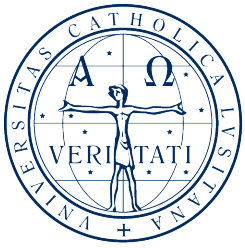


CATÓLICA
ESCOLA DAS ARTES

PORTO

MESTRADO EM SOM E IMAGEM

A reflection on the statistical method for trimbre synthesis

João DUARTE

UNIVERSIDADE CATÓLICA PORTUGUESA

MESTRADO EM SOM E IMAGEM

ESPECIALIDADE EM DESIGN DE SOM

A reflection on the statistical method for trimbre synthesis

Author:

João DUARTE

Supervisor:

Pedro PESTANA

Escola das Artes

January 10, 2022

Acknowledgements

Gostava de agradecer ao meu orientador, o professor Pedro Pestana, que durante este longo e tumultuoso processo sempre se mostrou compreensivo e disponível para ajudar e orientar.

Agradeço também a todas as pessoas que conheci e me acompanharam no meu mestrado, desde colegas de turma, a professores e colaboradores da Universidade Católica do Porto, que tornaram o primeiro ano muito especial. Em particular, ao Rafael Maia e Miguel Araújo, que estiveram comigo em todas.

A todos os meus amigos, por sempre me apoiarem e estarem presentes.

E à minha família, por tudo.

Dedicado aos meus pais.

Abstract

João DUARTE

A reflection on the statistical method for trimbre synthesis

Atualmente há uma lacuna no desenvolvimento de possiveis aplicações expressivas dos metodos desenvolvidos nas tecnicas de síntese natural.

Este trabalho teve como objetivo investigar os possiveis usos expressivos do método de síntese de Josh McDermott, que emprega uma metodologia de decomposição estatística e alcançou resultados promissores no campo da psicoacústica de texturas sonoras.

O estudo alterou o código computacional do método original de maneira testar a possibilidade de utilização do mesmo como um sintetizador de instrumentos e "morpher" entre samples instrumentais e texturas sonoras.

Os resultados demonstram que o metodo falha como um sintetizador instrumental devido a uma aparente incapacidade deste na desconstrução estatistica das qualidades harmônicas e tonais inerentes aos sons instrumentais, no entanto, este demonstrou bons resultados como um morpher com um alto grau de parametrização.

No seu estado atual, a técnica de síntese pode ser utilizada como uma ferramenta expressiva uma vez que traz qualidades distintas de texturas sonoras aos sons instrumentais.

Palavras-Chave: Síntese Expressiva, Síntese Estatística, Sintetizador de Instrumentos, Textura de Som, Timbre

Abstract

João DUARTE

A reflection on the statistical method for trimbre synthesis

There is relatively little effort placed into exploring potential expressive usage of natural synthesis techniques.

This research aimed to investigate possible expressive applications of Josh McDermott's natural synthesis method, which employs a statistical decomposition methodology and has shown promising findings in the field of sound texture psychoacoustics.

The study modified the computational code of the original method to examine if it could be used as an instrument synthesizer and morpher between instrumental samples and sound textures.

The findings demonstrate that the approach fails as an instrumental synthesizer due to an apparent inability to statistically deconstruct the harmonic and tonal qualities inherent in instrumental sounds, however, it performed well as a morpher with a high degree of parametrization.

In its current state, the synthesis technique can be used as an expressive tool, as it adds distinct qualities of sound textures to instrumental sounds.

Keywords: Expressive Synthesis, Statistical Synthesis, Instrument Synthesizer, Morpher, Sound Texture, Timbre

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The technique of generating sound using electronic hardware or software is known as sound synthesis. It can be used for a variety of purposes, including the replication of real-world acoustic events, the generation of unique sounds that cannot be produced acoustically or the advancement of system automation, such as text-to-speech software. It is a technique used widely in the arts, ranging from the electronic musical instruments known as synthesizers to the modern sound generation used in movies. Even though this method has been a prevalent since the genesis of modern western media (Corbella and Windisch, 2013), it is still evermore relevant nowadays (Okamoto et al., 2019).

However, due to the wide amount of practical uses and multidisciplinary nature of sound synthesis, in addition to the quick improvements with the help of modern computer technologies, and large number of methodologies and approaches employed, the potential of each sound synthesis advancement may not be explored to the full extent. For example, many sound synthesis methods used in movies and video games are of the natural variety, striving for realism and practicality, may not be considered for expressive uses of sound synthesis, such as in music.

This research aims to investigates various methods of computational texture synthesis, one of the most active areas of development and interest

in the sound and music computing landscape, with a particular focus on statistical synthesis.

This chapter will provide an introduction to the study by first discussing the background and context, followed by the research objectives and questions, contemplating on its significance and limitations.

1.1 Background

Art is (and has always been) intrinsically linked to and dependent on the stage of technical advancement. The artist's relationship and work are inextricably linked to the limitations of what can be understood and constructed. Of course, technological advancements bring with them new options for artistic expression.

Beginning in the 1990s, with the introduction of affordable computer technology, digital audio signal processing methods provided greater control over the creative process. Rather than working directly with the electrical signal, it was feasible to work mathematically with a digital representation of sound, allowing one to computationally construct what would be exceedingly difficult to reproduce by analogue means with relative ease. With today's knowledge and technology, digital sound signal processing enables a wide range of applications, from speech recognition to noise suppression to sound synthesis.

Sound synthesis, the process of artificially creating sounds through simulation or modelling, has piqued the interest of many in recent years because it not only stands on its own but also complements other artistic disciplines such as cinema, theatre, and video games, as well as aids in the development of new tools and instruments in music. The subject itself is interdisciplinary, embracing areas such as sound wave physics, computational signal processing, and sound psychoacoustics.

There are several kinds of sound synthesis methods. At the moment, procedural sound generation approaches are of particular interest, for example, for developing models of object sounds and natural events (Selfridge, Moffat, and Reiss, 2017b; Bahadoran et al., 2018), as well as for the synthesis of human speech. Synthesis techniques also often use physical modelling (Serra, 1993; Farnell, 2007) and corpus-based approaches (Schwarz, 2006; Einbond et al., 2021). The work of Josh McDermott (2009; 2011; 2013), which utilizes a succession of statistical analysis on a weighted implementation of cochlear-based spectral filters in sound, is likewise relevant and will be described in detail further in this dissertation.

Many cutting-edge methods are also incorporating the most recent machine and deep learning technologies. Machine learning is widely sought after due to its efficient approach of managing enormous volumes of data, wide range of applications, and potential scope for advancement. However, the disadvantages of these methods, such as generalization issues, learning biases, and limited control over the algorithm, are also present.

1.2 Research Overview

1.2.1 The Problem

Today's audio synthesis systems can successfully imitate natural sounds; for example, WaveNet can generate high-fidelity speech (Oord et al., 2018) and AutoFoley is able to generate believable foley-type audios when given a certain video as input (Ghose and Prevost, 2020).

There is no doubt that audio synthesis is well established in the artistic and creative fields as well: one only needs to open Digital WorkStation to have a plethora of synthesizers, effects, and VSTs at one's disposal.

Depending on the usage and purpose synthesis method, we can distinguished synthesis in two categories: natural and expressive (Schwarz, 2011). The following is the definition of both categories applied to sound textures, since they are going to be the of relevant in this study:

- Natural Texture Synthesis - it attempts to synthesize realistic textural sound as part of a wider soundscape. It focuses on realism and recognizability, drawing heavily on computer vision algorithms and having applications in the actual world of movies and video games. The majority of recent methods that have been developed are for this aim.
- Expressive Texture Synthesis - it aims to be able to separate created sound texture that is determined primarily by timbre rather than pitch or rhythm. The goal is to produce sound interactively for music creation and sound arts.

However, there appears to be very little link between the last two approaches outlined. Natural sound synthesis research rarely strives for or even explains possible expressive components in their outcomes. The common usage of deep learning on the current advancements, while achieving outstanding results for natural synthesis, also tends to create the "blackbox" problem, where there is little control over the design of such networks, which is usually desirable for expressive use.

Until recently, there has been no systematic investigation of the useable parameter space (Di Scipio, 1999) for expressive sound synthesis, with some newer works (H. v. Coler, 2019; H. v. Coler, 2021) attempting to introduce a more informed manner for this sort of application.

Since there is a considerable body of documented developments in natural sound synthesis methods, the lack of a bridge between them and expressive synthesis means a significant loss of potential for educated applications in the artistic domains.

1.2.2 Aims and Objectives

At a high theoretical level, the goal of this document is to examine and comprehend the majority of the audio synthesis systems now in use. However, because the range of sound synthesis methods is so vast, the choice was made to concentrate on a specific form of sound creation: sound textures synthesis.

Sound textures are defined as sounds that are produced by the superposition of several small scaled events, having stability on a greater time scale. Common examples include the sounds of fire, rain, and wind.

Among the existing methods, the one developed by Josh McDermott that employs a statistical filtering approach is of particular interest. The method has produced great results in the field of psychoacoustics, remarking on the nature of our perception of sound textures. From this hypothesis resulted the development of a sound synthesis technique that convincingly recreates the majority of traditional sound textures.

However, as previously said, there is very little research on the possible expressive usage of this technology. As a result, the main goal of this dissertation is to analyse the various possible expressive applications, highlighting their strengths and limits.

The research will look for any possible direct applications of the approach as a music instrument synthesizer, following a possible parallel between the concepts of sound texture and an instrument's timbre. If the method could achieve the recreation of an instrument's timbre independently of the pitch and rhythm can use, by definition, as a tool for expressive synthesis.

In addition, it will also look at more nuanced and supportive aspects in the synthesis role that it can take. This was performed by testing the possibility of transferring certain characteristics from sound texture to instrument samples through the manipulation of the methods variables.

Therefore, the dissertation's objectives are to comprehend and deconstruct the statistical synthesis process in order to understand and develop tools for testing the expressive alternative. In summary, it aims to answer the following questions, relating to Josh McDermott's method:

- Can this method be used, and if so how,as an instrument synthesizer?
- Can this method transform an instrument sound into sound texture in a controllable manner?

The methodology used an experimental approach, starting with an already existing method developed by the original developer as a foundation, then constructing a control data set and tweaking with its parameter inputs in order to test the questions stated above to the best of its abilities.

The investigation's usefulness lies in the production of an informed introduction to a possible new approach for expressive sound synthesis. As a consequence, it will also comment and expand in usage the method itself. Furthermore, by bridging a natural synthesis process to an expressive method, it will reflect on the approach to such a task and hopefully be used as a guide for future attempt.

1.2.3 Scope and Limitations

This approach has no intention of improving or expanding theoretically on an existing method. It will concentrate solely on the possibilities as an expressive synthesis tool, rather than the original study's aims.

This document will only address one method, a rather narrow scope in the larger topic of connecting natural and expressive methods. This was done to fully understand and explore the method's full potential given the resource constraints, with the understanding that this type of singular study

would produce better results than a broader attempt to connect several methods.

This investigation proved to necessitate of a significant amount of computer resources, and having access to more computational power would undoubtedly generate more results and, as a consequence, a more robust final data set. As such, the tests that were carried out were those that the author believed would yield the possibility of the most interesting results.

There is also the issue of generability. Because synthesis methods differ so greatly from one another, even when attempting the same tasks, it is not possible to fully transfer the knowledge and conclusions from this study to another attempting similar feats but employing a different synthesis method. It is, however, a credit to the broad and impressive body of work being done presently in sound synthesis, taking advantage of various disciplines of inquiry, from the ever extending list of computational approaches to new knowledge in psychoacoustics, that such generalization is not attainable.

1.3 Structural Outline

This dissertation is divided into seven chapters:

- Introduction
- Theoretical Concepts
- State of the art
- Methodology
- Results
- Discussion
- Conclusion

The second chapter contains the basic theory and analytical concepts for understanding computer-aided sound synthesis while also exploring the different properties of sound, from the physical phenomenona to its perception and digital representation. It also presents in detail the statistical sound synthesis method of interest, since information about its parameter and approach is essential for understand the investigation subsequent chapters. The objective is to introduce the necessary information to fully engage with further material.

The following chapter presents the literature review of sound synthesis. It looks at the various techniques to make sound, presenting them in chronological order, from the early methods of sound synthesis to the state of art approach for sound texture synthesis, while contextualizing the statistical synthesis method.

The fourth chapter covers the methodology. It outlines the methods that were employed to test your study's hypotheses, containing details on the procedures, data preparation, parametrization and result analysis.

The fifth chapter discusses the study's results in relation to each of the methods mentioned in the preceding chapter, as well as provide some brief comments on the data as it is presented.

In the sixth chapter, the results are reviewed and contextualized, making a more in-depth statement on what contribution the study makes to the literature as a whole, outlining the implications of your study's findings, identifying the limitations of the study and presenting your recommendations for future research.

The document ends in a brief chapter with the conclusion of the study.

Chapter 2

Theoretical Concepts

Sound plays out an integral part of most people's lives. The auditory system receives stimuli constantly, even when we are asleep. We can sense a large amount of information about our environment just by acknowledging with differences in air pressure, from discerning whether a train is approaching or leaving to speculating about the size of a room from an audio recording.

From speech to art, sound is an essential facet of our communication, thus it is no wonder that comprehending sound as a physical and psychoacoustic phenomenon is critical to the development and study of digital audio technologies.

2.1 Sound, physically

In physical sense, sound is a disturbance that takes the shape of a mechanical pressure wave that propagates across a material medium. A series of alternating compressions and rarefactions caused by a vibration propagating through a transmission medium like air or another gas, liquid, or solid. As illustrated in figure 2.1, sound is a longitudinal wave, which implies that the vibration happens in the same direction as the wave propagates across space.

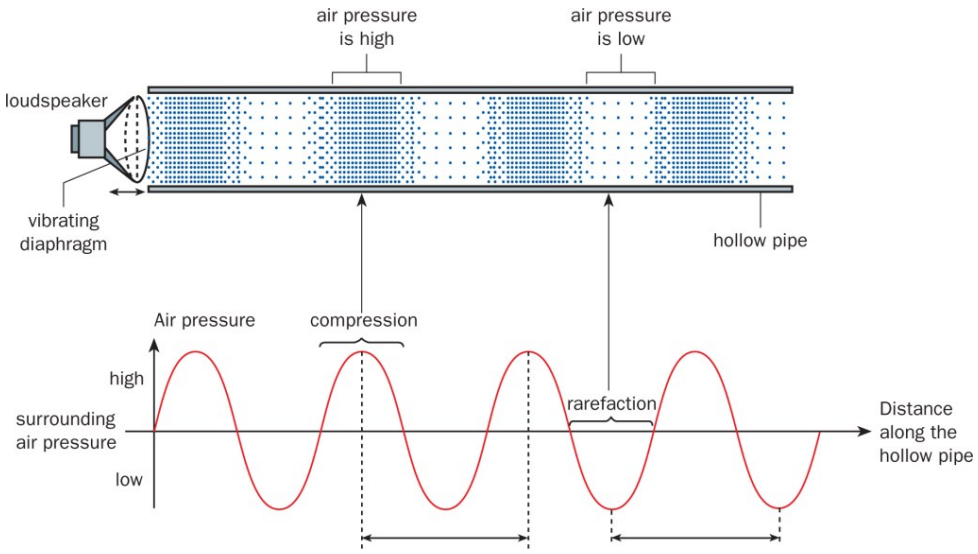


FIGURE 2.1: Visual depiction of compressions and rarefactions in a sound wave, as well as their parallel representation in a 2D pressure vs time graph, shown below. From *Sound 101* 2017

As a result, we may construct a representation of the wave through space using a simple two-dimensional graph, allowing us to readily identify two of its essential properties:

- The amplitude, A, is the highest local compression or rarefaction displacement from the average pressure value.
- The wavelength, *lambda*, is the physical distance between two consecutive corresponding points, that is, points in the same phase.

There will necessarily be a time delay, T, between two subsequent equivalent locations as the wave passes through a medium. This temporal measure is particularly significant because it determines the frequency, f, which is an important attribute in sound perception, namely the number of cycles of waves travelling past a location in space in one second.

$$f = 1/T \tag{2.1}$$

The wavelength-frequency connection is determined by the speed of sound in a medium, s. As a result, higher-frequency sound waves have shorter wavelengths (and periods).

$$s = f\lambda \tag{2.2}$$

It is relevant to note that the descriptions being used are oversimplified: Sound waves are not planar; they travel spherically across the medium and are often composed of a complicated mix of frequencies that fluctuate over time, but the fundamental qualities illustrated remain.

2.2 Digital Audio

To take use of the computing capabilities available, sound has to be converted into a digital representation of itself. The benefit of a digital version is that the computer can readily alter and manipulate the information, which in some cases would be difficult or impossible in an analog media.

To do this, sound must first be transformed into an analog signal, which is accomplished by devices, often microphones, that catch sound vibrations and convert them into a changeable electrical voltage. Analog signals consist of continuous range of values, as opposed to the computer world's discrete mathematical structure.

Computers function using binary information, which is stored in values of "1" and "0", therefore the information in sound must be translated into a binary language before it can be comprehended and processed. In contrast, digital information stored in computers must be translated into an analog language before it can be repeated through a speaker. As a result, there must be two types of conversions: analog to digital conversion (ADC) and digital to analog conversion (DAC), as seen in 2.2.

