and Epochs**: Each configuration was trained for 100 epochs using a batch size of 64, ensuring sufficient coverage of the dataset.

The model was trained on single-cycle waveforms from the AKWF dataset, and the brightness of each waveform was used as the conditioning parameter to guide the generation process. The different configurations allowed for a comparison of how varying the latent dimension and brightness weight impacted both the reconstruction quality and the perceptual alignment with the brightness parameter.

3.3.3 Normalization of Brightness Values

As part of the brightness loss, the predicted brightness values were normalized based on the maximum brightness in the dataset, which was set to a constant value of 5000 for this experiment. The brightness predictions for the reconstructed waveforms were clamped between 0.0 and 1.0 to ensure they remained within a valid range.

3.3.4 Validation

To monitor the model's performance during training, a validation set consisting of 20% of the data was used. The validation loss, consisting of the same three components (reconstruction, KL divergence, and brightness), was computed after each epoch to track overfitting. Early stopping was applied based on the validation loss to prevent overfitting and ensure generalization.

Additionally, qualitative assessments were performed by visually inspecting the reconstructed and generated waveforms and conducting listening tests to ensure that the generated waveforms aligned with their conditioning parameters.

At the end of training, the model's ability to generate new waveforms was tested by sampling from the latent space and conditioning on different brightness values. These generated waveforms were compared both visually (in terms of waveform shape) and aurally to confirm that the CVAE was successfully producing sonically meaningful waveforms aligned with the input conditions.

Chapter 4

Results

This chapter presents the outcomes of the experiments conducted to evaluate the performance of the CVAE in generating sonically meaningful wavetables. The results are organized into three sections: model performance, the characteristics of the generated wavetables, and a detailed evaluation of the generated sounds.

4.1 Model Performance

The performance of the CVAE model was evaluated based on four primary metrics: the **reconstruction loss** (MSE), **KL divergence**, **brightness prediction loss**, and the **total loss**. These metrics were used to assess how well the model was able to learn and reconstruct waveforms, as well as the quality of the latent space learned during training and its ability to predict brightness.

4.1.1 Training and Validation Loss

During the training process, the model's loss functions—reconstruction loss (Mean Squared Error), KL divergence loss, brightness prediction loss, and total loss—were monitored for both the training and validation datasets. Figures 4.1, 4.2, 4.3, and 4.4 show the progression of these metrics over the course of training for several different configurations.

As seen in Figures 4.1 to 4.4, the model shows strong convergence across all metrics. The brightness loss decreases rapidly during the early steps of training, suggesting that the model learns to predict the brightness of generated waveforms accurately from the start. The KL divergence stabilizes after some initial fluctuation, indicating that the model maintains a well-formed latent space for effective waveform generation. Both the total loss and MSE converge steadily, confirming the model's ability to minimize reconstruction errors over time.

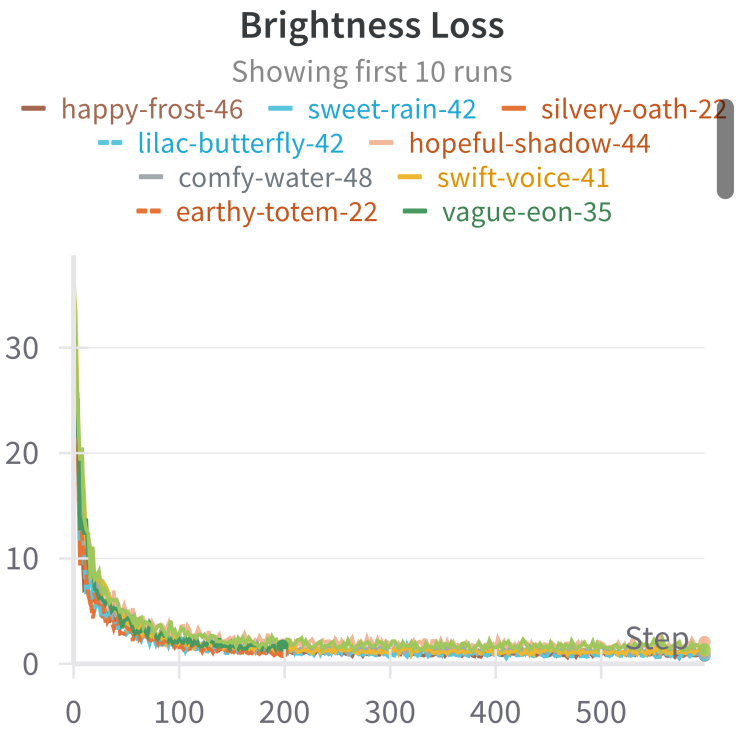

Figure 4.1: Brightness Loss over 500 steps for multiple runs of the CVAE model. The loss converges quickly, indicating the model's ability to align the generated waveforms with the target brightness values.

