An Acousmatic Approach to Neural Audio Synthesis

Max Ardito

Department of Music Research (Music Technology)
Schulich School of Music

McGill University
Montréal, Québec, Canada

December 2024

A thesis presented for the degree of Masters of Arts in Music Technology

Abstract

This thesis proposes a novel framework connecting Pierre Schaeffer's acousmatic philosophy of sound with modern machine learning concepts. It critiques the prevailing trend in neural audio synthesis, which often confines neural audio frameworks to the modeling of preexisting musical forms, and argues for a reinterpretation that embraces Schaeffer's concept of the 'sound itself.' By establishing parallels between acousmatic music and neural audio synthesis—both operating within "black box" environments that emphasize the invariant properties of sound—our approach formalizes these connections through the lens of group representation theory by arguing that the concept of the Schaefferian sound object is best modeled as a latent differentiable manifold whose underlying structure forms a Lie group. We then explore existing geometric representations of deep learning architectures such as the *Scattering Transform*, and introduce new geometric interpretations of canonical neural audio models, such as *Differentiable Digital Signal Processing* (DDSP). Using these geometric deep learning frameworks, we then introduce a new method for analyzing and synthesizing acousmatic sound in a way that aligns with Schaeffer's notions of *typomorphology*. This method disentangles a set of spectrotemporal audio descriptors in order to find the most characteristic control parameters with respect to the given dataset of sounds. We find that this method allows one to condition a DDSP model in a way that uniquely resembles the nature of the dataset while yielding a more timbrally expressive output than preexisting DDSP models. The effectiveness of this method is demonstrated through applications on synthetic and percussion sound datasets, and in a typomorphological neural audio synthesizer that morphs sounds across latent space based on their spectrotemporal control parameters. This work marks a significant advancement in merging philosophical concepts with modern audio technology, enhancing artistic engagement and creativity.

Résumé

Cette thèse propose un cadre novateur reliant la philosophie acousmatique du son de Pierre Schaeffer aux concepts modernes d'apprentissage automatique. Elle questionne la tendance dominante en synthèse audio neuronale, qui restreint souvent l'approche neuronale à la modélisation de formes musicales préexistantes, et plaide pour une réinterprétation qui englobe le concept Schaefferien du son lui-même. En établissant des parallèles entre la musique acousmatique et la synthèse audio neuronale—toutes deux opérant dans des environnements de type 'boîte noire' qui mettent en avant les propriétés invariantes du son—notre approche formalise ces connexions à travers la théorie de la représentation des groupes en soutenant que le concept de l'objet sonore Schaefferien est mieux modélisé comme une variété différentielle latente dont la structure sous-jacente forme un groupe de Lie. Nous explorons ensuite les représentations géométriques existantes des architectures d'apprentissage profond telles que le *Scattering Transform*, et introduisons de nouvelles interprétations géométriques des modèles neuronaux audio canoniques, tels que le *Differentiable Digital Signal Processing* (DDSP). Cette méthode désenchevêtre un ensemble de descripteurs spectro-temporels audio afin de trouver les paramètres de contrôle les plus caractéristiques d'un ensemble de sons donné. Nous constatons que cette méthode permet de conditionner un modèle DDSP qui caractérise de manière unique la nature de l'ensemble de données tout en produisant un résultat plus expressif en termes de timbre que les modèles DDSP préexistants. L'efficacité de cette méthode est démontrée à travers des applications sur des ensembles de sons synthétiques et de percussions, et pour un synthétiseur audio neuronal typomorphologique qui transforme les sons à travers l'espace latent en fonction de leurs paramètres de contrôle spectrotemporels. Ce travail représente une avancée significative dans la fusion du certains concepts philosophiques avec des technologies audio récentes, tout en renforçant l'engagement artistique et la créativité.

Acknowledgements

I'd like to first thank Philippe Depalle not only for advising me on this thesis, but also for fundamentally changing the way I think about sound. As somebody who arrived at abstract mathematics by way of experimental music, I am grateful for his support and patience in the interdisciplinary nature of my research.

Likewise, I'd like to thank Carmine-Emanuele Cella for generously agreeing to be the external reviewer on this thesis.

For the dataset used in this work, I thank my collaborator and dear friend Stuart Jackson for two days spent recording numerous takes of friction percussion. Thank you also to those at the *Centre for Interdisciplinary Research in Music Media and Technology* (CIRMMT)—especially Yves Méthot—for facilitating the recording.

Thank you to Julian Neri and Meesh Sara Fradkin for making long days of work in the SPCL room entertaining and enlightening.

Finally, I'd like to thank Ezra Teboul and Selin Altuntur for a number of Friday afternoons spent at Thompson house, and the extended residents of 4252A Rue Saint-Urbain (Hannah, Hannah, Hannah, Amanda, Maddy, and the rest of our Sunday night dinner club) for their love and support.

Contents

List of Notations

Groups and Actions

$\mathfrak{G}$	Group
$\mathfrak{g}, \mathfrak{h}, \mathfrak{i} \in \mathfrak{G}$	Transformation elements of symmetric group
$\mathfrak{e}$	Identity element in group
$\circ$	Binary group operation
$\times$	Cartesian product
$\langle \mathfrak{g} \rangle^{\mathfrak{G}}$	Group generator
$\mathfrak{D}_N$	Dihedral group
$\mathfrak{H} \leq \mathfrak{G}$	Subgroup relation
$\triangleright$	Group action
Hom	Group homomorphism
End	Group endomorphism
$\rho(\mathfrak{g})$	Group representation of $\mathfrak{g}$
$\mathfrak{T}_\tau$	Translation group
$\mathfrak{A}_\tau$	Affine group
$\mathfrak{H}_\tau$	Heisenberg group
$\mathfrak{W}_\tau$	Weyl-Heisenberg group
$\mathfrak{a}$	Affine transformation
$\mathfrak{r}$	Reflection
$\mathfrak{s}$	Rotation
$\mathfrak{m}$	Modulation
$\mathfrak{t}$	Translation
$\mathfrak{p}$	Phase shifting
$\lvert\mathfrak{g}\rvert$	Magnitude of group element
$\tilde{\mathfrak{g}}$	Approximate group element
$\mathfrak{so}(2)$	Lie algebra for SO(2)
$\mathfrak{h}$	Lie algebra for $\mathfrak{H}_\tau$

Sets and Fields

$\mathbb{Z}$	Set of integers
$\mathbb{R}$	Set of real numbers
$\mathbb{Q}$	Set of rational numbers
$\mathbb{C}$	Set of complex numbers
Ω	Set or domain
$u \in \Omega$	Point in set or domain
$\subset$	Subset
$\subseteq$	Subset or equals
$\emptyset$	Null set
$\cup$	Set union
$\cap$	Set intersection

Linear Algebra and Operators

$L^2(\mathfrak{G}_\tau)$	Space of square integrable functions over $\mathfrak{G}_\tau$
$\mathrm{GL}(n, \mathbb{F})$	General linear group of $n \times n$ matrices over a field $\mathbb{F}$
$\mathrm{GL}(V)$	General linear group over a vector space V
$\oplus$	Direct sum
$\otimes$	Tensor product
$\rho\|_W$	Subrepresentation of $\mathrm{GL}(V)$ where $W \leq V$
id_Ω	Identity operator on elements in Ω
Φ_x	Generic contractive operator

Manifolds and Topology

τ	Topology
Ω_τ	Topology over the set Ω
$\mathfrak{G}_\tau$	Topological group
$\mathcal{M}_\tau$	Manifold

Measures and Norms

μ	(Haar) measure
$\|f\|_p$	pth norm of f
$\langle \cdot, \cdot \rangle_{\mathfrak{G}_\tau}$	Inner product in $\mathfrak{G}_\tau$

Audio and Control Parameters

t	Time
A	Amplitude
θ	Angle
ω	Angular frequency
ϕ	Phase
f_0	Fundamental frequency
c	Centroid
Q	Filter quality factor
$\mathcal{X}(\mathfrak{G}_\tau)$	'Sound object' space
$x \in \mathcal{X}(\mathfrak{G}_\tau)$	Point (sound) in the 'sound object' space
$\tilde{x} \in \mathcal{X}(\mathfrak{G}_\tau)$	Resynthesized sound
$\mathcal{F}$	Function space defined over the 'sound object' space
$\mathbf{x}$	Vector of audio
$\mathbf{X}$	Matrix
$\mathbf{v}$	Control parameter vector
$\tilde{\mathbf{v}}$	Resynthesis parameter vector
$\mathbf{V}$	Control parameter matrix
$\mathcal{V}_\tau$	Control parameter manifold
W_l	Control parameter subspace
Γ	DDSP analysis operator
$\tilde{\Gamma}$	DDSP resynthesis operator
$\mathscr{T}$	Sound matching operator

Time-Frequency Analysis

$u, \xi \in \Lambda$	Time and frequency coordinates
$a, b \in \Lambda$	Scale and shift coordinates (wavelet transform)
$j, k \in \mathbb{Z}$	Discrete scale and shift coordintaes
$\chi_{u,\xi}$	Time-frequency atom
$\mathcal{X}$	Time-frequency dictionary
w	Window function
H	Hop size
λ	Log-frequency
Φ_x^{MFCC}	MFCC Extractor
$\Phi_x^{\mathcal{F}}$	Short-Time Fourier Transform (STFT)
$\Phi_x^{\mathcal{W}}$	Continuous Wavelet Transform (CWT)
Φ_x^{JTFS}	Joint time-frequency scattering representation
Φ_x^{MSS}	Multiscale Fourier representation
$\psi(t)$	Wavelet function
$\phi(t)$	Scaling function
Ψ_x	2D wavelet function
E_x	Generic time-frequency energy
E_x^{MSS}	Multiscale spectrogram
$E_x^{\mathcal{F}}$	Spectrogram
$E_x^{\mathcal{W}}$	Scalogram
$E_x^{\mathcal{S}}$	Scalogram derived from scattering transform
E_x^{JTFS}	Scalogram derived from JTFS transform
$\mathcal{S}_x$	Wavelet scalogram (log frequency)
$\mathcal{S}_x^1$	First-order scattering transform
$\mathcal{S}_x^2$	Second-order scattering transform
J, Q, T	Scattering transform coefficients
θ	Scattering transform spin parameter
$T \in \mathcal{T}$	Window in set of MSS windows

Learning and Training

$f \in \mathcal{F}$	Neural network in the neural audio space
$\mathcal{D} = \{\tilde{x}_n, x_n\}_{n=1}^{N}$	Sound pairs (dataset)
$\mathcal{L}^{\text{MSS}}$	MSS loss
θ_i	ith training iteration
∇	Gradient
ν	Learning rate

Glossary

Acousmatic Listening A mode of listening where attention is directed solely at the auditory characteristics of the sound, disregarding its origin or the physical processes that produce it, in line with Pierre Schaeffer's philosophy of *l'objet sonore* (the sound object).

Acousmatic Music A form of electroacoustic music where the source of the sound is intentionally hidden from the listener, allowing them to focus purely on the listening experience without associating the sound with its source.

Affine Structure A mathematical structure that preserves Euclidean geometric forms such as points, lines, and planes. In machine learning and signal processing, affine transformations include linear mappings followed by translations, which are useful in modeling data transformations that maintain Euclidean structural integrity.

Contraction A process in signal processing and machine learning where information is compressed or simplified while retaining its essential features, such as through the pooling layers of a CNN or the scaling transformations in a scattering network.

Convolutional Neural Network (CNN) A type of neural network specifically designed to process data with a grid-like topology, such as images. CNNs use convolutional layers to detect local patterns that are invariant to translation, making them highly effective for tasks like image and sound recognition.

Deep Neural Network (DNN) A type of machine learning model consisting of multiple layers of neurons, where each layer learns abstract representations of the input data. These models have been particularly successful in tasks such as image recognition, natural language processing, and audio synthesis.

Differentiable Digital Signal Processing (DDSP) A framework that utilizes deep neural networks for synthesizing audio signals through learned mappings between control parameters and synthesizer parameters.

Disentanglement The process of separating and isolating different factors of variation in data into linearly independent spaces to produce representations where each dimension

corresponds to a distinct, interpretable property of the input.

Erlangen Programme A framework introduced by mathematician Felix Klein, which classifies geometric structures according to their group of transformations.

Geometric Deep Learning A branch of machine learning that interprets deep learning with respect to assumed geometric structures underlying the data domain, allowing deep learning models to capture and respect symmetric and invariant properties of the data.

Geometric Group Theory A field of mathematics studying groups by interpreting them as geometric objects. These geometric interpretations often form *group representations* (see Section A.3).

Invariance (Informal) Characteristics of an object or signal that remain unchanged under certain transformations. In the context of both acousmatic listening and deep learning, these refer to features of sound that remain stable despite changes in the way that they are either presented to the listener or represented by a neural audio model. For a formal mathematical definition of invariance, see Section A.2.2.

Latent Space In machine learning, a latent space is a lower-dimensional space in which high-dimensional signals are represented. From a computational point of view, the latent representation typically captures essential features or patterns in the data, making it easier to manipulate, analyze, or generate new data.

Machine Learning A subset of artificial intelligence in which algorithms learn patterns from data and use these patterns to make predictions or decisions without being explicitly programmed for specific tasks.

Mesoscale Structure The intermediate scale of sound organization, between micro (fine detail) and macro (large-scale structure). Scattering networks are particularly well-suited to analyze and synthesize audio at this scale.

Microscale Structure The fine-grained, detailed components of a sound or signal, which are often represented using a windowed Fourier transform. DDSP networks are well-suited for learning the timbral properties of audio by utilizing a collection of sounds captured at the microscale.

Morphology (of Sound) In Schaeffer's theory, morphology refers to the detailed, microstructural characteristics of sound, including texture and timbral qualities. This thesis models morphology using Differentiable Digital Signal Processing (DDSP), which represents the microstructure of sound.

Musique Concrète A musical practice developed by Pierre Schaeffer that uses recorded sounds as raw material. It contrasts with traditional music composition by focusing on manipulating recorded sounds rather than creating music through notated performance.

Neural Audio Synthesis The use of neural networks to generate or transform audio signals, allowing for new forms of sound synthesis that mimic or create non-linear auditory structures through machine learning models.

Perceptual Parameters Features of sound that directly correspond to human perception, such as pitch, loudness, and centroid.

Scattering Network A generalization of the Convolutional Neural Network (CNN) that uses wavelet transforms to create a multiscale representation of sound, capturing structural invariants.

Sound Object (*l'objet sonore*) A central concept in Schaeffer's writings, referring to the phenomenological object of listening. It is considered independently from its source, focusing purely on its auditory characteristics.

"Sound Itself" A Schaefferian notion that refers to the experience of sound as an autonomous entity, independent of its source or meaning. It emphasizes the phenomenological properties of sound that are directly perceived by the listener.

Spectrotemporal Audio Descriptors Features that describe both the spectral and temporal aspects of an audio signal, such as spectral centroid, spread, and harmonic energy. These descriptors are used to control DDSP models, capturing perceptually relevant sound characteristics.

Typology (of Sound) Schaeffer's concept of typology relates to the classification and understanding of sound objects based on invariant properties. In the thesis, this is modeled using scattering networks that capture mesostructural features of sound.

Wavelets Mathematical functions used in signal processing to decompose signals into components that vary by scale and location. Wavelets are integral to scattering networks, providing a time-frequency analysis that is invariant to time shifts and frequency transpositions.

Chapter 1

Introduction

Historically speaking, there have been two distinct perspectives on the development of contemporary music technologies. The first and more common perspective views technology as a fundamentally positivist medium for composers and musicians, where the inner workings of the tool remain opaque to its users. While this approach may lighten the intellectual burden on the artist caused by the complexity of the tool, it reinforces a separation between the artist and the tool's inner modalities, limiting deeper engagement and creativity. This division of intellectual labor arguably alienates the artist from the creative process by distancing them from the representational and epistemological frameworks embedded within the technology—frameworks that could be crucial to artistic expression. [1]

In contrast to this perspective, an alternative approach sees music technology itself not merely as a means to present the composer's work, but rather as a unique lens through which the work is revealed [Kan14] [Pal98]. Unlike the former, this latter approach has resulted in a number of radical and unprecedented formal shifts in both contemporary art music and contemporary popular music. In this work, we situatate our research around the philosophy of acousmatic music—a practice of electronic music composition that arose in France during the mid-20th century where sound is experienced without a visible source through the utilization of spatialized loudspeakers, manipulation of recorded sound, and implementation of audio synthesis and signal processing techniques. Likewise, the

[1] This type of critique stems heavily from Marx's notion of alienation and critique of commodity. See the *Economic and Philosophic Manuscripts of 1844* [Mar78].

development of this musical practice was deeply intertwined with parallel advancements in engineering and signal theory, enabling the composer to imagine unforseen ways to manipulate and present sound to a listener. These new frameworks in both composing and listening were thus the result of close collaborations between engineers and composers—particularly at institutions like the *Groupe de recherches musicales* (GRM)—which led to the reformalization of not only compositional techniques, but also the philosophical notions surrounding sound itself [Sch66].

Compared to these radical mid-century developments, it would appear as though many areas of music technology research in the 21st century are lacking such a level of transparent collaboration between engineers and composers. Contemporary collaborations between engineers and composers—more often than not—serve to imbibe modes of thought implicit in the intended use case of the technology, instead of imagining complete reformalizations of sound as a result of the technology. This phenomenon is especially resonant within the area of *Machine Learning* (ML), which has greatly influenced contemporary music technology research and has led to the development of the emerging field of *Neural Audio Synthesis*. Neural audio synthesis leverages the use of *Deep Neural Networks* (DNNs), which serve as universal function approximators [HSW89] able to generate and manipulate audio signals by learning patterns directly from audio data. However, due to the complexity of domain specific knowledge needed in order to understand various DNN models, a majority of neural audio research limits the artist's engagement with the distinctive modes of thought that the neural network presents, reducing these new frameworks to mechanisms for modeling preexisting musical forms rather than exploring entirely new dimensions of sound.

In this thesis, we take a different approach by reformalizing neural audio synthesis as a framework prefaced on the philosophy of acousmatic sound, which originated in the research and music of Pierre Schaeffer [Sch66]. We justify this approach by observing three fundamental similarities between acousmatic music and neural audio synthesis. The first similarity is that a piece of acousmatic music, much like a neural audio model, exists in an uninterpretable "black box" environment in which the listener is barred from the tangeable sources of sound. The second similarity is that acousmatic music is fundamentally concerned with a listening practice that emphasizes the audition of *invariant* properties of sound in order to better understand what Schaeffer famously denotes as the *sound object*

(*l'objet sonore*) [Sch66]. Recent literature on neural networks interprets the layers of a DNN as functional components that perform a strikingly similar task: operators that enforce structural priors which preserve geometric invariance [FWW21], [MFSL19], [BSL13], [Rav20]. This novel interpretation of the neural network reinforces one final connection to acousmatic music: that Schaeffer's philosophy of sound can be epistemologically linked to the mathematical philosophy surrounding the study of invariant geometric forms, more specifically the study of geometric group theory and representation theory pursued by mathematician Felix Klein's *Erlangen Programme* [Kle72].

After formalizing these geometric frameworks, we propose an approach that models the Schaefferian sound object as a continuous manifold with Lie group structure. We then argue in this thesis that Schaeffer's notion of the sound object's *typology* is well described by a generalization of the *Convolutional Neural Network* (CNN) called the *Scattering Network* [Mal12b], which yields a representation of sound that captures musical structure on the mesoscale [CHC+23]. A scattering network uses wavelets to iteratively separate and contract audio signals such that they form a multiresolution affine representation that is time-translation invariant, frequency-transposition invariant, and geometrically stable to time and frequency-warping [ALM19]. Likewise, we propose an interpretation of Schaeffer's notion of the sound object's *morphology* as a concept best represented using *Differentiable Digital Signal Processing* (DDSP) [EHGR20] models, which use neural networks to map from audio control parameters to the parameters of a synthesizer [VNWD14] while remaining equivariant to affine and time-warping transformations at the microstructural level.

Finally, we propose a practical method for analyzing and synthesizing sound objects using neural audio synthesis. This method reinforce's Schaeffer's notion of *sound itself* by disentangling the most perceptually independent parameters within a dataset of sounds such that morphological control at the microscale does not compromise typological strucure at the mesoscale [LYY23]. The resulting parameters form a subset of time-varying spectral audio descriptors [PGS+11] that can be used as a set of control parameters for a DDSP model. We furthermore demonstrate this method on a dataset of recordings of synthetic sounds and a dataset of friction-based percussion sounds. We also implement a neural audio synthesizer that morphs sounds in latent space based on their spectrotemporal control parameters. This method serves as the first instance, to our knowledge, that a DDSP model has utilized an

augmented control space of spectrotemporal audio descriptors, as well as the first time that a method has been proposed for finding control parameters most suitable for an arbitrary DDSP dataset.

In the remainder of this introduction, we first present some brief but necessary background concerning acousmatic music and neural audio synthesis, and then present a broad outline of the work.

1.1 What is Acousmatic Music

During the mid-20th Century, massive technological developments in 1940s post-war Europe allowed for the emergence of early tape recording, filtering, signal processing, and synthesis technologies that catalyzed developments and experiments in *musique concrète*—a term coined by Pierre Schaeffer, the first director of the GRM. Schaeffer's development of a musique concrète came as a result of consistent experimentation with the use of recorded and electronically processed sound as compositional material. In doing so, Schaeffer additionally played the role of both composer and music technologist, as he collaborated directly with psychologists, sociologists, technicians, and engineers in his musical endeavors [Ter15]. Furthermore, as a result of his work with early recording and signal processing technologies, he began to lay the groundwork for novel listening practices. Schaeffer produced formalized philosophical writings such as the *Traité des objets musicaux* (Treatise on Musical Objects) which reckoned specifically with how these emerging audio technologies could aid in revealing unique modalities of *sound itself* rather than how they might translate historical assumptions about sound still lingering from past formats like acoustic instruments and musical scores. Schaeffer even argued that these emerging technologies for working with audio required an entirely new taxonomy for the compositional analysis of sound in a way that—according to Michel Chion, a former assistant of Schaeffer's at the GRM—"broaden[ed] the descriptive range of sounds and instruments which might be limited to the identification of physical parameters (frequency, amplitude, duration)" [Chi83]. This taxonomy came to be known as *typomorphology*, a sort of music theory of Schaeffer's so-called 'sound itself.'

What then resulted from these early experiments was *acousmatic music*, which evolved subsequently out of the GRM [Bat07]. Building off of the musique concrète developed by

Schaeffer, acousmatic music presented an entirely new framework for composition in which the composer writes works for spatialized loudspeakers, consisting of recorded and synthesized sounds that are to be diffused live. Along with this new compositional practice came the listening practice of *acousmatic listening* which involves—for the listener—the mental process of hearing and reconstructing the *unseen* sources of sound, hidden by the technology of the acousmonium[2] [Ter15]. Gaining significant attention across broader contemporary music communities, the acousmatic format attracted many pioneers of 20th century composition to write works at the GRM including Iannis Xenakis and Bernard Parmegiani [Gay09].

The musical practice of acousmatic music still continues, as organizations like the GRM still remain active in both comissioning acousmatic music and releasing records of electronic music of a derivative vein. *Portraits GRM*—a relatively new record label run out of the GRM by current director François J. Bonnet—features a number of 21st century composers reinventing and continuing the cannon of acousmatic music including Florian Hecker, Okkyung Lee, Felicia Atkinson, Laurel Halo, and Jim O'Rourke [Por]. Furthermore, aspects of acousmatic music such as the spatialization [Bro21], granularization [Dav16], and sampling [Rey21] of electronic sound have influenced club music, techno, rock, hip-hop, and studio music at large. Indeed, the survival of both acousmatic composing and acousmatic listening is a testement to the unconscious influence that the field of engineering exhibits over electronic music, as paradigms in sound technology and audio signal processing slowly change.

1.2 What is Neural Audio Synthesis

Over the past decade, composers and music technologists alike have seen yet another influx of emerging technological developments, originating first in the late 1980s as a result of the introduction of the DNN as a universal function approximator [HSW89]. In the domain of audio, this development has resulted in the birth of *Neural Audio Synthesis*. Neural audio synthesis leverages the use of DNNs equipt with structural priors such as convolution [LBBH98] and recurrence [Wer90] to learn invariant properties of audio signal

[2] A word referring to the spatialized loudspeaker arrays used in acousmatic music, which can be interpreted as analogous to the 'orchestra' in orchestral music.

representations. Trained neural networks can be used to parametrically control audio synthesizers and effects processing for the purpose of electronic music performance and composition. Perhaps the most notable model for this new paradigm in audio synthesis is *Differentiable Digital Signal Processing* (DDSP) [EHGR20] which combines the continuous differentiability of trained DNNs with classical signal processing techniques such as additive synthesis [BSAL11], phase-vocoder [AKZB11], and source-filter [AKZV11] modeling.

Despite the popularity of such audio models for control, synthesis, and composition, the use of DNNs has received notable criticism in the field of engineering. This criticism comes from the fact that DNNs are notoriously hard to interpret, and sometimes entirely uninterpretable [SZS+14]. The statistical distribution that a DNN learns, which greatly depends on a training dataset that is unknown to the user at inference time, has led many to describe the DNN as a "black box" [Lip16]. For instance, a neural audio model that generates sound given a set of input parameters might learn to generate two very different types of output during inference time, given its prior training on one set of audio signals versus another.

This lack of interpretability is seen by many engineers as a major flaw. [Mal12b] contextualizes this problem in a 2017 lecture at UCLA on the use of DNNs for classification: audio signals live on a highly irregular domain whose dimensionality is huge ($N \geq 10^6$) and our task is to approximate this domain with very few samples ($D \ll N$). In order for a DNN to generalize the variance in the domain without biasing towards the distribution formed by the D samples, error is simply unavoidable [GBC16]. A hypothetical example of this might involve an instance of sound matching [HLL23] in which a listener expects to hear a particular class of sound from the output of a neural audio model but instead hears a different class of sound. This error is a testement to the complexity of the problem at hand and the difficulty faced when interpreting the DNN as a means through which sound is represented to both the composer and the listener.

1.3 Motivation: Who is This Thesis For

The ideas presented in this thesis are highly specialized yet also highly interdisciplinary making the motivation and intended readers hard to decipher. This thesis is intended primarily for music technology researchers to present a reformalization of neural audio

synthesis through a higher level of mathematics that we hope will also reflect a philosophical mode of thinking about sound that has been forgotten in the majority of neural audio literature. As a result, we put these topics in dialogue with Schaeffer's philosophy of sound to introduce a wider breadth of creative output in composition, theory, and sound studies. We hope also that this thesis will be appreciated for the mathematically inclined electronic composer, sound studies theorist, or experimental music appreciator in broadening their understanding of what *acousmatic sound* might be.

1.4 Contributions

This thesis contains the following main contributions

1. Experimentation with spectrotemporal audio descriptors as control parameters for various DDSP models, resulting in more expressive control of output sound (Chapter 6)
2. A method for finding a subset of DDSP control parameters most reflective of a given dataset that utilizes the dataset's scattering representation (Chapter 6)
3. A framework for a typmorphological neural audio synthesizer, as well as a basic implementation (Chapter 6)
4. An acousmatic approach to the analysis of neural audio models that utilizes geometric group representation theory [BBCV21] and topology [VNWD14] (Chapter 3, 4, and 5).

Other contributions include:

- A historical link between Schaeffer's musical ideas and the geometric ideas of Felix Klein's *Erlangen programme* (Chapter 2)
- Practical experimentation with relatively new DDSP networks such as NoiseBandNet [BRC24] for sound synthesis, and scattering networks such as the *Joint Time-Frequency Scattering* transform [LEHR$^+$21] for the purpose of sound classification (Chapter 6).
- An overview of basic group representation theory and topology that includes examples relating to music and sound synthesis (Chapter 3 and Appendix A).

1.5 Structure of Thesis

In Chapter 2, we make the connection between Schaeffer's philosophy of sound and the study of geometric invariance using ideas from Brian Kane's book on acousmatic music *Sound Unseen* (2014). We then shift towards a more mathematically inclined discourse in Chapter 3 and lay out the fundamentals of group representation theory needed to discuss groups in the context of neural audio synthesis, and propose that the *Lie group* is a suitable representation for the problem of measuring and mapping the acousmatic sound object. This group theoretical foundation allows us to argue in Chapter 4 that Schaeffer's concept of typology is effectively represented by the Scattering Network, which provides a representation of sound that captures musical structure on the mesoscale. Likewise, in Chapter 5 we argue that Schaeffer's concept of morphology is most effectively represented through Differentiable Digital Signal Processing (DDSP) models. These models employ neural networks to translate audio control parameters into synthesizer parameters, while maintaining equivariance to affine and time-warping transformations at the microstructural level. These reformalizations culminate in Chapter 6, where we introduce a practical method for analyzing and synthesizing sound objects using neural audio synthesis. This approach enhances Schaeffer's notion of the sound object by disentangling perceptually independent parameters within a sound dataset, ensuring that morphological control at the microscale does not interfere with typological structure at the mesoscale. The resulting parameters—a subset of time-varying spectral audio descriptors—serve as conditional parameters for a DDSP model. Chapter 7 summarizes and concludes the work, providing some considerations for future research.

Chapter 2

The Sound Object and its Geometric Invariance

In this chapter, we introduce the notion of the sound object as it pertains to the musical philosophy of Pierre Schaeffer. We show how the sound object is a theoretical unit of sound that reveals itself to the listener through its musical *invariance* in the acousmatic setting—a presentation of recorded or processed sound decontextualized from its source. We then show that the idea of musical invariance used by Schaeffer is epistemologically linked to a mathematical notion of invariance by way of the influence that Husserlian phenomenology had on Schaeffer, and transitively by Husserl's influence from the mathematical philosophy of the Erlangen Programme. Given Schaeffer's place in history, epistemological context, and philosophical influence, we thus present the argument that the sound object is a musical construct best suited to be analyzed using the mathematics of the Erlangen programme and geometric group theory. We conclude by showing that implementations of the Schaefferian sound object as data are best fit to be interpreted using geometric approaches to machine learning.