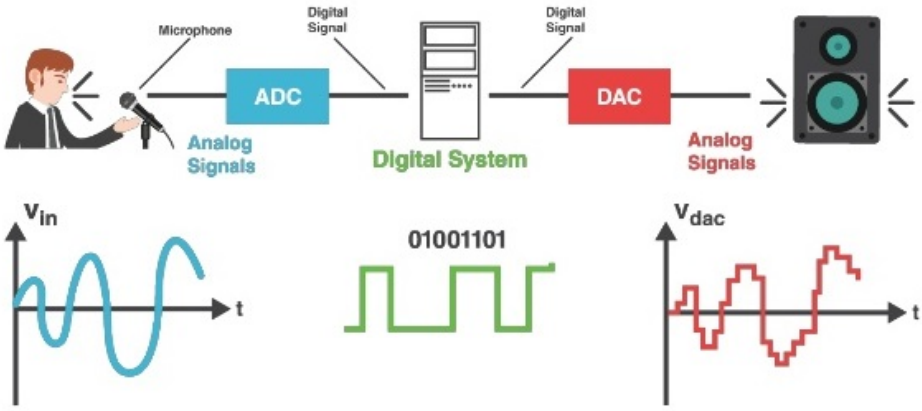


FIGURE 2.2: The capabilities of ADC and DAC, from sound capture to sound emission. From *0101000111 – Talk to the Computer* 2018

Analogue to Digital Conversion employs the sampling idea, which involves processing sound by measuring the amplitude (recorded as voltage) of a continuous stream at predetermined intervals. As seen in figure 2.3, the analog signal is split into discrete sets of values, samples, resulting in a binary representation of the fluctuation in voltage.

The sampling rate is the number of times the signal is split per second. The Nyquist-Shannon sampling theorem asserts that the sample rate must be at least two times greater than the higher frequency signal in order to adequately reproduce sound. As demonstrated in figure 2.4, the lower the sample rate in comparison to the maximum frequency, the more probable distortion will occur and information will be lost during conversion. This is known as aliasing distortion, and it happens

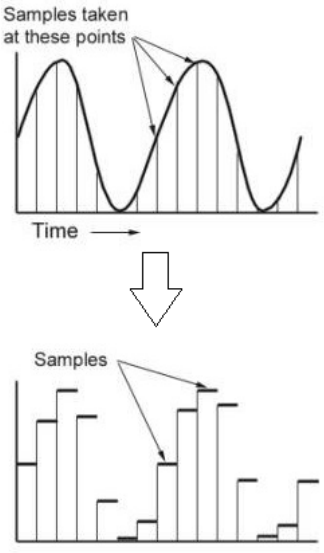


FIGURE 2.3: The sampling process. From *Digital Audio Basics: Audio Sample Rate and Bit Depth* 2021

when the sampling rate is insufficient
for the frequency of the original analog signal.

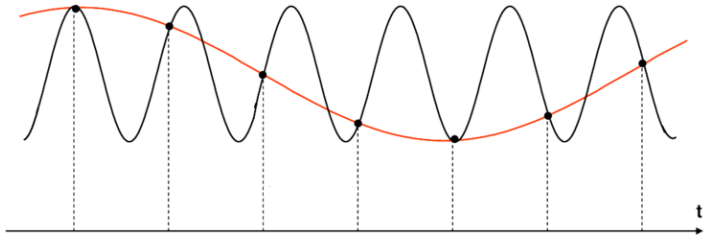


FIGURE 2.4: The aliasing effect. From *MT-002 TUTORIAL* 2016

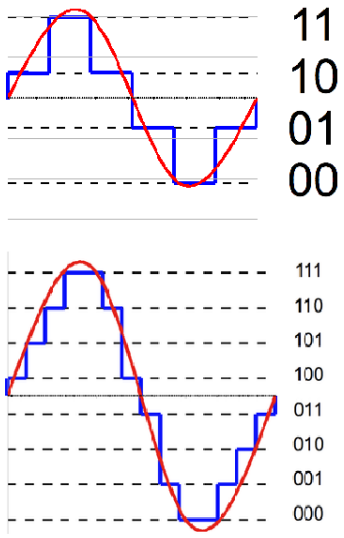


FIGURE 2.5: An example of amplitude quantization, with the difference between utilizing 2 bit and 3 bit. From *Difference Between Uniform and Nonuniform Quantization* 2018

The quantization of the amplitude values is another factor to consider during the ADC conversion. Each sample's amplitude is measured and assigned a value throughout the sampling process. As illustrated in figure 2.5, for each measurement, a range of values is allocated within which the amplitude value fits. The more extensive the range of values, the more precisely the connection between amplitude values is recorded, and the accompanying quantization noise is decreased.

2.3 Psychoacoustics

The study of the link between the physical qualities of sounds and how they are experienced is known as psychoacoustics. Sounds are primarily heard in the human body through the auditory system, which takes waves from the environment, interprets them, and transfers them to a higher brain level. Much of the information we can take up is included in the frequency relationship within the sound, allowing us to determine its character and quality.

2.3.1 Auditory System

As seen in the figure 2.6, the auditory system is split into three parts: the inner, middle, and outer ear.

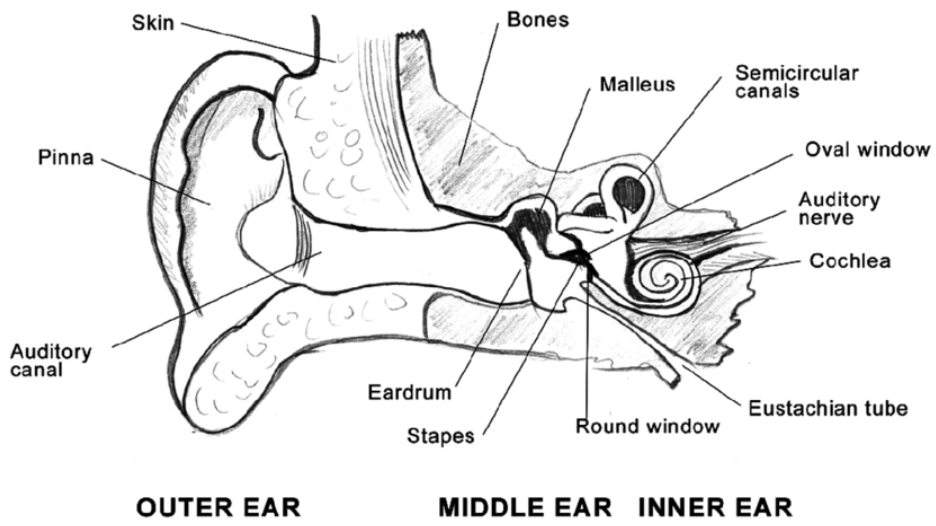


FIGURE 2.6: An Overview on the Auditory System. From *Unconventional 3D User Interfaces for Virtual Environments* 2021

- The outer ear, the external part of the auditory system, is further subdivided into:

- the pina, which serves as an important tool for sound capture and localization;

- the auditory canal, a structure that transfers energy through to the eardrum, amplifying certain frequencies (3kHz) and serving as a protector for the auditory system's inner structure (Fastl and Zwicker, 2007);

- The area between the eardrum and the oval window is referred to as the middle ear. The ossicles are made up of three microscopic bones: the malleus, incus, and stapes. This section's principal role is to match impedance as energy is transported through the middle ear via the ossicles and eventually turned into fluid movements in the inner ear (Moore, 2012).

- The inner ear is made up of two parts: the bone labyrinth (which houses the cochlea) and the membrane labyrinth. Its major function is to convert mechanical waves into electric impulses, which are then sent to the brain.

2.3.2 Auditory Transduction and Neuronal Encoding

The cochlea, a snail shell-shaped bonny structure filled with fluids, is where sound waves are converted into electric impulses (auditory transduction). As seen in the figure 2.7, its length is mostly made up of three canals, the scala vestibuli, media, and tympani, and two membranes, the basilar and Reiner's membrane.

The basilar membrane runs from end to end along the length of the cochlea. When pressure is applied on the membrane, it moves, as does the fluid in the neighboring chambers. The stiffness of the basilar membrane and

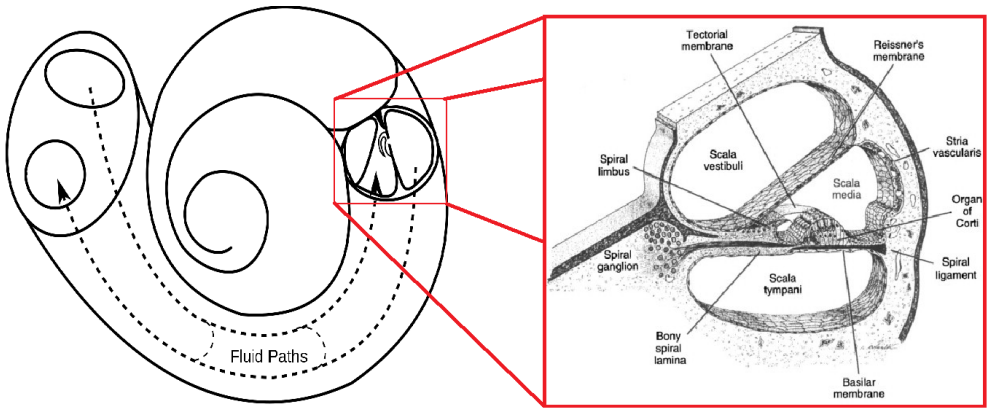


FIGURE 2.7: The cochlea. From *Cochlea From Wikipedia* 2010

the inertia of the surrounding fluids allow distinct portions of the cochlea to respond to various sound frequencies. High-frequency sounds are not ideal for moving fluid but can interact with the rigid basal end of the membrane, whereas low-frequency sounds can move fluid and interact with the softer edge at the apex, as seen in figure 2.7. This is referred to as basilar tuning.

The mechanical waves are converted into electrical impulses in the organ of Corti. The organ has microscopic protrusions from the surface known as hair cells, which come in two varieties: inner and outer hair cells. When inner hair vibrates, ion channels open, enabling positive ions to enter the cell, triggering the release of neurotransmitters to auditory nerve fibers that link to the brain.

The signal is deliberately transmitted to the primary auditory cortex, a section of the brain comprised of separate groupings of neurons that respond better to particular frequencies (similar to what occurs in the cochlea), allowing the signal's frequencies to be mapped and so discriminated in the brain. Tonotopy refers to the brain's spatial organization by frequencies.

Tonotopy generates insight for the use of filter banks or a Fourier Transform in sound encoders since it mimics the auditory system response

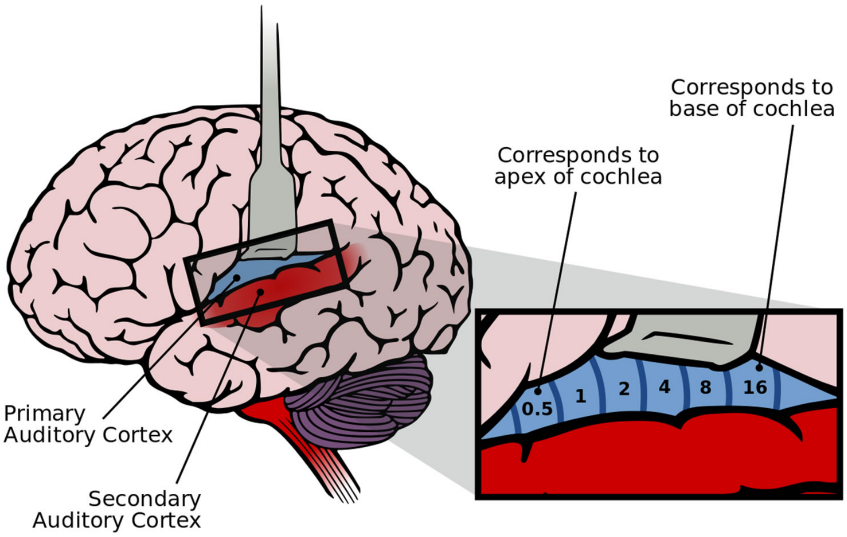


FIGURE 2.8: The Primary and Secondary Cortex, where neural encoding of sound if performed. From *Perception Space—The Final Frontier* 2009

to sound as a temporal event as well as its spectrum components in time (Lee et al., 2021).

2.3.3 Sound Perception

Loudness

The hearing system does not pick up on all vibrations; it is highly specialized to be sensitive to certain frequencies. The human ear can typically receive sounds with frequencies ranging from 20 to 20000 Hz, while lower frequency sounds can be perceived by touch, and the top limit decreases with age. Furthermore, not all frequencies are heard equally: as shown in figure 2.9, there is a non-linear connection between sound pressure level (SPL) and loudness, the subjective impression of the former, that varies with frequency and SPL itself.

Due to the previously described processes in the outer and middle ear, sensitivity is lowest at the limits of frequency perception and maximum

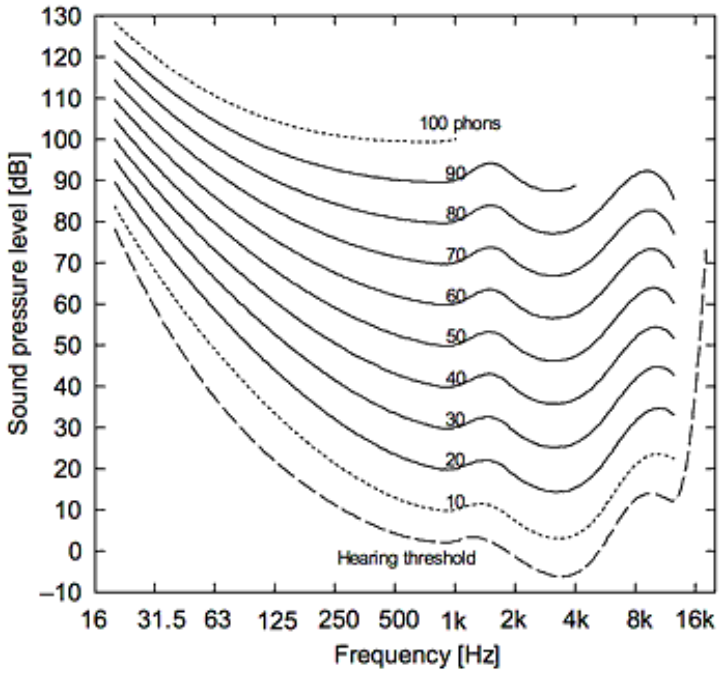


FIGURE 2.9: Fletcher–Munson curves, contours of equal loudness on an SPL versus frequency space. From *Equal-loudness-level contours for pure tones.* 2004

between 2 kHz and 5 kHz. As a result, the hearing threshold varies with frequency.

Critical Bands

The ear cannot indefinitely quantize the frequency range and identify every little fluctuation; this restricts how humans can discriminate between them, resulting in critical band bandwidths. The following is the connection between frequency, f, and bark, a constructed estimate for the distance between two successive critical bands, i.e., a frequency scale with equal value distance represents equal subject viewpoint difference:

$$bark = 13arctan((0.00076f) + 3.5arctan(\frac{f}{7500})^2) \tag{2.3}$$

2.4 Timbre

According to French composer Michel Chion (Chion, 2019):

"The emotional, physical, and aesthetic value of a sound is linked not only to the causal explanation we attribute to it but also to its own qualities of timbre and texture, to its own personal vibration. So just as directors and cinematographers (even those who will never make abstract films) have everything to gain by refining their knowledge of visual materials and textures, we can similarly benefit from disciplined attention to the inherent qualities of sounds."

Timbre is typically described in music as "the character or quality of a musical sound or voice that is unique from its pitch and intensity." While the principle is simple, the phrase has long been difficult, with meanings eluding a good explanation of how each sound source has its own unique "sonic fingerprint."

Timbre has been defined as the connection between the frequencies of a sound, most typically between the basic frequencies and overtones, as scientific and technical developments have transpired. Artists' interest and effort in exploring the potential of timbre expanded as their awareness of the phenomena grew.

Early twentieth-century composers, like as Arnold Schoenberg, were interested in researching timbre in compositions as part of their investigation of tonality, while futurists such as Luigi Russolo pushed the frontiers of what sorts of sound might be employed in music creation. However, due to technical restrictions, these artists were still severely constrained in what they could produce.

Since then, these limitations have been overcome, and modern methods allow for greater control over sound generation.

2.5 Sound Textures

Sound Textures are defined by the result of a superposition of several brief sound occurrences and are stable on a wider time scale. These occurrences follow a higher-level pattern that is revealed in a matter of seconds and might periodic and/or random (Schwarz, 2011). Rain, fire, wind, or crowd noises are examples of sound textures. These sounds are used in many types of art, from movies to sound installations to video games.

Strobl (Strobl, Eckel, and Rocchesso, 2006), divides common sound textures into five categories:

- Natural sounds, like fire, rain and wind
- Animal sounds, like crickets and humming
- Human utterances, like babble and chatter
- Machine sounds, like buzz, drone and traffic
- Activity sounds like rasp, rub and walking

However, there are still some disagreements over the entire definition of sound texture, which has resulted in the creation of various subcategories within the concept (McDermott and Simoncelli, 2011).

2.6 Statistical Sound Synthesis

In the late 2000's, Josh McDermott introduced a study (McDermott, Oxenham, and Simoncelli, 2009) that explored the hypothesis that the human auditory system detects sounds using statistical correlations. The major goal of the study was to learn how information is processed in the auditory system during the intermediate phases of hearing, as the beginning (peripheral

processing) and end stages (perceptual decision) are well established. The sound texture were employed for this investigation since kind of noise had the benefit of being rich enough to be clearly identified and distinguished. Simultaneously, it possessed a temporal homogeneity that minimized the complexity of the noise by focusing on texture rather than temporal aspects, making it well-suited to time-averaged statistics.