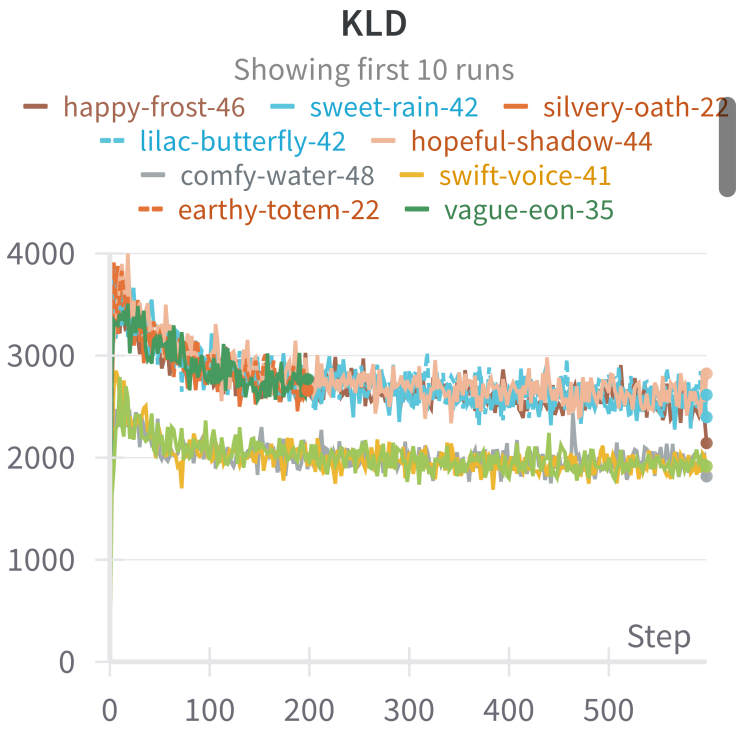

Figure 4.2: KL Divergence over 500 steps for multiple runs of the CVAE model. The latent space gradually stabilizes, as indicated by the declining KLD.

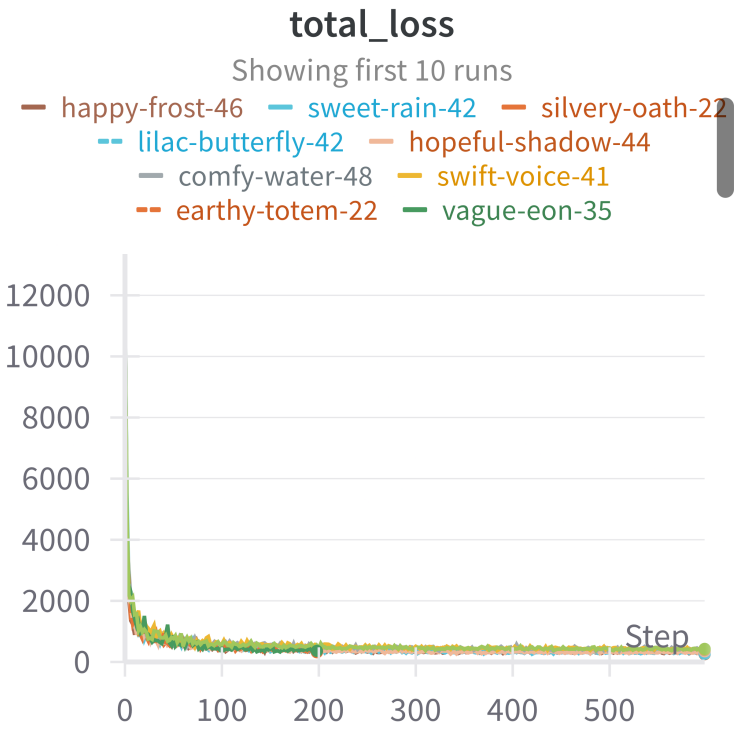

Figure 4.3: Total Loss over 500 steps for multiple runs of the CVAE model. The convergence of the total loss demonstrates the overall learning and stabilization of the model.