2.1 Philosophy of The Sound Object

For many composers and practitioners of computer music, the term *musique concrète* is often synonymous with the act of composing music using recorded sound, harkening back to the practice's historical genesis at Radio France and the GRM. But while the audio

signal processing and recording technologies of late 1940s post-war Europe may have aided Pierre Schaeffer in conceiving of musique concrète, Schaeffer's intentions in conceiving new musical practices were rooted less so in these new technologies themselves and more so in a philosophical inquiry surrounding these technologies. [Kan14] unravels the philosophical tradition from which Schaeffer takes influence using an early 1948 journal entry by Schaeffer titled *A la recherche d'une musique concrète* (In Search of a Concrete Music) formally published four years later.

> I have coined the term Musique Concrète for this commitment to compose with materials taken from "given" experimental sound in order to emphasize our dependence, no longer on preconceived sound abstractions, but on sound fragments that exist in reality, and that are considered as discrete and complete sound objects, even if and above all when they do not fit in with the elementary definitions of music theory.

Schaeffer was not simply a composer nor was he simply a researcher, but rather a philosopher of sound who produced extensive writings in tandem with his music. One such work titled the *Traité des objets musicaux* (Treatise on Musical Objects) reckons with how emerging audio technologies might reveal these unique modalities of sound that, as stated above, do not conform to historical assumptions about sound still lingering from traditional music theory [Sch66]. Schaeffer thus introduces a fundamental concept in the above quote aiding both the composer and listener: *sound object*. For Schaeffer, the sound object is a proposed theoretical unit of sound that is not defined by any prior construct in western music theory. Instead, the sound object is a notion of sound itself; a unit of sound that the listener recognizes as an objective entity through a myriad of variations generated by the manipulation or processing provided by a given piece of audio technologies [Kan14].

In order to fully understand the sound object as a philosophical construct, a number of Schaefferian concepts must be thoroughly introduced. In this section we will look at these concepts using [Kan14], namely those of the *Acousmatic Reduction* and *Invariance*. These concepts will later be contextualized within the broader historical context of the scientific and mathematical literature that influenced Schaeffer.

2.1.1 The Acousmatic Reduction

While it might be second nature today given the proliferation and availability of digital audio, the possibilities that the tape machine provided for the composer of the mid-20th century had grand philosophical implications. Arguably the most fundamental of these implications was the possibility of sound production without the original sound source physically present. For Schaeffer and many of composers of electronic music who followed, the reorientation of sound as raw musical material that could be recorded, manipulated, and played back warranted a totally novel compositional practice. What was originally called musique concrète soon became *acousmatic music*, reorienting Schaeffer's philosophy of the sound object towards a *compositional* unit of sound within the context of spatialized electronic music diffused via loudspeakers [CG19].

[Bat07] points out that it was the poet Jérôme Peignot who first suggested to Pierre Schaeffer the alternative name 'acousmatic music.' Peignot's claim was that the notion of the 'acousmate' gave a mystical dimension to the phenomenon of hidden sound. In the words of Peignot: "with sound technology one can transport or reproduce sound without its being associated with the material that produced it." This idea was then revitalized by François Bayle fifteen years later, when he applied it to the music of the GRM.

Despite the budding musical etymolygy of the term 'acousmatic' arising later in the century, it is crucial to recognize that the word's original meaning describes not so much a musical situation as it does a philosophical one. While acousmatic music itself might refer to a compositional practice, the acousmatic as *phenomenological situation* spans all forseeable interactions a listener might have with a sound divorced from its original physical context. This might indeed take place at a concert of acousmatic music, but also perhaps while listening to audio with headphones, or hearing a voice in a public space over a loudspeaker. Whether realized as music or not, Schaeffer argued that these acousmatic situations are the only instances in which the sound object can reveal itself clearly. [Kan14] reinforces this with another quote from Schaeffer from *Traité des objets musicaux*:

> In acoustics, we started with the physical signal and studied its transformations via electro-acoustic processes, in tacit reference to [...] a listening that grasps frequencies, durations, etc. By contrast, the acousmatic situation, in a general fashion, symbolically precludes any relation with what is

visible, touchable, measurable [Sch66].

[Kan14] interprets this quote by explaining that even though one can speculate about the causes of a sound source in the acousmatic situation, the acousmatic situation must always bar direct access to the visible, tactile, and physically quantifiable assessments in order to maintain this auditory speculation. It is therefore only in this acousmatic context in which a listener can understand the sound object, since they are unable to infer anything about the physical causes of the sound. The sound object is thus, Schaeffer claims, not the instrument that was played, nor is it the piece of music technology through which it is represented—e.g. a splice of magnetic tape. Instead, [Kan14] emphasizes that the sound object for Schaeffer exists solely in the act of listening: the ear must train itself to hear new musical values and formal devices that are always unique to the encountered sonic materials.

2.1.2 Variance and Invariance

Following this distinction, the new musical values and formal devices that help reinforce the sound object's identity are, from the composer's perspective, implemented using the parameters of a given piece of music technology. In Schaeffer's case, these technologies consisted of the tape player and early audio signal processing techniques. Much like the acousmatic situation, the piece of music technology that the composer works with is crucially noted by Schaeffer to be strictly pedagogical. In other words, there is nothing specifically technological about the "objectivity" of the sound object. A sound object could be demonstrated any number of ways within an acousmatic setting—not only using audio technology but also through the listener's own imagination [Kan14].

It then becomes innevitable that if the listener is barred from all tangeable physical cues which might aid in understanding the sound source, the sound object must reveal itself through its multiplicity of parametric variations, often organized formally by the composer in a piece of acousmatic music. The compositional organization of sonic material in acousmatic music thus becomes philosophically oriented—as Kane argues—around the concept of variation:

> By taking a sound and using electronic means to alter its qualities, Schaeffer pedagogically produces a set of variations with the aim of disclosing the sound

> object's invariant and essential features. The sound of a gong gently rolled with soft mallets is played twice, followed by variants: by adjusting the potentiometers, the envelope of the object is varied; by using low and high pass filters, the mass and grain of the object are varied; subtle shifts in volume create an object with more allure, or internal beating; and finally, a combination of techniques produces another variant. As a listener, not only do we recognize the different variations as variations, we also hear them as one and the same sound object. The objectivity of the sound object is intended to emerge across its various instances [Kan14].

For Schaeffer, the ability of a listener to recognize the sound object through its parametric manipulation is precisely what is made possible by the acousmatic situation. The music technology—whether it be a tape player or a computer—thus becomes the primary tool for this process, revealing the sound object's invariant properties to the listener through the expressive variance of its parametric span. The piece of technology aids the composer in presenting the listener with pedagogical variants of the sound object, and beyond these parametric variants of sonic material, it is the sound object that is thus identified as *invariant* to the listener.

2.2 The Influence of Philosophy and Mathematics

It is precisely this idea of the sound object's *invariance* that allows us to partake in a more detailed reading of Schaeffer's philosophy on sound, a philosophy rooted just as much in science and mathematics as in music itself. This section examines the link between the acousmatic invariance of the sound object and the idea of phenomenological invariance presented by philosopher Edmond Husserl. We then show that Husserlian invariance is a notion that exists in direct response to contemporaneous ideas concerning geometric invariance and group theory introduced by Felix Klein's *Erlangen Programme*. This epistemological link is crucial, as it provides a justification for the methodologies used to model the sound object in subsequent chapters.

2.2.1 Phenomenology of The Sound Object

The Schaefferian invariance of the sound object, Kane argues, stems from a Husserlian concept called *adumbration*. Like the listener's interaction with the sound object in an acousmatic setting, the concept of adumbration denotes the subject's reckoning with objects that are identified as the same across a variety of acts of consciousness. Schaeffer cites a specific example from Husserl in *Traité des objets musicaux* involving a table. The table, like any arbitrary object, can be perceived in the physical world but it can also be imagined. The subject can "narrate a story about [the table, and] hold various beliefs about its provenance" [Kan14]. Kane summarizes further by concluding that the subject thus reckons with the perceived qualities of the table in a physical setting, but only through the synthesis of these qualities are they able to posit the identity of the object as something that innevitably transcends the stream of adumbrations. This "identity" of the object's essence thus denotes its *invariance* learned through a multiplicity of imagined variations, which is precisely what the listener comes to terms with in the practice of acousmatic music [1].

Furthermore, the Husserlian connection to adumbration tells us something else about the subject's actual encounter with the object. The imagined variations of the object—or in the case of acousmatic music, the listener's encounter with parametrically manipulated variations of the sound object—innevitably warrants an argument for the object in question to be freed from its bonds to the physical world [Kan14]. It is thus only in this reduced acousmatic situation that the object decontextualized from the physical world can be perceived as having a transcendental and invariant identity.

2.2.2 The Influence of The *Erlangen Programme*

Husserl's method for examining the essence of objects by way of learning their imagined invariant properties reveals an important influence from similarly shifting ideas in the domain of mathematics. Indeed, Husserl's use of the term *invariance* was not arbitrary, but rather came from a decision to orient mathematics as his basis for formal ontology. [Mor91] points out that among Husserl's influences were Hilbert's *axiomatization of Euclidean Geometry*, Cantor's *Set Theory*, and—perhaps most importantly—the concept of geometric invariance

[1] Formally speaking, Husserl calls this the *Eidetic Reduction*. For clarity, we will simply use the term *invariance*.

from Felix Klein's *Erlangen Programme*.

The Erlangen Programme represented a departure from thinking about geometry as a field concerned with the structure of physical space towards thinking about geometry as a field concerned with abstract notions of structure, primarily centered around the concept of the mathematical *group* [Tie05] [2]. This reorientation of geometry around the notion of the group enabled geometry to "no longer [...] be considered as the theory of the structure of physical space but rather as the science of possible space forms" [Mor91]. Geometry thus came to be the investigation of everything that is invariant under the transformations of the given group, expressed by "the axioms, definitions and theorems that are or could be set up for each particular geometry" [Tie05]. [Tie05] traces the lineage of this shift in mathematics, remarking that the prioritization of invariant structure was made possible by a body of work in mathematics that has its historical origins in the theory of algebraic invariants of Cayley and Sylvester, as well as earlier work centering on the concept of invariance, along with the subsequent developments in geometry due to Grassmann, Riemann, Lie, Helmholtz, and others.

Husserl himself refers to these figures throughout his works, primarily in relation to his own notion of phenomenological invariance [Tie05], as they had a profound influence on him. [Rou23] suggests two ways in which these geometric concepts influenced Husserl's phenomenology. The first way considers an epistemic transition from the experienced shape to the geometric shape. In other words, there is "a gap between the imprecise shapes we experience and the geometrical shapes, [however] this gap does not entail that there is total separation between geometrical shapes and concrete sensuous intuitions". The second way considers how the relating of geometry to the physical experience of space invokes the act of learning invariance through interacting with the parametric variation of forms. For instance, we "arrive at geometrical notions, such as the notion of the circle, through both bodily movement and acts of imagining". Finally, [Rou23] notes that this process—one that reveals invariant geometries through our experiences in the physical world—is similar to the idea that geometries can be characterized in terms of their invariant properties under different transformations, as proposed by Felix Klein's Erlangen Program.

By proxy, this mode of thought undoubtably influenced the way Schaeffer writes about the invariance of the sound object. Indeed, we can understand a sort of "geometry of the

[2] See Appendix A. for a formal definition of the group.

sound object" similar to the way Klein attempts to understand the geometry of shapes through their invariant transformational properties, and likewise the way Husserl attempts to understand the essence of objects through their invariant imagined forms. These three interdisciplinary methodologies are epistemologically analogous, stemming from the very same historical and philosophical context.

2.3 Modeling Acousmatic Sound

Taking this mathematical influence into account, we return to the acousmatic sound object in a more practical context. Beyond the purely theoretical notions of invariance, Schaeffer laid the groundwork for what he called a *typomorphology* of the sound object. This taxonomical framework involved the analysis of synthesized and recorded sounds by their typology (*identification* and *classification*), and subsequently their morphology (*description* and *characterization*) [Chi83] for the purpose of acousmatic music composition. The project of typomorphology was notably continued extensively by Michel Chion, who described the typomorphological analysis of sound as "a broadening of the descriptive range of sounds and instruments which might be limited to the identification of physical parameters (frequency, amplitude, duration)" [Chi83].

Having previously drawn parallels between neural audio synthesis and the acousmatic situation, we finish this chapter by laying out a number of issues and strategies for modeling the Schaefferian sound object in the context of contemporary neural audio synthesis while following the most general guidelines of typomorphology laid out by Schaeffer and Chion. We first look at some first-hand descriptions of typomorphology from the *Traité*, and then propose situating a contemporary typomorphology within a group theoretical interpretation of neural audio synthesis. We use a body of literature that includes [BBCV21], [BSL13], and [Mal12a], all of which leverages the use of geometric priors and group symmetries to more clearly interpret the field of machine learning.

2.3.1 Typomorphology

Typomorphology consists of both a typology and a morphology of sound, each of which are concepts that are argued to be deeply intertwined. For instance, in the *Traité des objets*

musicaux, Schaeffer gives a definition of morphology:

> [H]aving abandoned any reference either to instruments or to accepted values, all we have is collections of disparate sound objects. All we can do is compare them with each other, in all sorts of ways, in their contexture or their texture. This activity is sound morphology [Sch66].

We can gather from this passage that Schaeffer's interpretation of typomorphological analysis is itself rooted in the acousmatic situation, given an implicit disregard for the sound's original source. The morphological analysis thus entails a comparison of disparate sound objects within an isolated environment, infering similarity and difference in both their texture and contexture. The use of the term 'contexture' here is particularly crucial, hinting to the fact that Schaeffer conceptualizes morphology as a weaving together of disparate sound objects into an interconnected structure.

All of these notions point towards strikingly analogous notions in neural audio synthesis. A neural audio model also deals with disparate sounds within an isolated environment, implemented as a *dataset* of sound. The neural audio model then infers similarity and difference in sound objects through a training procedure. This training procedure furthermore involves a weaving together of disparate sounds into a structure that one might call a 'sound object' by projecting data onto a space that is continuous and differentiable. This space is often called a *latent space* [GBC16], however we interpret the space in this work as a *manifold* (Section 3.1.1). The composer leveraging the acousmatic situation of neural audio generation thus abandons reference to prior causality of sound, subsequently guiding the structure of listening via the learned topology on which these sound representations exist.

In dialogue with this definition of morphology, we also come to a description of typology by Schaeffer in the same section of the *Traité*:

> [The musician's] invention has provided a good number of disparate objects in the material sense, [but] we still had to separate them from the continuums where they occurred and also classify them in relation to each other. If we took isolated objects, it comes down to the same thing: we are implicitly obeying rules of sound identification. What are they? They, too, can only be in response to an

> initial morphological approach. Typology, or the art of separating sound objects, identifying them, and if possible carrying out an initial crude screening, can only be based on morphological features [Sch66].

Typology of the sound object is thus directly informed by the contexture of morphology. In neural audio synthesis, we call this initial crude screening a *contraction* (Section 3.3.2), in which representations of sound in a dataset are projected to a lower dimensional space before training. A contraction thus facilitates the possibility for the classification of sounds based on some categorical taxonomy.

2.3.2 Geometric Invariance and Neural Networks

It is then key to figure out how, in a practical sense, a typomorphological analysis of sound could be executed given the advent of neural audio synthesis. The use of the Schaefferian analogy here is not coincidental. Much like Pierre Schaeffer's writings about sound, the field of machine learning owes much of its mode of thinking to the mathematics and philosophy of the Erlangen Programme that preceded it. While a large percentage of research in machine learning turns a blind eye to this epistemological connection [OC17], a few notable exceptions exist [BBCV21], [IL20], [Mal16], some of which have recently coined the term *Geometric Deep Learning* [BBCV21]. This alternative approach to writing about machine learning affords one the ability to more fluently interpret concepts in machine learning by relating them to ideas stemming from the Erlangen Programme. This approach also stands in constrast to approaches that are reliant on material concepts such as datasets, loss functions, and evaluation metrics, all of which might impose unforseen assumptions onto the domain in question.

A geometric approach proposes that the fundamental concern of machine learning is an analysis of groups, group transformations, and group invariance between various geometric domains of signal representation [BBCV21]. This approach benefits our analysis of Schaefferian sound as it provides us a link between the acousmatic situation of the neural audio model and the practice of typomorphology, and furthermore reinforces the connection between Schaeffer's sound philosophy to the Erlangen programme's philosophy of geometry. We can therefore propose that a typological interpretation of neural audio synthesis stands as a classification problem that reckons with the group *invariance* of the

sound object, and that a morphological interpretation of neural audio synthesis represents an optimization problem that learns a mapping between disparate geometric representations by way of an underlying group *equivarance* (see Appendix A: A.2.2).

In order to facilitate this study of the sound object, we review a number of concepts from topology and analysis to reinforce an intuition for geometrically informed sound processing in the following chapter. We also provide a procedural overview of group representation theory in Appendix A, which uses various examples from classical audio synthesis as conceptual aid. These two chapters provide all the necessary material for a typomorphological analysis of neural audio synthesis, which we return to in Chapter 4 and 5 in great detail.

Chapter 3

Geometrically Informed Sound Processing

In this chapter, we first introduce and define the *topological group* $\mathfrak{G}_\tau$ and argue that the topological group is a fitting mathematical construct for modeling the Schaefferian sound object. We then focus on $\mathfrak{G}_\tau$'s representation in L^p spaces and more specifically its measurability in L^2, introducing concepts such as separability, contraction, and deformation. Finally we introduce the Peter-Weyl theorem which elegantly links the subspace decomposition of $L^2(\mathfrak{G}_\tau)$ to a direct sum of irreducible representations, which allows us to interpret neural networks as layers of invariant and equivarant maps with respect to their linearly independent subspaces.

3.1 A Topological Domain for Sound

Groups are often constructed by starting with an unordered set, which we denote Ω. While making no assumptions about the structure of Ω might be desirable in the general study of group theory, we might consider adding some constraints to Ω in order to make its domain more suitable for the purpose of modeling sound.

[VNWD14] propose that a suitable domain for digital sound processing is that of the *topological space*. The justification for this can be seen in the need for a set that generalizes well to the task of *mapping*, which is a task fundamental to practially all computer music systems. Mapping is defined by the association of two parametric spaces for sound control.

The mapping of these two parametric spaces is often also linked to perception, associating the sensation of human intention or expectation with the sonic result of the mapping [VNWD14]. The proposed domain of a topological space thus allows for the constraint of continuous deformation between the control spaces—in other words, the topology implies an assumption of continuity, connectedness, and boundary in mapping associations between control parameters.

We start by laying out the basic properties of topological spaces using [Rud87] as our reference. By adding a few subsequent preliminary constraints, we arrive at the definition of a *manifold* which is needed for relating group theory and topology.

3.1.1 Topological Spaces and Manifolds

Definition: (Topological Space) A collection τ of subsets of a set Ω is said to be a topology on Ω if τ satisfies the following three properties:

1. *Trivial Elements*: $\emptyset \in \tau$ and $\Omega \in \tau$.
2. *Finite Intersections*: If $V_i \in \tau$ for $i = 1, \ldots, N$, then $\bigcap_{i=1}^{N} V_i \in \tau$.
3. *Arbitrary Unions*: If V_α is an arbitrary collection of members of τ (finite, countable, or uncountable), then $\bigcup_\alpha V_\alpha \in \tau$.

If τ is a topology on Ω, then we call the set a *topological space* (denoted Ω_τ), and the members of τ are called the open sets in Ω_τ.

Definition (Continuous Mapping) If Ω_τ and Ω'_τ are topological spaces and if f is a mapping from Ω_τ into Ω'_τ, then f is said to be *continuous* provided that $f^{-1}(V)$ is an open set in Ω_τ for every open set $V \in \Omega'_\tau$.

Definitions: (Compactness) A set $K \subset \Omega_\tau$ is *compact* if every open cover of K contains a finite subcover. More explicitly, the requirement is that if $\{V_\alpha\}$ is a collection of open sets whose union contains K, then the union of some finite subcollection of $\{V_\alpha\}$ also contains K. In particular, if Ω_τ is itself compact, then Ω_τ is called a *compact space*.

Definition: (Neighborhood) A *neighborhood* of a point $u \in \Omega_\tau$ is any open subset of Ω_τ which contains u.

Definition: (Haussdorf Space) We say that Ω_τ is a *Hausdorff space* if the following is true: If $u \in \Omega_\tau, u' \in \Omega'_\tau$, and $u \neq u'$, then u has a neighborhood U_τ and u' has a neighborhood U'_τ such that $U_\tau \cap U'_\tau = \emptyset$.

Corollary: (Local Compactness) Ω_τ is *locally compact* if every point of Ω_τ has a neighborhood whose closure is compact. Every space that is compact is locally compact.

Definition: (Manifold) A manifold $\mathcal{M}_\tau$ is a locally compact Haussdorf space where each point of Ω_τ has a neighborhood that is homeomorphic to an open subset of $\mathbb{R}^n$ (it is locally Euclidean).

Example: Fig. 3.1 demonstrates how a mapping between two manifolds can be used to model control parameter mappings, following [VNWD14]. The presented topological mapping f models a simple granular synthesizer. The domain $\mathcal{M}_\tau$ denotes a parameter space consisting of fundamental frequency (f_0), amplitude (A), and spectral centroid (c) values extracted from the audio signal $\mathbf{x}$ at windows of varying length. We can model a granular synthesizer as the topological mapping $f : \mathbb{R}^3 \to \mathbb{R}^2$, where $\mathcal{M}'_\tau \subset \mathbb{R}^2$ denotes the start (s) and end (e) indices of audio signal $\mathbf{x}$.

3.1.2 Topological Groups and Lie Groups

While a topological space might be suitable for general computer music systems, we might begin to think about how this domain differs when attempting to describe neural audio systems. We've previously argued that much like the acousmatic sound object, sound generated from a generative model can be best analyzed using the language of group representation theory. Invariance and equivariance can be analyzed within a topological space by introducing the notion of the *topological group*. Many topological groups thus aid the interpretation of sound synthesis from a Schaefferian perspective. We will introduce a couple of topological groups in this section, and borrow some definitions once again from [Wei22] and [GQ20].

Definition (Topological Group) A topological group $\mathfrak{G}_\tau$ is a group which is also a topological space, and for which the group operations are continuous. A representation of a topological group $\mathfrak{G}_\tau$ on a finite-dimensional vector space V is a continuous group homomorphism ρ :

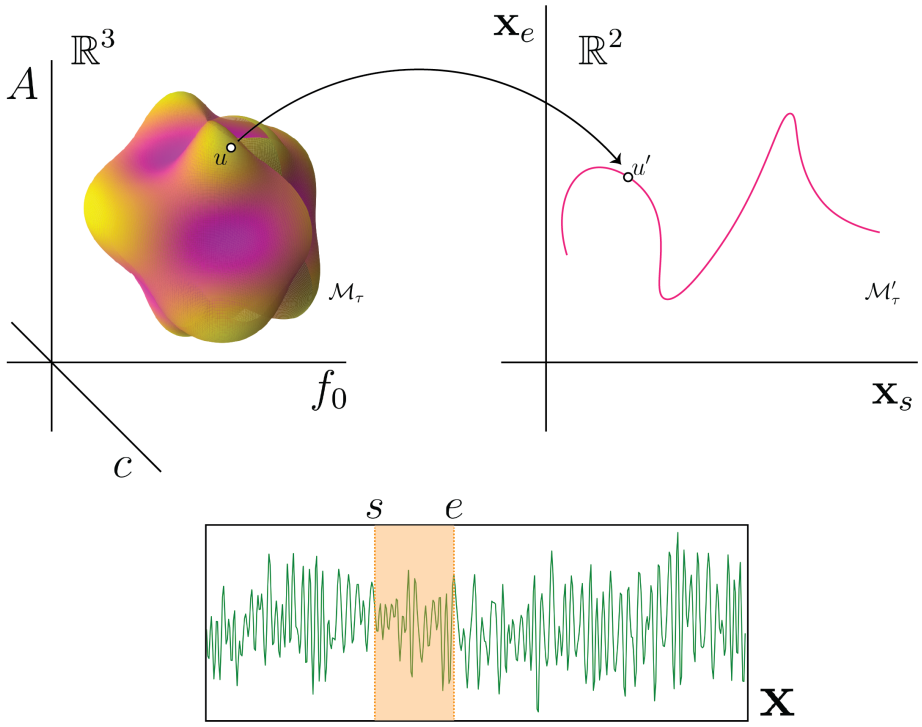


Figure 3.1: A visualization of a continuous mapping between two manifolds representing control parameters. Every point $u \in \mathbb{R}^3$ is associated with a point $u' \in \mathbb{R}^2$.

$\mathfrak{G}_\tau \to \mathrm{GL}(V)$, with the topology of $\mathrm{GL}(V)$ inherited from the space $\mathrm{Hom}(V, V')$ of linear self-maps (Appendix A: A.2.1).

Definition (Lie Group) A Lie group is a topological group $\mathfrak{G}_\tau$ whose topology is a differentiable manifold.

Remark Lie groups are named after Sophus Lie who was closely affiliated with the Erlangen Programme. They are an integral part of representation theory, since the general linear group $\mathrm{GL}(n, V)$ forms a Lie group when the vector space has a common basis such as $V = \mathbb{R}$ or $V = \mathbb{Z}$.

Remark (Lie Algebra) While we choose to skip its formal definition, the Lie algebra is a useful concept to observe simply by example, since it acts as the Lie group's generator. For instance, we can look at $\mathrm{SO}(2)$'s Lie algebra $\mathfrak{so}(2)$, which consists of the following matrix representing infinitesimal rotation.

$$\mathfrak{so}(2) = \begin{bmatrix} 0 & -\mathfrak{s} \\ \mathfrak{s} & 0 \end{bmatrix} \tag{3.1}$$

This infinitesimal rotation thus generates the entire group in the same manner that a finite discrete group can be generated from its actions. For an example of group generation in the context of a discrete group, we analyze the dihedral group $\mathfrak{D}_3$ and its group generators $\langle \mathfrak{r}, \mathfrak{s} \rangle^{\mathfrak{D}_3}$ in Appendix A (section A.1.2).

Example (Affine Group) We now introduce two Lie groups fundamental to audio signal representations. The first group is the affine group $\mathfrak{A}_\tau = \{\mathfrak{a}, \mathfrak{t}, \mathfrak{e}\}$, whose transformations adhere to the group axioms of associativity, identity, inverse, and composition respectively:

$$\begin{aligned} &((\mathfrak{a}_1, \mathfrak{t}_1) \circ (\mathfrak{a}_2, \mathfrak{t}_2)) \circ (\mathfrak{a}_3, \mathfrak{t}_3) = (\mathfrak{a}_1, \mathfrak{t}_1) \circ ((\mathfrak{a}_2, \mathfrak{t}_2) \circ (\mathfrak{a}_3, \mathfrak{t}_3)) \\ &(\mathfrak{a}, \mathfrak{t}) \circ (\mathfrak{e}, \mathfrak{e}) = (\mathfrak{a}, \mathfrak{t}) \\ &(\mathfrak{a}, \mathfrak{t})^{-1} = (\mathfrak{a}^{-1}, -\mathfrak{a}^{-1}\mathfrak{t}) \\ &(\mathfrak{a}_1, \mathfrak{t}_1) \circ (\mathfrak{a}_2, \mathfrak{t}_2) = (\mathfrak{a}_1\mathfrak{a}_2, \mathfrak{a}_1\mathfrak{t}_2 + \mathfrak{t}_1) \end{aligned} \tag{3.2}$$

This group's transformations consist of scaling $\mathfrak{a}$ and translation $\mathfrak{t}$ defined as such:

$$\begin{aligned} \mathfrak{a} \triangleright x &= \mathfrak{a}x \\ \mathfrak{t} \triangleright x &= x - \mathfrak{t} \end{aligned} \tag{3.3}$$

When dealing with the affine group's representation in $\mathrm{GL}(V)$, we can conveniently apply these group transformations to a matrix $\mathbf{X} \in \mathbb{R}^{m \times p}$ by interpreting $\mathfrak{a}$ and $\mathfrak{t}$ as linear transformations [GQ20].

$$\begin{aligned} &\rho(\mathfrak{a}, \mathfrak{t})\mathbf{X} = \mathfrak{a}\mathbf{X} + \mathfrak{t} \\ &\rho(\mathfrak{e}) = \mathfrak{e}\mathbf{X} = \mathbf{X} \\ &\rho(\mathfrak{a}, \mathfrak{t})^{-1}\mathbf{X} = \mathfrak{a}^{-1}\mathbf{X} - \mathfrak{a}^{-1}\mathfrak{t} \\ &\rho(\mathfrak{a}_1, \mathfrak{t}_1)\rho(\mathfrak{a}_2, \mathfrak{t}_2)\mathbf{X} = \mathfrak{a}_1\mathfrak{a}_2\mathbf{X} + \mathfrak{a}_1\mathfrak{t}_2 + \mathfrak{t}_1 \end{aligned} \tag{3.4}$$

where $\mathfrak{a} \in \mathbb{R}^{n\times m}$ and $\mathfrak{t} \in \mathbb{R}^{m\times p}$. Other common linear transformations such as rotation and shearing are also invariant to functions over representations of the affine group. The affine group's transformations are explored at length in the first chapter of Marcel Berger's *Geometry Revealed* [Ber10]. Berger commences the book in precise recognition of Klein: "If we want to characterize affine geometry according to the philosophy of Klein at the turn of the twentieth century, it is necessary to study its automorphisms, by which we mean the bijections that map the affine plane onto itself and preserve its structure."