To test the idea, a synthetic approach was created that pulls from computer vision concepts that textures may be recognized by their individual statistics, implying that sound textures with identical statistical values should sound similar. This also implies that imposing a set of statistical values on an existing sound should result in it sounding similar to the sound used to construct such a set, which is the method's primary notion.

2.6.1 Statistical Data Set

The synthesis approach uses a sequence of steps similar to the biological auditory system to generate the first statistical data set, which has been demonstrated to give superior outcomes to other methods. It aims to simulate sound decomposition in the ear and follows the processes described in 2.10:

- The original sound is initially processed with 30 bandpass cochlear filters uniformly dispersed on an ERB scale, a scale that approximates the bandwidths of the filters in human hearing (Moore and Glasberg, 2004), simulating cochlear decomposition.
- After that, each frequency band's sound is processed with amplitude envelopes (or low-pass filters). This permits information about the amplitude of the response to be maintained for each filter. These filters also take non-linear cochlear sound propagation into consideration.

- Each envelope is subsequently exposed to a bank of 20 bandpass modulation filters, which are comparable to cochlear filters but are tailored to information at considerably lower frequencies

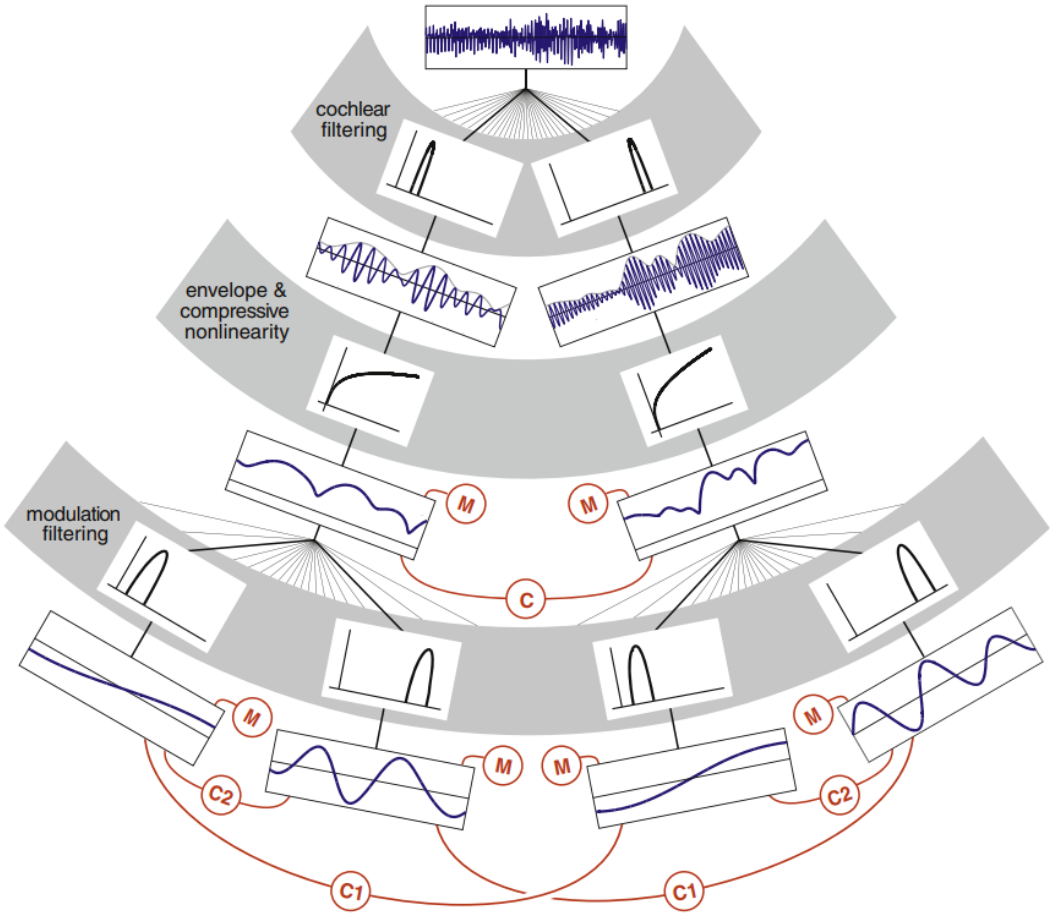


FIGURE 2.10: An overview of the numerous decompositions of the original sound that allow the necessary statistical metrics to be extracted. From McDermott and Simoncelli, 2011

2.6.2 The Statistics of Texture

The statistical classes used were the moments of the marginal distribution of the amplitude of each sub-band, M, and the pairwise correlations

between the Hilbert envelopes, C, which are important factors in the parallel study of computer vision and satisfied the condition of creating different data sets.

Moments

Moments are functions in mathematics that provide information on the shape of a function's graph. The following equation describes the moment m of the *nth* order:

$$\mu_n = \int_{\infty}^{\infty} (x - c)^n f(x) dx \tag{2.4}$$

These procedures are effective when applied to probability distribution functions, a function where x is the value of an event and $f(x)$ is the probability. In the image below (2.11), we can see how a basic plot of the evolution of sound pressure for three different files may be transformed into a probability histogram.

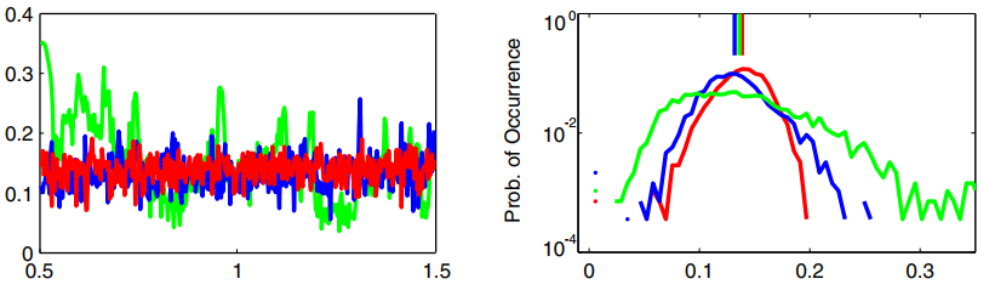


FIGURE 2.11: The sound pressure plot is on the left, and the histogram is on the right. From McDermott and Simoncelli, 2011

When we analyse the first crude moment (the one centred at zero, $c = 0$), we obtain $\mu_1 = \int x f(x) dx$, which is the total of all possible values weighted by their probability, i.e. the mean of the function for the original

function. When we centre the second moment on the previously computed mean, we obtain $mu_2 = \int (x - \mu_1)^2 f(x)dx$, which is the variance, which indicates how wide the original function's values are spread compared to its mean.

Following that, the third and fourth moments, when centred at the mean and suitably normalised by the deviation (σ), yield the skewness, $\mu_3 = \int \frac{(x-\mu_1)^3 f(x)}{\sigma_3 dx}$, the asymmetry of the curve relative to the mean, and the kurtosis, $\mu_4 = \int \frac{(x-\mu_1)^4 f(x)}{\sigma_4 dx}$, indicating the expression of a tail.

Although the moments in this approach are applied with minor alterations, it demonstrates that moments in principle are a strong means to appropriately define the structure of a distribution function.

The modulation band moments represent the fluctuation of the cochlear envelopes that distinguish rapid from slow modulated sounds while keeping part of the original sound's temporal information.

Pairwise Correlations

Marginal moments, on the other hand, are insufficient to provide a consistent representation of the sound because they only include limited time-averaged information within particular channels. It would be necessary to account for the connection between various channels. Correlation was a great tool for this since it shows a statistical link between variables. The following is the formula for a sample relation coefficient r:

$$r_{xy} = \frac{\sum(x_i - \mu_{1x})(y_i - \mu_{1y})}{\sigma_x \sigma_y} \tag{2.5}$$

There are three types of correlations taken into account:

- The correlation between distinct cochlear channels, C, provides for the

differentiation of events that occur with relationships to different channels, such as the sound of applause, from those that do not, such as the sound of water.

- Because modulation bands are used after cochlear filters, two sorts of relationships must be considered:

  - Between bands that are modulated at the same frequency but have distinct cochlear filters ($C1$). This helps us to distinguish between noises that are only linked in a subset of bands, such as wind, and those that are correlated in all bands, such as fire.

  - Those between bands used in the same cochlear filter but with distinct modulation frequencies ($C2$) enable for sound discrimination with abrupt onsets and/or offsets.

2.6.3 Synthesis Process

After obtaining the required statistics, the sound can be synthesized by superimposing these values over an original sound. This synthetic approach seeks to have a high degree of resemblance, i.e., most of the perceptually significant information will be kept, rather than to reproduce the sound, i.e., the created sound will be physically distinct from the original.

The sound to be altered (in this case, white Gaussian noise) is subjected to a systematic and iterative transformation to accomplish the synthetic transformation, as seen in figure 2.12.

To ensure that the values in a sub-band are accurately applied, the filters used to form that sub-band are applied again, which is not guaranteed if

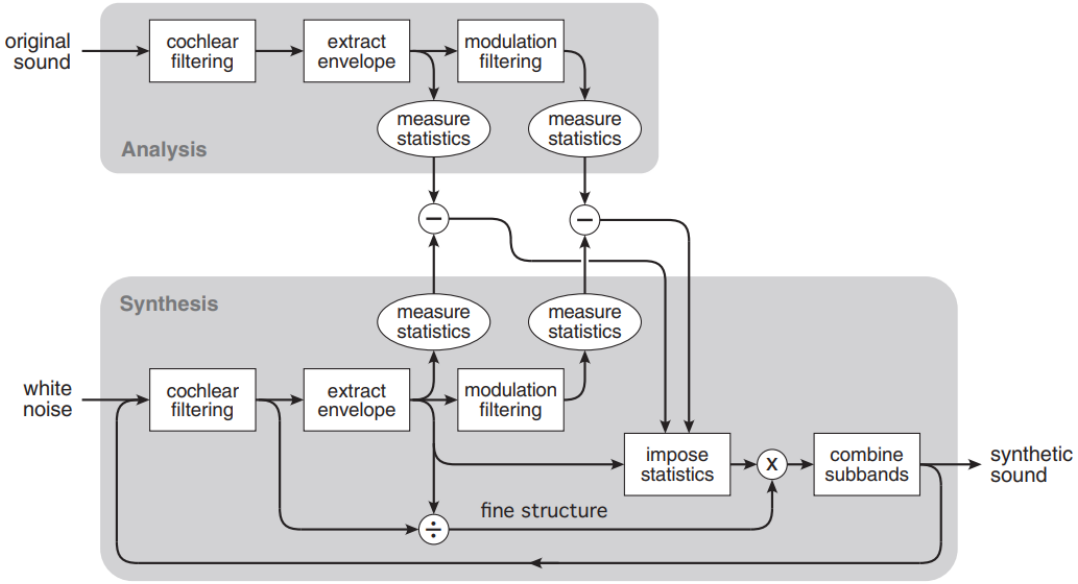


FIGURE 2.12: An overview over the synthesis architecture. From McDermott and Simoncelli, 2011

just once. However, because re-filtering and re-combining affect the respective statistical values of a sub-band, the procedure must be iterative in order to accurately impose the intended statistical values.

The gradient, r, is a mathematical quantity that forms a vector field whose vectors point in the direction with the quickest slope for a relative position when given a differentiable function. Each statistical value's gradient may be utilized to change its value in the direction that best matched the intended result. As a consequence, gradient descent was employed:

$$s_0 = s - lrf(s) \tag{2.6}$$

Where s and s_0 are the signals before and after the operation, respectively, l is a scalar step, and $f(s)$ is the statistical function for the signal. Gradient descent is an optimization process that is used to determine the values of a function's parameters that reduce a cost function as far as feasible, and in this case, it optimizes the similarity between sounds.

2.6.4 Conclusion

Despite the fact that this process appears to be somewhat limited, as it only relates to sound texture, it produces great results for its limited scope. It is possible to see that it obtains great degrees of resemblance between the pretended sound and the result in the reference studies (McDermott, Oxenham, and Simoncelli, 2009; McDermott and Simoncelli, 2011; McDermott, Schemitsch, and Simoncelli, 2013).

From all the marginal moments to all three types of correlations or even the iterative nature of the process, this approach has a multitude of variables, independent or semi-independent that make it promising to be examined in a parametrized fashion. It will be detailed in following chapters how this approach will be dismantled in order to fit a be test as an expressive synthesis method.

Chapter 3

Literature Review

The introduction of electrical devices in music allowed for more experimentation sound in music. In contrast to conventional instruments, which have not altered much in centuries of years, electronic instruments, such as synthesizers, have evolved virtually continually throughout recent history.

With today's easy access to synthesizers, interest in them is at an all-time high: the vast possibilities of these creative sonic tools have become an integral part of mainstream and underground art creation.

3.1 Brief History of Synthesizers

Synthesizers generates sound from electronic hardware or digital circuitry without the aid of any vibration of mechanical nature.

The advent of the vacuum tube amplifier in the early 1990s brought with it the creation of plethra of electronic instruments. The telharmonium, an early electrical organ developed by Thaddeus Cahill in 1896 and displayed in the photo 3.1, is commonly credited with being the first synthesizer.

Synthesizers began to take on the shape of the instruments we know

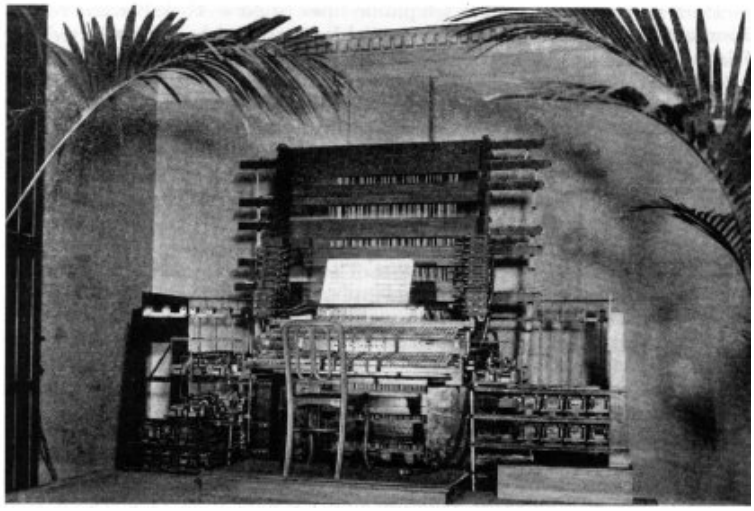


FIGURE 3.1: Teleharmonium

today in the 1950s. The RCA's Mark II, depicted in photo 3.2, was the first electronic sound production system with automated oscillators and modules attached to it, and was constructed and finished in 1957 by engineers Herbert Belar and Harry Olson.

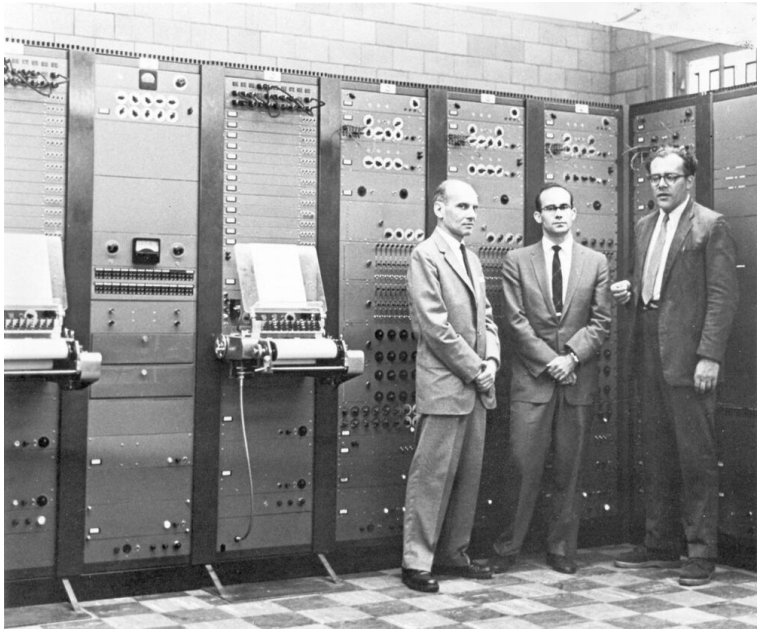


FIGURE 3.2: RCA's Mark II

Synthesizers gained popularity in the 1960s. The transistor, which eventually replaced vacuum tubes, was certainly a significant component in

the development of commercial synthesizers. Harold Bode, a German engineer, developed a modular concept synthesizer utilizing transistors, in which different pieces could be combined to form sounds. The idea was taken up by American engineer Robert Moog (shown in picture 3.3) and experimental composer Herbert Deutsch, who developed the first commercial synthesizer, the Moog Modular Synthesizer.

FIGURE 3.3: Bob Moog

Synthesizers began to embrace the new digital revolution in the late 1970s, with the arrival of the microprocessor. Digital synthesizers use digital signal processing techniques to simulate analogue sounds, as opposed to analogue synthesizers, which make music using true analogue equipment. This made synthesizers more accessible to the general public, and it was at this point that they began to play a larger part in the music industry, being at the forefront of then-new music genres such as disco and new wave.

Software synthesizers are becoming pretty prevalent. They make full use of computer capabilities to synthesize digital audio thought software, allowing them to perform tasks that formerly needed specialist hardware.