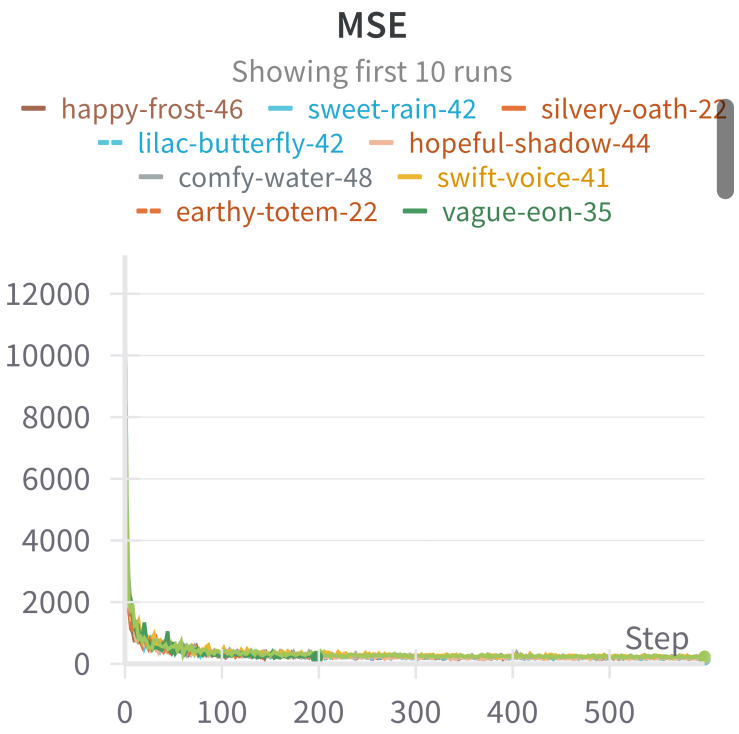

Figure 4.4: Mean Squared Error (MSE) over 500 steps for multiple runs of the CVAE model. The reconstruction error converges steadily, indicating the model's ability to accurately reconstruct input waveforms.

Chapter 5

Discussion

This chapter provides an interpretation of the results presented in the previous chapter, explores the limitations of the current approach, and discusses the broader implications of the findings for the field of wavetable synthesis and AI-driven sound design.

5.1 Analysis of Results

The results from the CVAE model demonstrate the potential of conditional generative models in creating sonically meaningful wavetables. The ability to condition the waveform generation on perceptual parameters such as brightness was a significant strength, as reflected in both the quantitative analysis and loss convergence metrics across multiple configurations.

5.1.1 Consistency Across Configurations

The training results showed a high degree of consistency across different configurations, even though the configurations varied in terms of brightness weight, KL divergence weight, and other hyperparameters. As demonstrated in Figures 4.1, 4.2, 4.3, and 4.4, the loss curves for all key metrics—including brightness prediction loss, reconstruction loss (MSE), KL divergence, and total loss—followed nearly identical trajectories across configurations.

This consistency is a positive indicator of the model's robustness, suggesting that the CVAE is capable of reliably learning the underlying waveform characteristics regardless of variations in the loss function weights. The model's ability to handle different configurations with similar performance indicates that it is flexible and stable across a range of parameter settings. This is particularly advantageous in practical applications, where exact hyperparameter tuning may not be feasible or desirable.

5.1.2 Impact of Brightness Weight and KL Divergence

The fact that the brightness weight (γ) varied from 1.0 to 4.0 without significantly affecting the final loss indicates that brightness is an easily learned feature for the model. This suggests that the model architecture naturally captures the brightness parameter well, even without a high weighting in the loss function. This could be due to the inherent structure of the input data, which may lend itself to straightforward representation of brightness-related features.