Example (Heisenberg Group) Another Lie group at the center of sound processing is the Heisenberg group $\mathfrak{H}_\tau = \{\mathfrak{m}, \mathfrak{p}, \mathfrak{t}, \mathfrak{e}\}$, which can be similarly defined by way of its group axioms

$$\begin{aligned}
&((\mathfrak{t}_1, \mathfrak{m}_1, \mathfrak{p}_1) \circ (\mathfrak{t}_2, \mathfrak{m}_2, \mathfrak{p}_2)) \circ (\mathfrak{t}_3, \mathfrak{m}_3, \mathfrak{p}_3) = (\mathfrak{t}_1, \mathfrak{m}_1, \mathfrak{p}_1) \circ ((\mathfrak{t}_2, \mathfrak{m}_2, \mathfrak{p}_2) \circ (\mathfrak{t}_3, \mathfrak{m}_3, \mathfrak{p}_3)) \\
&(\mathfrak{t}, \mathfrak{m}, \mathfrak{p}) \circ (\mathfrak{e}, \mathfrak{e}, \mathfrak{e}) = (\mathfrak{t}, \mathfrak{m}, \mathfrak{p}) \\
&(\mathfrak{t}, \mathfrak{m}, \mathfrak{p})^{-1} = (-\mathfrak{t}, -\mathfrak{m}, \mathfrak{t}\mathfrak{m} - \mathfrak{p}) \\
&(\mathfrak{t}_1, \mathfrak{m}_1, \mathfrak{p}_1) \circ (\mathfrak{t}_2, \mathfrak{m}_2, \mathfrak{p}_2) = (\mathfrak{t}_1 + \mathfrak{t}_2, \mathfrak{m}_1 + \mathfrak{m}_2, \mathfrak{p}_1 + \mathfrak{p}_2 + \mathfrak{t}_1\mathfrak{m}_2)
\end{aligned} \tag{3.5}$$

A group representation of $\mathfrak{H}_\tau$ can be defined over the set of 3×3 upper triangular matrices, representing the Heisenberg group's Lie algebra $\mathfrak{h}$ [Tel05].

$$\mathfrak{h} = \begin{bmatrix} 0 & \mathfrak{t} & \mathfrak{p} \\ 0 & 0 & \mathfrak{m} \\ 0 & 0 & 0 \end{bmatrix} \tag{3.6}$$

An interesting representation can then be derived with an exponential map that utilizes the Lie algebra, using the Taylor series expansion of a matrix exponential

$$e^{\mathfrak{h}} = \sum_{n=1}^{\infty} \frac{\mathfrak{h}^n}{n!} = \mathbf{I} + \mathfrak{h} + \frac{\mathfrak{h}^2}{2!} + \frac{\mathfrak{h}^3}{3!} ... \tag{3.7}$$

which yields the following matrix

$$\mathfrak{h} = \begin{bmatrix} 0 & \mathfrak{t} & \mathfrak{p} + \frac{\mathfrak{t}\mathfrak{m}}{2} \\ 0 & 0 & \mathfrak{m} \\ 0 & 0 & 0 \end{bmatrix} \tag{3.8}$$

The Lie algebra then describes the group's representation $\rho(\mathfrak{t}, \mathfrak{m}, \mathfrak{p})$ which can yield rather interesting actions on a vector $x \in \mathbb{R}$ by setting $e^{\mathfrak{h}} = \rho(\mathfrak{t}, \mathfrak{m}, \mathfrak{p})$

$$\rho(\mathfrak{t}, \mathfrak{m}, \mathfrak{p})x = e^{i\mathfrak{m}t+\mathfrak{p}}x(t - \mathfrak{t}) \tag{3.9}$$

This representation is an irreducible representation in which $\mathfrak{t}$ corresponds to the action of translation, $\mathfrak{m}$ corresponds to the action of frequency modulation, and $\mathfrak{p}$ corresponds to the action of phase shifting. This group has implications in audio synthesis given its association of group actions with the parametric properties of elementary sinusoids. Since each transformation maps to these sinusoidal parameters, we might correctly predict that $\mathfrak{H}_\tau$ is at work in the analysis and synthesis of sound. We will see in Chapter 5 how the Short-Time Fourier Transform (STFT) is itself a representation of a subgroup of $\mathfrak{H}_\tau$ called the Weyl-Heisenberg group.

Corollary (Continuous Action) Let $\mathfrak{G}_\tau$ be a topological group and let Ω be a topological space. An action $\mathfrak{g} \triangleright_{\Omega_\tau} : \mathfrak{G}_\tau \times \Omega_\tau \to \Omega_\tau$ is continuous (and $\mathfrak{G}_\tau$ acts continuously on Ω) if the map $\mathfrak{g} \triangleright_{\Omega_\tau}$ is continuous.

Corollary (Continuous Orbit) If a topological group action is continuous, then its orbit is also continuous.

Example The span of orbits can be interpreted as what is often called a *latent space*. [KPB+23] gives a number of examples of this in the context of different neural network architectures.

3.2 Measuring The Sound Object

In most areas of machine learning, an elementary geometric notion of *distance* is required. Distance is necessary to infer similarity and dissimilarity between representations of samples, and furthermore necessary in order to calculate loss. In the context of machine learning, loss can be interpreted as a rough evaluation regarding how well a certain model preserves the invariant symmetries of the underlying group [BBCV21]. In order to accommodate this, we must introduce the notion of *measurability* into our topological group.

In this section, we first introduce the idea of a *measure* and then introduce a translation invariant measure called the *Haar Measure* over $\mathfrak{G}_\tau$, allowing us to then define a norm within possible function spaces over $\mathfrak{G}_\tau$. We then introduce a 'sound object' space—denoted $\mathcal{X}(\mathfrak{G}_\tau)$—in which we will work exclusively for the remainder of the work, constraining our function space to the space of square-integrable functions $L^2(\mathfrak{G}_\tau)$ over an underlying topological group.

3.2.1 Haar Measure

In the previous section, we observed that constraining our sound representations to a topological group $\mathfrak{G}_\tau$ ensures that the orbits of our group actions are continuous. In other words, we can span the orbit of a given topological group's actions without having to worry about any non-linearities. This property results in the ability to define a notion of size or volume that is *translation invariant*, insofar as the underlying topology is continuous and locally compact. This construct is known as the Haar measure. We first present a definition of a measure from [Rud87] and then a definition of the Haar measure derived from [Bou60].

Preliminaries (Measure) A measure is a function μ that takes a set Ω[1] and returns a value denoting size. The measure μ adheres to the following properties, such that $\forall U_\tau \subseteq \Omega_\tau$:

1. *Non-Negativity*: $\mu(U_\tau) \geq 0$
2. *Null Measure*: $\mu(\emptyset) = 0$
3. *Countable Additivity*: For any collection of disjoint subsets $\{U_\tau^i\}$

$$\mu\left(\bigcup_{i=1}^{\infty} U_\tau^i\right) = \sum_{i=1}^{\infty} \mu(U_\tau^i) \tag{3.10}$$

Definition (Haar Measure) Let μ be a measure on the topological space Ω_τ. μ is said to be invariant under $\mathfrak{g} \in \mathfrak{G}_\tau$ if $\mathfrak{g} \triangleright \mu = \mu \quad \forall \mathfrak{g} \in \mathfrak{G}_\tau$

Example (Left Translation) Often the Haar measure is applied to the situation of a left translation invariant measure. To visualize this, we can begin to think of a common temporal

[1] Technically, a measure is defined over a σ-algebra, but treating the domain as a topological space Ω_τ is sufficient for our purposes

representation of a sound—$x(t)$ where $t \in \mathbb{R}$—as a representation of the *translation group*, a topological group we denote $\mathfrak{T}_\tau$ whose sole transformation is translation $\rho(\mathfrak{t})x = x(t - \mathfrak{t})$.

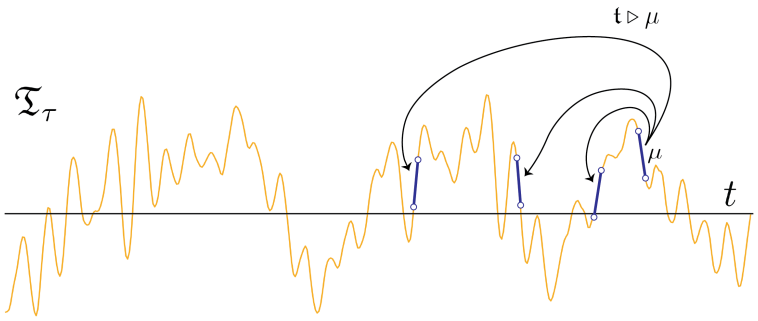


Figure 3.2: Left-translation Haar measure on a temporal representation of sound.

Fig. 3.2 denotes a measure μ on x that remains invariant to the action of left translation. The plot represents a causal real-time audio stream on which a Haar measure μ can be defined, so long as the stream is continuous and differentiable. The left shift of μ via an action $\mathfrak{t} \triangleright \mu$ is represented with the arrows, showing that the measure μ remains invariant regardless of the value $\mathfrak{t} \in \mathfrak{T}_\tau$. More casually speaking, if the Haar measure exists, then there is no neighborhood within the domain that will affect the measurement or volume of μ.

3.2.2 The *Sound Object* Space $\mathcal{X}(\mathfrak{G}_\tau)$

While the temporal representation is commonly used in many practical settings, it neglects the reinforcement of geometric priors that might hint towards a sound's invariant symmetries, useful for a more acousmatic investigation of sound. Taking inspiration from a proposed space of signals in [BBCV21], in this section we propose the Hilbert space as being a suitable space over which we might represent the sound object. We notate this space $\mathcal{X}(\mathfrak{G}_\tau)$, accentuating a geometrically agnostic representation that invokes only the constraints of an underlying topological group and an L^2 function space. We furthermore use [BBCV21]'s notion of a *hypothesis space* to describe neural audio models as functions over the 'sound object' space. We start by defining some basic properties of L^p spaces.

Definition (L^p Space) If $0 < p < \infty$ and if f is a measurable function over a locally compact topological group $\mathfrak{G}_\tau$ (i.e. a topological space Ω_τ equipt with the Haar measure μ), then the L^p space consists of all measurable functions $f : \mathfrak{G}_\tau \to L^p(\mathfrak{G}_\tau)$ where the integral of the pth norm $\|f\|_p$ is finite

$$\|f\|_p = \left(\int_{\mathfrak{G}_\tau} |f(x)|^p d\mu(x) \right)^{\frac{1}{p}} < \infty \tag{3.11}$$

Definition (Hilbert Space) Let $\mathfrak{G}_\tau$ be a locally compact topological group over the set Ω. A Hilbert space is a space of functions:

$$\mathcal{X}(\mathfrak{G}_\tau) = \{x : \mathfrak{G}_\tau \to L^2(\mathfrak{G}_\tau)\} \tag{3.12}$$

where, given two real scalars α and β and any arbitrary point $u \in \mathfrak{G}_\tau$, the space $L^2(\mathfrak{G}_\tau)$ exhibits addition and scalar multiplication for all $u \in \mathfrak{G}_\tau$ and for all $x_1, x_2 \in \mathcal{X}(\mathfrak{G}_\tau)$

$$(\alpha x_1 + \beta x_2)(u) = \alpha x_1(u) + \beta x_2(u) \tag{3.13}$$

and given an inner product $\langle v, w \rangle_{\mathfrak{G}_\tau}$ (implicit by association from the codomain $L^2(\mathfrak{G}_\tau)$) and a measure μ on Ω, we define an inner product on $\mathcal{X}(\mathfrak{G}_\tau)$

$$\langle x_1, x_2 \rangle = \int_{\mathfrak{G}_\tau} \langle x_1(u), \overline{x_2(u)} \rangle_{\mathfrak{G}_\tau} d\mu(u) \tag{3.14}$$

Remark We propose that this slightly unconventional method for defining a Hilbert space is rather suitable for representing a 'sound object' space. This is in part due to the invariant symmetries of the sound being unknown in the acousmatic setting—in other words, the topology of $\mathfrak{G}_\tau$ is most likely nonlinear. In the presentation of such a space, we must then linearize $\mathfrak{G}_\tau$ while still attempting to preserve these invariant symmetries. In other words, the justification for a representation of a sound $x \in \mathcal{X}(\mathfrak{G}_\tau)$ versus something more canonical like a continuous temporal representation $x \in \mathbb{R}$ or a Fourier representation $x \in \mathbb{C}$ is based on the need to not only emphasize Schaeffer's acousmatic situation but also to emphasize the geometric priors intended to be represented. Representations of sound in this format reinforce a sound's invariant geometric transformations, and precisely their preservation of geometric invariance by functions that map x onto their group representation in L^2 space.

We will subsequently see later how we can define the Schaefferian 'sound object' as a low-dimensional manifolds in the space $\mathcal{X}(\mathfrak{G}_\tau)$ allowing us to model the geometric invariance of acousmatic sound, given an underlying Lie group structure.

3.2.3 The Neural Audio Space $\mathcal{F}(\mathcal{X})$

Definition (Neural Audio Space) The neural audio space consists of functions $f \in \mathcal{F}$ where $\mathcal{F} : \mathcal{X}(\mathfrak{G}_\tau) \to \mathcal{X}(\mathfrak{G}'_\tau)$.

Remark In a geometric interpretation of neural audio synthesis, a neural audio model can be thought of as a function f belonging to a 'hypothesis' space [BBCV21], denoted $\mathcal{F}$. The model treats sounds $x \in \mathcal{X}(\mathfrak{G}_\tau)$ and $\tilde{x} \in \mathcal{X}(\mathfrak{G}'_\tau)$ as its training data, such that a dataset $\mathcal{D} = \{\tilde{x}_n, x_n\}_{n=1}^N$ can be constructed and a mapping $\tilde{x} = f(x)$ can be learned. Note that the dataset $\mathcal{D}$ does not itself form a continuous manifold, rather a continuous manifold is learned with the optimization of a neural audio model f resulting in a latent space that resembles the acousmatic sound object.

Remark The underlying groups representing the domain and the codomain are not necessarily the same. We elaborate this claim in Appendix A (A.2.2), but emphasize that this distinction affects our understanding of which geometric priors $\mathfrak{g} \in \mathfrak{G}_\tau$ should be retained when selecting $f \in \mathcal{F}$. The interpretation of these geometric priors motivates the subject of the next section.

3.3 Group Symmetries Over $\mathcal{F}$

In an ideal situation, the space $\mathcal{X}(\mathfrak{G}_\tau)$ is constructed in such a way that it causes a (Lipschitz) *separation* of functions $f \in \mathcal{F}$, thus allowing for the possibility of invariant and equivariant mappings between representations of topological groups. A more practical interpretation of this property is that f successfully *learns* the transformational symmetries $\mathfrak{g} \in \mathfrak{G}_\tau$ as a result of their group representations $\mathcal{X}(\mathfrak{G}_\tau)$. Further reducing the complexity of this problem can be achieved by introducing a *contractive* operator that acts on our sound object representations $x \in \mathcal{X}(\mathfrak{G}_\tau)$ by ensuring a space that is *Lipschtiz Continuous*. We define and formalize these aforementioned concepts borrowing terminology from [Mal16].

3.3.1 Separability

Definition (Lipschitz Separation) The 'sound object' space $\mathcal{X}$ *separates* functions $f \in \mathcal{F}$ by imposing structure from $\mathfrak{G}_\tau$ onto the functions in $\mathcal{F}$. We say that this separation is a *Lipschitz Separation* if

$$\begin{aligned} &\forall (x, x') \in \mathcal{X}(\mathfrak{G}_\tau) \quad \exists \epsilon > 0 \quad \text{s.t.} \\ &\|x - x'\| \geq \epsilon |f(x) - f'(x)| \end{aligned} \tag{3.15}$$

Corollary This can be seen as the margin condition for specifying minimum distance across sounds, shown in Eq. 3.16 where ϵ is reworked to denote the minimum distance across sounds $x \in \mathcal{X}(\mathfrak{G}_\tau)$.

$$\begin{aligned} &\forall (x, x') \in \mathcal{X}(\mathfrak{G}_\tau) \quad \exists \epsilon > 0 \quad \text{s.t.} \\ &\|x - x'\| \geq \epsilon \quad \text{if} \quad f(x) \neq f(x') \end{aligned} \tag{3.16}$$

Remark The Lipschitz separation of f is a necessary precursor to dimensionality reduction techniques for $x \in \mathcal{X}(\mathfrak{G}_\tau)$. We might reiterate here the high dimensional and irregular nature of the domain at hand. It is in most cases desirable to introduce a contractive operator $\Phi : \mathcal{X} \to \mathcal{X}$ that reduces the range of variability in x with respect to the group actions $\mathfrak{g} \in \mathfrak{G}_\tau$. We can then rely on Φ to separate functions $f \in \mathcal{F}$, where a lower dimensional vector $\Phi(x)$ will locally linearize the actions $\mathfrak{g} \in \mathfrak{G}_\tau$.

3.3.2 Contraction

Definition (Lipschitz Continuity) The contractive operator Φ is *Lipschitz Continuous* if

$$\begin{aligned} &\forall (x, \mathfrak{g}) \in \mathcal{X}(\mathfrak{G}_\tau) \times \mathfrak{G}_\tau \quad \exists C > 0 \quad \text{s.t.} \\ &\|\Phi(\rho(\mathfrak{g})x) - \Phi(x)\| \leq C|\mathfrak{g}|\|x\| \end{aligned} \tag{3.17}$$

where $|\mathfrak{g}|$ denotes the magnitude of the group element $\mathfrak{g}$ (i.e. measure of the group action), and C is some constant.

Corollary The Lipschitz continuity of Φ reflects a property known as *geometric stability* [BSL13], which can be visualized in Fig. 3.3. Sounds in the inner space $\mathcal{X}(\mathfrak{G}_\tau)$ are linearized and contracted in $\Phi(\mathcal{X}(\mathfrak{G}_\tau))$. Distances between sounds in the space are thus predictably bounded to the contracted Φ-space.

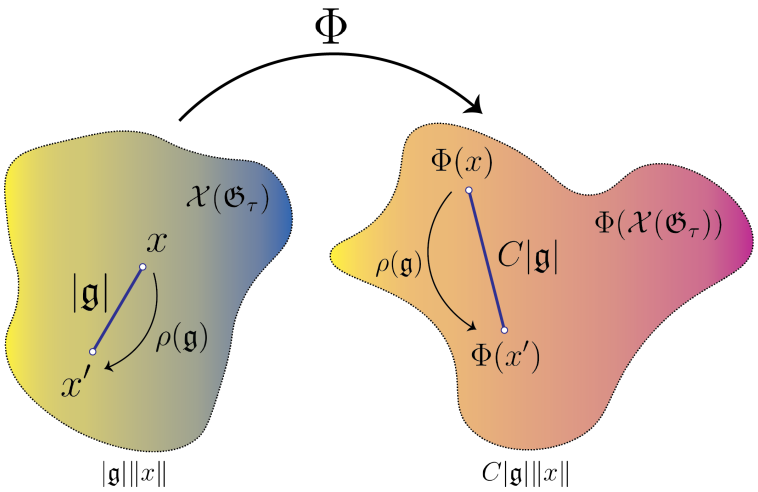


Figure 3.3: Boundary conditions for geometric stability in measures of distance between sounds x and x' both in their representational space $\mathcal{X}(\mathfrak{G}_\tau)$ and their Φ-space.

Example We might say that the operator Φ yields what is commonly known as the *feature vector*. For instance, an example of a contractive operator might be the calculation of *Mel-Frequency Cepstral Coefficients* (MFCCs), which are coefficients derived from the STFT that generalize the energy distribution of the magnitude spectrum. Training a neural audio model f on MFCC representations $\Phi^{\text{MFCC}}(x)$ is one of many common approaches to contracting the 'sound object' space.

3.3.3 Deformation

The formal properties of invariance and equivariance are used to describe group transformations under abstract mathematical settings (see Appendix A: A.2.2). When applied to real world phenomena such as sounds, we've seen that these ideal properties become difficult to analyze due to the complexity of the underlying domain. Because of this complexity, a notion of deformation stability—often called *approximate invariance*—is

thus introduced to account for the type of invariance and equivariance we might be referring to in a practical settings [BBCV21]. This notion of an approximate invariance stems from the fact that the constant C can stand as a threshold for invariance with respect to the actions of the group. The Lipschitz Continuity property defined in Eq. 3.17 alludes to the fact that f is locally invariant to the actions of $\mathfrak{G}_\tau$ if $|\mathfrak{g}| < C$ [Mal16].

Definition (Approximate Invariance) A function $f \in \mathcal{F}$ is approximately invariant if

$$\|f(\rho(\tilde{\mathfrak{g}})x) - f(x)\| \leq Cc(\tilde{\mathfrak{g}})\|x\| \quad \forall x \in \mathcal{X}(\mathfrak{G}_\tau) \tag{3.18}$$

where $\tilde{\mathfrak{g}}$ is a small perturbation of x that may or may not be an element of $\mathfrak{G}_\tau$ and c is a "complexity measure" where $c(\tilde{\mathfrak{g}}) = 0$ if $\tilde{\mathfrak{g}} \in \mathfrak{G}_\tau$. The complexity measure generalizes invariance in $\mathfrak{G}_\tau$ by defining a domain specific function. One such function is the elasticity function (Eq. 3.19), introduced in [BBCV21] alongside their definition of approximate invariance.

$$c(\tilde{\mathfrak{g}}) = \int_{\mathfrak{G}_\tau} \|\nabla\tilde{\mathfrak{g}}(u)\|^2 d\mu(u) \tag{3.19}$$

3.4 Mapping The Sound Object

Our approaches for mapping representations of sound thus far have involved linearizing $\mathcal{X}(\mathfrak{G}_\tau)$ through contractive dimensionality reduction. While this accounts for capturing local group symmetries, it does not take into account symmetries at different scales [Mal89] [BBCV21]. To mitigate this issue, we might imagine a compositional chain of linear projections $f_{(1)} \circ f_{(2)} \circ ... \circ f_{(N)}$ where each layer $f_{(i)} \in \mathcal{F}$ accounts for the preservation of different group symmetries. This compositional chain describes the construct of the *Equivariant Neural Network*, notably studied by [FWW21], [BSL13], [Rav20], [MFSL19] and others.

In this section, we introduce the equivariant neural network by first discussing its relationship to representational disentanglement via the *Peter-Weyl Theorem*, and then analyze invariant and equivariant networks using commutative diagrams.

3.4.1 The Peter-Weyl Theorem

The selection of an L^2 function space as our representational space for sound objects follows [BBCV21] and their application of L^2 space to a proposed space of signals. L^2 space is a suitable function space for many reason, but can perhaps be best described using the Peter-Weyl Theorem.

Theorem (Peter-Weyl) The space of square-integrable functions on $\mathfrak{G}_\tau$ is the direct sum over finite-dimensional irreducible representations V, denoted as a direct sum of endomorphisms (see Appendix A: A.2.1) of V:

$$L^2(\mathfrak{G}_\tau) \cong \bigoplus_{l=1}^{L} \mathrm{End}(V) \tag{3.20}$$

Remark What the Peter-Weyl theroem implies is that the compact topological group $\mathfrak{G}_\tau$ allows us to compose functions $f \in \mathcal{F}$ where each function corresponds to a subspace $W \subseteq V$ responsible for representing a certain group symmetry $\rho(\mathfrak{g})|_W \leq \rho(\mathfrak{g})|_V$. In the language of [HAP+18] we *fully disentangle* the vector space V (Appendix A: A.4.2) which we use to represent our topological group $\mathfrak{G}_\tau$ using a direct sum of irreducible components.

3.4.2 Equivariant and Invariant Networks

The result of the Peter-Weyl Theorem gives way to the notion of the *equivariant neural network*. Equivariant neural networks are neural networks constructed through crafting a complete disentanglement of a representation into its irreducible components, and can be interpreted as a chain of linear projections $f \in \mathcal{F}$ in which a group symmetry is preserved in each layer of the chain. We introduce the equivariant architecture first, followed by the invariant architecture [Wei22].

Definition (Equivariant Network) An equivariant neural network is a sequence of equivariant layers each preserving a different linear group action (group representation) $\rho(\mathfrak{g})^l$

$$\begin{array}{ccccccccc}
\mathcal{X}(\mathfrak{G}_\tau^{(1)}) & \xrightarrow{f_{(1)}} & \mathcal{X}(\mathfrak{G}_\tau^{(2)}) & \xrightarrow{f_{(2)}} & \mathcal{X}(\mathfrak{G}_\tau^{(3)}) & \xrightarrow{f_{(3)}} & \cdots & \xrightarrow{f_{(L)}} & \mathcal{X}(\mathfrak{G}_\tau^{(L)}) \\
\downarrow{\scriptstyle \rho(\mathfrak{g})^{(1)}} & & \downarrow{\scriptstyle \rho(\mathfrak{g})^{(2)}} & & \downarrow{\scriptstyle \rho(\mathfrak{g})^{(3)}} & & & & \downarrow{\scriptstyle \rho(\mathfrak{g})^{(L)}} \\
\mathcal{X}(\mathfrak{G}_\tau^{(1)}) & \xrightarrow{f_{(1)}} & \mathcal{X}(\mathfrak{G}_\tau^{(2)}) & \xrightarrow{f_{(2)}} & \mathcal{X}(\mathfrak{G}_\tau^{(3)}) & \xrightarrow{f_{(3)}} & \cdots & \xrightarrow{f_{(L)}} & \mathcal{X}(\mathfrak{G}_\tau^{(L)})
\end{array} \tag{3.4.1}$$

Definition (Invariant Network) An invariant neural network is a sequence of equivariant layers followed by an invariant mapping

$$\begin{array}{ccccccc}
\mathcal{X}(\mathfrak{G}_\tau^{(1)}) & \xrightarrow{f_{(1)}} & \cdots & \xrightarrow{f_{(L)}} & \mathcal{X}(\mathfrak{G}_\tau^{(L)}) & \xrightarrow{\Phi} & \\
\downarrow{\scriptstyle \rho(\mathfrak{g})^{1}} & & & & \downarrow{\scriptstyle \rho(\mathfrak{g})^{L}} & & \mathcal{X}^*(\mathfrak{G}_\tau^{(1)}) \xrightarrow{f^*_{(1)}} \cdots \xrightarrow{f^*_{(L)}} \mathcal{X}^*(\mathfrak{G}_\tau^{(L)}) \\
\mathcal{X}(\mathfrak{G}_\tau^{(1)}) & \xrightarrow{f_{(1)}} & \cdots & \xrightarrow{f_{(L)}} & \mathcal{X}(\mathfrak{G}_\tau^{(L)}) & \xrightarrow{\Phi} &
\end{array} \tag{3.4.2}$$

Remark These interpretations of the neural network treat each layer as a mapping f between function spaces $\mathcal{X}(\mathfrak{G}_\tau^{(l)})$ whose underlying groups may differ. Group transformations $\rho(\mathfrak{g})$ act as a function from a given space to itself that can traverse the geometric orbit of each space, once again reflecting [KPB+23]'s interpretation of a latent space.

Chapter 4

Neural Audio Typology

This section explores neural audio synthesis from the perspective of Schaeffer's concept of *typology*. We've observed thus far that a typology of sound must entail a classification based on a broad range of parameters that deviate from traditional musical parameters. With a continued emphasis on the sound object's geometry, we show in this section that time-frequency analysis can provide a group representation that parameterizes the sound object based on invariant group transformations. We first look at the Weyl-Heisenberg group and its representation using the windowed Fourier dictionary, and then observe its affinized representation using the wavelet dictionary. We show that the wavelet dictionary can be parameterized in such a way that produces a locally stable representation of the sound object, as well as a representation that is invariant to time-translation and time-warping.

This invariant representation is known as the *Scattering Transform* [Mal12a], a representation proven to be geometrically analogous to the convolutional neural network [Mal16]. We subsequently show how the scattering transform can be further extended to produce a frequency-transposition invariant representation using *Joint Time-Frequency Scattering* (JTFS), and look at examples of how the JTFS representation can be used for a typological analysis of mesostructures.

4.1 Typological Sound Representations

Parametric estimation is at the crux of a wide variety of problems in the domain of audio signal processing. Certain analyses of audio might represent perceptually relevant features

as well-localized points in Euclidean space [PGS+11]. In a group theoretical context, one can interpret these as representations of the affine group (see 3.1.2) due to their preservation of affine transformations across functions defined over the space. Other parametric analyses might yield representations that describe low-level sinusoidal interaction in a signal, which might be less perceptually descriptive when dealing with non-stationary signals but more practical for the reconstruction or "resynthesis" of signals over time [AKZB11]. These analyses serve to represent the Weyl-Heisenberg group, a subgroup of the Heisenberg group from section 3.5 that is invariant to time-translation and frequency modulation.

In this section, we use the group representational language established in the previous chapter to describe the typology of the sound object. We introduce the Weyl-Heisenberg group $\mathfrak{W}_\tau$ whose group representation $\mathcal{X}(\mathfrak{W}_\tau)$ forms a set of elementary functions called a *dictionary* [Mal09]. We show how the *Short-Time Fourier Transform* is derived from a dictionary that serves as one of many possible representations of sounds $x \in \mathcal{X}(\mathfrak{W}_\tau)$. We then show that by using an alternate dictionary such as the wavelet dictionary, an affine multiresolution representation of sounds $x \in \mathcal{X}(\mathfrak{A}_\tau)$ can be derived, contracting the 'sound object' space into a domain that is well-localized in both space and frequency [Mal89]. We later review Schaeffer's writings on typology and argue that these multiresolution approaches model a typology of the sound object.

4.1.1 Fourier Representations and The Weyl-Heisenberg Group

Dennis Gabor showed that the problem of time-frequency localization is closely related to our perception of sound [Gab46]. Since the work of Gabor, the parameterization of the time-frequency plane has remained integral to tasks such as sound reconstruction and sound matching [AKZB11] [HLL23]. Following our group theoretical treatment of the neural network in Chapter 3, a group theoretical interpretation of time-frequency analysis further augments our understanding of how the Schaefferian sound object can be represented using neural audio models. [Cel17] shows that in the context of sound synthesis, time-frequency representations often entail the use of high-dimensional group representations that usually relate to geometric transformations of fairly low abstraction (e.g. translation) that can be defined over Lie groups. We therefore start by showing how time-frequency representations can be thought of as group representations.