3.2 Early Methods of Sound Synthesis

Robert Moog's and Donald Buchla's discoveries in the 1960s radically changed the sonic landscape. With one being from the east coast and other the west, respectively, their synthesizers created two distinct frameworks for the philosophical and technological growth that are seen in present synthesizer theory.

The East Coast technique focused efficiency and pragmatism, centered on the standard concept of subtractive synthesis, while the West Coast was focused on non-traditional and experimental ideas of manipulating recorded sounds, embracing the use of additive synthesis, wave shaping and frequency modulation synthesis (Tony J. Rivas, 2016).

3.2.1 Subtractive Synthesis

As the name implies, subtractive synthesis generates new sounds by removing unwanted information. This approach starts with rich sound sources like noise and pulse waveforms, and then removes the unwanted spectral content to produce the output signal.

The most frequent way architecture in early analog synthesizers followed an oscillator-filter-amplifier design, and it is still employed in many current digital synthesizers.

The technique is commonly made up of the following modules, as seen in figure 3.4:

- Oscillators create a signal from a selected waveform, which is generally spectrally rich and has a wide dynamic and frequency range.

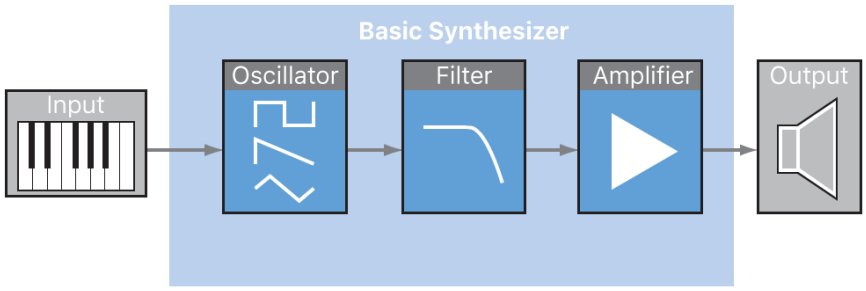


FIGURE 3.4: Subtractive Synthesis Architecture. From *The Fundamentals of Subtractive Synthesis.* 2018

- Filters are used to selectively eliminate information from the frequency spectrum, which is where subtraction happens.
- The term Amplifier refers to the device used to adjust the level of sound. Envelopes with a controlled attack, decay, sustain, and release can likewise be used to alter the sound in this case.

In the East Coast approach, these synthesizers were usually paired with ADSR envelopes, which control certain aspects of each signal's transience, such as the attack, decay, sustain, and release.

The West Coast method employed filters in a unique manner. Instead of filters to eliminate information, it employed waveshapers, which were devices that used mathematical functions to change the shape of the waveform. It also embraced the use of low pass filters, which worked as amplifiers.

3.2.2 Additive Synthesis

The earliest kind of synthesis, additive synthesis, creates sound by layering simple waveforms together to create a more complex timbre. Because the component waves are usually basic, such as sinusoidal, it is a simple approach that may be modulated with a simple equation:

$$s(t) = \sum_{i=1}^{N} A_i(t) sin(\phi_i(t)) \tag{3.1}$$

The output signal is $s(t)$, the instantaneous amplitude is $A(t)$, the instant phase is $\phi_i(t)$, and the total number of oscillators is N. The primary disadvantage of this technology is that it requires a large number of synthesizers to provide complex enough sounds.

3.2.3 Frequency Modulation Synthesis

Frequency modulation synthesis employs a modulator on a carrier wave, which, as the name indicates, changes the frequency of the input sound. By quickly raising or reducing the carrier wave frequency, additional frequencies, called side-band frequencies, develop that are not part of the original oscillators.

$$FM(t) = A_c sin(2\pi f_c(t) - \frac{\Delta f}{f_m} cos(2\pi f_m(t))) \tag{3.2}$$

Where A_c and f_c are the amplitude and frequency of the carrier wave, respectively, and f_m is the modulator's frequency.

3.3 Sound Texture Synthesis Methods

The use of oscillators, synthesizers, and other analogue hardware as conventional tools for electronic music composers is fast being replaced by the use of digital approaches for producing electronic sounds. Digital circuitry and programming are not only more adaptable and accurate, but they are also much less expensive.

While conventional analogue approaches progressed and were eventually transferred to a digital approach, and then to computer software, the introduction of this means also allowed for the birth of previously unattainable synthesis methods.

There have been several earlier efforts on the subject of synthesizing sound textures (Schwarz, 2011). Here, they are divided into three groups in this groups: physical-model synthesis, wavelet synthesis, and granular/corpus-based synthesis.

3.3.1 Physical-Model Synthesis

Most synthesizers take a signal-based approach to sound reproduction, meaning they aim to recreate the output signal, with no regard for the behaviour of the acoustic event itself.

Picking a string on a guitar, for example, might produce varying results depending on whether the stroke occurred on a fixed or already moving string, closer or farther away from the neck, or the angle of contact between the pick and the string.

These techniques examine the physical processes that resulted in the sound texture and then use the results as a starting point for developing algorithms (Bilbao, 2009), thus, it is possible to analyse the physical process of a certain type of sound texture to then build high-fidelity synthesis.

The ability to generate sound textures in real time has now been achieved by the most recent synthesis models. Siegfried (Selfridge, Moffat, and Reiss, 2017a) constructed a real time physical model in which the geometries of the propeller involved are employed in sound synthesis, and he also built a model for the same effect but for the development of aeroacoustic sound effects, such as swings (Selfridge, Moffat, Avital, et al., 2018). Lui (Liu, Cheng, and Tong, 2019) presented a physically-based statistical simulation method for synthesizing realistic rain sound, which divides the sound into two phases, the initial impact and subsequent pulsation, to generate sound textures based on a specially constructed signal decomposition and reconstruction model.

Hsu (Hsu, 2019) developed a physically informed, abstract synthesis approach for real-time parametric synthesis of percussive sounds via loopback frequency modulation

Finite-Difference Time Domain techniques are used in some of the most exact physical models (Bilbao et al., 2019). The limitation that this kind of models is that they are highly specialized for a certain sound, thus a model must be created for each type of sounds to be synthesized.

Although automatic parameter estimation approaches have been established, some methods have also lately contemplated or exploited the benefits of supervised convolutional machine learning to enhance the outcomes (Gabrielli et al., 2017; Gan et al., 2020; Françoise, Schnell, and Bevilacqua, 2013; Hawley, Chatziiannou, and Morrison, 2020).

3.3.2 Wavelet Synthesis

Signals commonly show slowly fluctuating osculations that are punctuated by transients. This rapid transition is typically the most important element of data, specially perceptually.

An extra approach was required to be added to the conventional FFT decomposition of the signal, as it does not completely reflect sudden changes due to the representation of a mixture of sine wave functions that do not localize them in time and space. As a result, wavelets were introduced, which are quickly decaying oscillations with zero mean that persist for a short time. This is the preferred approach for current synthesis of time-dependent sounds such as speech (Kronland-Martinet, 1988; Koguchi and Sagayama, 2018; Al-Radhi et al., 2021; Salah Al-Radhi et al., 2021).

They are used to decompose a signal into a wavelet coefficient tree, which is then resampled by changing the order of the pathways along the tree structure. The inverse wavelet transform is then used by each route to

resynthesise a small bit of signal.

With the proper parameter set, synthesizers can produce environmental sound textures from a small set of simple sounds. Bruna (Bruna and Mallat, 2011) proposed a novel wavelet technique that improves texture parametrization while collecting high-order statistics while Kersten (Kersten, Purwins, et al., 2012) research intended to re-create the sound of fire crackling.

These techniques are inspired by computer vision methods, namely texture analysis, and they attempt to characterize temporal and hierarchical connections between different levels of the multi-level tree representation they utilize (Morala-Argüello, Barreiro, and Alegre, 2012). The statistical method described in the previous chapter is also inspired by the connection between computer vision analysis and sound texture generation. More recently, Lostanlen (Lostanlen and Hecker, 2019) developed a category of audio transformations stated in the domain of time–frequency scattering coefficients, an approach converges wavelets and deep convolutional networks.

In fact, recent approaches been using deep convolutional networks for signal classification to improve wavelet functionality such as introducing control over perceptual parameters such as roughness or vibration (Lostanlen and Mallat, 2016; Lostanlen and Hecker, 2019), generate new sounds from a database (Narváez and Percybrooks, 2020; Babaei and Geranmayeh, 2009) and even for sound qualification (Qian et al., 2019).

3.3.3 Granular and Corpus Based Synthesis

Granular synthesis is a method of processing audio via granulation, which involves splitting the original signal into small slices of audio ranging from one to one hundred milliseconds (Roads, 2004). These grains can then be played in a different sequence from the original, superimposing on each other to create sound textures. This method can be supplemented with

statistical analysis to reduce repetition.

Frojd (Fröjd and Horner, 2009) used many sound snippets clipped from and original sound texture then reorganized and cross-faded them into a new one. Fascinani (Fasciani, 2018) employed a method in which sound grains are processed in the frequency domain and merged at the spectral level, contributing to sound synthesis solely with their magnitude spectrum.

Schwarz (2004) presented a descriptor-driven, corpus-based approach to sound texture synthesizing. Corpus-based synthesis may be thought of as a progression of granular synthesis. It allows the construction of slice of audio by traversing through a space where each one is put according to its sonic character in terms of audio descriptors, such as loudness, sharpness, roughness or any high level- meta-data ascribed to them. He afterwards (Schwarz and Schnell, 2010) also presented a technique for modelling the descriptors that employs a histogram and a Gaussian mixture model. In a more recent method, Bitton (2020) combined generative neural networks with granular synthesis by replacing the audio descriptor basis with a probabilistic latent space learnt with a Variational Auto-Encoder.

3.3.4 Deep Learning

As can be observed from the previously described methods, the majority of cutting-edge approaches employ some form of machine or deep learning method to supplement or build new methods.

Machine learning is a branch of artificial intelligence and computer science that focuses on using data and algorithms to mimic how people learn, progressively improving its accuracy. Deep Learning is an subset of machine learning dealing with artificial neural networks, which are algorithms inspired by the structure and function of the brain.

Although Arthur Samuel coined the term machine learning in back

in 1959, there has been a resurgence in interest in deep learning. Although it originally gained popularity in image processing (Purwins et al., 2019), it was then applied to audio owing to its success in voice recognition (Hinton et al., 2012) and audio synthesis. Modern deep learning employ various strategies depending on the wanted results, mostly deep feed forward neural networks, convolutional neural networks, and long short-term memory.

Some methods focus on expanding these new concepts in use (Tian, Xu, and D. Li, 2019; Huzaifah and Wyse, 2020), such as Antognini (2019)'s introduction of two new names for techniques in neural networks based on matching the Gram matrix of feature activations in order to keep rhythm and uniqueness in the sounds created. For speech texture synthesis, Chorowski (2018) developed a proof-of-concept method based on an approximation inversion of the representation learnt by a speech recognition neural network and neuron activation statistics matching.

Others methods concentrate on incorporating these trends into tried-and-true methods, like the ones mentioned previously(Gan et al., 2020; Hawley, Chatziiannou, and Morrison, 2020; Lostanlen and Hecker, 2019; Lostanlen and Mallat, 2016; Narváez and Percybrooks, 2020; Babaei and Geranmayeh, 2009; Bitton, Esling, and Harada, 2020). In order to improve on McDermott's method, Carcalla used convolutional neural networks on compressed real-imaginary spectrograms, improving results (Caracalla and Roebel, 2019; Caracalla and Roebel, 2020).

With the significant advances in concurrent research of deep produced visual textures, and there is some cross compatibility it and audio texture synthesis (Md Shahrin and Wyse, 2020). The notion of style transfer to audio was the most recent resurrection from this compatibility. In images, the goal was to transfer the "style" of one image, like the post-impressionist style of a Van Gohg's painting, to another image, such as a photograph. The same

is being attempted in audio, with the goal of transferring whatever characterizes style in an audio, such as the evolution of spectral density in time, transient cadence or psychoacoustic metris, to another (Grinstein et al., 2018; Verma and J. O. Smith, 2018; Tomczak, Southall, and Hockman, 2018; Lin, n.d.). Notably, this approach differs from most other deep learning methods in that,due to its inherent nature, it pays special attention to the expressive aspects of its results.

Although most deep learning approaches produce the some of greatest results for audio synthesis in general, they have a disadvantage when it comes to artistic and creative applications. Their neural network functions are in practice "black boxes", in which high-level behaviour emerges from the intricate interactions of hundreds or millions of tiny computational units rather than being explicitly programmed (Wyse, 2019; J. Smith and Freeman, 2021), and this lost of control over the basic control over the creative process is not ideal to optimal levels of creativity (Kaufman and Sternberg, 2010).

3.4 Conclusion

Josh Mcdermott's statistical synthesis method (McDermott, Oxenham, and Simoncelli, 2009) is inspired by several of the approaches discussed above. Because statistical analysis is used to noise filtering, it uses a subtractive method as its foundation. It has remarkable resemblance to various wavelet approaches because to its closeness to computer vision methods and use of specific statistical mathematical analyses such as autocorrelations. Although it is not a physical model, it is a model-based method to sound synthesis in that it uses theoretical textures to construct the model. As previously indicated, more recent research use a deep learning strategy to enhance the method, however this is not suitable for a study of the system's expressive character.

As a result, it is crucial to understand the parallel studies and where they shine and where they stumble. The comparable model-based approach, for example, makes it a reliable method for sound texture generation, but it also limits its applicability to anything outside the model's parameters.

There was also no research done on the method's expressive potential. As a result, there is no clear path to follow in order to ensure the study's success. In fact, the development of a guideline for an expressive method is considered part of the research and hence part of the conversation.

Chapter 4

Methodology

The fundamental purpose of this dissertation, as stated in the introduction, was to analyse the possible expressive applications of McDermott's statistical sound texture synthesis method, highlighting their strengths and limitations. The study attempted to postulate the following applications:

- Use this synthesis method to create an instrument synthesizer. If the technique could capture the timbre of the instrument regardless of pitch or rhythm, it could be employed as a tool for expressive synthesis.
- Examine this method's capacity to gradually morph an instrumental sound into a sound texture. If this were conceivable, with the statistical methodology of forcing certain variables into sounds, it would be possible to assess how the process transfers sound characteristics from one sound to another and whether they can be selected.

The chapter will begin with an introduction to the research philosophy, followed by a detailed discussion of the study design, including data preparation, method parametrization, and the approach utilized to answer to both postulates. It also discusses the method's shortcomings towards the end.

The code that was used to run the tests, as well as some representative results, may be found at the following link.

4.1 Research Philosophy

A duality in the nature of this investigation must first be acknowledged. The goal of this study was to look at the expressive application, which is at least partially qualitative in character, of a natural synthesis technique, which is primarily a quantitative approach to sound synthesis.

Due to the general underlying subjective character of the study's application, taking a completely objective look at data without regard for the observer's interpretation would have been a reductive take on the results. As a consequence, a positivist approach with interventionist considerations was used, i.e., the methodology constructed mainly to attempted to test the objective metrics that could be extracted while also taking into account there is a subsequent subjective interpretations on the results.

For this objective, a deductive experimental technique was employed. Starting with a control sample, a set of tests were performed in order to provide a comparison for subsequent outcomes. The high number of the method's inputs allowed for manipulation of a specific variable or collection of factors, and the resulting results were compared with the control group, allowing for the testing of causality between those variables. To complement the subjective nature of the results, the objective data will be a presented with a individual subjective comments.

4.2 Research Design

The first step in testing the technique was to get access to the method itself. Fortunately, it is publicly available on the website of the Massachusetts Institute of Technology (McDermott, Josh, 2013). The code is written in MATLAB and comprises the whole process of the synthesis technique, as well as the ability to adjust the various parameters. Additional controller scripts

were written that did not interfere with the synthesis process but made the desired parametrization more accessible for this tests.

The synthesis method with revised naming for the sake of clarity is shown in figure 4.1.

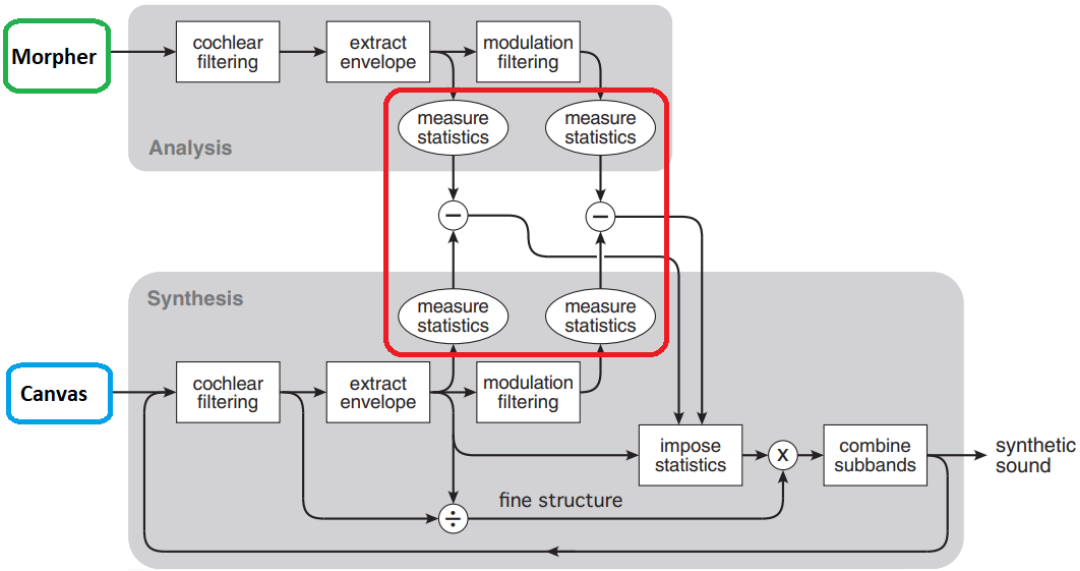


FIGURE 4.1: Synthesis process with revised naming.