Similarly, the KL divergence loss curves indicate that the model's latent space stabilized consistently across different KL weights ($\beta = 0.1$ and $\beta = 0.05$). This shows that the model is well-regularized and that slight changes in the weight of the KL divergence term do not drastically impact the learned latent space. The latent space regularization is effective enough to prevent overfitting while maintaining sufficient capacity for waveform diversity.

5.1.3 Evaluation of Overall Performance

The convergence of both the total loss and individual loss components (brightness loss, MSE, and KL divergence) confirms the model's ability to balance multiple objectives—reconstruction accuracy, latent space regularization, and brightness prediction. The similarity in the results across configurations suggests that the model is not overly sensitive to changes in loss weighting, which further supports its robustness.

However, the lack of variation between configurations could also indicate that certain parameters, such as the brightness weight, may not need as much emphasis in future experiments. This opens the possibility for simplifying the model by reducing the complexity of the loss function or adjusting other hyperparameters, such as the latent dimension or learning rate, to explore more diverse behaviors.

Key observations include:

- The CVAE consistently generated waveforms that aligned with the desired brightness while maintaining a stable latent space, as indicated by the steady KL divergence.
- Increasing the brightness weight did not significantly alter the brightness prediction performance, suggesting that brightness is an inherently learnable feature of the input waveforms.
- The model performed well across different configurations, demonstrating that it is robust and capable of maintaining good performance regardless of variations in loss weighting.

These findings suggest that the CVAE could be an effective tool for wavetable generation, as it provides reliable performance without requiring extensive hyperparameter tuning.

5.2 Computational Complexity

One of the strengths of the current CVAE model is its efficiency in generating new waveforms. The model was able to generate waveforms in less than a second, making it highly suitable for real-time sound design applications. This rapid generation time ensures that the model can be integrated into music production workflows without significant delays, allowing for near-instant feedback when adjusting parameters such as brightness.

However, while the waveform generation process is fast, the training phase still requires substantial computational resources, especially when working with large datasets of waveforms or fine-tuning model parameters. As a result, training the CVAE model on standard hardware may still be a limiting factor for some users without access to high-performance computing systems. Future work could focus on optimizing the training process, for instance, by reducing the dataset size or leveraging more efficient model architectures, while maintaining the high speed of waveform generation.

5.3 Limitations

Despite the promising results, several limitations of the current approach should be acknowledged. These limitations highlight areas for future improvement and suggest possible extensions of the work.

5.3.1 Lack of Sensitivity to Hyperparameters

While the consistency in performance across configurations is a strength, it also suggests that the model may not be particularly sensitive to the weighting of different loss terms. This could mean that the model is overly dependent on the reconstruction loss (MSE) and less influenced by the brightness prediction or KL divergence terms. Further investigation into the impact of these loss components is necessary to ensure that the model is capturing a broad range of sonic features, rather than focusing predominantly on waveform reconstruction.

5.3.2 Dataset and Generalization

One of the primary limitations is the scope of the dataset used for training the CVAE. While the dataset was curated to include a variety of single-cycle waveforms, it may not fully capture the diversity of waveforms encountered in real-world sound design scenarios. The model's ability to generalize to new and unseen waveforms, particularly those with highly complex harmonic content or noise components, may be limited by the diversity of the training set.

5.3.3 Latent Space Exploration

While the KL divergence loss suggests that the latent space is well-regularized, further exploration of the latent space's structure could reveal additional insights. Evaluating how well the model interpolates between different waveforms and how diverse the generated sounds are would provide more information about the richness of the learned latent space.

5.4 Implications

The findings from this research have several important implications for the future of wavetable synthesis and the integration of AI into sound design workflows.

5.4.1 Democratizing Sound Design

One of the most significant implications of this work is the potential to democratize the sound design process. By allowing users to condition the generation of waveforms on high-level perceptual parameters, the CVAE model removes much of the complexity traditionally associated with creating custom wavetables. This could open up the world of sound design to a broader audience, including hobbyist producers who may not have the technical expertise to manually design waveforms.

5.4.2 Expanding Creative Possibilities

The ability to smoothly interpolate between waveforms in the latent space also presents new creative possibilities for musicians and sound designers. Producers could potentially explore novel sounds by navigating the latent space of the model, discovering new timbres that might not have been possible through traditional synthesis methods. This could lead to more experimental and personalized sound design approaches, pushing the boundaries of what is possible in digital music production.