In time-frequency analysis, it is common to introduce the construct of a dictionary, which is a set of functions $\mathcal{X} = \{\chi_{u,\xi}\}_{u,\xi\in\Lambda}$, that acts as the group representation $\mathcal{X}(\mathfrak{G}_\tau)$ for x. We can thus project a sound x onto the set of functions in the dictionary by defining an operator Φ_x, shown below in Eq. 4.1:

$$\Phi_x(u,\xi) = \int_{t\in\mathbb{R}} x(t)\overline{\chi_{u,\xi}(t)}dt = \langle x, \chi_{u,\xi}\rangle \tag{4.1}$$

where each function $\chi_{u,\xi} \in \mathcal{X}$ is called a *time-frequency atom*, an elementary function indexed by $u, \xi \in \Lambda$ [Mal09] onto which we decompose the sound. [Fla99] shows that energy conservation resulting from the signal's projection can be directly related to the Haar measure on the group $\mathfrak{G}_\tau$. Let $\mu_{\mathfrak{G}_\tau}$ denote the Haar measure on $\mathfrak{G}_\tau$ and let the set of atoms $\mathcal{X}$ act as the *group generator* for $\mathfrak{G}_\tau$. The conservation of energy E_x of the sound x can then be derived from the Haar measure as such.

$$E_x = \iint_{u,\xi\in\mathbb{R}^2} |\Phi_x(u,\xi)|^2 d\mu_{\mathfrak{G}_\tau}(u,\xi) \tag{4.2}$$

As a measure on the group $\mathfrak{G}_\tau$, E_x is directly related to the underlying group transformations $\mathfrak{g} \in \mathfrak{G}_\tau$ [Fla99], which are furthermore informed by the uncertainty principle in time-frequency localization shown in [LC04]. As a continuous signal, this makes the possible representations of $\mathfrak{G}_\tau$ extremely redundant. A fundamental example of this can be shown in the relationship between the windowed Fourier atom and the windowed Fourier Transform [Mal09]. The windowed Fourier Transform is a collection of atoms that yield the following projection, which we denote $\Phi_x^{\mathcal{F}}$.

$$\begin{aligned} \chi_{u,\xi}(t) &= e^{i2\pi\xi t} w(t-u) \\ \Phi_x^{\mathcal{F}}(u,\xi) &= \int_{t\in\mathbb{R}} x(t) w(t-u) e^{-i2\pi\xi t} dt \end{aligned} \tag{4.3}$$

where w is a window function of arbitrary size. Here, the operator Φ_x fully defines the *Short-Time Fourier Transform* (STFT). Following the energy conservation established by Eq. 4.2, an energy spectral density can also be derived by squaring the magnitude. We might define a more specific operator $E_x^{\mathcal{F}}$ that operates on a sound x to derive its energy density.

$$E_x^{\mathcal{F}}(u,\xi) = \left| \int_{t\in\mathbb{R}} x(t)w(t-u)e^{-i2\pi\xi t}dt \right|^2 \tag{4.4}$$

Previously we introduced the Heisenberg group $\mathfrak{H}_\tau$ in Example 3.5—a subgroup of $\mathrm{GL}(3,\mathbb{R})$ that can be conveniently represented as a complex exponential (Eq. 3.9). The Weyl-Heisenberg group is a subgroup of the Heisenberg group that only includes the transformations of translation $\mathfrak{t}$ and modulation $\mathfrak{m}$. This modification allows us to interpret the windowed Fourier transform as a representation of $\mathfrak{W}_\tau$. This can be shown by interpreting the set of atoms in $\mathcal{X}$ as group generators belonging to $\mathfrak{W}_\tau$. This also implies that all functions $f \in \mathcal{F}$ are invariant to the Weyl-Heisenberg group's transformations of translation and modulation, which we continue to denote $\rho(\mathfrak{t})x(t) = x(t-\mathfrak{t})$ and $\rho(\mathfrak{m})x(t) = x(t)e^{i2\pi\mathfrak{m}t}$ following Eq. 3.9. The Weyl-Heisenberg representation of a sound $x \in \mathcal{X}(\mathfrak{W}_\tau)$ ensures that for any window function $w \in L^2(\mathbb{R})$ and any shift factor $\tau \in \mathbb{R}$:

$$\begin{aligned} \langle x(t-\tau), \rho(\mathfrak{t})w\rangle &= \langle x, \rho(\mathfrak{t}-\tau)w\rangle \\ \langle x(t)e^{i2\pi\tau t}, \rho(\mathfrak{m})w\rangle &= \langle x, \rho(\mathfrak{m}-\tau)w\rangle \end{aligned} \tag{4.5}$$

We redirect readers to texts such as [Won02] and [Jan98] for more formal derivations of the Weyl-Heisenberg group's transformations.

4.1.2 Wavelet Representations and the Affine Group

While the Weyl-Heisenberg group might suitably characterize the low-level structures of a sound through its group transformations, its energy distribution is proven to be unstable under small deformations. [BBCV21] shows this by evaluating affine translations on the spectrogram. Working from our spectrogram $E_x^{\mathcal{F}}(u,\xi)$, we now define an affine coordinate system $u, v \in \Lambda$ where v represents affine coordinates over the frequency axis such that the spectrogram is now indexed as $E_x^{\mathcal{F}}(u,v)$. Working from these new coordinates, let the actions $(\mathfrak{t}_T, \mathfrak{t}_F) \in \mathfrak{A}_\tau$ be the actions of translation on both axes of the plane, such that $\rho(\mathfrak{t}_T,\mathfrak{t}_F)E_x^{\mathcal{F}} = E_x^{\mathcal{F}}(u-\mathfrak{t}_T, v-\mathfrak{t}_F)$.

To evaluate the geometric stability of the spectrogram, [BBCV21] then introduces the notion of *approximate translation*. Approximate translation is a function of both time $\tilde{\mathfrak{t}}_T(u)$ and frequency $\tilde{\mathfrak{t}}_F(v)$ that measures the difference between any 'approximate' translation $\tilde{\mathfrak{t}}$

and true translation $\mathfrak{t}$ by subtracting the approximate translation from its orginal coordinate. For instance, true translation over time can be put in terms of approximate translation by declaring $\mathfrak{t}_T(u) = u - \tilde{\mathfrak{t}}_T(u)$. The same definition can be derived for frequency, such that $\mathfrak{t}_F(v) = v - \tilde{\mathfrak{t}}_F(v)$. Looking at the maximum value of the gradient of each action can then be put in terms of its approximate action such that $\|\nabla \mathfrak{t}_T\|_\infty = \sup_{u\in\Omega}\|\tilde{\mathfrak{t}}_T(u)\| \leq \epsilon_T$ and $\|\nabla \mathfrak{t}_F\|_\infty = \sup_{v\in\Omega}\|\tilde{\mathfrak{t}}_F(v)\| \leq \epsilon_F$, where ϵ_T and ϵ_F each represent an upper bound on each gradient's maximum value.

Importantly, this relationship states that the closer $\mathfrak{t}_T$ or $\mathfrak{t}_F$ is to a shift in time or frequency, the smaller the upper bounds ϵ_T and ϵ_F become, therefore making implicit a notion of geometric stability for the joint action $\rho(\mathfrak{t}_T, \mathfrak{t}_F)$. [BBCV21] then use this to show that the application of $\rho(\mathfrak{t}_T, \mathfrak{t}_F)$ onto a spectrogram $E_x^{\mathcal{F}}(u, v)$ yields an unstable result—i.e. a result that is not bound by ϵ_T and ϵ_F:

$$\frac{\|\rho(\mathfrak{t}_T, \mathfrak{t}_F)E_x^{\mathcal{F}} - E_x^{\mathcal{F}}\|}{\|x\|} = \mathcal{O}(1) \tag{4.6}$$

where $\mathcal{O}$ represents magnitude of the deformation at point u, v, which in this case is a constant value instead of a value proportional to the upper bounds of $\mathfrak{t}_T$ and $\mathfrak{t}_F$. In other words the Weyl-Heisenberg group yields an energy distribution that causes non-rigid transformations by functions $f \in \mathcal{F}$ [BBCV21], resulting in a lack of stability and localization of energy in both time and frequency.

A remedy for this issue is to use a different time-frequency dictionary, since not all dictionaries are representations of the Weyl-Heisenberg group. For instance, the wavelet dictionary bridges this geometric gap by defining a representation that is localized in both time and frequency [Mal89]. Wavelets are defined by the following dictionary:

$$\mathcal{X} = \left\{ \frac{1}{\sqrt{a}}\, \psi\left(\frac{t-b}{a}\right) \right\}_{a,b\in\Lambda} \tag{4.7}$$

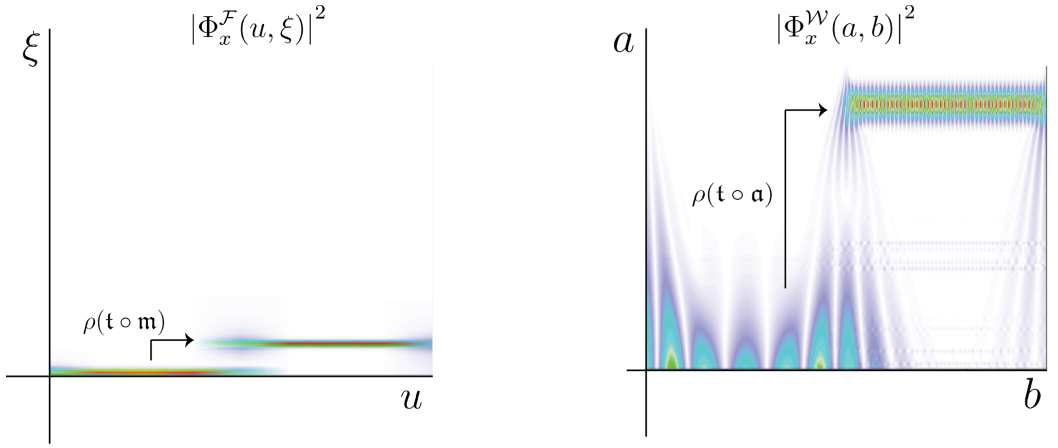
where now instead of time and frequency coordinates $u, v \in \Lambda$ we use coordinates denoting scale a and shift b. The wavelet representation can be derived similarly to the Fourier representation in Eq. 4.3:

$$\Phi_x^{\mathcal{W}}(a, b) = \frac{1}{\sqrt{a}} \int_{t\in\mathbb{R}} x(t)\psi^*\left(\frac{t-b}{a}\right) dt \tag{4.8}$$

This representation similarly yields an energy distribution called the wavelet *scalogram*, which we denote $E_x^{\mathcal{W}}(a,b) = |\Phi_x^{\mathcal{W}}(a,b)|^2$. [BBCV21] notes that this allows for a decomposition that is *approximately equivariant* to deformations, as shown in [Mal12a]. In this case, we can measure approximate translations once again and show the following:

$$\frac{\|\rho(\mathfrak{t}_T, \mathfrak{t}_F)E_x^{\mathcal{W}} - E_x^{\mathcal{W}}\|}{\|x\|} = \mathcal{O}(\epsilon_T, \epsilon_F) \tag{4.9}$$

where the magnitude of deformation of ϵ_T and ϵ_F is in proportion to the joint action $\rho(\mathfrak{t}_T, \mathfrak{t}_F)$, unlike the scenario in Eq. 4.6. Importantly, this implies that $E_x^{\mathcal{W}}$ is a contractive linear operator that is also Lipschitz continuous (Eq. 3.17) [Mal16], yielding a representation of x that contains a family of locally stable features. These locally stable features are the result of dialating and translating the atoms $\psi \in \mathcal{X}$. Such affine modifications to the Weyl-Heisenberg group representation can also be demonstrated visually in the resulting scalograms (Fig. 4.1).



(a) $\mathfrak{W}_\tau$ group transformations of translation and modulation on its STFT representation

(b) $\mathfrak{A}_\tau$ group transformations of translation and scaling on its Mel-Scalogram representation

Figure 4.1: Group actions over two contrasting time-frequency representations.

With the introduction of an additional scaling function acting as a low-pass filter, we can further augment this affinized representation of the Weyl-Heisenberg group to form an orthonormal basis in L^2 space. Importantly, this means that the resulting representation of x can also act as a disentangled representation (Eq. A.15), which we have seen in Section 3.4

is desirable for neural audio synthesis due to its inherent association of group actions with representational subspaces. [Mal89] describes the orthonormality conditions for the wavelet basis functions parameterized in terms of translation and dilation. This is best shown by discretizing the parameters of the wavelet function ψ such that $j, k \in \mathbb{Z}$ now represent discrete dilation and translation respectively:

$$\psi_{j,k\in\mathbb{Z}}(t) = 2^{\frac{j}{2}}\psi(2^j t - k) \tag{4.10}$$

We once more introduce the additional a scaling function ϕ that similarly abides by the same translation and dilation properties as the wavelet atom ψ

$$\phi_{j,k\in\mathbb{Z}}(t) = 2^{\frac{j}{2}}\phi(2^j t - k) \tag{4.11}$$

The function ϕ is associated with nested subspaces $\cdots V_{j-1} \subset V_j \subset V_{j+1} \cdots$ for each scale factor j. Translations k then span each subspace such that the space is dense and complete:

$$\begin{aligned} \bigcup_{j=-\infty}^{+\infty} V_j &= L^2(\mathbb{R}) \\ \bigcap_{j=-\infty}^{+\infty} V_j &= \{0\} \end{aligned} \tag{4.12}$$

At each scale j, the wavelet function ψ is responsible for capturing the local details of the space. Furthermore, each subspace $V_j \in L^2(\mathbb{R})$ can be interpreted as a direct sum $V_j = V_{j-1} \oplus W_{j-1}$, meaning that W_j contains complementary information to analyze x at different resolutions. Assuming that V_0 captures the coarsest detail of x, we can see that these functions provide the following disentangled representation of $L^2(\mathbb{R})$:

$$L^2(\mathbb{R}) = V_0 \oplus \bigoplus_{j=0}^{\infty} W_j \tag{4.13}$$

As a correlary, this also shows that $\phi_{j,k}$ and $\psi_{j,k}$ form an orthonormal basis, meaning that $\langle \phi_{j,k}, \phi_{j',k'} \rangle = \delta_{jj'}\delta_{kk'}$ and $\langle \psi_{j,k}, \psi_{j',k'} \rangle = \delta_{jj'}\delta_{kk'}$. In practice, this means that a temporal representation of a sound $x \in \mathcal{X}(\mathfrak{T}_\tau)$ can be decomposed onto this subspace and—through application of this additional scaling function ϕ—yield a time-frequency representation of

x that is not only *translation invariant* like the Fourier transform, but also geometrically stable and separable.

4.2 Wavelet Scattering

The multiresolution properties of the wavelet basis allow for the derivation of many possible affine representations of x, ideal for a geometric analysis of the Schaefferian sound object. In this section, we introduce one notably flexible representation derived using *Wavelet Scattering* [Mal12b] [AM14]. The scattering transform yields a representation of x that is not only time-translation invariant but also stable to time-warping deformations—a property that allows the transform to capture structural properties such as *amplitude modulation* (AM). The scattering transform can also be interpreted as an invariant neural network, where the weights are fixed instead of learned [AM14], making it functionally analogous to the convolutional neural network [Mal16]. We elaborate by also introducing the *Joint Time-Frequency Scattering* (JTFS) transform [ALM19], which extends the scattering transform to the frequency axis to produce a representation that is additionally frequency-translation invariant and stable to frequency-warping, allowing the transform to capture structural properties such as *frequency modulation* (FM). We finish by reviewing [CHC+23] which introduces a loss function that measures the square of the Euclidean distance between JTFS representations of sounds, an inference framework that we argue is fundamentally typological.

4.2.1 Scattering Networks

A scattering transform of x first involves a projection onto the wavelet basis parameterized by time $t \in \mathbb{R}$ and log-frequency $\lambda \in \mathbb{R}$. Using a complex Morlet wavelet $\psi_\lambda = 2^\lambda \psi(2^\lambda t)$, we similarly decompose x by convolving across time and taking the modulus of the result.

$$\mathcal{S}_x(t, \lambda) = |x * \psi_\lambda(t)| \tag{4.14}$$

Equation 4.14 defines a wavelet scalogram, similar to $E_x^{\mathcal{W}}$ except derived in this instance from a complex wavelet and adjusted to log-frequency scale. As seen in the previous section, utilizing a lowpass filter ϕ can provide a representation that is invariant to time-translation:

$$\mathcal{S}_x^1(t,\lambda) = |x * \psi_\lambda(t)| * \phi_T(t) \tag{4.15}$$

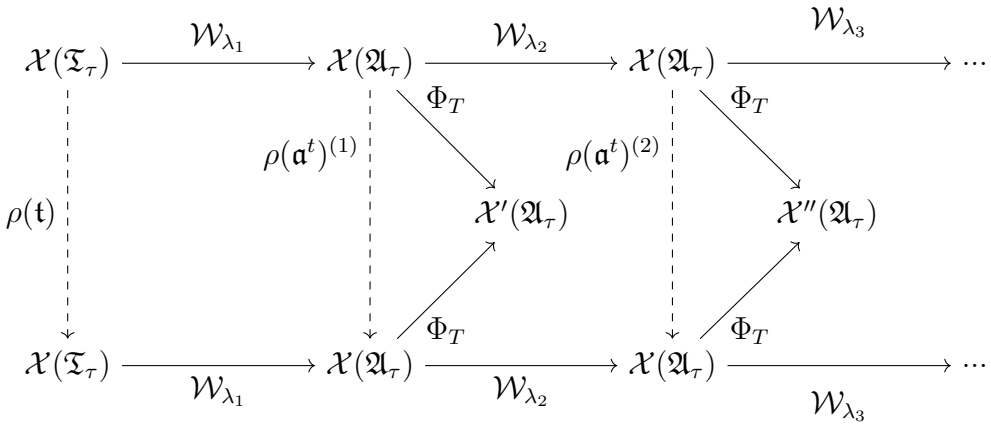
In this case, ϕ is scaled by some constant duration T where $\phi_T(u) = T^{-1}\phi(T^{-1}\phi(u))$. This representation is analogous to the mel-spectrogram—a variation on the spectrogram that maps frequencies to the mel-scale to reflecting our non-linear perception of pitch—however, we follow [AM14] by instead calling this equation a 'first-order' scattering transform. The first-order scattering transform is invariant to time-translation and approximately invariant to time-warping deformations, by way of adjusting the duration parameter T. [AM14] additionally shows that convolving by another wavelet and low-pass filter can further represent variability in the signal by capturing amplitude modulations. This augmentation is then referred to as the 'second-order' scattering transform:

$$\mathcal{S}_x^2(t,\lambda_1,\lambda_2) = ||x * \psi_{\lambda_1}(t)| * \psi_{\lambda_2}| * \phi_T(t) \tag{4.16}$$

The process of wavelet scattering has been shown in many works to be a functional analogue to the CNN [Mal16] [BSL13]. Using the invariant network diagram from Chapter 4, we can interpret the scattering transform as a series of equivariant layers followed by an invariant contraction. In order to do this, we first construct the following operators for both the wavelet and low-pass filter operations:

$$\begin{aligned} \mathcal{W}_{\lambda_n}\{x\} &= |x * \psi_{\lambda_n}(t)| \\ \Phi_T\{x\} &= |x * \phi_T(t)| \end{aligned} \tag{4.17}$$

where $\mathcal{W}_{\lambda_n}$ represents an equivariant layer of order n and Φ_T represents an invariant contraction. A scattering transform of arbitrary order can then be written as a commutative diagram:



(4.2.1)

In diagram 4.2.1, translation equivariance $\rho(\mathfrak{t})$ between the original sound $x \in \mathcal{X}(\mathfrak{T}_\tau)$ and the resulting scalogram are captured by the first-order operator $\mathcal{W}_{\lambda_1}$. Affine transformations on the time axis denoted $\rho(\mathfrak{a}^t)(n)$ are subsequently made invariant by the contractive operator Φ_T, resulting in sounds $x \in \mathcal{X}'(\mathfrak{A}_\tau)$. The process is repeated at the second-order to capture AM, resulting in a second space $x \in \mathcal{X}''(\mathfrak{A}_\tau)$. While the commutative diagram above shows that the process could be repeated for higher-order scattering transforms ($\mathcal{W}_{\lambda_3}$), it has been shown that these are negligible in the context of sound classification [Wal17]. Therefore, a suitable scattering representation for sound is commonly defined simply as a concatenation of the first-order and second order scattering coefficients.

$$E_x^{\mathcal{S}}(t, \lambda_1, \lambda_2) = \mathcal{S}_x^1(t, \lambda_1) \oplus \mathcal{S}_x^2(t, \lambda_1, \lambda_2) \tag{4.18}$$

The complete scattering representation (Eq. 4.18) can thus capture time-warping invariance through its ability to capture AM via affine scaling along the time axis. Fig. 4.2 shows an example of the energy distribution derived from the first-order scattering transform computed using the Kymatio library for Python [AAE+22]. The scattering transform was performed on a recording of a large aluminium triangle approximately 50cm in length on each side. The triangle was struck once while suspended in air with fishing wire, naturally decaying until fully damped, thus creating low-frequency amplitude

modulations as a result of its rotations in space[1].

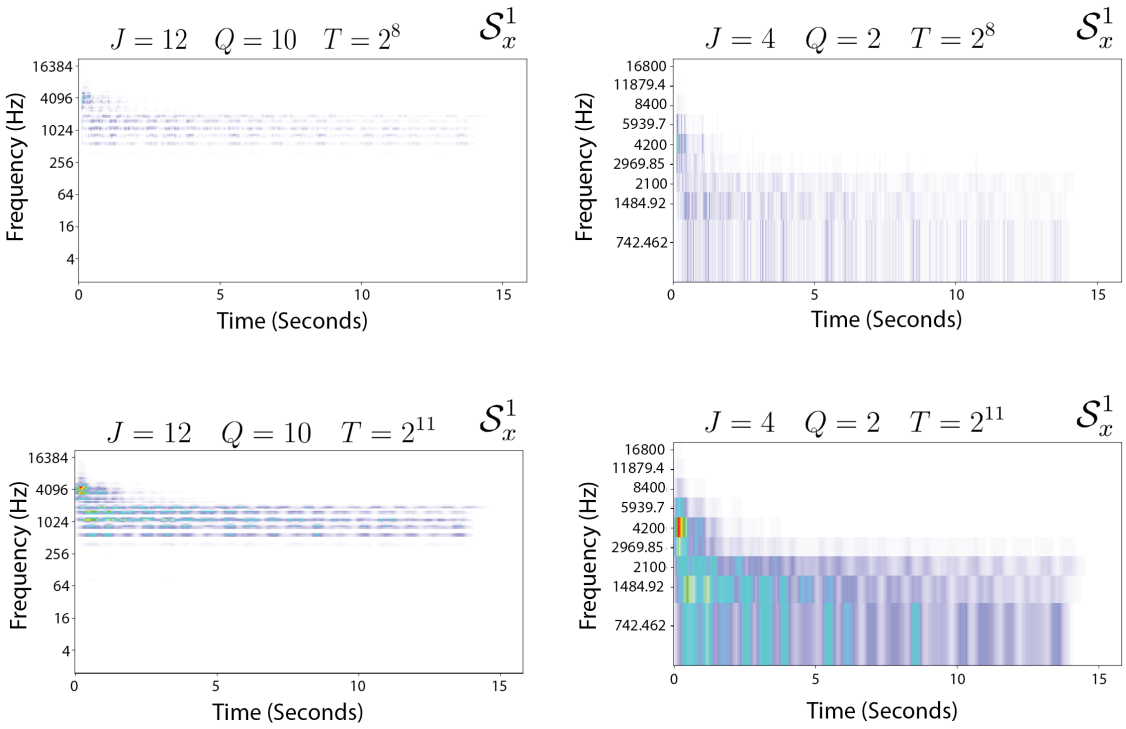


Figure 4.2: First-order scattering transform of a large aluminum triangle at two different resolutions. Increasing T produces a representation that is more geometrically stable to translation.

In practice, the scattering transform is calculated using different wavelet resolutions with the coefficients J and Q, from which we derive the wavelet scale factor λ. In this case, J dictates the number of octaves in our wavelet filter bank, while Q dictates the number of filters per octave. Comparing the left and right columns in Fig. 4.2, the resulting scalograms on the left contain much finer details due to the greater numbers for J and Q. We also observe that increasing the duration parameter T augments the scale of time-invariance. AM thus becomes globally clearer in the scalograms when increasing the duration parameter from 2^8 to 2^{11}.

While a geometrically stable representation of the triangle can be captured with the first-

[1]This recording is presented as the first preliminary example on the companion site for this thesis: `https://acousmatic-ddsp.netlify.app/`

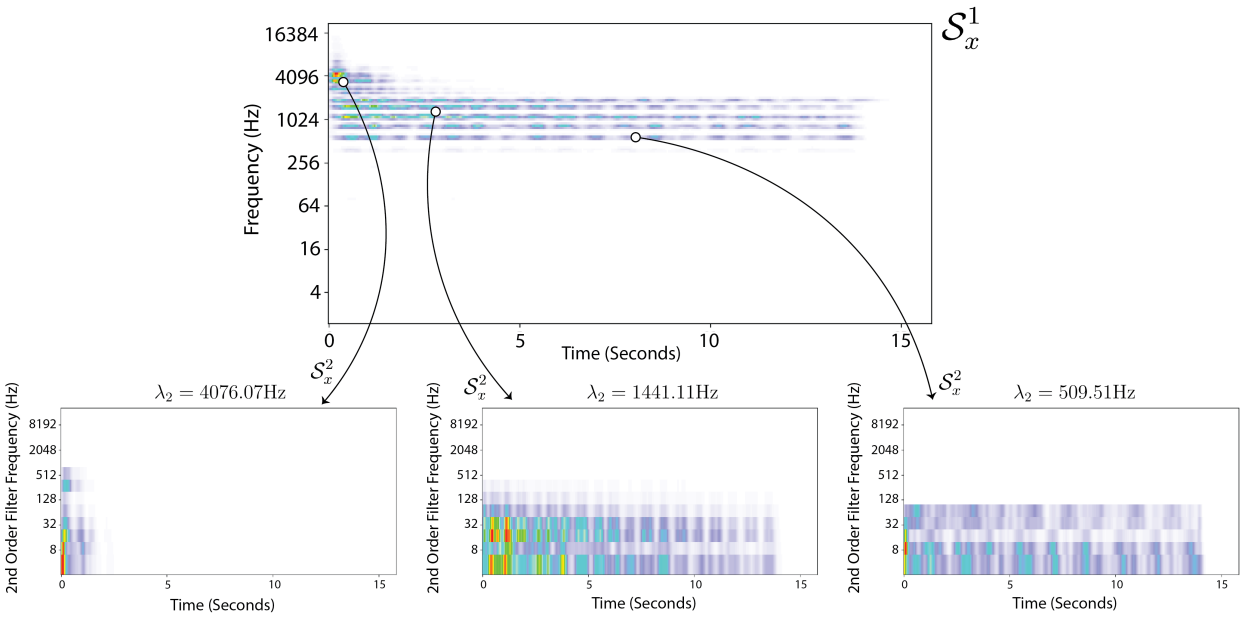


Figure 4.3: Second-order scattering transform of a large aluminum triangle at three different resolutions. AM caused by the rotation of the triangle is captured when centering the second-order filters around different frequencies.

order scattering transform, greater structural details such as the triangle's low-frequency amplitude modulations can be captured using a second-order scattering transform[2]. In this example, we derive three different second-order energy distribution where the second-order filters are centered around three different frequencies from the first-order transform. At each of these first-order frequency values, we are able to derive a corresponding second-order energy distribution in which we can observe a better representation of the amplitude modulations active around that given frequency region. Fig 4.3 shows this process and how it captures amplitude modulations ranging from approximately 0Hz – 1024Hz. Most importantly, the second-order transform captures low-frequency AM patterns resulting from the suspended triangle's rotations in space at a better resolution. This can be seen most clearly in the distribution where $\lambda_2 = 509.51$Hz, where low-frequency oscillations occur

[2]Technically these rotations are identical to $\mathfrak{r}$, the same as those of the group $\mathfrak{D}_3$ from Section A.1.2. While we won't formally prove it, one could infer that the time-warping invariance described in second-order scattering is related to the transformations $\rho(\mathfrak{r}), \rho(\mathfrak{s}) \in \mathcal{X}(\mathfrak{D}_N)$.

periodically around 0Hz − 8Hz. Finally, the distribution where $\lambda_2 = 4076.07$Hz provides an informative representation of the signal's transient.

4.2.2 Joint Time-Frequency Scattering

With a slight modification of the second-order transform $\mathcal{S}_x^2$, we can augment the structural properties captured by the scattering transform simply by extending its second-order translations and dilations into the frequency dimension [ALM19] [MVW+22], thus yielding the Joint Time-Frequency Scattering (JTFS) transform.