For sake of clarity, this following terms are going to be used to distinguish between the input sound of the method:

- The Morpher, the input sound texture used to collect statistical data for imposing on the Canvas (shown in green in the image 4.1);
- The Canvas, in input sound that is utilized for the statistical data that is gathered and to be imposed on (shown in blue in the image 4.1);

To test the method, there was a need to have control over the following variables:

- Both input sounds, Canvas and Morpher. This would offer a straightforward way to test the instrument synthesizer hypothesis by using instruments samples as Morphers, as well as test how the method operates with the instrument as Canvas, to test the hypothesis of the incremental morphing expressive synthesis approach;
- The mathematical variables of the statistical synthesis process, i.e. all marginal moments and correlations. While control over them is not strictly essential for the instrument synthesizer, it is necessary to evaluate how each variable effects the sound by forcing or skipping the statistical imposition process while employing incremental morphing expressive synthesis (in red in the image 4.1);
- The number of iterations, in order to figure out how control one has over the transformation process over each iteration cycle (in red in the image 4.1).

Finally, being able to choose both the input sound and the metrics that are enforced, as well as the number of iterations, would provide the required tools for testing the desired applications of these tests. Since the Morpher sound is always the desired outcome in this procedure, the output file can also be rated using the signal to noise ratio (SNR) calculation, which is the metric used on the original studies as well. SNR is a scientific statistic that compares the level of a desired signal to the level of undesirable sound; the higher the SNR, the closer the output signal is to the Morpher sound. SNR may be applied to each parameter separately to observe how they all add up to the eventual outcome. It can be summarized in the following equation:

$$SNR = \frac{P_{Signal}}{P_{Noise}} \tag{4.1}$$

Where P_{Signal} is the power of the desired signal and P_{Noise} the undesired signal's power. In this study, it was calculated as the ratio of a statistic

class's squared error (summed across all statistics in the class), to the sum of that class's squared statistic values (McDermott and Simoncelli, 2011).

This measure can also be supplemented by a subjective sound interpretation. Although it was effective in the initial technique, since the tests are expanding the use of this method, it may or may not be adequate, hence subjective recognizability assessments should be supplied. When deconstructing the process, subjective perceptions like catching the turbulence of a sound may not be adequately reflected in the SNR score, therefore recognizability may not be the only potential good conclusion from the testing. This evaluation will be done by the writer and complemented when relevant.

It should be noted that the tests conducted used the most recent state of the synthesis process, which inherited all the previously outlined accomplishment possibilities as well as any potential existing limitations. The experiments were designed to investigate the possibilities and potential uses, not necessarily to improve on it, however conjecture about the findings and future applications will be discussed in further chapters. For a detailed information of the process, there is a detailed description on chapter 2.

4.2.1 Data Preparation

In order to use other sounds as inputs on the method, it was first necessary to define the method's bounds in order to guide the tests into its powerful suites. Since the technique works, by definition, with sound texture, it should be concerned with more time-redundant or atemporal features of the sounds rather than relevant time-related phenomena such as onsets and transients.

As a result, before starting with the synthesis testing, it was necessary to confirm that the sound sources used for the instruments satisfy the idea of sound texture as close as possible. The recordings of each instrument were

adjusted to maintain temporal homogeneity by removing time-related occurrences and compressing, stitching, and fading the original sound to make it constant in time as possible, as seen in the figure 4.2.

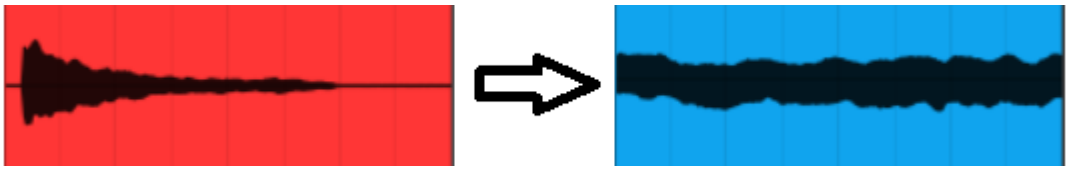


FIGURE 4.2: Before and after of the waveform of the piano sample used.

As for the selected instrument samples used, a theremin, a violin, a guitar, a piano, and a saxophone were chosen in order to provide recognizable timbres from a variety of sources. The theremin has a clear, crystalline tone in comparison to most other instruments, while the remaining selections provide a wide yet familiar variety of complex tones, switching between wind and string instruments, single and multiple note instruments.

4.2.2 Parametrization

This synthesis method, as mentioned in the previous chapter, is complex and includes several variables that can be accounted for; as such, it was important to focus on the most notorious and plausible variables that might greatly influence the result.

Because this process is iterative, one of the key considerations was to manage the number of iterations in each experiment, with the idea that managing the number of iterations would result in a sound with characteristics that were intermediate between the Morpher and the Canvas.

Another important aspect would be deciding which statistical measurements would be applied to the final result. For example, theoretically

the correlation indicators are known to aim to preserve specific statistical relationships inside the sounds, such as C2 preserving the sharp onsets and/or offsets on sounds, which means that deciding whether to impose or not may also grant or deny that specific characteristic on the final result (McDermott and Simoncelli, 2011). This would also allow for a more accurate understanding of the impact of each statistical moment, M, which may not be obvious at first glance.

The table 4.1 shows an overview of the parameters and variables that were controlled and adjusted between tests, and the way they were organized.

Parameters
Canvas
Morpher
Max number of iterations
Statistical metrics impositions

TABLE 4.1: Base table for with overview of parameters

"Statistical metrics" refers to the collection of statistical variables (moments, correlations) used in the synthesis technique; for example, if mean is the only one employed, it is the only one used from the Morpher sound that is forced on the morphed input, and also the only one used in SNR calculations.

4.2.3 Control Group Tests

To comprehend the outcomes of this experiment, it was helpful to first have a general understanding of the data that will be extracted and how those results appear in the reference conditions, i.e., using the default settings. Using this data set as a reference allows us to evaluate how the SNR evolves, knowing that the synthesized sound was consistently recognized as

the original to the extent of the original research McDermott and Simoncelli, 2011 . This stage also served to investigate how the amount of iteration affects the final outcome and whether any intermediate stages of transformation between Morpher and Canvas could be achieved. The parameters utilized were the same as in the previous table, with the exception of changing the number of iterations to recognize its impact, as indicated in 4.2.

Parameters	Inputs
Canvas	White noise
Morpher	Default sound textures
Max number of iterations	60
Statistical metrics impositions	Full set

TABLE 4.2: Control Group Set

White noise was the Canvas used in the original tests since it has a consistent spectral density across the whole spectrum. The default sound textures are the samples of sound textures that are provided when the method is downloaded; they include samples of bubbles, applauses, and pen writing, which provide a good variety in themselves containing different sound spectral densities and occurrence and duration of sharp offset. Sixty is also the default max number of iteration of the method. Full set means that all of the previously described variables, all correlations and statistical moments, are taken into account.

4.2.4 Instruments as Morphers

The concept of sound texture, a sound defined by properties that remain constant over time, and timbre, the character or quality of a musical sound or voice as distinct from its pitch and intensity, seem to have some parallel concepts between them, i.e., the timbre of a musical instrument may

have distinct statistical distributions over its spectrum that allow us to differentiate it from other instruments. If that is the case, there was a possibility that the synthesis method would be able to synthesizer instrumental timbre as well, enable the development of an it into a instrument synthesizer tool.

To test this, the instrumental samples refereed in the data preparation were used as the Morpher's input while keeping the white noise as the canvas to be transformed. Because the purpose was to investigate if it was possible to synthesize instrumental sound, the default maximum number of iterations was utilized, as well as experiments with more than 60, because there is yet no evidence that the approach would converge. The full set of metric impositions was designed to converge as many metrics as possible to assure the best chance of resemblance between output and intended sound.

The following table 4.5 summarizes the tests, highlighting the differences with the control group tests.

Parameters	Inputs
Canvas	White noise
Morpher	Selected Instrument Samples
Max number of iterations	60 | Higher
Statistical metrics impositions	Full set

TABLE 4.3: First Experiments, introducing the instrument sample as the Morpher sound.

4.2.5 Instruments as Canvas

The alternative strategy was to use the instrumental sound as the Canvas source while preserving the method's previously established sound textures as Morphers. In this synthesis method, it is known that the default sound texture samples can turn white noise into identifiable imitations of themselves. This implies that the properties that allow us to recognize those

sounds may be transmitted to the output file. While the instrumental samples differ statistically from the original canvas, white noise, due to the former's preference for certain frequency relationships and the latter's uniform distribution over them, there is no clear reason why the approach would be unable to do so. If it is possible, it would mean that it was successfully transferring those noticeable characteristics of the sound texture over the instrument, allowing for some possible expressive applications.

The following table 4.4 summarizes the tests, highlighting the differences with the control group tests.

Parameters	Inputs
Canvas	Selected Instrument Samples
Morpher	Default sound textures
Max number of iterations	60
Statistical metrics impositions	Full set

TABLE 4.4: First Experiments, introducing the instrument sample as the Morpher sound.

Since there was the possibility of testing how the process would behave with the usage of a subset of statistical metrics, a new set of experiments were idealized if the first ones were successful. In these, the objective was to test how each statistical metric influences the final sound, by forcing or bypassing the imposition of a preselected set of them over the output file. If the method is able to independently apply them, it allows to see how each metric influences sound and give greater control over the final product. Thus, the following alterations were done, shown on table 4.4 summarizes the tests.

Because the most recent tests were based on trial and error, they were updated when the findings were available, but the general framework remained unchanged.

Parameters	Inputs
Canvas	Selected Instrument Samples
Morpher	Default sound textures
Max number of iterations	60
Statistical metrics impositions	Subset

TABLE 4.5: First Experiments, introducing the instrument sample as the Morpher sound.

4.3 Limitations

During the planing of this work, there were some compromises that had to be done.

The data preparation process, which aims to get instrumental sounds as near to sound textures as feasible, is not a standardized procedure. Because no strategy for totally isolating the timbre of a note independent of the transient was discovered during the investigation, this was the way that seemed to make the most logical sense given the resources available.

There is also only one concrete objective metric for measuring expressive sound properties, and there are no validation procedures for subjective evaluation of the results. This was a difficult endeavour to do since there is no clear way to assess expressive character using a synthesis approach, which resulted in a paucity of extensive research on the subject. The original study's validation tests looked for recognizability, which is an objective observation of the sound, and that portion is already accounted for in the SNR connection, based on the assumption that the approach can already synthesize identifiable sound textures.

Chapter 5

Results

This chapter presents the findings of the experimental methodological study that was carried out to address the following research questions:

- Can this method be used, and if so how, as an instrument synthesizer?
- Can this method transform an instrument sound into sound texture in a controllable manner?

The primary goal of this dissertation was to examine the expressive applications of McDermott's statistical sound texture generation approach.

Recalling the previous discussed methodology, an experimental approach was built, with the ability to modify both the essential input variables to test: both input sounds, the maximum number of repeats, and the statistical metrics imposed on the final sound, based on the original synthesis MATLAB code. The code used to conduct the tests, as well as some typical results, are available at the following link.

The main analytical success metric employed was SNR, which was also used in the original study. The original research was selected for these tests because it met the goal of effectively replicating sound textures utilizing SNR as an internal variable that was used as a success criterion. This measurement will be displayed in the form of a graph, with a dedicated values corresponding to each statistical metric, generally with a correlation with

the number of iterations. There will also be observations using additional metrics when necessary, such as showing the statistical distribution development of specific elements, to better define the synthesis outcome. Only a selected representative graphs are provided for clarity; more can be found in the shared folder.

The findings of the control group test will be provided first. This information can be used to predict how successful testing will turn out. It will also provide an opportunity to delve further into each measure, with parallel graphs depicting the evolution of specific statistical metrics to supplement the single SNR number assigned to each iteration.

The first question tests will be shown in the second section, "Instruments as Mophers," where the instrument synthesizer technique was explored and the results were compared to the control group.

Finally, the results of the test addressing the second question, the morphing between instrument and sound texture, will be revealed in "Instruments as Canvas."

All of the procedures are described in further detail in the methodology section of the preceding chapter.

5.1 Control Set Results

The summary of the inputs for this section are displayed on the following table 5.1:

Let us consider the value 20 dB as a success reference point for this method, since it is the default threshold value specified in the synthesis method that terminates the procedure once it is reached. All specified measures should have an associated SNR higher threshold for this to happen. While

Parameters	Inputs
Canvas	White noise
Morpher	Default sound textures
Max number of iterations	60
Statistical metrics impositions	Full set

TABLE 5.1: Controll Group Set

the greater the value, the better, the default threshold value has shown to be sufficient for determining if the sound texture strategy worked; nonetheless, a subjective evaluation of the output file is always necessary, as this hypothetical limit may not be applicable to all other sounds.

With enough iteration, all of the synthesized sounds utilizing reference files converged, having a signal-to-noise ratio higher than 20. When charting the evolution of the SNR of each parameter with the number for each consecutive iteration, we may observe certain patterns, as shown in the image 5.1.

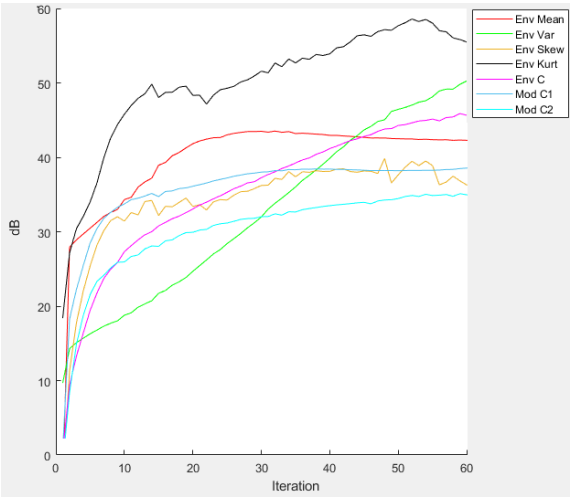
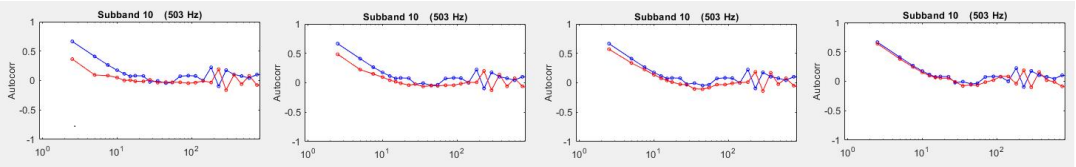


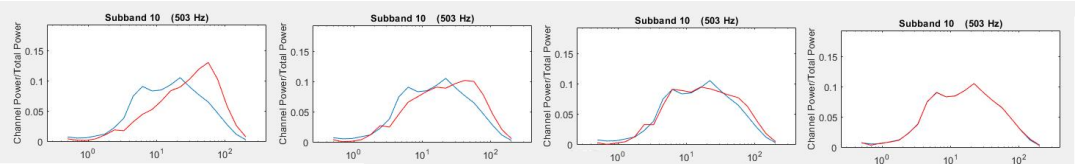
FIGURE 5.1: Plot illustrating the evolution of each parameter's SNR values, with statistical moments at the top and correlation factors below.

As can be seen, the first iteration has a considerable impact on the SRN value for some statistical measures, namely statistical moments, which rapidly increase their value in few iterations, while the correlation measurements, on the other hand, often take more repetition increase the similarity. The latter is shown when plotting the correlation and the modulation power, a measure that captures envelope correlations of the modulation bands through time, a may be seen in the figure 5.2. The original sound's statistics are represented in blue, whereas the synthesized sound

is represented in red. As can be seen, the red graphs become increasingly identical as the iterations progress.



(A) Envelope correlation for 1, 5, 10 and 60 iterations, from left to right respectively.



(B) Modulation Power for 1, 5, 10 and 60 iterations, from left to right respectively.

FIGURE 5.2: Plot of envelop correlation and modulation power with various amounts of iterations.

This is in contrast to the evolution of statistical moments while plotting the envelope histogram, as seen in figure 5.3. Despite the fact that the synthesized sound becomes increasingly close to the original, the first iteration's histogram already indicates a degree of similarity in shape, which is directly related to the statistical moments, as discussed in section 3.2 of this document.

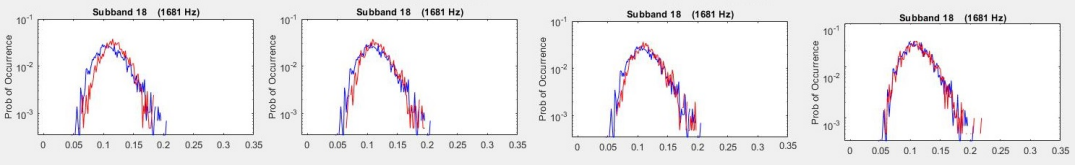


FIGURE 5.3: From left to right, histogram comparing the original and synthesised sound for 1, 5, 10, and 60 iterations, respectively.