Chapter 6

Conclusion and Future Work

This chapter concludes the thesis by summarizing the key findings of the research and discussing potential avenues for future work. The contributions made through the application of a Conditional Variational Autoencoder (CVAE) for generating sonically meaningful wavetables are highlighted, along with recommendations for future improvements and extensions.

6.1 Summary of Findings

The primary objective of this thesis was to explore the use of Conditional Variational Autoencoders (CVAEs) for automating the creation of wavetables in wavetable synthesis, conditioned on perceptually relevant sonic parameters such as brightness. The following are the key findings from this work:

- The CVAE model successfully generated a variety of novel and musically useful waveforms, conditioned on high-level parameters such as brightness. The ability to condition waveform generation on user-specified values enabled an intuitive and controllable interface for sound design.
- Despite testing several configurations with varying weights for the brightness loss, KL divergence, and reconstruction loss, the model's performance was robust and consistent across all configurations. This suggests that the model architecture is flexible and well-suited for waveform generation, with minimal sensitivity to hyperparameter adjustments.
- The CVAE's latent space provided smooth interpolation between waveforms, presenting creative possibilities for sound designers to explore new timbres by navigating the latent space. The stability and regularization of the latent space suggest that the model can generate diverse and sonically interesting waveforms.
- Several limitations were identified, particularly with regard to the diversity of the training data and the relatively minimal impact of varying certain

hyperparameters. While the model performed well overall, further refinement could enhance its precision in controlling specific sound qualities and improve its generalization to more complex or unseen waveforms.

Overall, this thesis demonstrated that AI-driven techniques, particularly the use of CVAEs, have strong potential to enhance the creative possibilities in wavetable synthesis, making sound design more accessible and personalized. The model's robustness and flexibility highlight its suitability for real-world applications in music production.

6.2 Future Work

While the CVAE model presented promising results, several areas for future research and development were identified throughout the thesis. These future directions are aimed at addressing the limitations of the current work and exploring new opportunities to further enhance AI-driven sound synthesis.

- **Expanding the Dataset:** A key area for improvement involves expanding the diversity of the training data. The current dataset, while effective, may not fully capture the range of waveforms encountered in professional sound design. Incorporating more complex or hybrid sounds, including waveforms with noise components or unconventional harmonic structures, could improve the model's ability to generalize to a broader range of sonic possibilities.
- **Exploring Latent Space Dynamics:** While the CVAE exhibited consistent performance in learning and regularizing the latent space, further exploration of the latent space could yield valuable insights. Future research could focus on analyzing how smoothly the latent space interpolates between waveforms and how diverse the generated outputs are. This could help in ensuring that the latent space is not only well-regularized but also meaningfully structured for creative sound exploration.
- **Refining Parameter Control:** Although the current model successfully conditioned on brightness, additional sonic characteristics such as warmth, sharpness, or harmonic richness could be incorporated to offer finer control over the generated waveforms. By introducing multi-dimensional parameter spaces, the model could allow for more granular conditioning inputs, providing users with a more comprehensive and intuitive set of controls for sound design.
- **Investigating Other Generative Models:** Future work could investigate the potential of combining different generative models to improve the

diversity and quality of the generated waveforms. For instance, combining Conditional Variational Autoencoders with Generative Adversarial Networks (GANs) could help generate higher-quality and more human pleasing waveforms. Additionally, exploring the use of noise as an additional timbral dimension using Diffusers could

- **Investigating Other Generative Models:** Future research could explore the combination of different generative models to further enhance the diversity, quality, and perceptual appeal of the generated waveforms. For example, integrating Conditional Variational Autoencoders (CVAEs) with Generative Adversarial Networks (GANs) could leverage the strengths of both models—using the CVAE for structured, controllable latent spaces and the GAN to refine waveform realism and produce more musically pleasing outputs. Additionally, investigating the use of Diffusion Models, which gradually refine noise into structured outputs, could introduce noise as an additional dimension for timbral control, enabling more sophisticated sound shaping and texture generation. This approach could unlock novel timbral characteristics, especially for more complex and evolving sounds.