Taking a JTFS transform involves defining a two-dimensional wavelet, commonly denoted Ψ. The two-dimensional wavelet is parameterized by the tensor product of two one-dimensional wavelets, one parameterized by a temporal scale parameter—now denoted λ_2^T—and one parameterized by an additional frequency scale parameter λ_2^F as well as a spin parameter $\theta \pm 1$ that represents the slope of oscillation.

$$\Psi_{\lambda_2^T,\lambda_2^F,\theta}(t,\lambda_1) = (2^{\lambda_2^T}\psi(2^{\lambda_2^T}t)) \otimes (2^{\lambda_2^F}\psi(\theta 2^{\lambda_2^F}\lambda_1)) \tag{4.19}$$

We thus modify the second-order scattering transform to derive the JTFS transform:

$$\mathcal{S}_x^2(t,\lambda_1,\lambda_2^T,\lambda_2^F,\theta) = ||x * \psi_{\lambda_1}(t)| \overset{t,\lambda_1}{*} \Psi_{\lambda_2^T,\lambda_2^F,\theta}| \overset{t,\lambda_1}{*} \Phi_{T,F}(t,\lambda_1) \tag{4.20}$$

The introduction of translations and dilations in the frequency axis thus requires an additional constant F to be defined for our contractive operator Φ, corresponding to the constant T on the time axis. This also results in a two dimensional convolution, denoted $\overset{t,\lambda_1}{*}$, allowing for an invariance to translation along the frequency axis, as well as for an additional stability to frequency-warping. It's noted in [MVW+22] that omitting the frequency axis convolution $\overset{\lambda_1}{*}$ might also be desirable in order to preserve equivariance to frequency-transposition. Deriving the full tensor of scattering coefficients now becomes a concatenation of the output of both the first-order scattering transform and the second order JTFS transform, shown below:

$$E_x^{\text{JTFS}}(t,\lambda_1,\lambda_2^T,\lambda_2^F,\theta) = \mathcal{S}_x^1(t,\lambda_1) \oplus \mathcal{S}_x^2(t,\lambda_1,\lambda_2^T,\lambda_2^F,\theta) \tag{4.21}$$

With the introduction of frequency-translation and frequency-warping invariance, the

JTFS representation is now beneficial for modeling higher-order structures in the sound object given its ability to not only capture amplitude modulation in sound, but also capture frequency modulation. Fig. 4.4 shows a selection of scalograms derived from the JTFS transform performed using a recording of two strikes of a large suspended spring coil. Each strike was passed through a frequency shifting effect, the first one ascending in frequency and the second one descending in frequency. The selected scalograms represent subspaces of the JTFS transform where λ_2^F is tuned to different center frequencies, enabling the scalograms to represent the effects of FM at different frequency bandwidths. At $\lambda_2^T = 16.4$Hz and $\lambda_2^F = 16.8$kHz, frequency shifting is somewhat visible in the lower bandwidth of the scalogram. Higher tunings allow us to visualize the resonating frequencies of the strike, and slight differences in the magnitude of ascending versus descending energy. These slight differences are circled in Fig. 4.4, demonstrating the effect of setting $\theta = +1$ and $\theta = -1$.

4.2.3 Mesostructural Distance

'Mesostructure' is a term coined by Iannis Xenakis used to describe mid-level structural elements of music, in contrast to *micro* and *macro* structures [Xen92]. [Roa14] further elaborates on mesostructures as the structures that emerge from the grouping of sounds and their complex spectrotemporal evolution. This evolution can be dictated by a myriad of structural elements, spanning from traditional pitch and rhythmic sequences to parametric features of the frequency spectrum such as centroid, loudness, and harmonic energy.

The JTFS transform's ability to capture higher-order frequency and amplitude patterns has notably deemed it a *mesostructural* representation of sound by [CHC+23]. This interpretation further fits into our group theoretical framework of the sound object, in that the JTFS transform contracts and linearizes sounds by projecting them into an affine space in which Euclidean distance between different sounds $x \in \mathcal{X}(\mathfrak{A}_\tau)$ denotes mesostructural similarity. More formally, we might develop a distance metric to compare the Euclidean distance between mesostructural representations of two different sounds. This can be derived simply by taking the magnitude difference between two JTFS tensors:

$$\| E_{\tilde{x}}^{\text{JTFS}} - E_x^{\text{JTFS}} \|_2^2 \tag{4.22}$$

We can interpret Equation 4.22 in two different ways. The first way follows our group

theoretical approach, where it represents the Euclidean distance $\mathrm{d}(\cdot,\cdot)$ between sounds. This relates to our predefined notion of geometric stability (Eq. 3.17), in that the contractive nature of the JTFS transform in both time and frequency ($\Phi_{T,F}$) yields a domain that follows the boundary condition $C|\mathfrak{a}|\|x\|\ \forall x \in \mathcal{X}(\mathfrak{A}_\tau)$ and $\forall \mathfrak{a} \in \mathfrak{A}_\tau$. Assuming such a condition for geometric stability allows us to compare sounds based solely on affine geometry. Another interpretation of 4.22, however, relates directly to neural audio synthesis, in that this distance can serve as a metric or *loss function* that evaluates how well a neural audio model has 'learned' the mesostructures of a collection of sounds as a unified continuous *sound object*. This interpretation allows one to put the notion of mesostructural geometric stability into practice, given a collection of sounds.

4.2.4 JTFS as Schaefferian Typology

The JTFS representation has demonstrated its effectiveness across several practical applications, where the Euclidean distances derived from different JTFS representations suggest its potential as an implementation of Schaefferian typology. For instance, in [LEHR+21], short recordings of orchestral extended techniques were clustered based on distances between their JTFS representations. These clusters aligned closely with perceptual studies in which participants grouped the recordings according to perceived timbral similarity. Additionally, [LYY23] provides compelling evidence for the stability of the JTFS representation at the mesoscale. In this sound matching experiment, recordings from a 'chirplet' synthesizer were compared against their ground-truth parameters, demonstrating the JTFS transform's ability to capture mesostructural similarities between sounds and their generative models.

These examples underline the JTFS transform's capability to generalize the structure of acousmatic sound—sonic material without an identifiable source—thereby providing a concrete foundation for Schaeffer's concept of the sound object and its perceived invariance. In the following chapter, we will further develop the notion of the sound object by exploring morphology, which similarly involves the geometric localization of sonic similarities, except at scale of the microstructure, allowing for the possibility of a continuous control of acousmatic sound.

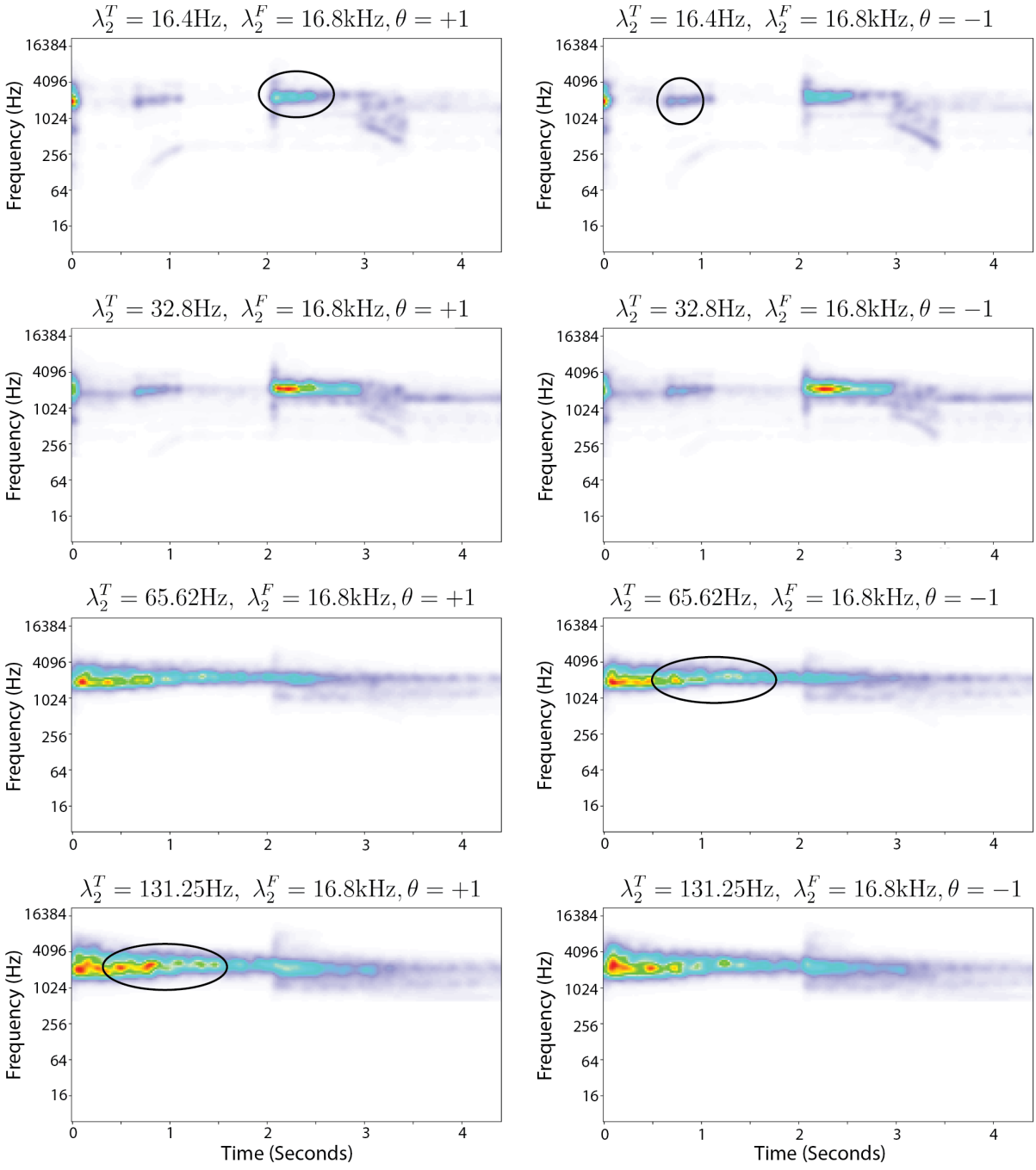


Figure 4.4: Selections of the JTFS transform performed on a recording of a large suspended spring coil with added ascending and descending frequency shifts. The frequency scale parameter λ_2^T is tuned to various different center frequencies while the spin parameter captures minor differences in ascending and descending frequency.

Chapter 5

Neural Audio Morphology

This section explores the Schaefferian concept of *morphology* in the context of neural audio synthesis. Following the previous section on typology, we shift our focus from the *mesostructural* to the *microstructural* by constructing representations that enable disparate sounds to seamlessly 'morph' into one another. This approach reinforces our claim that the Schaefferian 'sound object' should be represented as a continuous and differentiable manifold with the underlying structure of a Lie group. We thus introduce another affinized representation of the Weyl-Heisenberg group called the *Multiscale Spectrogram* (MSS) [EHGR20] [SM23], which adequately represents sound on the microscale while still acting as a contractive operator. When used in tandem with the parameters of a synthesizer, the MSS can serve as a distance metric or 'loss function' for optimizing a neural audio model $f : \mathcal{X}(\mathfrak{A}_\tau) \to \mathcal{X}(\mathfrak{A}_\tau)$, learning a continuous differentiable mapping between control parameters. The learned mapping is equivariant to affine transformations, as well as approximately equivariant to small time-warping transformations.

This morphological paradigm is more commoly known as *Differentiable Digital Signal Processing* (DDSP). We argue in this section that DDSP can be interpreted from the perspective of group representation theory as a mapping whose codomain approximately spans the orbit of the latent space $\Phi(\mathcal{X}(\mathfrak{G}_\tau))$. Ideally this span can be reached by choosing a set of control parameters that disentangles the latent space through an estimate of approximate perceptual independence [PGS+11].

5.1 Morphological Sound Representations

As stated in the *Traité*, Schaeffer proposes that a typology of the sound object is inherently linked to its own morphological features. This means that a morphology of sound object—i.e. an interpolation over the features of various sonic material—must be based off of a set of parameters resulting from typological analysis. More technically speaking, any contractive operator Φ used to project a set of sounds into an affine space will yield a parametric space that also reflects the conditions for a subsequent 'morphology' of those given sounds. In this section, we show that the typological distribution of sounds resulting from Φ can also be used to form a morphological representation of the Schaefferian 'sound object' simply by switching time-scale from the mesostructural level to the *microstructural* level [CHC+23].

5.1.1 Parametric Extractors

Microstructures can be thought of as structural elements of sound at a time-scale somewhere between a few samples and a few milliseconds [CHC+23]. These structures are inherently linked to computer music processes such as granular synthesis and wavetable synthesis, both of which deal with sonic material on the microscale [Roa01]. Due to their incredibly short-term nature, sounds on the microscale are often represented using only the Weyl-Heisenberg STFT representation. Given a windowing function w_T, where T is the length of a chosen microscale, we can reparameterize the STFT representation from Eq. 4.3 based on our microscale T.

$$\Phi_x^{\mathcal{F}}(u, \xi, T) = \int_{t \in \mathbb{R}} x(t + uH) w_T(t) e^{-i2\pi\xi t} dt \tag{5.1}$$

The reparameterization in Eq. 5.1 allows us to analyze x as a representation of the translation group $\mathcal{X}(\mathfrak{T}_\tau)$ iteratively over time, where w_T slides across time via some translation factor H, often called a *hop size*. As noted in the previous chapter, the STFT is not a geometrically stable representation of sound. However, much like the wavelet scalogram, one can apply a contractive operator on the STFT representation such that a well-localized statistical representation can be extracted at the microscale, ideally reflecting the spectrotemporal evolution of x over time in a low-dimensional space when concatenated across the time axis.

Treating the window function w_T as a sort of microstructural analogue to the low-pass filter ϕ_T from the scattering transform (Eq. 4.15), we begin by assuming for the remainder of the chapter that the Weyl-Heisenberg representation of a sound $x \in \mathcal{X}(\mathfrak{W}_\tau)$ is implicitly constrained to a given window $x \cdot w_T$ where $T \leq 8192$. From this assumption, a more stable representation of x can be derived by constructing a vector of low-dimensional parameters, each of which relate to the microsound's perceptual features such as harmonic and noise content (e.g. harmonic energy, noisiness), or to the statistical distribution of its frequency spectrum (e.g. centroid, kurtosis). A formalized list of such parameters can be found in [PGS+11], who compile a set of time-varying audio descriptors that describe the spectrotemporal evolution of sound (see Appendix B). This set of audio descriptors contains a list of parameters shown to be least correlated with one another, a result derived from various perceptual studies. The proposed perceptual independence of these spectrotemporal descriptors allows us to make the assumption that they might represent x from approximately independent subspaces. We can convey both their independence and their stability through the definition of an operator Γ that takes a microsound $x \cdot w_T$ and extracts a vector containing a subset of the time-varying audio descriptors.

$$\Gamma : \mathcal{X}(\mathfrak{W}_\tau) \to \bigoplus_{k=1}^{K} W_k(\mathfrak{A}_\tau) \tag{5.2}$$

The assumed perceptual disentanglement of the space $\bigoplus_{k\in\mathbb{Z}+} W_k(\mathfrak{A}_\tau)$ allows for a maximally expressive representation of the microsound $x \cdot w_T$ through the disentanglement of group actions $\rho|_{W_k}(\mathfrak{a})$ associated with each subspace W_k. The resulting codomain of Γ thus ideally yields a K-dimensional disentangled parametric space. Figure 5.1 represents this parametric extraction visually using an example in which the codomain contains a subspace of three parameters. We denote a point in this disentangled control space as $\mathbf{v} \in \bigoplus_{k\in\mathbb{Z}+} W_k(\mathfrak{A}_\tau)$. The manifold $\mathcal{V}_\tau$ shown within the control space furthermore represents a situation in which a collection of microsounds $x \in \mathcal{X}(\mathfrak{W}_\tau)$ forms a set of points $\bigcup_{n\in\mathbb{R}} \mathbf{v}_n$ that resembles a locally compact Hausdorff space.

Note that $\mathcal{V}_\tau$ resembles an abstract model of the Schaefferian sound object proposed in Chapter 3 as a manifold that forms a Lie group. In this regard, we now interpret the notion of the sound object as a sort of topology for sound control [VNWD14] in which a function $\mathbf{v}(t)$ might define the reconstruction of microsounds *morphologically* by traversing the surface

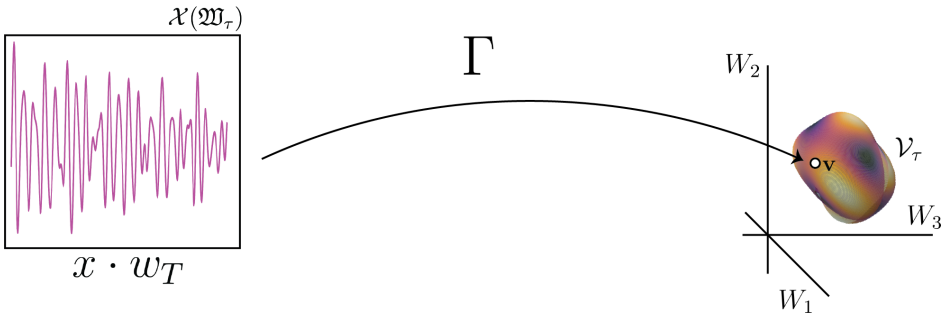


Figure 5.1: Parametric extraction at the microscale resulting in points on a manifold of possible microsounds.

of the manifold over time. This puts emphasis on the notion that the sound object is not any one 'concrete' piece of sonic material, but rather an abstract acousmatic phenomenon that resembles the full span of parametric variations [Kan14].

5.1.2 Parametric Synthesizers

Taking this into acount, we now invite the reader to imagine that this manifold resembles a *latent space*, denoting the set of all possible inputs values to a neural audio model f. While we formalize the intuition behind this in the following section (5.2), it will be sufficient to simply imagine this space as a continuous and differentiable representation of control parameters, where the representation has been interpolated to 'span' the orbit of each group action $\rho|_{W_k}(\mathfrak{a})$ (see Section 3.1.2). Taking into account these newly generated points that make our representation continuous, we must now define a method to resynthesize latent control parameters back into microsounds.

In order to facilitate such a resynthesis of latent microsounds, we define another operator $\tilde{\Gamma}$ which takes a set of synthesizer parametres $\tilde{\mathbf{v}}$ and plugs them into a pre-existing synthesizer suitable for reconstructing each original microsound x with adequate resolution. Some examples of these resynthesis parameters might be the set of amplitude values over a noise-driven filter bank [AKZV11] (Fig. 5.2), or the set of amplitude values over a harmonic plus noise synthesizer [SS90], both of which can resynthesize a given microsound $x \in \mathcal{X}(\mathfrak{W}_\tau)$ with higher precision depending on the harmonic nature of the microsound.

$$\tilde{\Gamma} : \bigoplus_{l=1}^{L} W_L(\mathfrak{A}_\tau) \to \mathcal{X}(\mathfrak{W}_\tau) \tag{5.3}$$

The operator $\tilde{\Gamma}$ thus represents a synthesizer which takes L parameters—denoted as another point of parameters $\tilde{\mathbf{v}} \in \bigoplus_{l\in\mathbb{Z}+} W_l(\mathfrak{A}_\tau)$—and resynthesizes them back into a microsound. Much like each disentangled control space W_k, the resynthesis parameters represented in each space W_l are assumed to be geometrically affine, facilitating a similarly stable low-dimensional representation of x.

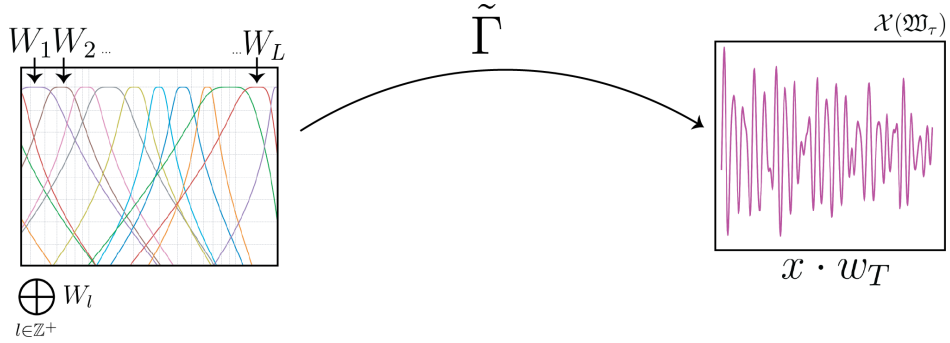


Figure 5.2: Parametric resynthesis at the microscale. An L dimensional space represents the amplitude values of L noise-driven bandpass filters.

5.2 Differentiable Digital Signal Processing

In this section, we argue that with the addition of two components, namely a neural audio mapping $f \in \mathcal{F}$, and an affinized representation of the STFT spectrogram, the morphological model constructed thus far forms the fundamental building blocks of what's commonly known as *Differentiable Digital Signal Processing* (DDSP) [EHGR20]. DDSP is a framework that uses a neural network to learn a mapping f from spectrotemporal parameters to resynthesis parameters by using an affinized representation of the Weyl-Heisenberg group called the *Multiscale Spectrogram* (MSS) [SM23] as a distance metric to ensure stability in the network. These components are often introduced using terminology oriented around computational implementation, where f is called an 'autoencoder' and the MSS distance metric is called a 'loss function.' We furthermore show that the DDSP network allows one to morphologically

control the space $\mathcal{X}(\mathfrak{W}_\tau)$ through various audio effects such as extrapolation and timbre transfer, which leverage group equivariant relationships between different representations of sound.

5.2.1 The Multiscale Spectrogram

Initially proposed by [EHGR20] and expanded upon by [HSF+24], DDSP is a proposed methodology in which classical signal processing techniques such as additive synthesis [BSAL11], phase-vocoder [AKZB11], and source-filter modeling [AKZV11] can be utilized by neural networks in order to learn a differentiable synthesizer for the purposes of parametric estimation and generative audio modeling. Perhaps the most straightforward use case for a DDSP model is the task of sound matching, which is described at length in [HLL23]. Sound matching consists of finding an ideal parameter set such that a sound object x can be adequately reconstructed. The task can be described in even greater detail using the aforementioned operators Γ and $\tilde{\Gamma}$, slightly augmenting the framework described in [HLL23].

Given a sound $x \in \mathcal{X}(\mathfrak{T}_\tau)$, the task of sound matching involves finding a set of parameters $\tilde{\mathbf{v}}$ such that $\tilde{x} = \tilde{\Gamma}\{\tilde{\mathbf{v}}\}$, where $\tilde{x} \approx x$. The way in which we might find such a set of parameters is through the training of a neural audio model $f \in \mathcal{F}$. While [HLL23] describe an approach in which the parameters $\tilde{\mathbf{v}}$ are learned directly from the original sound x, it is more common in neural audio effects such as those described in [EHGR20] and [BRC24] to instead learn an intermediary mapping from control parameters to resynthesis parameters. We can visualize this more clearly using a commutative diagram.

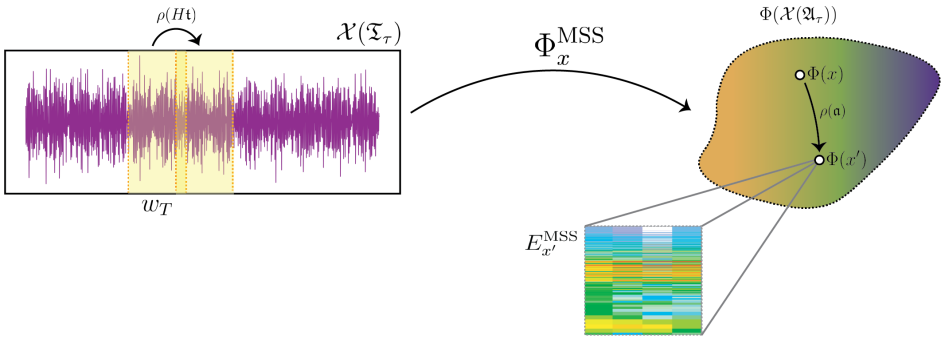
$$\begin{array}{ccc} \mathcal{X}(\mathfrak{W}_\tau) & \xrightarrow{\Gamma} & \bigoplus_{k=1}^{K} W_k(\mathfrak{A}_\tau) \\ \Big\downarrow \mathscr{T} & & \Big\downarrow f \\ \mathcal{X}(\mathfrak{W}_\tau) & \xleftarrow[\tilde{\Gamma}]{} & \bigoplus_{l=1}^{L} W_l(\mathfrak{A}_\tau) \end{array} \tag{5.2.1}$$

In diagram 5.2.1, we represent the sound matching operation as $\tilde{x} = \mathscr{T}\{x\} = (\tilde{\Gamma} \circ f \circ \Gamma)\{x\}$. In order to properly evaluate how closely $\tilde{x}$ resembles x, we must use another affinized time-frequency representation, similar to those used in the previous chapter. One way to achieve this is by extracting the STFT representation at multiple timescales, which effectively performs multiresolution analysis on the microscale. This modification is known as the *Multiscale Spectrogram* (MSS) [EHGR20] [SM23], shown below in Eq. 5.4 as an energy distrubution.

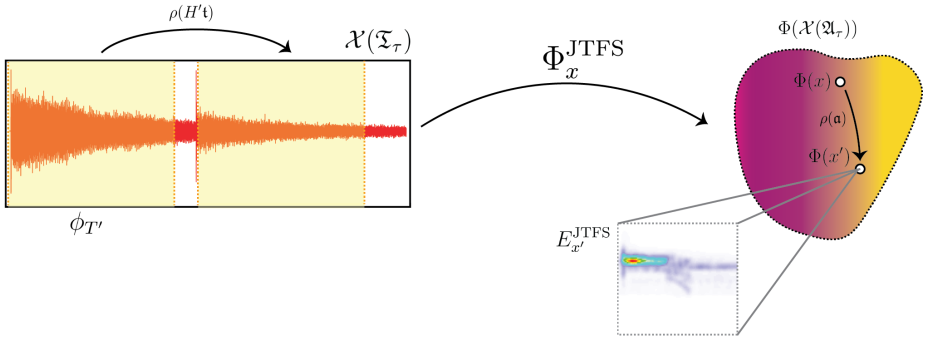
$$E_x^{\text{MSS}}(u, \xi, T, H) = \bigoplus_{T \in \mathcal{T}}^{N} \left| \int_{t \in \mathbb{R}} x(t + uH) w_T(t) e^{-i2\pi\xi t} dt \right|^2 \tag{5.4}$$

The MSS representation yields a representation that concatenates STFT representations at multiple time-scales $T \in \mathcal{T}$. This provides a representation that circumvents the time-frequency resolution tradeoff implicit in the STFT representation [LC04]. For this reasons, the MSS representation has been noted to be approximately equivalent to the wavelet scalogram $\Phi_x^{\text{MSS}} \approxeq \mathcal{S}_x$ [CHC+23]. This equivalence furthermore makes Φ_x^{MSS} another *affinized* representation of the Weyl-Heisenberg group acting at the microscale, contracting sound into a space that is better localized than $\Phi_x^{\mathcal{F}}$. The similarities between this contraction at the microscale and the JTFS transform's contraction at the mesoscale can be observed in Figure 5.3, where T represents the largest window of a MSS analysis and T' represents the larger time-scale from a JTFS analysis. Likewise, H and H' denote the respective hop sizes from a micro and mesostructural analysis.

Given a sound $x \in \mathcal{X}(\mathfrak{T}_\tau)$, we can declare an equivariant relationship between the action of time-translation $\rho(\mathfrak{t})$ and affine deformation $\rho(\mathfrak{a})$ of microsounds. Let H be the translation factor (hop size) and N be the number of resulting of multiscale spectrograms derived from the smallest window size $T \in \mathcal{T}$. Since $\Phi_x^{\text{MSS}} \approxeq \mathcal{S}_x$, it also follows that Φ_x^{MSS} is approximately equivariant to time-translation deformations, as is the case for $\mathcal{S}_x$ (Eq. 4.9). In diagram 5.2.2 below, $\rho(H\mathfrak{t})$ acts on a window $w_T(t)$ such that $\rho(H\mathfrak{t})w_T(t) = w_T(t - H\mathfrak{t})\ \forall T \in \mathcal{T}$. The equivariant nature of this operation can be shown by the fact that time-translations in $\mathcal{X}(\mathfrak{T}_\tau)$ yield affine transformations in the MSS domain $\mathcal{X}(\mathfrak{A}_\tau)$, i.e. $\Phi^{\text{MSS}}(\rho(H\mathfrak{t})^{(n)}x) = \rho(\mathfrak{a})^{(n)}\Phi^{\text{MSS}}(x)$.



(a) Affine representation at the microscale using the multiscale spectrogram.



(b) Affine representation at the mesoscale using the JTFS representation.

Figure 5.3: Contractive spaces $\Phi(\mathcal{X})$ at both the microscale and the mesoscale.

$$\begin{array}{ccccccccc}
\mathcal{X}(\mathfrak{W}_\tau) & \xrightarrow{\rho(H\mathfrak{t})^{(1)}} & \mathcal{X}(\mathfrak{W}_\tau) & \xrightarrow{\rho(H\mathfrak{t})^{(2)}} & \mathcal{X}(\mathfrak{W}_\tau) & \xrightarrow{\cdots} & \mathcal{X}(\mathfrak{W}_\tau) & \xrightarrow{\rho(H\mathfrak{t})^{(N)}} & \mathcal{X}(\mathfrak{W}_\tau) \\
\downarrow{\scriptstyle \Phi_x^{\mathrm{MSS}}} & & \downarrow{\scriptstyle \Phi_x^{\mathrm{MSS}}} & & \downarrow{\scriptstyle \Phi_x^{\mathrm{MSS}}} & & \vdots & & \downarrow{\scriptstyle \Phi_x^{\mathrm{MSS}}} \\
\mathcal{X}(\mathfrak{A}_\tau) & \xrightarrow{\rho(\mathfrak{a})^{(1)}} & \mathcal{X}(\mathfrak{A}_\tau) & \xrightarrow{\rho(\mathfrak{a})^{(2)}} & \mathcal{X}(\mathfrak{A}_\tau) & \xrightarrow{\cdots} & \mathcal{X}(\mathfrak{A}_\tau) & \xrightarrow{\rho(\mathfrak{a})^{(N)}} & \mathcal{X}(\mathfrak{A}_\tau)
\end{array} \tag{5.2.2}$$

The equivariant relationship observed between time-translations and affine deformations on the microscale thus crucially hints at the fact that the MSS representation has the potential to facilitate a continuous morphology of sound. Consider two sounds x and $\tilde{x}$, whose mesostructural distance might be substantially large. Despite a large distance on the mesoscale, the same two sounds analyzed on the microscale might exhibit windows of negligible distance that hint at near-identical microstructures. The microstructural distance after applying a window w_T onto x and $\tilde{x}$ is given by:

$$\mathcal{L}^{\text{MSS}}(\tilde{x}, x) = \sum_{T \in \mathcal{T}} \|E^{\text{MSS}}_{\tilde{x}}(u, \xi, T, H) - E^{\text{MSS}}_{x}(u, \xi, T, H)\|_2^2 \tag{5.5}$$

In the context of DDSP, this is called the MSS loss, which appropriately describes the MSS representation's use case in our original task of sound matching. For a sound $x \in \mathcal{X}(\mathfrak{T}_\tau)$, we can form a commutative diagram that evaluates the geometric stablity of a reconstruction of a microsound. This can be shown as an extension of diagram 5.2.1.

$$\begin{array}{ccccc}
\mathcal{X}(\mathfrak{A}_\tau) & \xleftarrow{\Phi^{\text{MSS}}_x} & \mathcal{X}(\mathfrak{W}_\tau) & \xrightarrow{\Gamma} & \bigoplus_{k=1}^{K} W_k(\mathfrak{A}_\tau) \\
\Big\downarrow{\scriptstyle \rho(\mathfrak{a})} & & \Big\downarrow{\scriptstyle \mathscr{T}} & & \Big\downarrow{\scriptstyle f} \\
\mathcal{X}(\mathfrak{A}_\tau) & \xleftarrow[\Phi^{\text{MSS}}_x]{} & \mathcal{X}(\mathfrak{W}_\tau) & \xleftarrow[\tilde{\Gamma}]{} & \bigoplus_{l=1}^{L} W_l(\mathfrak{A}_\tau)
\end{array} \tag{5.2.3}$$

In the modification above (5.2.3), two additional nodes represent the MSS spaces for both x and $\tilde{x}$. An affine group action $\rho(\mathfrak{a})$ can be used to denote an operation which maps points in the original space to the reconstruction space via a linear transformation, where $|\mathfrak{a}| = \mathcal{L}^{\text{MSS}}(\tilde{x}, x)$.