5.2 Instruments as Morpher

The first experiments were performed by incorporating the instruments as Morpher sounds while keeping all other parameters the same as those used in the reference data set, as shown in the table 5.2

Parameters	Inputs
Canvas	White noise
Morpher	Selected Instrument Samples
Max number of iterations	60
Statistical metrics impositions	Full set

TABLE 5.2: First Experiments, introducing the instrument sample as the Morpher sound.

The ideal outcome would have been for the synthetic sounds to have converging SNR results, similar to the reference data values previously shown.

This was not observed, and all of the data from the instrument was replicated inaccurately. This may be shown in figure 5.4 by comparing the SNR evolution while using the violin and theremin samples as the Morphers sound with the values from the control group set.

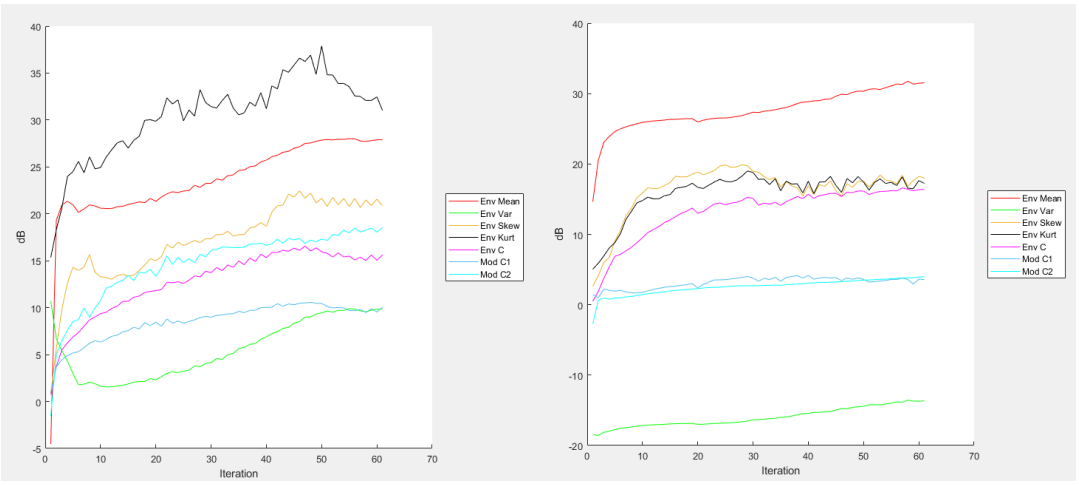


FIGURE 5.4: SNR vs number of iteration for the violin, on the left, and theremin, on the right.

While the SNR for statistical moments has a similar beginning behaviour, i.e., shown the strongest growth in the first iteration. Some moments, such as mean and skewness, even reach comparable values while other metrics, like variance, commonly had much lower values than the reference counterpart.

It could be hypothesised that, given enough repeats, the SNR will finally attain a value similar to the reference. However, the variance SNR does not achieve a reasonable value when employing a higher number of iterations, as demonstrated in the figure 5.5 which followed the tests from table 5.3, which represents the evolution of SNR for 240 iterations.

Parameters	Inputs
Canvas	White Noise
Morpher	Selected Instrument Samples
Max number of iterations	240
Statistical metrics impositions	Full set

TABLE 5.3: Test overview with increased number of iterations.

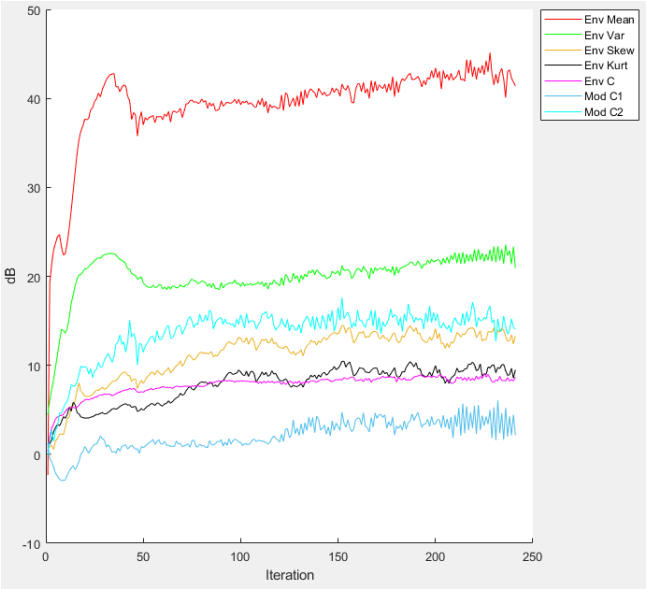


FIGURE 5.5: Plot illustrating the evolution of each parameter's SNR values for the piano sample, up to 240 iterations

To investigate if alterations on the Canvas and Morpher for this method would significantly generate better results, two further tests with slight alterations were be performed:

- Changing the Canvas sound: Since the initial studies employed white noise as the source, which compliments the spectral dispersion of the sound textures, sounds more resembling instruments could produce better outcomes.;
- Changing the Morpher sounds: providing the synthesis process an instrumental sound that covers and fills the spectral density more uniformly may produce better results;

A further discussion on the selection of this tests is presented on the discussion chapter.

Changing the Canvas

Since the instruments used for the samples play a single note, for example, C_4 for the theremin and D_5 for the piano and violin samples, those fundamental frequencies were used as inputs. As a result, the choice was to use a sound file with simply a sinusoidal frequency, 261.63 and 587.33 Hz, respectively, while keeping all other parameters the same as in the previous test, as shown in the table 5.4.

Parameters	Inputs
Canvas	Fixed Sinusoidal Frequency / Selected Instrument Samples
Morpher	Selected Instrument Samples
Max number of iterations	60
Statistical metrics impositions	Full set

TABLE 5.4: Tests further alternatives of the Morpher source on the instrument synthesizer tests.

All of the experiments, however, yielded similar results to the preceding ones. The synthetic sounds did not resemble the intended ones. This is demonstrated when visualizing the SRN data, as shown in the figure 5.6, which shows the results for synthesizing the piano sample with the D_5 sinusoidal as its canvas. Even the theremin sample, which timbre is closer to the pure frequency, did not give satisfactory results.

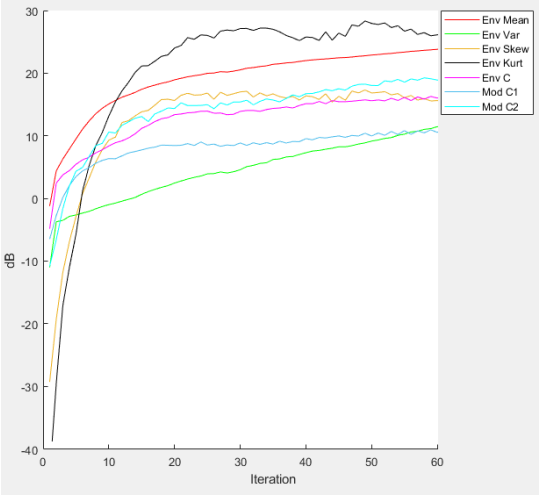


FIGURE 5.6: SNR evolution with each iteration for the piano sample as Morpher and pure sinusoidal as canvas.

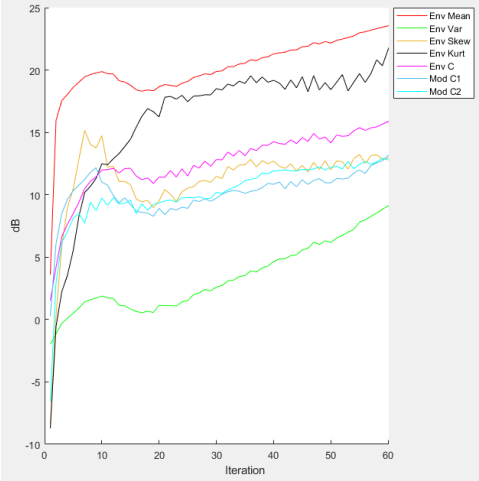


FIGURE 5.7: Results of synthesizing a piano sample from violin sample with the same base note

Considering a different "frequency-focused" sound would be interesting in order to have a more robust answer on the potential properties of the input morphed sound for the synthesis of instruments, since the problem with using pure frequencies could stem from the lack of spectral distribution on the morphed source. As such, the violin and piano samples, which that have the same root note, were combined. However, the

results are similar to those obtained earlier, with the sound generated being substantially different from the original, but also, as seen in image 5.7, with SNRs for moments and correlation factors not being high enough.

Changing the Morpher

These tests investigated the procedure by imposing an even spectral distribution on the instrument samples. Two strategies were employed to put this to the test, using the following alterations on the piano sample:

- Make a single sample out of all piano notes performed at the same time and sustained for the entire period;
- Create a sample that is made up of multiple rapid and short piano sample notes that are played in a random order;

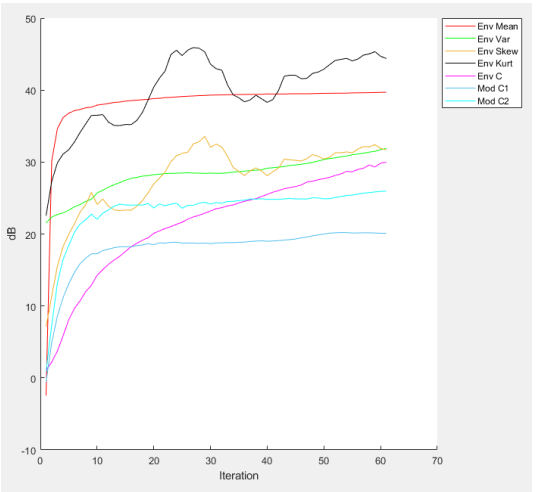


FIGURE 5.8: SNR evolution with each iteration while using an altered piano sample

Although the findings were not as positive as those from the reference sound source, both demonstrated a substantial improvement when compared to the original piano sample, with SNR values for statistical moments and correlation being significantly closer to the values from the reference sound source, as shown in figure 5.8. It is also feasible to identify some similarity of the tonality qualities of each one by listening to them, even though they were not competitively recreated and are not indistinguishable from the original files.

5.3 Instruments as Canvas

This section shows the results from utilizing the instrument samples as a canvas template to imprint the sound qualities from the already established sound textures, as shown in table 5.5.

Parameters	Inputs
Canvas	Selected Instrument Samples
Morpher	Default sound textures
Max number of iterations	60
Statistical metrics impositions	Full set

TABLE 5.5: Changing the Canvas Source on the instrument synthesizer tests.

The first test aimed to expose the instrument recordings to the entire transformation into sound textures, using a complete set of statistical metrics and several iterations, to determine if the Morpher sound could be replicated in this manner. This time, the sounds were completely replicated using this method, with SNRs for all metrics exceeding the 20dB threshold and recreating the recognizable properties of the Morpher sound, as can be seen in figure 5.9.

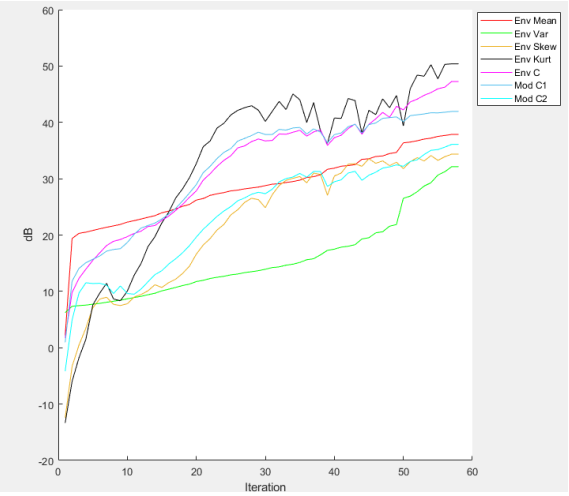


FIGURE 5.9: SNR evolution using an instrument sample as the Canvas source.

This raises the question of whether there are any discernible qualities from the morpher sound that are applied with each unique statistical metric. For this, a new set of tests were performed, following what is describable in table 5.6.

Parameters	Inputs
Canvas	Selected Instrument Samples
Morpher	Default sound textures
Max number of iterations	60
Statistical metrics impositions	SubSet

TABLE 5.6: Overview on the reduced number of statistical metrics impositions.

5.3.1 Marginal Moments

Mean

Since the statistical mean only indicates the average value for a specific frequency channel, it primarily reflects the channel's average sound power. As a result, when this measure is considered separately, the only difference in the result file is mostly a potential increase in sound power.

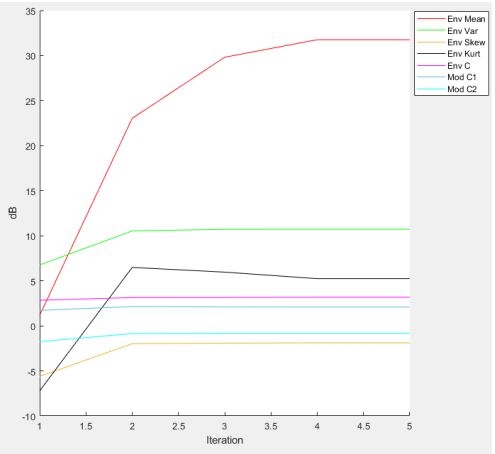


FIGURE 5.10: SNR evolution with iterations when only mean is imposed

As a result, converging mean values is a relatively fast process, achieving high SNR values with few iterations while also interfering slightly with other statistics, mostly the other moments, as shown in figure 5.10, but this could simply be the result of adding a spectrum filled noise to a frequency dedicated sound while attempting to converge to a less frequency focused sound.

Variance, Skewness and Kurtosis

As previously demonstrated, these parameters captured the breadth, asymmetry, and existence of a long positive tail on the histogram of a certain frequency channel. These moments indicates the pace at which the envelopes fluctuate, allowing us to distinguish between quickly and slowly modulated sounds.

Moments are theoretically not independent, but in the scope of this tests, applying the imposition on single moments barely altered the SNR value of the ones that were not selected. As it is demonstrated by graphs 5.10 and 5.11, even if barely, the lower the rank of the statistical moment individually imposed, the greater the effect it has on other succeeding moments.

All the moments were able to reach the threshold with few iterations, usually fewer than the ones needed when trying to converge all at once. They also showed to have distinct yet recognizable results between each other.

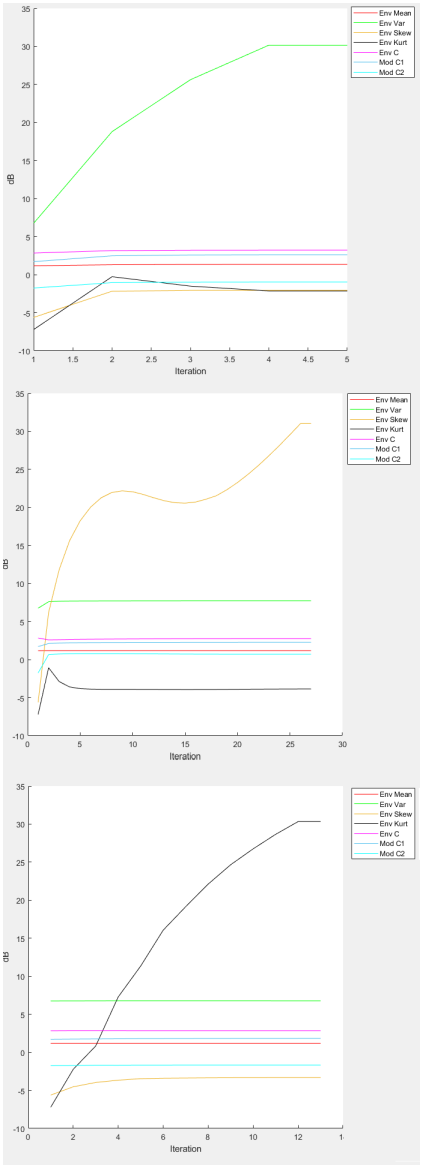


FIGURE 5.11: SNR evolution with iterations when only variance, skewness and kurtosis are imposed, respectively

5.3.2 Cochlear and Modulation Correlations

Cochlear correlations (C) may discern relationships between different cochlear bands. They can, for example, discriminate textures with broadband events that activate multiple channels at once, such as hitting drums, from those that create nearly independent channel responses, such as the previously mentioned water, allowing to capture dependencies over time and frequency.

This correlation is computed in tandem with modulation correlations: cross-channel modulation correlations, $C1$, computed on a specific modulation band of each cochlear channel, allowing to distinguish sound, such as wave (focused on low modulation-frequency bands) from fire (present across all bands), and within-channel modulation correlations, $C2$, allowing to distinguish sounds with sharp onsets or offsets.

The evolution of the SNR for all tests can be seen in the image 5.12. As it is possible to see, they tend to have a much higher influence on the other unselected metrics when compared to the marginal moments.

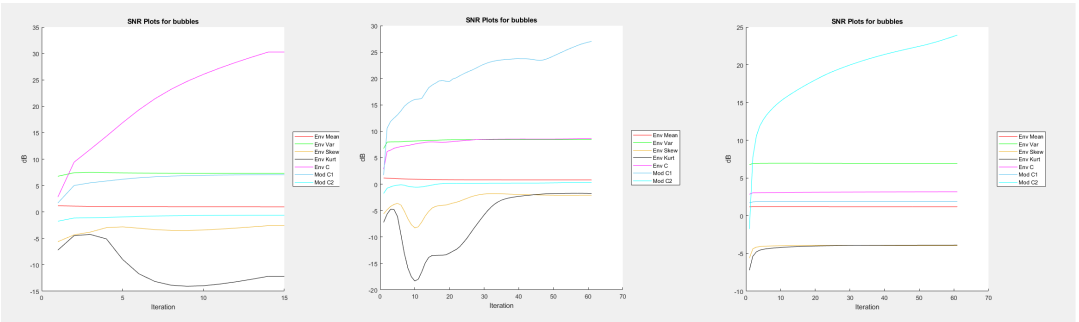


FIGURE 5.12: SNR evolution with iterations when only cochlear correlations, cross-channel modulation correlations and within-channel modulation correlations are imposed, respectively

Chapter 6

Discussion

The results and findings reported previously are discussed in this chapter.