- **User Interface and Accessibility:** While the CVAE demonstrates strong technical performance, its practical application in a user-friendly environment is still an open challenge. Developing graphical user interfaces (GUIs) or integrating the model into existing digital audio workstations (DAWs) as plugins would make it more accessible to a wider audience. By simplifying the interaction with the model, sound designers could benefit from its capabilities without needing deep technical knowledge.

- **Advanced Evaluation Techniques:** Expanding the evaluation methods to include more subjective and objective metrics could provide a more comprehensive understanding of the model's performance. Psychoacoustic measures, for example, could quantify how listeners perceive the quality of the generated waveforms, while more in-depth comparisons with human-designed wavetables could offer insights into the model's creative potential. Listening tests involving diverse audiences would also help validate the model's applicability in real-world sound design contexts.

6.3 Conclusion

In conclusion, this thesis has demonstrated that Conditional Variational Autoencoders (CVAEs) offer a powerful and flexible approach for automating wavetable generation. The model's consistent performance across configurations, combined with its ability to generate perceptually meaningful waveforms, highlights the potential for AI to transform sound design workflows. The CVAE enables musicians and producers to create custom wavetables without extensive technical

expertise, potentially democratizing sound design and opening new avenues for creative exploration.

However, the field of AI-driven sound synthesis is still evolving, and there are many opportunities for further innovation. By expanding the training dataset, refining parameter control, exploring new generative models, and enhancing the user experience, future work can build on the foundations established by this research. With continued advancements in AI, the future of sound design holds exciting possibilities for both professional and amateur creators alike.

Bibliography

[1] R. A. Moog, "Voltage-controlled electronic music modules," in *AES Convention 16*, no. 346, Audio Engineering Society. Trumansburg, NY: AES, October 1964.

[2] T. Pinch and F. Trocco, *Analog Days: The Invention and Impact of the Moog Synthesizer*. Harvard University Press, 2002.

[3] J. M. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *Journal of the Audio Engineering Society*, vol. 21, no. 7, pp. 526–534, 1973.

[4] T. Holmes, *Electronic and Experimental Music: Technology, Music, and Culture (4th ed.)*. Routledge, 2012.

[5] B. Frei, *Digital Sound Generation*, Baslerstrasse 30, CH-8048 Zürich, Switzerland, 2024, available online at http://www.icst.net.

[6] R. Bristow-Johnson, "Wavetable synthesis 101, a fundamental perspective," 01 1996.

[7] U. Andresen, "A new way in sound synthesis," in *AES Convention*, no. 1434. Hamburg, W-Germany: Palm Productions Germany, March 1979, available online: https://aes2.org/publications/elibrary-page/?id=2920.

[8] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Nsynth: Neural audio synthesis," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[9] J.-P. Briot, G. Hadjeres, and F. Pachet, "Deep learning techniques for music generation–a survey," *arXiv preprint arXiv:1709.01620*, 2017.

[10] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[11] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[12] P. Verma and J. O. Smith, "Neural representation of sound for transformation and synthesis," in *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2019.

[13] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[14] AIVA Technologies, "Aiva: Artificial intelligence virtual artist," 2020. [Online]. Available: https://www.aiva.ai/

[15] C. Hawthorne, A. Roberts, I. Simon, C. Raffel, and D. Eck, "Enabling factorized piano music modeling and generation with the maestro dataset," *arXiv preprint arXiv:1810.12247*, 2018.

[16] Z. Xiong, W. Wang, J. Yu, Y. Lin, and Z. Wang, "A comprehensive survey for evaluation methodologies of ai-generated music," 08 2023.

[17] J.-P. Briot and F. Pachet, "Deep learning for music generation: challenges and directions," *Neural Computing and Applications*, vol. 32, no. 4, pp. 981–993, 2020. [Online]. Available: https://doi.org/10.1007/s00521-018-3813-6

[18] Acoustical Society of America Standards Secretariat, *Acoustical Terminology ANSI S1.1–1994 (ASA 111-1994).* New York, USA: American National Standard, 1994.

[19] E. Schubert, J. Wolfe, and A. Tarnopolsky, "Spectral centroid and timbre in complex, multiple instrumental textures," in *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC)*, 2004, pp. 654–657.

[20] E. Schubert and J. Wolfe, "Does timbral brightness scale with frequency and spectral centroid?" *Acta Acustica united with Acustica*, vol. 92, pp. 820–825, 2006.