5.2.2 The Differentiable Synthesizer

In practice, the model f learns a continuous differentiable mapping based on the reconstruction loss between $\tilde{x}$ and x. This differentiability is a result of 'training' the neural network model f using backpropagation and stochastic gradient descent (SGD) [GBC16]. These are two common methods used in neural network training that iteratively update the parameters of the network f such that they are optimized to minimize the loss function. SGD is performed iteratively, and in the context of DDSP involves minimizing the perceptual distance between $\tilde{x}$ and x using MSS representations such that an adequate mapping is learned from time-varying control parameters $\mathbf{v}$ to resynthesis parameters $\tilde{\mathbf{v}}$.

By slightly updating our notation, we now represent the neural audio model as f_θ where θ denotes the parameters associated with the network, following [HLL23]. SGD thus entails the following iterative process:

$$\theta_{i+1} \leftarrow \theta_i - \nu \nabla \mathcal{L}_\theta^{\text{MSS}}(\tilde{x}, x) \tag{5.6}$$

in which i is the current iteration of training, ν is a variable learning rate, and $\nabla \mathcal{L}_\theta^{\text{MSS}}$ represents the gradient of the loss function with respect to θ. Computation of the gradient involves backpropagation, in which the partial derivative $\frac{\partial \mathcal{L}}{\partial \theta_j}$ is calculated for each j, often implementing the chain rule iteratively due to the compositional nature of the neural network.

In practical implementations of DDSP networks such as [EHGR20] and [BRC24], networks are often composed of three sections. First, a short time-series of control parameters belonging to each space $\bigoplus_{k=1}^{K} W_k(\mathfrak{A}_\tau)$ are passed through K separate networks in parallel, each of which learn an embedded representation equivariant to affine transformations. These networks are implemented as fully-connected *Multi-Layer Perceptrons* (MLP) [GBC16], which are known to be universal function approximators [HSW89]. Next, the network learns a single representation by concatenating the outputs of each MLP and passing the concatenated tensor through a *Gated Recurrent Unit* (GRU), a type of network that learns sequential structure shown to be invariant to time-warping by [BBCV21]. Finally, the resulting representation is passed through a final MLP plus a time-interpolation function to yield the synthesizer parameters $\tilde{\mathbf{v}}$ over each time-step.

Combined all together, the DDSP network learns an equivariant mapping from the space of spectrotemporal control parameters to the space of resynthesis parameters. Let $\rho(\mathfrak{a}^k) \in$

$\bigoplus_{k=1}^{K} W_k(\mathfrak{A}_\tau)$ denote the group actions that control spectrotemporal parameters in each disentangled representational space W_k and $\rho(\mathfrak{a}^l) \in \bigoplus_{l=1}^{L} W_l(\mathfrak{A}_\tau)$ be the same group actions that denote resynthesis parameters in each space W_l. The DDSP neural audio model f_θ is equivariant to the group actions, such that $f(\rho(\mathfrak{a}^k)\mathbf{v}) = \rho(\mathfrak{a}^l)f(\mathbf{v}) \quad \forall k \forall l$.

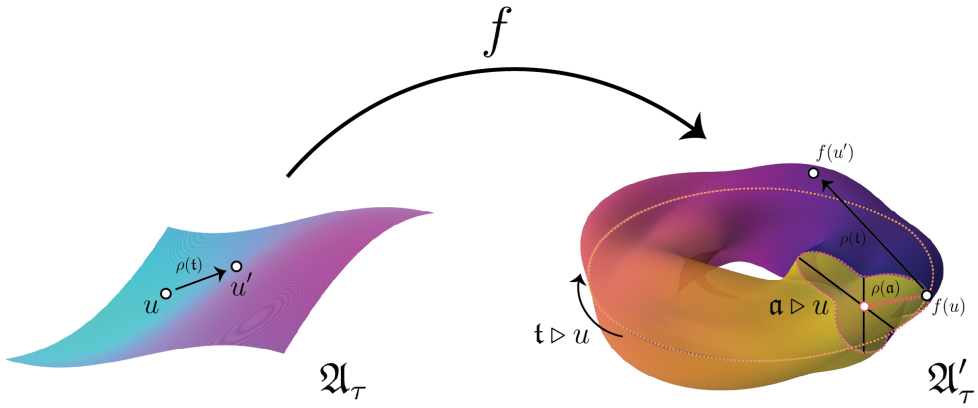


Figure 5.4: A topological interpretation of the DDSP harmonic autoencoder from [EHGR20] as a mapping $f : \mathbb{R}^N \to \mathrm{T}^3$ from a N-dimensional parameter space to a torus representing the phase space of an additive synthesizer. See [ABH+24] for a more detailed interpretations of tori in time-frequency analysis.

This equivariant mapping is augmented by the choice of both the spectrotemporal operator Γ and synthesizer $\tilde{\Gamma}$, which sometimes change the topology of the manifold. Fig. 5.4 denotes a neural audio mapping in which a two-dimensional control space is mapped to the parameters of an additive harmonic synthesizer, following part of the approach described in [EHGR20]. The translation of points in the control space is mapped to a latent space that forms a three-dimensional torus, in which a cycle around the major radius $\mathfrak{t} \triangleright u$ represents the fundamental frequency f_0 of the harmonic synthesizer, and a cycle around the minor radius $\mathfrak{a} \triangleright u$ represents the phase space of the waveform resulting from the harmonic partials of the additive synthesizer at each f_0 value. As parameters in the space $W_1(\mathfrak{A}_\tau) \oplus W_2(\mathfrak{A}_\tau)$ change from points u to u', so in turn does the fundamental frequency value of the synthesizer $f(u)$ to $f(u')$, as shown using the composition of actions $\rho(\mathfrak{t})$ and

$\rho(\mathfrak{a})$. Translations $\rho(\mathfrak{t})$ thus represent the harmonic synthesizer's morphing of f_0 values across time, while affine actions $\rho(\mathfrak{a})$ denote the morphing amplitudes of the harmonic oscillator's partials—represented by amplitude scaling in the phase space. The scaling of harmonic partials at a given f_0 value along the major radius are thus dependent on the volume and contour of the torus. The torus can be imagined as a sort of geometric model for a learned DDSP latent space, given the use of a harmonic additive synthesizer for $\tilde{\Gamma}$.

5.2.3 DDSP as Schaefferian Morphology

In this chapter, we have observed that DDSP can provide a convincing representation for Schaefferian morphology in its ability to model the time-varying parametric control of a collection of sounds at the microstructural level. The latent control space that DDSP provides thus allows a composer or performer to start with 'sound itself' by currating datasets of sonic material in order to construct invariant representations of acousmatic sound. This approach reflects the nature of acousmatic composition, since the chosen sonic material might often be agnostic to any one sound source, but rather might reflect a myriad of different sources whose sounds contain perceptually similar features.

Finding a suitable control space for such acousmatic material then becomes a question of constructing a latent control space which generates sounds that are typologically invariant (i.e. perceptually similar) to the sounds in the dataset. In the following chapter, we turn our focus towards practical implementations of typomorphology using both scattering networks and DDSP networks. We will use both of these tools to estimate an appropriate latent control space for a given DDSP model in a way that best reflects the mesostructural typology of the dataset.

Chapter 6

Typomorphology In Practice

In this chapter, we implement and evaluate different methods for the typomorphological analysis and synthesis of acousmatic sound using neural audio synthesis to model our proposed manifold representation of the Schaefferian sound object. In these implementations, we utilize the scattering network defined in Chapter 4 for a 'typological' analysis of the acousmatic sound object and the DDSP network defined in Chapter 5 for a 'morphological' analysis of the acousmatic sound object. We implement a method that disentangles a set of spectrotemporal audio descriptors in order to find the most characteristic control parameters with respect to the chosen dataset of sounds. The method we implement outperforms the expressivity of DDSP models that use a more common set of conditional parameters (e.g. pitch, centroid, loudness). We also implement a method that morphs audio between two sound types using a single DDSP model, which more closely resembles Schaefferian typomorphology. This contribution serves as the first time, to our knowledge, that spectral audio descriptors have been used to condition a DDSP model, as well as the first time JTFS representations have been used to aid the training of a DDSP model for computer music composition. It also marks the first time that Schaefferian approaches to sound have been used to completely guide the implementation of a generative audio model. A companion site[1] available online hosts the audio examples accompanying the experiments covered in this chapter.

[1] `https://acousmatic-ddsp.netlify.app/`

6.1 Experiments

This section presents an overview of our 'acousmatic' neural audio experiments. Our initial research question involves augmenting the available control parameters for a given DDSP model to the time-varying spectral audio descriptors introduced in [PGS+11]. Approaching this question with a Schaefferian philosophy in mind, we attempt to find a small subset of these audio descriptors that suitably disentangle the latent space derived from training a differentiable synthesizer on collection of sounds. We propose a method that involves selecting three descriptors that are most correlated to the dataset's scattering representations, ensuring that the model produces sound objects that are geometrically stable at the microscale—i.e. across each control parameter's group orbit—while also remaining geometrically stable at the mesoscale. This method reinforces the treatment of 'acousmatic' sound, seeing as control parameters are selected strictly based on the nature of the dataset in question.

6.1.1 Disentanglement Hypothesis

Given a dataset of sounds, our goal is to extract the optimal microstructural control parameters that are best suited to control and condition the dataset during the training of a DDSP model. We restrict the available control parameters to the set of time-varying spectral audio descriptors laid out in [PGS+11] (see Appendix B). Recall from the previous chapter that this problem can be interpreted from a group theoretical perspective, such that for a given microsound $x \in \mathcal{X}(\mathfrak{W}_\tau)$, we define an operator Γ that projects x into K approximately independent subspaces $\bigoplus_{k=1}^{K} W_k(\mathfrak{A}_\tau)$. This projection disentangles the space of control parameters [HAP+18], ensuring that group actions are approximately orthogonal such that $\rho|_{W_k}(\mathfrak{g})$ constrains $\mathfrak{g}$ to act only on the space W_k. In terms of DDSP, this results in the construction of a latent space in which each control parameter exhibits perceptually independent control of the output.

At first glance, we might choose to implement Γ with the help of a technique such as *Principal Component Analysis* (PCA). This approach would allow us to extract the control parameters that display the maximum amount of variance across sounds in the dataset. While this approach would indeed disentangle parameters with respect to the control space, this particular disentanglement only takes into account variance at the

microstructural level. A more suitable implementation of Γ might instead involve the dataset's *mesostructural* representations. This approach would maintain equivariance between DDSP's control parameters and resynthesis parameters at the microscale, while still yielding an output that remains approximately invariant to larger timbral structures at the mesoscale. In Schaefferian terms, this proposed method would ideally generate an expressive latent *morphology* of sounds in the dataset, while still remaining invariant to the *typology* of sounds in the dataset.

We therefore propose a method in which we find a small subset of spectrotemporal control parameters most correlated with the features of a low-dimensional projection of the dataset's JTFS representation. [LYY23] implements a model suitable for this methodology, using the isomap algorithm to greatly reduce the dimensionality of the JTFS representation. The isomap algorithm is a graph-based algorithm for manifold learning that plots high-dimensional representations onto a lower-dimensional space such that the resulting dimensions are orthogonal without sacrificing the preservation of small Euclidean distances in the feature space. The model proposed in [LYY23] yields a three-dimensional feature space where points represent short sounds, distances represents an estimate of perceived timbral similarity [LEHR+21], and the dimensions of the space correlate to the distributions of ground-truth control parameters of the sounds. These properties are shown using a synthetic dataset of chirps generated by spanning two different parameter spaces consisting respectively of FM/AM control parameters and additive harmonic synthesizer parameters.

Compositionally speaking, the isomap model proposed in [LYY23] makes a connection between acousmatic typology and morphology, namely that microstructural control parameters at the time-scale of an STFT window w_T correlate with the distribution of mesostructural representations at the larger time-scale of the low-pass scattering filter ϕ_T. The correlation implies a perceptual disentanglement innate to the action of each control parameter, due to the Euclidean distance of points in the JTFS domain representing a mesostructural perceptual distance [LEHR+21].

6.1.2 Methodology

Following this logic, we might imagine a dataset of sounds for a DDSP model to act in place of this dataset of chirps. We make the hypothesis that given a dataset of short sounds

(200ms - 1000ms), an ideal set of control parameters $\mathbf{v} \in \bigoplus_{k\in\mathbb{Z}+} W_k(\mathfrak{A}_\tau)$ for a potential DDSP model will correlate with the dataset's JTFS isomap distribution. Our experiment thus attempts to select the most disentangled control parameters by selecting the three spectral audio descriptors most correlated with the three dimensions of the dataset's JTFS isomap representation.

DDSP's compositional capabilities are often demonstrated using two methods of audio generation: extrapolation and 'timbre transfer.' Extrapolation generates audio from a trained DDSP model using direct inference from a matrix of control parameters $\mathbf{V} \in \mathbb{R}^{K\times N}$, where K is the number of control parameters and N is the number of time-steps specified for audio generation. Timbre transfer first extracts the matrix of control parameters from an input audio signal and then passes these parameters to the DDSP model, creating a style transfer effect [EHGR20]. We utilize both of these methods for audio generation in order to qualitatively and quantitatively evaluate each model.

In our main experiment, we train three independent DDSP models on recordings of three types of friction percussion techniques (see Section 6.2.1 for details). We first compute the JTFS isomap of each dataset and plot the sounds as points in three-dimensional space, following [LYY23]. We then extract the entire set of time-varying audio descriptors from [PGS+11] and calculate the mean value of each descriptor with respect to each recorded sound. We then use the Pearson correlation coefficients to find which three audio descriptors correlate most to the distribution of sounds across each dimension of the JTFS isomap, witholding loudness which we include as a default parameter in all three models. For each dataset, we train a DDSP model conditioned on the three resulting parameters, as well as a baseline DDSP models trained using the highest scoring parameter from a PCA performed on the set of all spectrotemporal audio descriptors. This baseline model allows us to compare the difference between the spectrotemporal parameters that exhibit the most variance on the microscale with the set of disentangled parameters which exhibit parametric correlation over the mesoscale.

6.1.3 Evaluation

After training, we first evaluate each model objectively using extrapolation techniques. This is done by generating 100 new sounds—each 1 second in duration—that span the parametric control space of each model. This generated set of sounds is projected once more using the

JTFS isomap in order to calculate two objective metrics that measure the geometric stability of the resulting sounds on the mesoscale. These metrics reward the adjacency of generated sounds in the JTFS isomap's space while penalizing stark outliers.

First, the point cluster is evaluated as a smooth and continuous manifold interpolated by creating a tangent space using the convex hull of the JTFS isomap. Then, the volume of the manifold is calculated. We perform this interpolation both with the disentangled parameters and the baseline PCA parameters, estimating that the model using our disentanglement method will yield a more compact manifold that contains less volume. Finally, we compare the sum total of distances between sounds in the JTFS isomap by treating the point cluster as a fully connected graph. We compute the graph for both the baseline PCA model and the disentangled model and perform both L1 and L2 regularization on their cumulative pair-wise distances respectively, predicting once again that the disentangled model will yield a smaller cumulative distance between points than the baseline model.

Subjective metrics are then evaluated using the timbre transfer algorithm, in which we evaluate the baseline model and the model trained on disentangled parameters in their ability to reconstruct sound from the dataset. We analyze the various timbral properties of the resulting reconstructions, which we estimate will be augmented in the disentanglement method's reconstructions.

6.2 Implementation Details

This section presents an overview of various implementation details concerning the experiments performed. We give a detailed description of the friction percussion datasets, the DDSP synthesizer, and hyperparameters used for both the DDSP model and the scattering transform.

6.2.1 Datasets

Prior to Schaeffer's foray into electronic sound, many early experiments in musique concrète were performed using unconventional physical objects as compositional tools [Kan14]. Situating our research in the spirit of acousmatic music, the sounds used for our experiments utilize three datasets consisting of short musical gestures derived from three

recordings of friction percussion improvisations. Friction percussion is an extended technique that utilizes the rubbing and scraping of nontraditional objects on the surface of a drum head, notably used in the David Tudor piece *COEFFICIENT* [Tud91]. We recorded three datasets at the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) each consisting of around 20-30 recordings between 200ms and 1s. Each dataset represents a different object being played on the membrane of a floor tom. These objects consisted of a spring coil (type A), a threaded rod (type B), and a small piece of styrofoam (type C). This particular selection of objects generated three timbrally unique recordings that lended well to the construction of three distinct datasets of short acousmatic sounds.

6.2.2 Scattering Model

The JTFS isomap used to plot each acousmatic sound onto a three dimensional space follows the ISMIR 2023 tutorial on GEAR [VML23]. The isomap is learned using JTFS transforms performed with parameters $J = 13$, $Q = 8$, and T set to the duration of the sound in the dataset. Since the GEAR model was initially evaluated by the authors using a dataset of generated chirps along with their corresponding parameter set, we simply replaced the set of chirps with each friction dataset along with their corresponding set of time-varying spectral audio descriptors.

6.2.3 DDSP Model

Our choice in DDSP model must reflect the need for a synthesizer that generalizes well to the types of sonic material used in acousmatic music. This involves two important requirements: the ability to flexibly add and remove conditional parameters, and the use of a synthesizer that can model both harmonic and stochastic sounds. We thus chose to use the filterbank model laid out in [BRC24], which learns a mapping f between control parameters and time-varying amplitudes of a large ($M \approx 2048$) white-noise filterbank.

The DDSP model in [BRC24] follows the design pattern mentioned in the previous chapter in which time-invariance is learned using a combination of MLPs and GRUs in sequence. The model in [BRC24] also allows for the addition of custom conditional parameters, making it perfect for experimentation with time-varying spectral audio

descriptors. For each training session, a batch size was chosen to reflect the number of training samples processed simultaneously during each training iteration. The batch size parameter is often chosen for the purpose of balancing between memory limitations and gradient stability, ensuring smoother learning that utilizes a more diverse selection of data during each iteration. Additionally, a learning rate ν was selected for the MSS loss (Eq. 5.6), in order to better stabilize large oscillations in the convergence of the loss function. For each experiment, we trained the DDSP model for 10k epochs (full passes on each dataset) while using a batch size of 16 and a learning rate of 0.001. Each parameter follows the configuration of the original experiments performed in [BRC24].

Finally, time-varying conditional parameters were extracted from the datasets using an STFT with a window size of 1024 samples, which allowed for detailed frequency resolution, and a hop size of 128 samples, ensuring an overlap that preserves temporal continuity while capturing subtle spectral changes.

6.3 Results

This section reviews the results from our experiments in modeling the sound object. We start by going over some preliminary experiments which confirm that the use of spectrotemporal parameters can aid the training of DDSP models from a typomorphological perspective. We then review the results of the proposed disentanglement method, which chooses spectrotemporal parameters based on the mesostructural properties of the dataset.

6.3.1 Preliminary Experiments

Before working with disentangled parameters and friction percussion sounds, we first attempt to demonstrate the preliminary hypothesis that the spectrotemporal parameters from [PGS+11] can aid the timbral expressivity of a DDSP model. We test this claim using two different categories of audio descriptors: parameters related to frequency content, and statistical 'moments' in the spectrogram.

To first show the effect of conditioning a DDSP model on parameters related to frequency content, we trained a model using the recording of the suspended triangle from Fig. 4.3 as our

training data. Using a timbre transfer procedure to reconstruct the original triangle strike, we extract loudness, centroid, harmonic energy, and noise energy from the original signal to be used as conditional parameters for our reconstruction. By then manually defining the harmonic and noise energy parameters in the reconstruction, we can evaluate how well the reconstructed audio adapts to the changes in the control parameters.

According to [PGS+11], harmonic energy is the energy of the signal over harmonic partials, obtained by summing the energy of the partials detected in a given STFT window. Noise energy is simply the remaining energy, extracted by subtracting the harmonic energy from the total energy of the signal. Fig. 6.1 depicts the corresponding spectrograms generated from setting harmonic energy and noise energy to their maximum values of 0.0 and 1.0 in inverse relation to each other, hypothetically extracting the noisiest and most harmonic reconstructions of the triangle strike. The results of this training are promising, as denoted in Fig. 6.1 which shows clear and prominent partials in the spectrogram of the harmonic reconstruction, while the noise reconstruction shows energy spread more stochastically around the spectrogram. These differences are furthermore apparent when listening to each recording.

A clear perceptual difference between these recordings can be heard in the fact that the noise energy = 1 reconstruction captures much more high-frequency content when compared to the harmonic energy = 1 reconstruction. Intuitively this can be explained by a stronger presence of stochasticity in the high-frequency part of the spectrum. While it should be noted that a maximum value of noise energy does not fully eliminate the harmonic partials of the signal, it does increase the presence of stochastic components in the signal. Likewise, a maximum value of harmonic energy does not eliminate stochastic components entirely, but certainly increases the prominence of harmonicity in the signal.

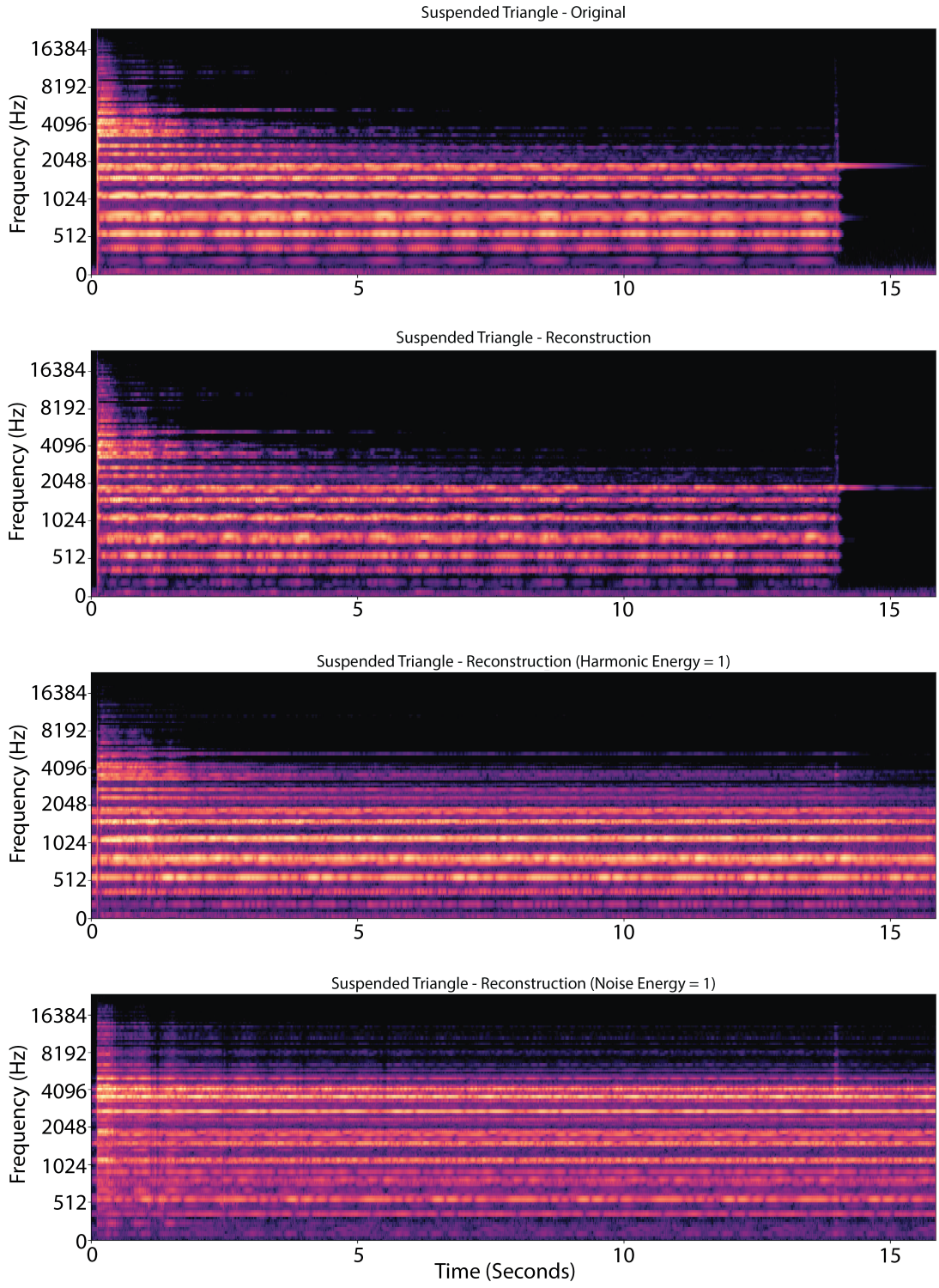


Figure 6.1: DDSP filterbank model reconstructions of suspended triangle recording with hardcoded values for harmonic energy and noise energy.

For demonstrating the effect of conditioning on statistical descriptors, we then trained the DDSP model on a synthetic dataset of 'noise chirps.' These signals were synthesized using white noise passed through a bandpass filter whose center frequency sweeps the range of the frequency spectrum from zero to the Nyquist frequency over a given time interval, in this case 5 seconds. The dataset consists of 8 noise chirps, each of which is synthesized using quality factors $Q = [0.05, 0.5, 1, 2.5, 5, 10, 25, 50]$.

For the noise chirp dataset, we conditioned the model on loudness, centroid, kurtosis, and decrease. First we evaluate how well the DDSP model learns kurtosis, a metric that measures the flatness of the spectrogram around its mean value [PGS+11]. Much like the bell example, we reconstruct one of the noise chirps from the dataset ($Q = 1$) and set the kurtosis parameter to its normalized minimum and maximum values of 0.0 and 1.0 while leaving the other parameters according to their values extracted from the audio. An expected result from setting kurtosis to a constant value of 0.0 would be a flattening of the spectrogram, resulting in a signal whose chirp contains a very wide bandwidth (low Q value), while setting the kurtosis parameter to 1.0 would result in a very thin bandwidth (high Q value). The modified bandwidth can be heard in each resulting reconstruction, especially the kurtosis = 1.0 reconstruction which has the perceptual quality of a sin wave being stochastically modulated. These assumptions were furthermore confirmed based on the resulting spectrograms, shown in Fig. 6.2 and 6.3.

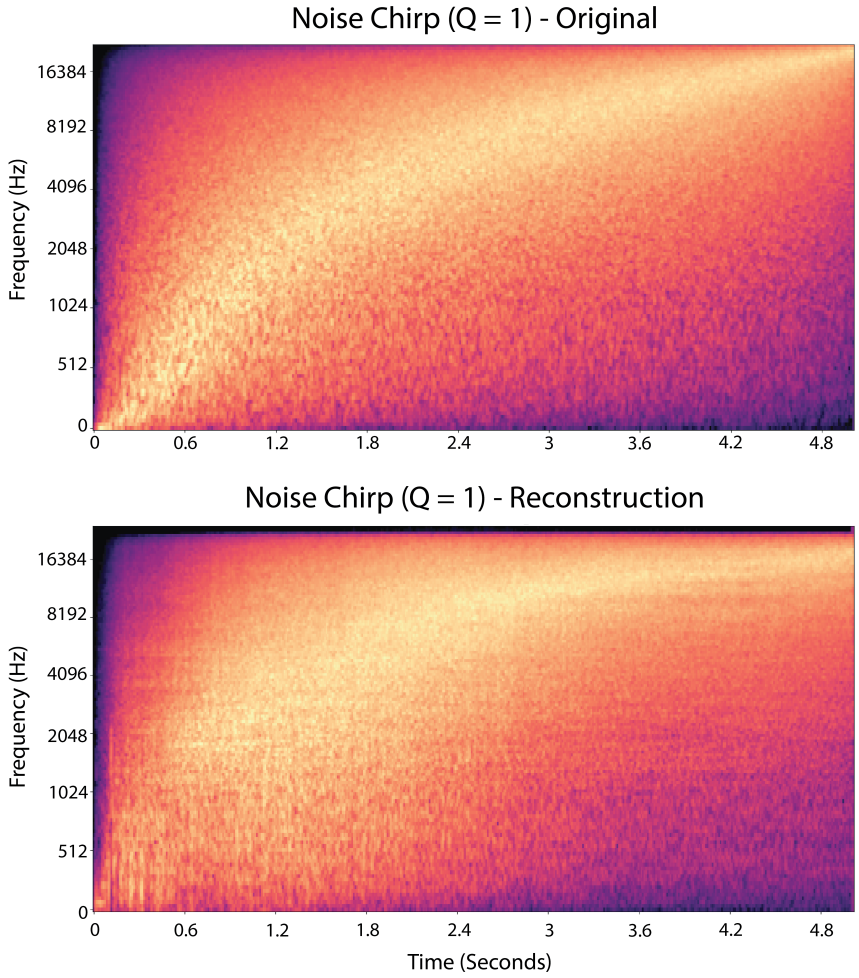


Figure 6.2: (Top) Original recording of noise chirp with $Q = 1$ and (Bottom) DDSP timbre transfer reconstruction performed without modifying conditional parameters.