It is worth to repeat the fundamental purpose of this dissertation. The aim was to investigate the expressive uses of McDermott's statistical sound texture generating method, with the following problems in mind:

- Can this method be used, and if so how, as an instrument synthesizer?
- Can this method transform an instrument sound into sound texture in a controllable manner?

The answer to the first question is negative, indicating that the method fails in its direct use as an instrument synthesizer. However, the results of the tests conducted to answer the second question were much more positive, indicating that the method succeeds as a morpher between instrument sounds and sound textures.

6.1 Synthesis Method as Instrument Synthesizer

When utilized directly as an instrument synthesizer, the approach failed in terms of both quantitative statistics, such as SNR, and the recognizability

of the sounds it created. It is reasonable to say that the instrument's input sounds lack the necessary characteristics to be used as Morpher sounds in this manner.

This disparity may be due to the implied origin of sound textures, which contrasts sharply with instrument notes since textures are formed through the superposition of several sound sources, rather than discrete events like instruments.

While the idea of the similarity between textures and timbre at the beginning of this chapter may be viable, there is still some differences that may arise to cause the failure. For example, the superposition of events may provide additional necessary features for the original sound, most notably a sufficiently equal distribution of spectral density. The occurrence of various sound similar yet distinct sound events does not allow for the creation of clear tonalities in the sound. In contrast, the spectral distribution of instruments is more centred on a few fundamental and complementary frequencies, as shown in figure 6.1.

FIGURE 6.1: FFT of a sound texture, in green, compared to an instrumental sample, in red.

Two further tests were conducted to test this idea, the first using a tonal canvas sound source and the second with a morpher sound with a more uniform spectral density.

The first , which used pure frequencies and instrument samples that were comparable to the ones being transformed as canvas, yielded unsatisfactory results in terms of SNR, as shown in figure 5.6, and recognizability. This seams to indicate the method does not work well with tonal sounds as morpher inputs. In reality, when compared to other instruments, highly tonal samples, such as the theramin and saxophone samples, whose FFT is displayed in figure 6.2, where the majority of the sound capture was centred on single or few frequencies, obtained the lowest SNR for variation.

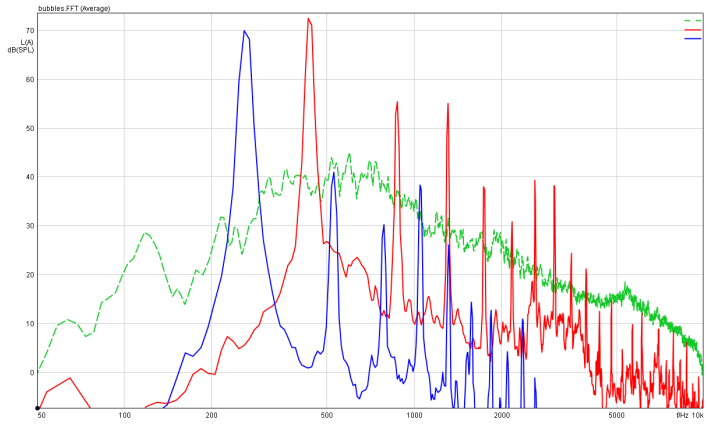


FIGURE 6.2: FFT of a sound texture, in green, compared to an instrumental sample, saxophone, in red, and theremin, in blue.

The results, on the other hand, appear to improve significantly when the Morpher sound is made less tonal, as seen in figure 5.8. The adjusted samples, which produced superior results, are significantly less tonal, as can be seen in figure 6.3.

The approach fails to recreate pitched noises, as stated in the original document (McDermott and Simoncelli, 2011), although it did not try to replicate altered instrumental samples. The original study does not come to any firm conclusions, stating that it is not evident that such sounds would cause synthesis failures because they all have stable spectral and temporal

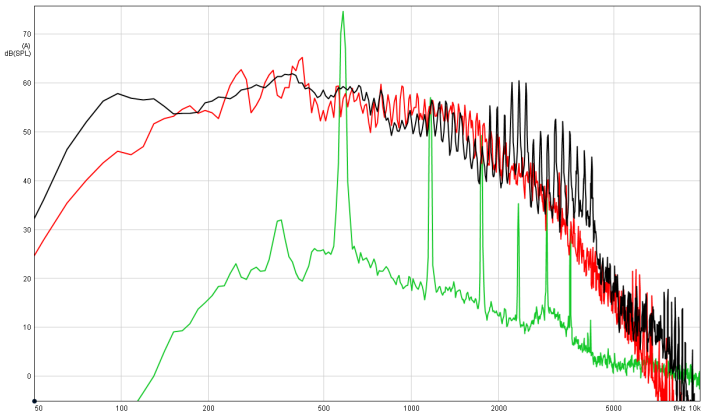


FIGURE 6.3: FFT of a the original piano sample, in green, the altered sample with multiple simultaneous notes, in red, and multiple non simultaneous notes, in black.

characteristics.

This seems to imply that pitch and the relationship of harmonics on instrument samples are not capture by the statistical measurements applied by the method.

This could mean one of two things: either the method is not forcing the right amount of correlations between harmonic information, such as the relationship between fundamental and complementary frequencies, that are far apart in the frequency spectrum, like it is discussed in (Caracalla and Roebel, 2019); or that the human hearing system uses a mechanism other than the time-averaged statistical relationships to recognize pitched sounds. This is a possible study to be performed in the future to gain further information on the method.

In conclusion, there are clear constraints on the kind of morpher sounds that may be employed for the approach to succeed, therefore ruling it out as an instrumental synthesizer in the current state.

6.2 Synthesis Method as Morpher Synthesizer

The tests that used instrumental recordings as a canvas in the synthesis methods produced good results, both in terms of SNR metrics (as shown in figure 5.9) and by replicating sounds with recognizable elements from the required morpher.

Listening to outputs, it is recognizable that an instrument was used as a canvas, as there is a lingering frequency that is audible even with high number of iterations, mostly the fundamental frequency from the original file. However, most of the recognizable features of the sound texture selected as morpher are also listenable, and the Morpher sound is clearly recognizable, in contrast to the approaches that attempted to do the inverse conversion. Since the process is interactive, it gradually transforms one sound into another, allowing to have some control over the amount of Morpher sound that is placed as canvas.

It came to question if the properties from each statistical entity use, the statistical moments and correlations, could be isolated and if that would result in transferring specific properties to the output sound.

6.2.1 Marginal Moments

As discussed in the previous chapters, marginal moments represent the distribution of amplitudes for individual frequency channels. They are often theoretically associated with the concept of sparsity, which relates to systems having few pairwise interactions. As a result, marginal moments are highly effective in conveying sounds with sparsely dispersed and band independent acoustic occurrences.

Sounds that met this description in one of the one McDermott studies (2009) were described as "water-like" sounds, the only sort of sounds that

sounded realistic when generated only from cochlear moments. As a result, the bubble sound texture was the one that generated the best results when only using marginal moments, while the other textures did not produce such recognizable results.

Mean

The statistical mean mostly represents the channel's average sound power because it only displays the average value for a given frequency channels.

As a consequence, when this moment is imposed individually, the only variation in the result file is mostly an increase in sound power. Since it lacks the other statistical relationships, the process works by matching the sound intensity incrementalist over all the frequency spectrum, as it is possible to see in figure 6.4. Because the instrumental sample is harmonic rich compared with the average natural sound texture, this overall adjustment in all spectrum frequencies results in an overall sound that is "muddled" given that the relative harmonic information being reduced, while the other frequencies are increased, also giving it a ghastly and windy vibe.

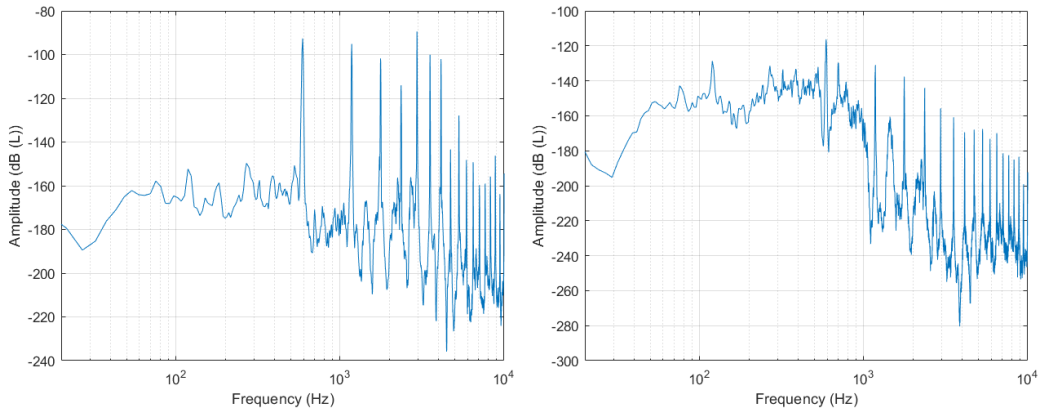


FIGURE 6.4: FFT of the original instrumental sample, on the left, compared with FFT of the output sample using only the statistical imposition of the mean, on the right.

Variance, Skewness and Kurtosis

While using these isolated marginal statistical moments of a higher order, it is possible to recogniz some distinguishable properties of that sound texture imposed on the output file. With the increase in the order of the moment used, the effect that occurred in the use of mean, the "muddliness" of the sound, seemed to dissipate.

While only imposing variance, there is still some sound power variation, similar to the use of mean, causing the to have the greatest impact on the volume modulation of the three sounds. Although it occurs mainly on in the begging of the sample, thus it is not noticeable on time averaged representation, like in the graph 6.5. Using the bubble sample as an example, it also contributed by transferring some distinguishable aspects to the sound, resulting in "bigger" occurrences on the sound texture, such as lower-end bubble explosions and turmoil.

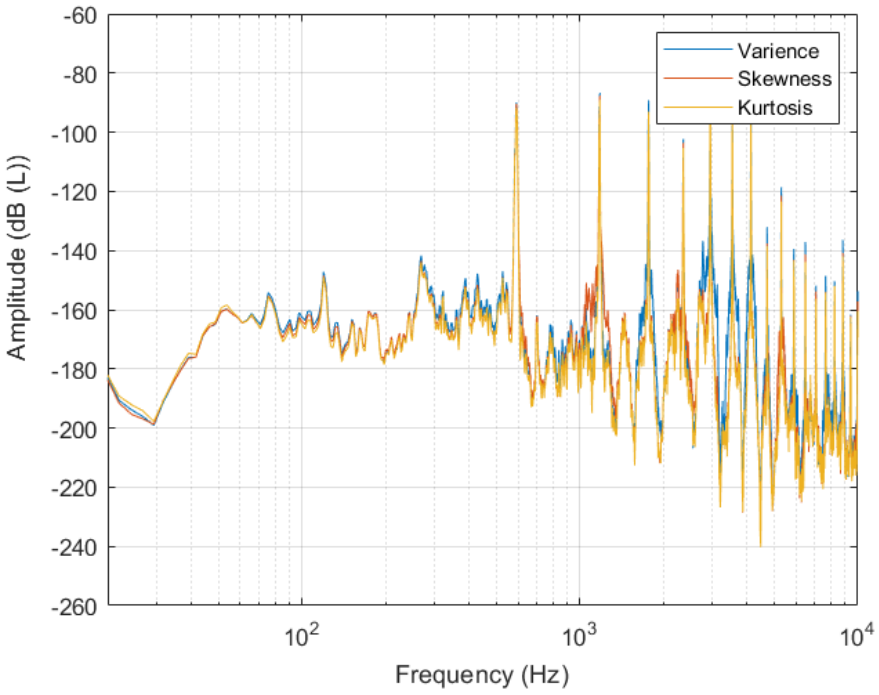


FIGURE 6.5: FFT of the output sample imposing only the variance, skewness and kurtosis, as indicated.

When only the skewness was forced, it did not cause as much muddliness of the sound as variance. This time, it appeared that several finer parts of the sound were captured, in the case of the bubble sample, the first impact of the bubble rupture as well as the turbulent sense of the sound itself, were better captured.

Using only kurtosis, the muddliness is mostly gone, it had the least impact on the instrument sample, on a sound power level. It did, however, appear to yield a stripped-down version of the bubble sound, not detectable in any one event but plainly in the tumultuous mood of the original sample's moment, with the emergence of "crack"-like noises.

These metrics seem to have continuity between them. The lower the order, the most effect it has on overall sound power, with the presence of the "muddle" effect and the presence of more lingering and long events of the texture, while higher order moments seem to unaffected the instrumental sample on a broader spectral level, embracing more the small events and noises of the texture. For practical usage, variance and skewness usually give the most interesting results, having a good balance between the higher order and lower order output effects.

6.2.2 Correlations

Recalling, correlations discern relationships between different cochlear bands, cross-channel and within-channel modulations. This allows the method to account for certain features in the sounds, such with sharp onsets or offsets.

However, in this method these metrics take a supportive role. Themselves, without the influence of the alteration to the spectrogram that are made with the marginal moments, do not have much theoretical weight. As

such when these statistical metrics are used solely, they do not seem to provide as many recognizable features from the original sound as the moments did. In fact, the appearance of unrelated noises was common.

As a result, they should be considered as a complementary metric to be used in conjunction with the changes made by the marginal moments. Using this approach, the inherent correlations of non-water like "features" are more prevalent on the final output.

6.3 Expressive Synthesis Tool

With the results in hand, it is possible to consider how the method can be used as an expressive synthesis tool and what can be tested in future works.

Inputs

As a method based on a mathematical model, it has all of the advantages and disadvantages that such approaches have. It is highly specialized, similar to other model-based synthesis methods, such as physical model approaches, to be very effective in a narrow scope, with the drawback of having minimal transferability to things beyond that scope.

The tests performed tried to expand was possible to use of the inputs of this method, highlighted in the figure 6.6

As the first findings indicate, the sound sources that may be utilized as morphers are quite limited; the only ones that succeeded were the natural sound textures that were already evaluated in the original research. Aside from the superposition of several sounds, an even spectral distribution seemed to be an important factor from these natural sources. This rules

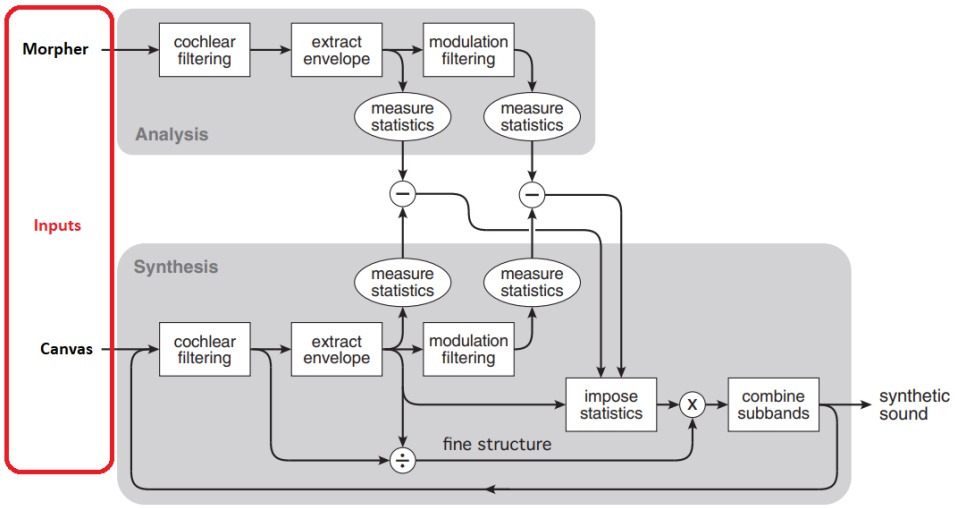


FIGURE 6.6: Inputs of the synthesis method.

out using a synthesizer to create any harmonically rich sounds, such as instruments.

The second experiments, on the other hand, show that similar restrictions do not always apply to the canvas sources that can be employed. This might be due to the fact that the whole information stored in the canvas file, including the harmonics relationships in its frequency distribution, is already available in the early stages of the synthesis process, as opposed to the morpher spectrum information that must be imposed on the canvas. Indeed, the presence of harmonic information while attempting to synthesize sound textures from instrument samples, even with a large number of iterations, may indicate that such harmonic information is not captured by the technique approach.

Parametrization

Some of the experiments were conducted to see how much control of the statistical metrics retrieval and enforcing may have on the synthesis process, highlighted shown in figure 6.7.

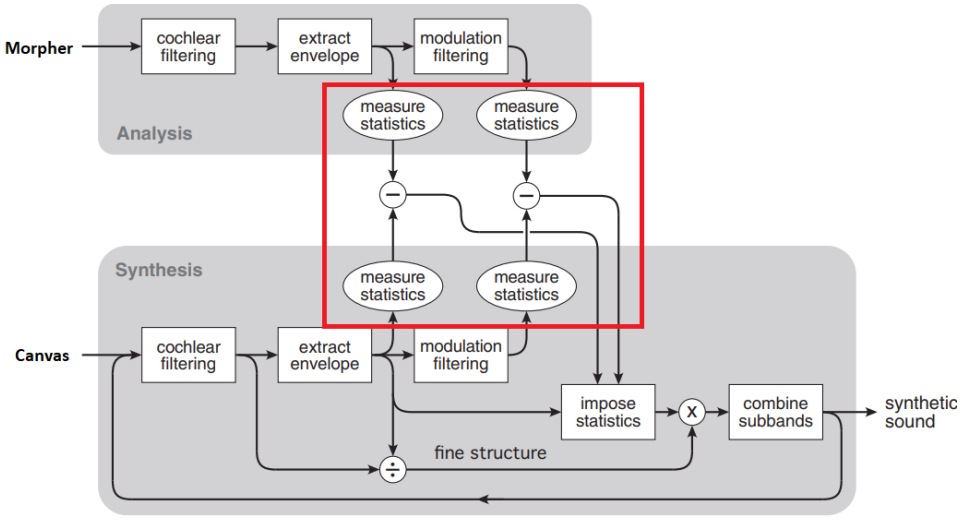


FIGURE 6.7: Statistical metrics retrieval and enforcing stages highlighted in the synthesis method.

The experiments employing a subset of statistical metrics revealed that they are, for the most part, substantially independent of one another for practical purposes: the majority of time measures were able to converge with minimum impact on the other metrics. This implies that the characteristics that each brings to the sound may be applied separately.

The experiments revealed that each statistic had controllable, although abstract, impact on the output sound, rather than any discernible sound qualities that are generally sought for expressive synthesis. This has both advantages and disadvantages: on the one hand, there is no previous body of knowledge that can be directly transferred to streamline the usage of this method, and as a result, there is a great deal of learning to take full advantage of it; on the other hand, this method appears to bring an original approach to parametrizing sound features, allowing for future experiments and applications.

Future Research

Since this is the first attempt to investigate the expressive usage of this approach, additional study might be conducted to improve the process and

eliminate the limitations found.

It would be beneficial to be aware of the method's limits while using harmonically rich sounds. To begin, some tests could be run to see if the method can capture such harmonic relationships within the scope of statistical metrics by expanding the use of new statistical metrics, like for example, correlations which account for and guarantee the preservation of relationships between different tonal frequencies of sound. If this is shown true, the entire process might be updated by adding the new statistical measurements, with only minor changes to the remaining process. If not, the process in its current state could be complemented in parallel synthesis process, a non statistical modelled method that would safeguard the limitations of the approach and would be implemented in a symbiotic process to synthesize the output sound.

It would also be beneficial to investigate the parametrization of this procedure to its greatest extent. Given that the statistical measures may be used independently, it is feasible to dig further into each one and investigate the results of each potential combination of them. This testing might be done in collaboration with artists who could provide input on the outcomes of each isolated statistic and their combinations of, with the goal of developing a terminology are roadmap of the conceivable outputs this technique can produce. In practice, this may be implemented as a practical synthesis tool in which each input controls a linear combination of statistical measures that are applied to the output sound, for example, as a knob in a effects tool that is associated with a certain term in the terminology developed.

The approach currently has a great limit in terms of its practical utility due to its use of computer resources. The procedure is rather rapid when only a few statistical measurements are applied to the output sound, and the sound is synthesized in a matter of seconds. When a large number of statistical measurements are applied at simultaneously, however, the process might

take a long time to synthesize the final product, lasting more than an hour with enough iterations. Since the output is only listenable once the procedure is complete, the method's practical potential is severely limited as it is, and some optimizations would considerably enhance future experiments.

Chapter 7

Conclusion

This chapter will wrap up the study by summarizing the important research results in connection to the aims and questions proposed, examine the study's limitations as well as its worth and contributions.

In summary, the main objective was to study the possible expressive application of McDermott's statistical synthesis method (2011) by investigating the possible application this this process as an instrumental synthesizer and a morpher between instrument and sound textures.

For the usage of this method as an instrumental synthesizer, the results seems indicate that is inept to do such task, mainly due to the processes apparent inability to retrieve tonal and harmonic relationships in the sounds in its statistical deconstruction of them.

The outputs in the second test, which employed an instrument sample as a canvas for a morph with a sound texture, were significantly better, synthesizing a discernible morph between the inputs. The parametrization of the procedure was also successfully tested throughout this tests.

The overall goal of this research was to bridge the gap between natural and expressive synthesis, where there is typically little attempt to integrate the former's large body of work with the latter's. In these studies, a natural synthesis approach was shown to be feasible and practical for usage as an expressive synthesis tool, generating unique synthesized outputs not seen in

other expressive synthesis methods.

This work did not completely explore the expressive potentials of this approach, allowing the entire set of possibilities to be investigated in the future. Future research with better computational resources and greater sample sizes, both sound patterns and instrument samples, might produce a more solid set of results and resolve some of the problems raised. The subjective analysis given was also done alone by the researcher, leading to possible bias with few counterpoints, however, some objective discoveries were made that could serve as a springboard for future investigation attempts.

Although the findings cannot be applied to every relationship between natural and expressive synthesis techniques, they demonstrated that such approaches can provide favourable results and might be utilized as a model for future worlds. The findings also indicated that this strategy merits additional research and refinement for expressive purposes.

In its current state, the synthesis method is already possible to be used as an expressive tool: since it bring to instrumental sounds unique properties of sound textures, it can be used an sound effect tool and a morpher. Although not the most robust method, it as potential in future research experiments, more detailed described in the studies discussion, that could flesh the method as much less limited synthesis tool that could possible be used in its on merit or as a complement to the the various tools and effects used already in expressive sound synthesis.

Bibliography

0101000111 – Talk to the Computer (2018). URL: https://maker4robot.wordpress.com/2018/11/03/0101000111-talk-to-the-computer/.

Antognini, Joseph M, Matt Hoffman, and Ron J Weiss (2019). "Audio texture synthesis with random neural networks: Improving diversity and quality". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3587–3591.

Babaei, Sepideh and Amir Geranmayeh (2009). "Heart sound reproduction based on neural network classification of cardiac valve disorders using wavelet transforms of PCG signals". In: *Computers in biology and medicine* 39.1, pp. 8–15.

Bahadoran, Parham et al. (2018). "Fxive: A web platform for procedural sound synthesis". In: *Audio Engineering Society Convention 144*. Audio Engineering Society.

Bilbao, Stefan (2009). *Numerical sound synthesis: finite difference schemes and simulation in musical acoustics*. John Wiley & Sons.

Bilbao, Stefan et al. (2019). "Physical modeling, algorithms, and sound synthesis: The NESS project". In: *Computer Music Journal* 43.2-3, pp. 15–30.

Bitton, Adrien, Philippe Esling, and Tatsuya Harada (2020). In: *arXiv preprint arXiv:2008.01393*.

Bruna, Joan and Stéphane Mallat (2011). In: *2011 IEEE 10th IVMSP Workshop: Perception and Visual Signal Analysis*. IEEE, pp. 99–104.

Caracalla, Hugo and Axel Roebel (2019). "Sound texture synthesis using convolutional neural networks". In: *arXiv preprint arXiv:1905.03637*.

Caracalla, Hugo and Axel Roebel (2020). "Sound texture synthesis using RI spectrograms". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 416–420.

Chion, Michel (2019). "2. The Three Listening Modes". In: *Audio-vision: Sound on screen*. Columbia University Press, pp. 22–34.

Chorowski, Jan et al. (2018). "On using backpropagation for speech texture generation and voice conversion". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2256–2260.

Cochlea From Wikipedia (2010). URL: https://en.wikipedia.org/wiki/Cochlea#/media/File:Cochlea.svg.

Coler, Henrik von (2019). In: *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx), Birmingham, UK*.

— (2021). "A system for expressive spectro-spatial sound synthesis". In.

Corbella, Maurizio and Anna Windisch (2013). "Sound Synthesis, Representation and Narrative Cinema in the Transition to Sound (1926-1935)". In: *Cinémas: revue d'études cinématographiques/Cinémas: Journal of Film Studies* 24.1, pp. 59–81.

Di Scipio, Agostino (1999). "Synthesis of environmental sound textures by iterated nonlinear functions". In: *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, pp. 109–117.

Difference Between Uniform and Nonuniform Quantization (2018). URL: https://www.differencebetween.com/difference-between-uniform-and-nonuniform-quantization/.

Digital Audio Basics: Audio Sample Rate and Bit Depth (2021). URL: https://www.izotope.com/en/learn/digital-audio-basics-sample-rate-and-bit-depth.html.

Einbond, Aaron et al. (2021). "Instrumental Radiation Patterns as Models for Corpus-Based Spatial Sound Synthesis: Cosmologies for Piano and 3D Electronics". In: *Proceedings of the International Computer Music Conference*.

Equal-loudness-level contours for pure tones. (2004). URL: https://d3i71xaburhd42.cloudfront.net/39b0734b89ad5595a7e93b88bec21643182032f8/11-Figure11-1.png.

Farnell, Andy (2007). "An introduction to procedural audio and its application in computer games". In: *Audio mostly conference*. Vol. 23. Citeseer, pp. 1–31.

Fasciani, Stefano (2018). "Spectral granular synthesis". In.

Fastl, Hugo and Eberhard Zwicker (2007). "Loudness". In: *Psychoacoustics*. Springer, pp. 203–238.

Françoise, Jules, Norbert Schnell, and Frédéric Bevilacqua (2013). "A multimodal probabilistic model for gesture–based control of sound synthesis". In: *Proceedings of the 21st ACM international conference on Multimedia*, pp. 705–708.

Fröjd, Martin and Andrew Horner (2009). "Sound texture synthesis using an overlap–add/granular synthesis approach". In: *Journal of the Audio Engineering Society* 57.1/2, pp. 29–37.

Gabrielli, Leonardo et al. (2017). "Introducing deep machine learning for parameter estimation in physical modelling". In: *Proceedings of the 20th International Conference on Digital Audio Effects*.

Gan, Chuang et al. (2020). "Threedworld: A platform for interactive multimodal physical simulation". In: *arXiv preprint arXiv:2007.04954*.

Ghose, Sanchita and John J Prevost (2020). "Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning". In: *IEEE Transactions on Multimedia*.

Grinstein, Eric et al. (2018). "Audio style transfer". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 586–590.

Hawley, Scott H, Vasileios Chatziiannou, and Andrew Morrison (2020). "Synthesis of musical instrument sounds: Physics-based modeling or machine learning". In: *Phys. Today* 16, pp. 20–28.

Hinton, Geoffrey et al. (2012). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal processing magazine* 29.6, pp. 82–97.

Hsu, Jennifer (2019). *Physically-informed Percussion Synthesis with Nonlinearities for Real-time Applications*. University of California, San Diego.

Huzaifah, Muhammad and Lonce Wyse (2020). "Mtcrnn: A multi-scale rnn for directed audio texture synthesis". In: *arXiv preprint arXiv:2011.12596*.

Kaufman, James C and Robert J Sternberg (2010). Cambridge University Press.

Kersten, Stefan, Hendrik Purwins, et al. (2012). In: *Proc. Int. Conf. on Digital Audio Effects DAFx*.

Koguchi, Junya and Shigeki Sagayama (2018). "Composite wavelet model for stability-oriented speech synthesis from cepstral features". In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 1697–1701.

Kronland-Martinet, Richard (1988). "The wavelet transform for analysis, synthesis, and processing of speech and music sounds". In: *Computer Music Journal* 12.4, pp. 11–20.

Lee, Woo Seok et al. (2021). "Fast frequency discrimination and phoneme recognition using a biomimetic membrane coupled to a neural network". In: *Bioinspiration & Biomimetics* 16.2, p. 026012.

Lin, Kevin (n.d.). In: ().

Liu, Shiguang, Haonan Cheng, and Yiying Tong (2019). In: *ACM Transactions on Graphics (TOG)* 38.4, pp. 1–14.

Lostanlen, Vincent and Florian Hecker (2019). "The shape of RemiXXXes to come: audio texture synthesis with time-frequency scattering". In: *arXiv preprint arXiv:1906.09334*.

Lostanlen, Vincent and Stéphane Mallat (2016). In: *arXiv preprint arXiv:1601.00287*.

McDermott, Josh H, Andrew J Oxenham, and Eero P Simoncelli (2009). "Sound texture synthesis via filter statistics". In: *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 297–300.

McDermott, Josh H, Michael Schemitsch, and Eero P Simoncelli (2013). "Summary statistics in auditory perception". In: *Nature neuroscience* 16.4, pp. 493–498.

McDermott, Josh H and Eero P Simoncelli (2011). "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis". In: *Neuron* 71.5, pp. 926–940.

McDermott, Josh (2013). *Sound Texture Synthesis Toolbox*. https://mcdermottlab.mit.edu/downloads.html, Last accessed on 2021-11-30.

Md Shahrin, Muhammad Huzaifah bin and Lonce Wyse (2020). "Applying visual domain style transfer and texture synthesis techniques to audio: insights and challenges". In: *Neural Computing and Applications* 32.4, pp. 1051–1065.

Moore, Brian CJ (2012). *An introduction to the psychology of hearing*. Brill.

Moore, Brian CJ and Brian R Glasberg (2004). "A revised model of loudness perception applied to cochlear hearing loss". In: *Hearing research* 188.1-2, pp. 70–88.

Morala-Argüello, Patricia, Joaquıén Barreiro, and Enrique Alegre (2012). "A evaluation of surface roughness classes by computer vision using wavelet transform in the frequency domain". In: *The International Journal of Advanced Manufacturing Technology* 59.1, pp. 213–220.

MT-002 TUTORIAL (2016). URL: https://www.researchgate.net/figure/Aliasing-in-the-Time-Domain_fig3_265323668.

Narváez, Pedro and Winston S Percybrooks (2020). In: *Applied Sciences* 10.19, p. 7003.

Okamoto, Yuki et al. (2019). "Overview of tasks and investigation of subjective evaluation methods in environmental sound synthesis and conversion". In: *arXiv preprint arXiv:1908.10055*.

Oord, Aaron et al. (2018). "Parallel wavenet: Fast high-fidelity speech synthesis". In: *International conference on machine learning*. PMLR, pp. 3918–3926.

Perception Space—The Final Frontier (2009). URL: https://en.wikipedia.org/wiki/File:Auditory_Cortex_Frequency_Mapping.svg.

Purwins, Hendrik et al. (2019). "Deep learning for audio signal processing". In: *IEEE Journal of Selected Topics in Signal Processing* 13.2, pp. 206–219.

Qian, Kun et al. (2019). "Deep wavelets for heart sound classification". In: *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, pp. 1–2.

Al-Radhi, Mohammed Salah et al. (2021). "Continuous Wavelet Vocoder-based Decomposition of Parametric Speech Waveform Synthesis". In: *arXiv preprint arXiv:2106.06863*.

Roads, Curtis (2004). *Microsound*. MIT press.

Salah Al-Radhi, Mohammed et al. (2021). "Continuous Wavelet Vocoder-based Decomposition of Parametric Speech Waveform Synthesis". In: *arXiv e-prints*, arXiv–2106.

Schwarz, Diemo et al. (2004). "Data-driven concatenative sound synthesis". In.

Schwarz, Diemo (2006). In: *Journal of New Music Research* 35.1, pp. 3–22.

— (2011). "State of the art in sound texture synthesis". In: *Digital audio effects (DAFx)*, pp. 221–232.

Schwarz, Diemo and Norbert Schnell (2010). In: *Sound and music computing (SMC)*, pp. 510–515.

Selfridge, Rod, David Moffat, Eldad J Avital, et al. (2018). In: *Journal of the Audio Engineering Society*.

Selfridge, Rod, David Moffat, and Joshua D Reiss (2017a). In: *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, pp. 1–8.

— (2017b). "Sound synthesis of objects swinging through air using physical models". In: *Applied Sciences* 7.11, p. 1177.

Serra, Xavier (1993). "Spectral modeling synthesis: Past and present". In: *Proceedings of DAFX London*.

Smith, Jason and Jason Freeman (2021). "Effects of Deep Neural Networks on the Perceived Creative Autonomy of a Generative Musical System". In: vol. 17. 1, pp. 91–98.

Sound 101 (2017). URL: https://www.mrthou.com/sound-101/.

Strobl, Gerda, Gerhard Eckel, and Davide Rocchesso (2006). In: *Sound Music Computing*.

The Fundamentals of Subtractive Synthesis. (2018). URL: https://theproaudiofiles.com/the-fundamentals-of-subtractive-synthesis/.

Tian, Yapeng, Chenliang Xu, and Dingzeyu Li (2019). "Deep audio prior". In: *arXiv preprint arXiv:1912.10292*.

Tomczak, Maciek, Carl Southall, and Jason Hockman (2018). "Audio style transfer with rhythmic constraints". In: pp. 45–50.

Tony J. Rivas (2016). *Sound Texture Synthesis Toolbox, Fact Sheet N°282*. https://reverb.com/news/the-basics-of-east-coast-and-west-coast-synthesis, Last accessed on 2021-11-30.

Unconventional 3D User Interfaces for Virtual Environments (2021). URL: https://www.researchgate.net/figure/The-auditory-system_fig8_242414463.

Verma, Prateek and Julius O Smith (2018). "Neural style transfer for audio spectograms". In: *arXiv preprint arXiv:1801.01589*.

Wyse, Lonce (2019). "Mechanisms of artistic creativity in deep learning neural networks". In.