Likewise, we perform the same reconstruction experiment by manually setting the decrease parameter, which measures the slope of energy in each STFT frame with an emphasis on lower frequencies [PGS+11]. We set decrease to its minimum and maximum values, while leaving the kurtosis values to automatically be extracted from the input audio. The resulting spectrograms are shown on the right of Fig. 6.3. The spectrograms show that when setting decrease to 0.0, the ascending noiseband of the chirp remains almost entirely unnoticable, and even tapers out completely when the center frequency

reaches the very high part of the spectrum (approximately above 16kHz). Setting decrease to 1.0 results in the opposite effect, yielding a reconstruction that instead prioritizes the higher frequencies of the chirp. Since decrease is a measure of slope that emphasizes the lower frequencies of the signal, we estimate that this parameter acts as a filter of sorts, prohibiting the reconstruction of spectral energy that does not adhere to the manually configured slope of the spectrogram. In this regard, the energy completely tapering out above 16kHz when setting decrease to 0.0 may be due to a lack of representative data for such a specific combination of decrease, kurtosis, and loudness values.

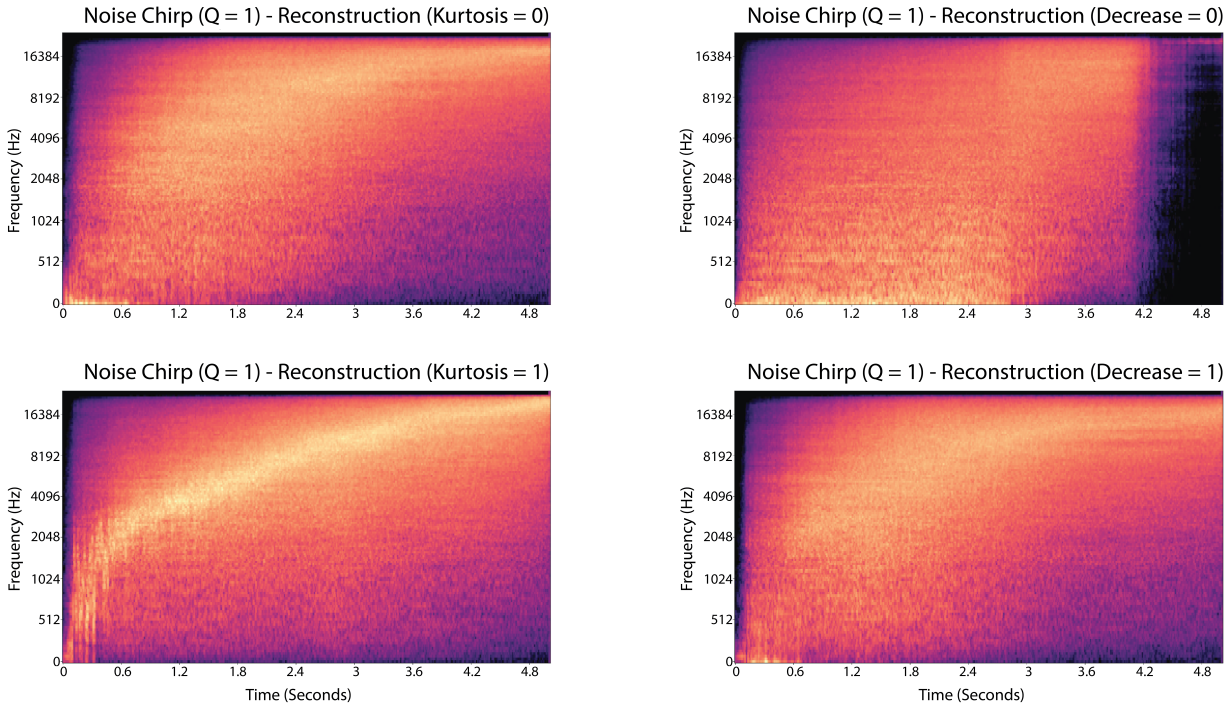


Figure 6.3: DDSP reconstructions with hardcoded values for spectral kurtosis (left) and spectral decrease (right) with parameters set to min/max values.

6.3.2 Typomorphological Experiments

Next, we experiment with the spectrotemporal audio descriptors on real-world sounds by implementing a method for typomorphological synthesis. The driving idea here is to 'morph' between two different sound types by training a single DDSP model on the union of two

datasets that each reflect the sound types in question. When conditioned on a sufficient number of spectrotemporal parameters, a sound from the initial sound type would be able to seamlessly morph into a sound from the terminal sound type over a certain number of iterations. This can be represented using a series of points in the scattering isomap, which form a pathway connecting the iterated sounds extrapolated from morphing between the initial and terminal sound types. The trajectory thus outlines a continuous affine group action on the control parameter vector $\rho(\mathfrak{a})\mathbf{V}$ which morphs the microstructural parameters of the initial sound into those of the terminal sound while retaining a suitable mesostructural topology in the scattering domain. We implement the morphology of control parameters from the initial sound to the terminal sound using a simple linear interpolation such that the intial sound's control matrix $\mathbf{V}_1$ morphs into the terminal sound's control matrix $\mathbf{V}_2$. The typomorphological synthesizer was trained on two different dataset pairs. The first pair consisted of environemntal sounds: field recordings of rainfall and recordings of applause. The second pair of datasets consisted of a collection of dog barks along with a collection of snare drum strikes. Both datasets were compiled using creative commons (CC) licensed sounds available on `https://freesound.org/`.

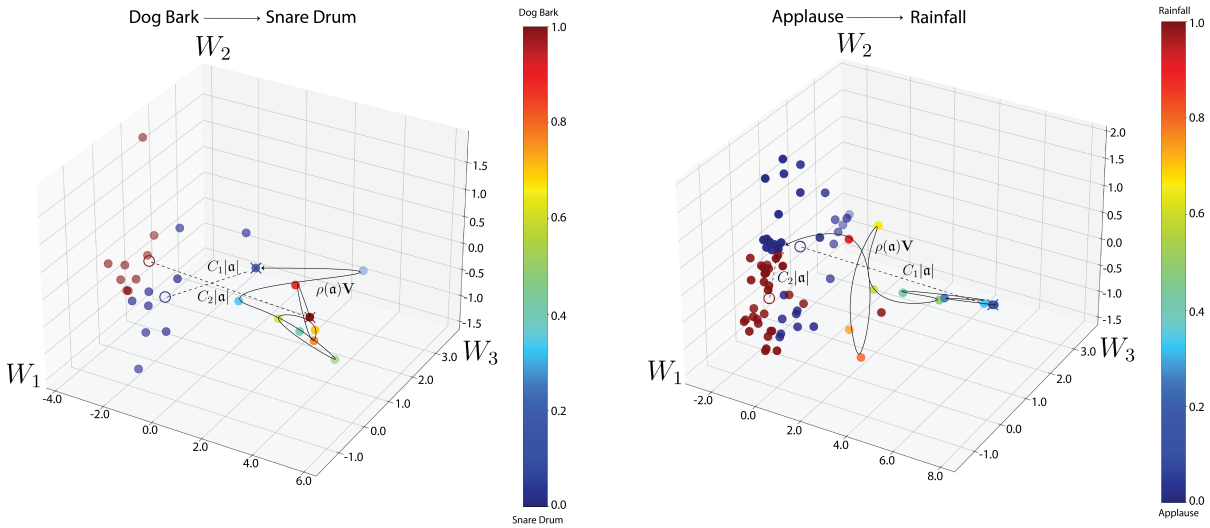


Figure 6.4: Typomorphological neural audio synthesis performed on two different pairs of sound types.

The action of the typomorphological synthesizer is represented in Fig. 6.4 which also shows the suitability of the scattering isomap representation for contextualizing Schaefferian typomorphology in neural audio synthesis. Points on the plots represent the sounds found in the dataset, while sound types are denoted using contrasting red and blue colors. Reconstructed sounds generated from the linear interpolation of control parameters are denoted along the vector from an initial reconstruction to a terminal reconstruction through the action $\rho(\mathfrak{a})\mathbf{V}$. Finally, the dashed lines represent both the initial and terminal sound's distance $C|\mathfrak{a}|$ from the center of mass of each respective sound type's cluster, measuring the approximate perceptual deviation and geometric stability (see Eq. 3.17) of the reconstruction from the sound type.

From the plots, we can observe how far each of the iterations of morphing sounds deviate from each type's cluster, thus allowing us to evaluate a DDSP model's fidelity on the mesoscale. This leads to important conclusions we can make about training the model from [BRC24] on two different sound types at once—more specifically, conclusions that are not readily availible by analyzing the MSS loss. One important observation we can make from both these plots and from listening to the reconstructions is that the model often leverages the timbral qualities of one sound type over another. For instance, in the dog bark $\rightarrow$ snare drum experiment, the initial dog bark produces a reconstruction not far from the cluster of dog barks in the dataset, however the terminal snare drum reconstruction deviates significantly from the cluster of snare drums. Likewise, in the applause $\rightarrow$ rainfall example, the initial applause example is perceptually far from its dataset cluster, but the terminal rainfall reconstruction ends up just around the rainfall cluster.

Despite these metrics, the reconstructions are still somewhat perceptually convincing even if they do produce ambiguous timbral qualities. For instance, there is a general increase in harmonic energy in the dog bark $\rightarrow$ snare drum model, even though the dog bark reconstruction more acurately takes after the original recording in its spectral envelope. In the environemnental model, we hear that the rainfall resembles its original recording much more accurately than the applause, which ends up sounding sharper. Upon listening to the morphological trajectories and observing the scattering plots, we might conclude that these models provide interesting situations in which the acousmatic ambivalence between two sound types is exploited, in that the resulting models often favor the spectral characteristics of one sound type at the expense of the other. While a more

accurate overall reconstruction might be achieved using a more intricate DDSP synthesis operation, this ambivalence also has the potential to be creatively exploited in an acousmatic compositional setting.

6.3.3 Disentanglement Results

Following the preliminary experiments which confirm that conditioning DDSP on spectrotemporal audio descriptors augments the timbral control of the resulting output, we now turn to the evaluation of our disentanglement method which involves finding a set of disentangled spectrotemporal parameters that are most correlated with the mesostructural properties of the dataset. This approach defines a novel method that follows [LYY23] by estimating the best suited spectrotemporal parameters for a dataset of sounds by measuring parametric correlation against a low-dimensional projection of their scattering representations. Instead of relying on prior information about the sound sources, or simply measuring which parameters demonstrate the maximum amount of variance across the dataset, this approach bridges a gap between parametric evolution on the microscale—related to acousmatic morphology—and parametric evolution on the mesoscale—emphasizing acousmatic typology. In this regard, the proposed method focuses on reinforcing the generation of sounds that mesostructurally resemble sounds in the dataset.

Table 6.1 show the results of our disentanglement method for type A, B, and C of friction percussion data. We denote the three prospective conditional parameters for our DDSP model as W_1, W_2, and W_3 following their notation as disentangled group representations. As noted earlier in the chapter, these selections do not always reflect the parameters resulting from the PCA, whose values are similarly shown in Table 6.2. For instance, the most disentangled parameters for type A are spectral slope, inharmonicity, and noise energy, while the top three parameters that account for the most variance in the dataset are spectral centroid, spectral slope, and spectral crest. The differences between 6.1 and 6.2 thus demonstrate how these metrics for variance differ between the microscale and the mesoscale of the sounds.

	Type A [Spring Coil]: Top Ranked Correlated Features		
Dimension	**1st**	**2nd**	**3rd**
W_1	**Slope**	Centroid	Harmonic Spectral Deviation
W_2	**Inharmonicity**	Harmonic Energy	Noisiness
W_3	**Noise Energy**	Decrease	Crest

	Type B [Threaded Rod]: Top Ranked Correlated Features		
Dimension	**1st**	**2nd**	**3rd**
W_1	**Decrease**	Crest	Noisiness
W_2	**Slope**	Centroid	Flatness
W_3	**Inharmonicity**	Odd/Even Ratio	Centroid

	Type C [Styrofoam]: Top Ranked Correlated Features		
Dimension	**1st**	**2nd**	**3rd**
W_1	**Harmonic Spectral Deviation**	Inharmonicity	Decrease
W_2	**Flatness**	Noise Energy	Slope
W_3	**Noisiness**	Harmonic Energy	Crest

Table 6.1: Spectrotemporal parameters most correlated with isomap JTFS representations for each friction percussion type.

Type A [Spring Coil]: Principal Component	**Explained Variance**
Centroid	0.517
Slope	0.327
Crest	0.074

Type B [Threaded Rod]: Principal Component	**Explained Variance**
Centroid	0.506
Slope	0.175
Decrease	0.130

Type C [Styrofoam]: Principal Component	**Explained Variance**
Slope	0.526
Decrease	0.264
Flatness	0.089

Table 6.2: Spectrotemporal parameters that account for the most variance in each friction percussion type. Parameters in bold are also predicted in the disentanglement method's selection.

Qualitatively, we can analyze these results in accordance to the timbral features of each sound type. Type A contains potentially the most diverse range of features, in which the spring coil not only produces resonant sounds that are both low and high in frequency, but also produces a fair amount of distinct stochastic components. The disentanglement method's selection of slope, inharmonicity, and noise energy seems to capture this diverse range of features, as this collection of descriptors contains features that deal with complex harmonic structure (inharmonicity), stochasticity (noise energy), and control of the spectrum based on statistical distribution of energy (slope). Type B contains very sharp and noisey transients that take up most of the frequency spectrum below 16kHz, caused by the threaded rod scraping up against the rim of the floor tom. As a result of these sharp transients, the

resonant modes of the tom are also apparent in the signal. For this type of playing technique, the model's selection of both slope and decrease may reflect the need for a refined slope metric in both the high and low frequency range of the spectrum. This could be due to the signal's combination of high-frequency transients caused by the periodic scraping of the threaded rod on the tom rim, and low-frequency resonance that results from this action in its periodic exciting and damping of the drum membrane. Inharmonicity then allows the model to capture the resonant components produced by the threaded rod. Finally, type C contains potentially the most stochastic sounds, as the effect of rubbing styrofoam against the drum membrane often produces a clear pitch while still containing a substantial amount of noise. This is clearly reflected in the model's choice of parameters, which include both flatness and noisiness measures for distinguishing between harmonic peaks and noise energy, while harmonic spectral deviation measures the deviation of amplitudes between the harmonic peaks.

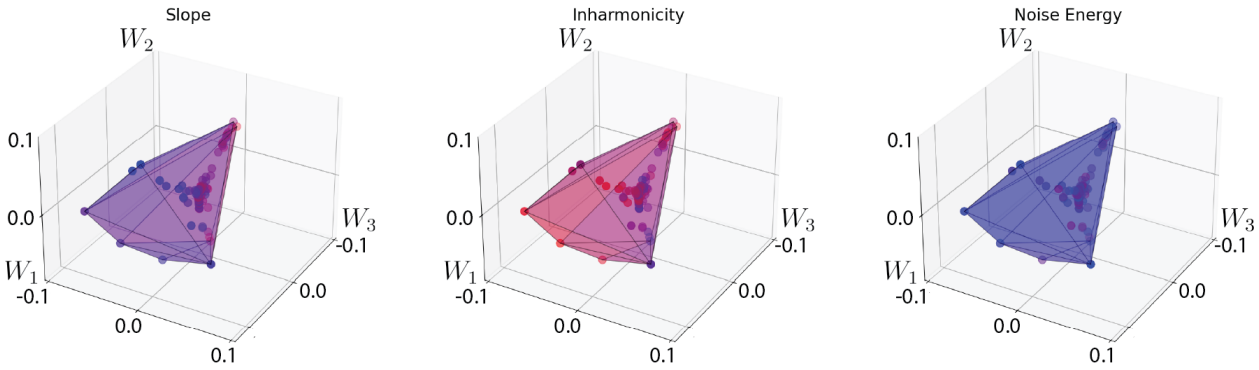
6.3.4 Extrapolation Results

We first evaluate the disentangled methods quantitatively by extrapolating over the span of conditional parameters and measuring the mesostructural similarity between generated sounds. This is done by first approximating each model's latent space by generating 100 new sounds—each 1 second in duration—whose extrapolation parameters span the entire parameter space. We then plot each sound in the isomap space using a color hue for each sound to denote a certain spectrotemporal parameter's value, following the same procedure in [LYY23] in order to better visualize correlation between parameters and dimensions of the isomap space. The hypothesis in this experiment is based off of the assumption that the isomap space maps timbrally similar sounds closer together in distance—an effect of the JTFS transform's ability to model timbral similarity [LEHR+21]. We thus treat the set of 100 points in this space as a geometric object by defining a convex hull over the outer-most points in the space. This creates a hypothetical manifold resembling the Schaefferian 'sound object' containing all of the generated sounds.

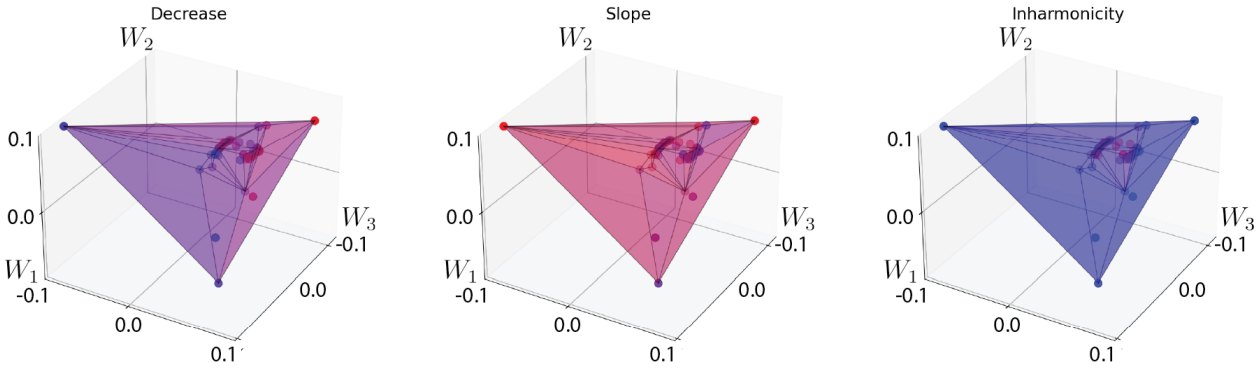
The manifolds for each model are depicted in Fig. 6.5, where the color hues on each face of the mesh represent the mean parameter value across the corresponding vertices. With this visualization, we can confirm that the generated sounds from type A and type B strikingly correlate with the dimensions of the isomap space. This parametric correlation hints towards

the ability of these two models to adapt well to both the microstructural and mesostructural properties of the sound in the dataset. Type C adheres less so to this correlation at the generation stage, as demonstrated in the multicolor meshes produced for dimensions W_2 and W_3. This is potentially due to the high-frequency and noisy nature of the styrofoam sounds.

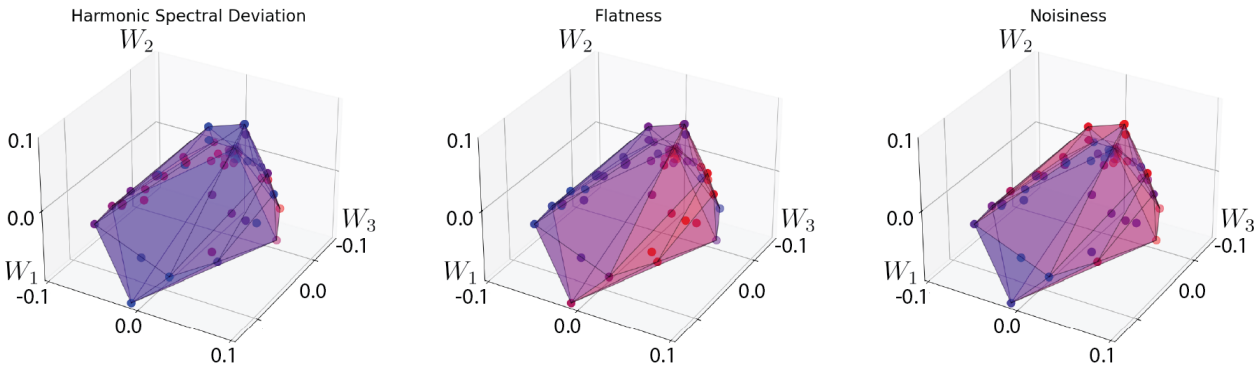
The resulting manifold plots also depict properties concerning the geometric stability of the model from a group theoretical perspective. We equate a smaller manifold volume with a more timbrally homogeneous space, such that a more contained cluster of points would resemble not only a smoother and more continuous latent space across the span of spectrotemporal parameters, but also a more coherent and contained typology of the sound object with less chance of generating sounds that do not adhere to the mesostructural properties of the dataset. In order to evaluate the geometric stability of the disentangled DDSP models, we compare these manifolds to similar manifolds generated using DDSP models conditioned only on loudness and the top scoring parameter from the principal component analysis in Table 6.2. We refer to these models as the 'baseline' models, shown in Fig. 6.6.



(a) Type A [Spring Coil]



(b) Type B [Threaded Rod]



(c) Type C [Styrofoam]

Figure 6.5: Manifolds from each disentangled model generated by applying a convex hull to the outer-most sounds in the isomap JTFS space. Color hues depict the mean value of the spectrotemporal parameter associated with each plot.

Visually comparing the manifolds in Fig. 6.6 first shows that points within the disentangled manifolds are generally much more contained around a precise center of mass in the isomap JTFS space, with the exception of a few extreme outliers. This is in contrast to the baseline models, where the selection of a single parameter makes the resulting sounds more evenly distributed across the space. Likewise, this compactness can be quantified by calculating the L1 and L2 sum of pair-wise distances between each of the models' generated point clusters. Table 6.3 shows that nearly all of the disentangled models contain sets of points that are more locally compact than those of the baseline models. Comparing the volumes of each manifold show that the disentangled manifolds are generally smaller than the baseline manifolds, with the exception of type C which once again could be due to its relatively more stochastic sounds.

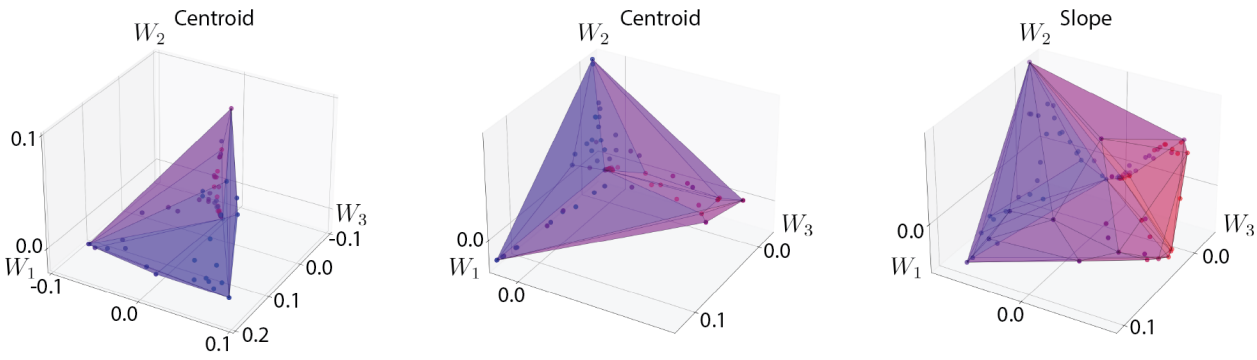


Figure 6.6: Type A (Left), Type B (Center), and Type C (Right) manifolds generated from each baseline model. Baseline models were trained solely on the parameter that accounts for the most variance in each dataset.

6.3.5 Timbre Transfer Results

We then evaluate the ability of both the disentangled models and baseline models to reconstruct friction percussion sounds using a timbre transfer algorithm. We do this by taking a 10 second recording from each sound type and passing it through each model, subsequently evaluating its perceptual similarity to the original recording while also evaluating the corresponding spectrograms to glean more information. We furthermore

	Type A [Spring Coil]: Extrapolation Metrics		
Model	**Sum of Pairwise Distances (L1)**	**Sum of Pairwise Distances (L2)**	**Convex Hull Volume**
Baseline	0.4413	0.3286	0.0463
Disentangled	**0.2525**	**0.1991**	**0.0341**

	Type B [Threaded Rod]: Extrapolation Metrics		
Model	**Sum of Pairwise Distances (L1)**	**Sum of Pairwise Distances (L2)**	**Convex Hull Volume**
Baseline	0.3134	0.2568	0.0167
Disentangled	**0.1245**	**0.0900**	**0.0162**

	Type C [Styrofoam]: Extrapolation Metrics		
Model	**Sum of Pairwise Distances (L1)**	**Sum of Pairwise Distances (L2)**	**Convex Hull Volume**
Baseline	0.4937	0.3541	**0.0675**
Disentangled	**0.4479**	**0.3170**	0.0786

Table 6.3: Volume and distance metrics for each disentangled manifold and baseline manifold.

evaluate the effect of each spectrotemporal parameter on the resulting output, taking each parameter's unique treatment of the spectrogram into account.

Fig. 6.7 - 6.9 show the spectrograms of the different sound excerpts for each type, including their original 10 second recordings, their timbre transfer reconstructions using the disentangled parameters, and their timbre transfer reconstructions using the baseline parameters. The most notable qualitative improvements from the disentangled model can be heard from type B, in which the resonant responses from the scraping of the threaded rod are captured with the disentangled model but not with the baseline model. We believe that this is due to the interaction between the inharmonicity and the slope/decrease parameters, which create a latent control space suitable for modeling the harmonic resonance of the tom with respect to a wide range of distributions of energy across the frequency spectrum. Furthermore, despite the disentangled model better capturing some harmonic partial interaction between the ranges of 2048Hz - 4096Hz, it can be seen that a wide spread of high-frequency partials is present in the original spectrogram from

approximately 7-10 seconds that is not accounted for in either of the reconstructions. If parameters were instead handpicked, this issue could potentially be improved by augmenting the parameter set with the addition of something akin to harmonic-to-noise ratio or harmonic energy.

An improvement can also be heard from type C's reconstruction of partials, which appear to evolve more smoothly across time. These partials are also more present in the last few seconds of the disentangled reconstructions, which capture the resonant responses of the sharp transients more accurately than in the baseline model. This improvement also makes the development of the partials much smoother in the disentangled version. Additionally, the overall distribution of energy across the spectrogram of the disentangled reconstruction much better resembles the magnitude of energy in the original recording, while the baseline parameters produce a generally quieter reconstruction. This causes some discrepencies in the baseline reconstruction, where for instance the first two prominent partials between 512Hz - 1024Hz are much better represented in the disentangled reconstruction than they are in the baseline reconstruction.

Finally, type A contains the most subtle differences between reconstructions, slightly improving the presence of the resonant frequency responses produced from the spring coil on the drum. However, the differences between the baseline and disentangled recontructions are generally very subtle and not immediately perceivable for this sound type. Furthermore, the energy distribution in both recontructions yields a dynamic variance much greater than that of the original, causing a number of important timbral qualities to be lost in both reconstructions. We predict that adding a parameter such as harmonic spectral variation might improve the general harmonic shape of the output, filling in the many gaps of energy missing from both reconstructions.

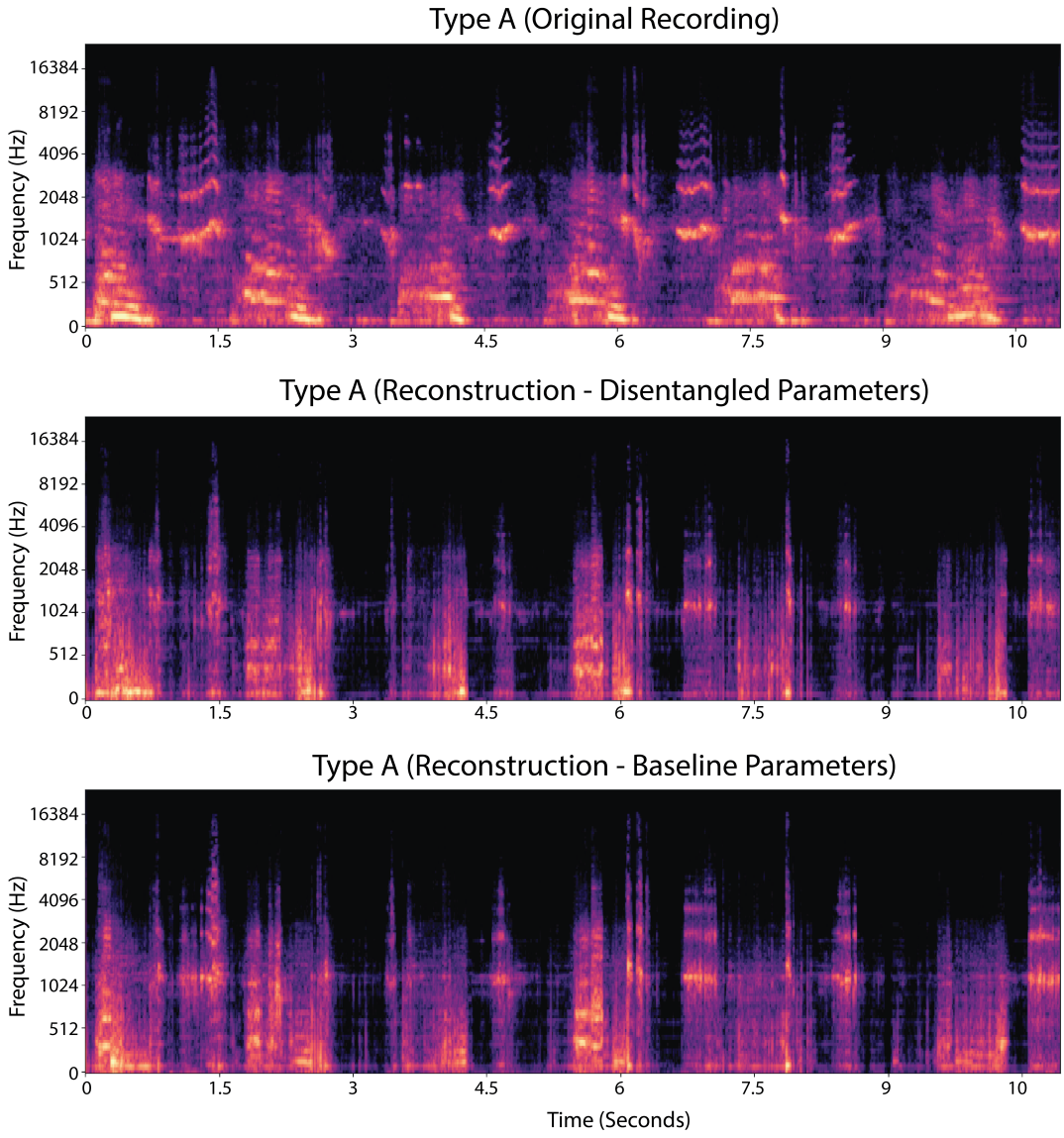


Figure 6.7: Type A [Spring Coil]: Ten second friction improvisation on floor tom using a spring coil. Spectrograms of original audio (Top), reconstruction using disentangled parameters (Middle), and reconstruction using baseline parameters (Bottom).

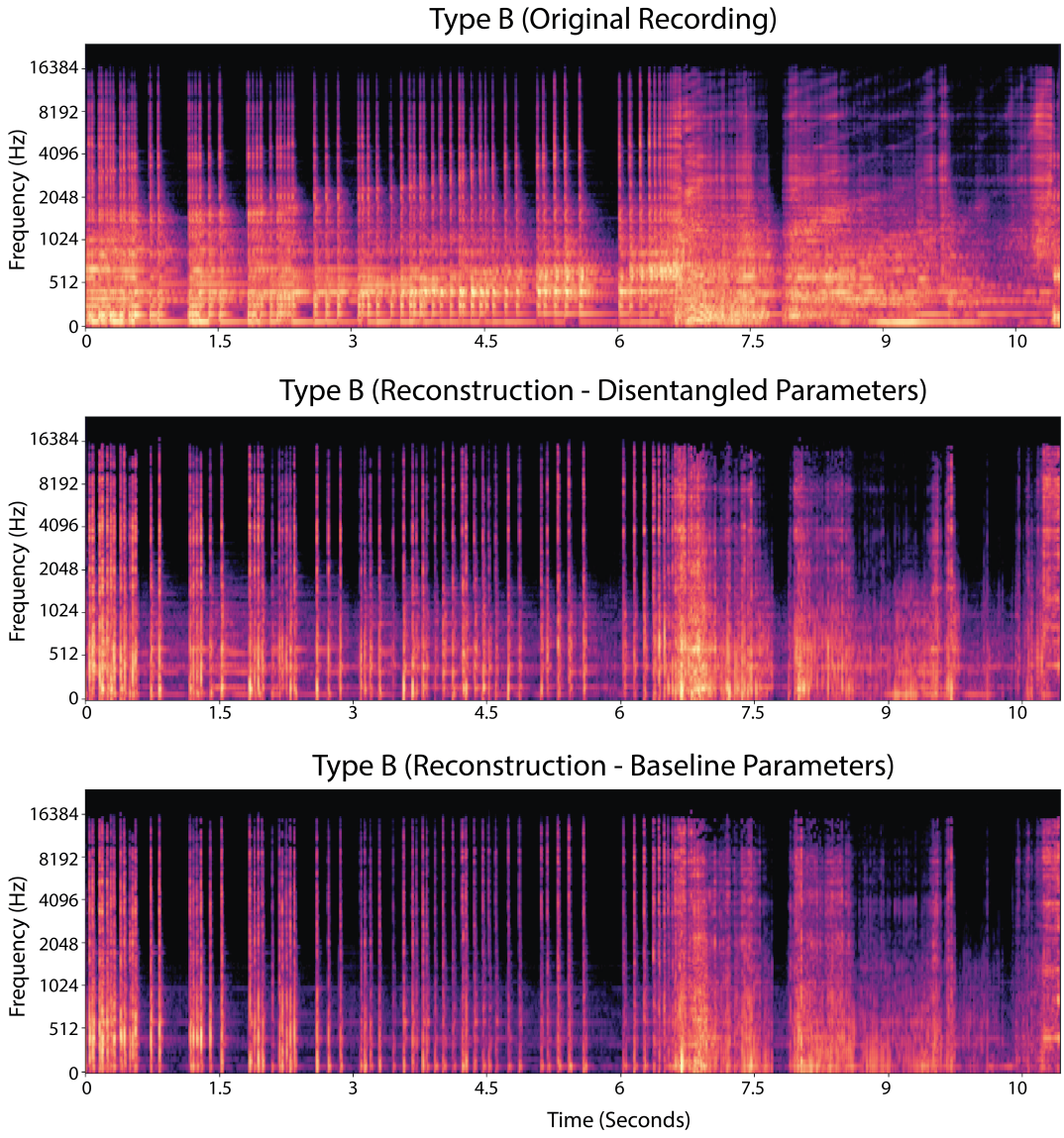


Figure 6.8: Type B [Threaded Rod]: Ten second friction improvisation on floor tom using a threaded rod. Spectrograms of original audio (Top), reconstruction using disentangled parameters (Middle), and reconstruction using baseline parameters (Bottom).

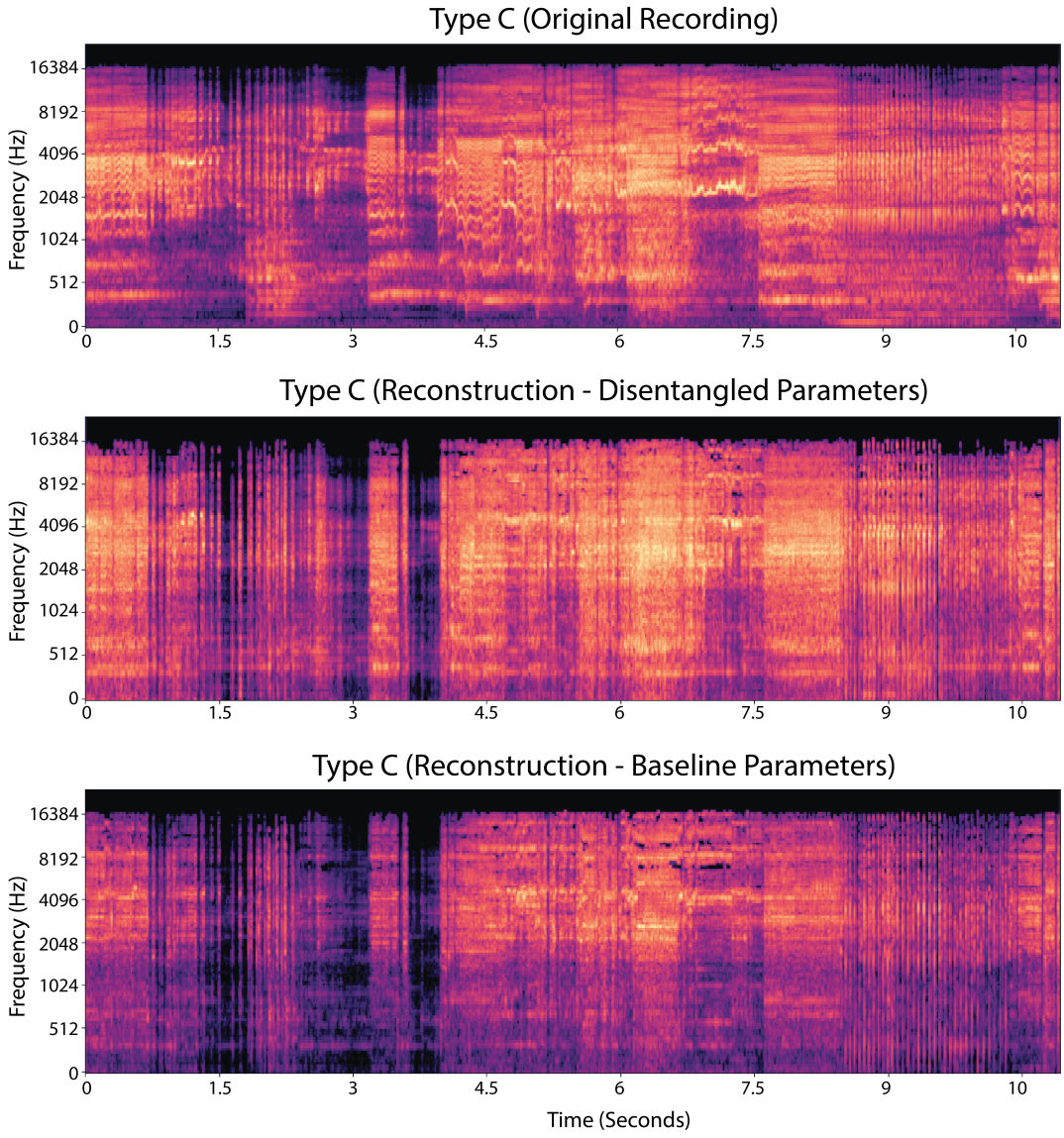


Figure 6.9: Type C [Styrofoam]: Ten second friction improvisation on floor tom using styrofoam. Spectrograms of original audio (Top), reconstruction using disentangled parameters (Middle), and reconstruction using baseline parameters (Bottom).

Chapter 7

Conclusion

7.1 Summary and Future Work

This thesis has explored a number of theoretical and practical methods towards modeling acousmatic sound using the framework of neural audio synthesis. Following the spirit of Schaeffer's philosophy of sound, our central contribution of this thesis has been the broadening of the analysis and synthesis of sound using DDSP by augmenting its typomorphological capabilities with the aid of spectrotemporal audio descriptors and joint time-frequency scattering representations.

We provided a unique methodology that reinforces Schaefferian approaches by focusing on the group invariant properties of sound and the time-frequency analysis of sound in terms of the structure of the topological group. We exercised these methodologies in terms of Schaefferian typology by looking at how scattering networks can be used to classify sounds on the mesoscale strictly by comparing their affine geometries on the time-frequency plane. We followed this by demonstrating a similar approach to Schaefferian morphology by observing how DDSP networks can be used to learn mappings between control parameters and synthesizer parameters that reconstruct the sounds on the microscale by similarly comparing affine time-frequency representations derived from the multiscale spectrogram. We then demonstrated an acousmatic approach to neural audio synthesis that chooses a select set of control parameters that perceptually disentangle the latent control space of time-varying spectrotemporal parameters based on information extracted using a mesostructural analysis of sounds in the dataset. This disentanglement

serves to represent both the typological and morphological modeling of the Schaefferian sound object.

While we experimented with these techniques using both synthetic datasets and real-world percussive datasets, these experiments also present the opportunity for lots of future work. Despite the flexibility of spectrotemporal audio descriptors, as well as their close ties to acousmatic music composition, an important question remains as to whether a suitable method can be constructed for optimizing a completely synthetic set of control parameters that best fit a given dataset's mesostructural distribution. Such a method would act as a more powerful disentanglement tool, but might risk lacking the stability of the perceptual priors found in the spectrotemporal audio descriptors [PGS+11]. Another important question is the feasibility of real-time timbre transfers. Our method for extracting disentangled control parameters takes place prior to audio reconstruction inference, however our choice of DDSP model from [BRC24] does not allow for real-time inference, and many of the harmonic spectrotemporal audio descriptors used in our experiments similarly cannot run in real-time. We don't find this to be an issue for the scope of our work considering that acousmatic composition is canonically in the form of fixed media. However, it would be worthwhile to further adapt other real-time DDSP models such as the one in [EHGR20] to be conditioned on a subset of real-time spectrotemporal audio descriptors.

Finally, we hope that this work influences further creative research in compositional practices involving neural audio synthesis. While we have provided some minimal examples on the companion site for this thesis that demonstrate the use of our methods in generating compositional material, we hope that our research inspires future exploration towards even more imaginative methods of neural audio composition.

Appendix A

Group Representation Theory

This appendix provides all of the prerequisites concerning group representation theory that are needed to comprehend this thesis. We first introduce the notion of the *group*, which was formulated by Galois, later to be adopted by Klein and the Erlangen Programme. Groups are particularly well suited for the analysis of geometric invariance [Bou60]. We then review the *group representation*, which can be roughly interpreted as a linearization of a group's actions. We then arrive at the notion of *representational disentanglement* which is an area of interest in both group representation theory and deep learning [HAP+18] that attempts to associate group actions with corresponding representational subspaces.

The preliminary concepts introduced in this Appendix will allow us to interpret neural audio synthesis models presented in this thesis as compositional chains of operators that each preserve different group actions. For this appendix, we assume only a prerequisite knowledge of linear algebra and set theory.

A.1 Group Theory

We start by introducing the notion of the group, which lends itself to the study of invariant geometric forms across a wide range of mathematical domains. We present some simple examples of groups borrowed from [Löh17] and focus our attention primarily towards the *actions* and *orbits* of a group. We then finish the section by introducing a formal definition of group invariance and equivariance.

A.1.1 Groups

Definition (Group) A group is a set $\mathfrak{G}$ along with a binary operation $\circ : \mathfrak{G} \times \mathfrak{G} \to \mathfrak{G}$ called composition (sometimes simply denoted $\mathfrak{g} \circ \mathfrak{h} = \mathfrak{gh}$) satisfying the following axioms

- *Associativity*: $(\mathfrak{gh})\mathfrak{i} = \mathfrak{g}(\mathfrak{hi})$ for all $\mathfrak{g}, \mathfrak{h}, \mathfrak{i} \in \mathfrak{G}$
- *Identity*: There exists a unique $\mathfrak{e} \in \mathfrak{G}$ such that $\mathfrak{eg} = \mathfrak{ge} = \mathfrak{g}$ for all $\mathfrak{g} \in \mathfrak{G}$
- *Inverse*: For each $\mathfrak{g} \in \mathfrak{G}$ there is a unique inverse $\mathfrak{g}^{-1} \in \mathfrak{G}$ such that $\mathfrak{gg}^{-1} = \mathfrak{g}^{-1}\mathfrak{g} = \mathfrak{e}$
- *Closure*: The group is closed under composition (for all $\mathfrak{g}, \mathfrak{h} \in \mathfrak{G}$ there exists a $\mathfrak{gh} \in \mathfrak{G}$)

Example (Algebraic Groups) Groups are commonly introduced in the context of abstract algebra (see introductory texts such as [Art11]). For instance, the set of integers $\mathbb{Z}$ along with the operation of addition $+ : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ form a group called the *additive group of integers* $(\mathbb{Z}, +)$. Likewise, the real numbers $\mathbb{R}$ along with the operation of multiplication $\cdot : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ form the *multiplicative group* $(\mathbb{R}, \cdot)$.

Example (Symmetric Groups) Some groups denote the collection of all *symmetries* within a set. These groups are called *symmetric groups*, and contain a set of *transformations* along with the operation of composition $\circ : \mathfrak{G} \times \mathfrak{G} \to \mathfrak{G}$. A symmetry can be thought of as a transformation that leaves a property of a certain mathematical object unchanged [BBCV21].

Definition (Subgroup) A subset $\mathfrak{H} \subseteq \mathfrak{G}$ of a group $\mathfrak{G}$ forms a subgroup if it is closed under composition and is able to take inverses. We denote the group $\mathfrak{H}$ as a subgroup of the group $\mathfrak{G}$ by writing $\mathfrak{H} \leq \mathfrak{G}$.

Remark [BBCV21] points out that subgroups often corelate to rather intuitive subsets. For instance, the additive group of integer vectors $(\mathbb{Z}^N, +)$ is a subgroup of the additive group of real vectors $(\mathbb{R}^N, +)$, each of dimension N. The 2-dimensional special orthogonal group SO(2)—to be defined in the next section—is a subgroup of the 3-dimensional special orthogonal group SO(3). Finally, each group has a subgroup that contains only the identity $\mathfrak{e} \leq \mathfrak{G}$ and a subgroup that contains only itself $\mathfrak{G} \leq \mathfrak{G}$. These are known as the *trivial* subgroups. Often the symmetric groups are subgroups of larger groups, and we say that they are *generated* from small subsets of these larger groups.

A.1.2 Generators

Definition (Generators) A group generator is a set of group elements $\langle \mathfrak{g}, \mathfrak{h} ... \rangle^{\mathfrak{G}}$ that can derive all elements of $\mathfrak{G}$ using only the elements of the set and the group operation $\circ$.

Example (Dihedral Group) Let's borrow an example from [Löh17] which looks at the *dihedral groups* $\mathfrak{D}_N$. A dihedral group consists of a set $\mathfrak{D}_N$ that we say is *generated* using the set $\langle \mathfrak{r}, \mathfrak{s} \rangle^{\mathfrak{D}_N}$ whose elements consist of the transformations of reflection ($\mathfrak{r}$) and rotation ($\mathfrak{s}$) along with the operation of composition $\circ : \mathfrak{D}_N \times \mathfrak{D}_N \to \mathfrak{D}_N$. Consider the equilateral triangle in Fig. A.1 (a), which we might informally say *represents* $\mathfrak{D}_3$. Rotation and reflection of the triangle provide an intuitive geometric interpretation of the group $\mathfrak{D}_3$ since these transformations along with their composition abide by all of the group axioms.

Proof More rigorously, we can check each of the group axioms by constructing what's called a Cayley table [Fig. A.1 (b)]. This table provides information that pertains to every possible permutation of the group elements. For the dihedral groups, this would contain one identity $\mathfrak{e}$ (which does nothing), one reflection $\mathfrak{r}$, and N rotations—in this case, 3. We then check for the existence of each axiom.

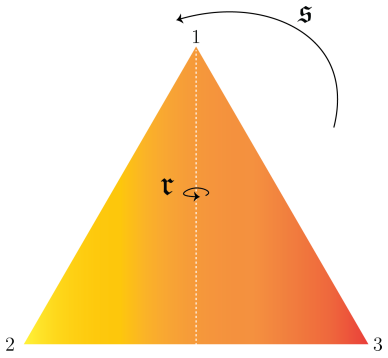
- *Associativity*: Any symmetrical shape's transformations are inherently associative
- *Identity*: We denote $\mathfrak{e}$ as the identity, which is a transformation that does nothing on the triangle
- *Inverse*: An inverse can always be reached with the following composition of transformations: $\mathfrak{r} \circ \mathfrak{s} \circ \mathfrak{r} = \mathfrak{r}^{-1}$
- *Closure*: The table contains transformations and group elements that are exclusive to the group $\mathfrak{D}_3$

□

Remark In the Husserlian or perhaps Schaefferian sense, we can better understand the essence of the triangle since we see that the object is the same, or *invariant* throughout the span of its transformations shown in the Cayley table. We will thus primarily be working with symmetric groups, since they focus on this very notion of invariance through a formalization of geometric symmetry.

$\circ$	$\mathfrak{e}$	$\mathfrak{s}$	$\mathfrak{s}^2$	$\mathfrak{r}$	$\mathfrak{rs}$	$\mathfrak{rs}^2$
$\mathfrak{e}$	$\mathfrak{e}$	$\mathfrak{s}$	$\mathfrak{s}^2$	$\mathfrak{r}$	$\mathfrak{rs}$	$\mathfrak{rs}^2$
$\mathfrak{s}$	$\mathfrak{s}$	$\mathfrak{s}^2$	$\mathfrak{e}$	$\mathfrak{rs}^2$	$\mathfrak{r}$	$\mathfrak{rs}$
$\mathfrak{s}^2$	$\mathfrak{s}^2$	$\mathfrak{e}$	$\mathfrak{s}$	$\mathfrak{rs}$	$\mathfrak{rs}^2$	$\mathfrak{r}$
$\mathfrak{r}$	$\mathfrak{r}$	$\mathfrak{rs}$	$\mathfrak{rs}^2$	$\mathfrak{e}$	$\mathfrak{s}$	$\mathfrak{s}^2$
$\mathfrak{rs}$	$\mathfrak{rs}$	$\mathfrak{rs}^2$	$\mathfrak{r}$	$\mathfrak{s}^2$	$\mathfrak{e}$	$\mathfrak{s}$
$\mathfrak{rs}^2$	$\mathfrak{rs}^2$	$\mathfrak{r}$	$\mathfrak{rs}$	$\mathfrak{s}$	$\mathfrak{s}^2$	$\mathfrak{e}$

(a) Cayley table



(b) $\mathfrak{D}_3$ transformations

Figure A.1: (a) Cayley table consisting of the $\mathfrak{D}_3$ group generated by rotations $\mathfrak{r}$ and reflections $\mathfrak{s}$. (b) $\mathfrak{D}_3$ transformations represented on an equilateral triangle.

A.1.3 Actions

Definition (Action) A group action of $\mathfrak{G}$ on a set Ω is defined as a mapping

$$\triangleright : \mathfrak{G} \times \Omega \to \Omega \tag{A.1}$$

associating a group element $\mathfrak{g} \in \mathfrak{G}$ and a point $u \in \Omega$ with some other point $u' \in \Omega$ such that $u' = \mathfrak{g} \triangleright u$ (where $\mathfrak{g} \triangleright u$ denotes the group action) in a way that is compatible with the group axioms.

Example We turn our attention to the *special orthogonal group*, one of many groups defined over a set that forms a *field* $\Omega = \mathbb{F}$. Informally speaking, a field is a set equipt with an addition and multiplication operation, and common examples of fields include the set of real numbers $\mathbb{R}$, the set of rational numbers $\mathbb{Q}$, and the set of complex numbers $\mathbb{C}$. We denote the special orthogonal group $\mathrm{SO}(2, \mathbb{F})$ defined over some field $\mathbb{F}$. This group consists of the set of all 2×2 invertible matrices with $\det = 1$ of the form

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{A.2}$$

equipt with the binary operation of matrix multiplication $\cdot : \mathbb{F}^2 \times \mathbb{F}^2 \to \mathbb{F}^2$. Like the equilateral triangle in the previous example, the operation of rotation $\mathfrak{s}$ is *represented* through 2×2 matrices. We can thus use the matrix defined in Eq. A.2 to demonstrate the group SO(2)'s *action* on a set.

An example of this can be shown by applying SO(2)'s action of rotation onto the set of complex numbers $\mathbb{C}$. Consider a complex exponential phasor, commonly used to represent a sinusoid with frequency ω, amplitude A, and phase ϕ.

$$z(t) = Ae^{i(\omega t+\phi)} \tag{A.3}$$

The group SO(2)'s action over the complex numbers results in a phase-shift to the output

$$\mathfrak{s} \triangleright Ae^{i(\omega t+\phi)} = Ae^{i(\omega t+\phi+\theta)} \tag{A.4}$$

Proof The matrix in Eq. A.2 can alternatively be defined as an exponential map, as demonstrated in [GQ20].

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \equiv e^{i\theta} \tag{A.5}$$

When multiplied, this yields a phase-shift in the exponent:

$$e^{i\theta} \cdot z(t) = Ae^{i\theta}e^{i(\omega t+\phi)} = Ae^{i(\omega t+\phi+\theta)} \tag{A.6}$$

Replacing multiplication with the group action operation for clarity, we come back to a phase-shift that performs the action of rotation $\mathfrak{s}$.

$$\mathfrak{s} \triangleright z(t) = Ae^{i(\omega t+\phi+\theta)} \tag{A.7}$$

□

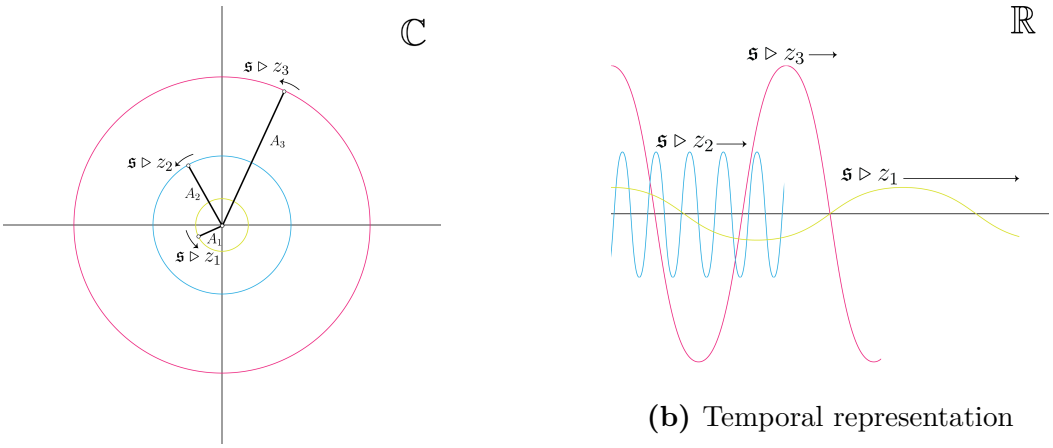
A.1.4 Orbits

Definition (Orbit) Let $\mathfrak{g} \triangleright u$ be an action of $\mathfrak{G}$ on a set Ω and consider any element $u \in \Omega$. The subset

$$\mathfrak{G} \triangleright u := \{\mathfrak{g} \triangleright u \mid \mathfrak{g} \in \mathfrak{G}\} \tag{A.8}$$

of Ω is then denoted as the *orbit* of u.

Corollary Given a constant amplitude A, the set of all points of a sinusoid can be thought of as SO(2)'s orbit on a complex exponential phasor. Fig. A.2 demonstrates the orbits of three complex exponential phasors z_1, z_2, and z_3, with respective constant amplitudes A_1, A_2, and A_3. The orbits can be visualized as rotations around the circle (a) or as sinusoids in the temporal domain (b).



(a) Phase representation

(b) Temporal representation

Figure A.2: Rotations $\mathfrak{s}$ acting on three different complex exponential z_n phasors of increasing amplitude. Orbits of each z_n trace out circles in $\mathbb{C}$ (a), and temporal sinusoids in $\mathbb{R}$ (b).

In this sense, the group's orbit can be interpreted as a *parametric space* for sound with respect to its actions. It defines certain geometric boundaries given a fixed set of parameters, such as the case with amplitude above. In this case, the set could be interpreted as an additive synthesizer constrained by a parameterization of phase interaction ($\mathfrak{s}$) between z_1, z_2 and z_3.

A.2 Functions Over Groups

We now define some basic functional terminology in group theory, including *homomorphisms* and *endomorphisms* [Löh17]. We then formally define the properties of group *invariance* and *equivariance*, which deal with the preservation of group actions over functions. These properties have previously been observed as fundamental to Schaeffer's philosophy of sound and the Erlangen Programme's philosophy of geometry.

A.2.1 Group Homomorphisms and Endomorphisms

Definition (Homomorphism) A map $\eta : \mathfrak{G} \to \mathfrak{H}$ is a group homomorphism if η is compatible with the composition of each group respectively, i.e. if

$$\eta(\mathfrak{g}_1 \cdot \mathfrak{g}_2) = \eta(\mathfrak{g}_1) \cdot \eta(\mathfrak{g}_2) \tag{A.9}$$

holds $\forall(\mathfrak{g}_1, \mathfrak{g}_2) \in \mathfrak{G}$. We denote $\mathrm{Hom}(\mathfrak{G}, \mathfrak{H})$ as the set of all group homomorphisms from $\mathfrak{G}$ to $\mathfrak{H}$.

Definition (Endomorphism) If a similar map $\eta : \mathfrak{G} \to \mathfrak{G}$ follows the same properties as A.9, the map is a group endomorphism. We denote $\mathrm{End}(\mathfrak{G})$ as the set of all group endomorphisms from $\mathfrak{G}$ to itself.

A.2.2 Group Invariance and Equivariance

Definition (Invariance) Let $\triangleright_\Omega$ be a group action on the set Ω. A function $f : \Omega \to \Omega'$ is $\mathfrak{G}$-invariant if it satisfies $f(\mathfrak{g} \triangleright_\Omega u) = f(u) \quad \forall \mathfrak{g} \in \mathfrak{G}, \forall u \in \Omega$.

Example The area S of an N-gon with vertices $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ can be calculated using the Shoelace theorem:

$$S = f(\mathbf{u}) = \frac{1}{2}\left|\sum_{i=1}^{N}(x_i y_{i+1} - y_i x_{i+1})\right| \tag{A.10}$$

where the input $\mathbf{u}$ is a vector of pairs representing our vertices

$$\mathbf{u} = \begin{bmatrix} (x_1, y_1) \\ (x_2, y_2) \\ \vdots \\ (x_N, y_N) \end{bmatrix} \tag{A.11}$$

Now let's introduce some non-linearity to our function f that only returns an output if our intput is above a certain threshold

$$\hat{f}(\mathbf{u}) = \begin{cases} 1 & \text{if } f(\mathbf{u}) \geq C \\ 0 & \text{if } f(\mathbf{u}) < C \end{cases} \tag{A.12}$$

where C is some constant representing an area threshold. Let's once again introduce the transformations of the group $\mathfrak{D}_N$ as group actions on the plane $\mathbb{R}^2$. The actions $\{\mathfrak{r}, \mathfrak{s}\} \in \mathfrak{D}_N$ leave the resulting area unchanged, such that $\hat{f}(\mathfrak{r} \triangleright_{\mathbb{R}^2} \mathbf{u}) = \hat{f}(\mathbf{u})$ and $\hat{f}(\mathfrak{s} \triangleright_{\mathbb{R}^2} \mathbf{u}) = \hat{f}(\mathbf{u})$. The function $\hat{f}$ is thus *invariant* to rotation and reflection.

Remark (Classification) Not only is $\hat{f}$ invariant to these actions, but so is the original Shoelace function f. The function $\hat{f}$, however, allows us to informally relate this notion of invariance to the task of *classification* [Wei22], in which a machine learning model is asked to specify to which k categories some input belongs [GBC16]. In the case of $\hat{f}$, $k = 2$. The need for $\hat{f}$ over f is necessary since in classification situations, the domain Ω is often much larger than the codomain Ω'. Ω' is usually a finite set representing different classes, derived as a result of a nonlinearity similar to that of $\hat{f}$. We relate invariance and classification to Schaeffer's notion of *typology* in Chapter 4.

Definition (Equivariance) Let $\triangleright_\Omega$ and $\triangleright_{\Omega'}$ be group actions on the sets Ω and Ω'. A function $f : \Omega \to \Omega'$ is $\mathfrak{G}$-equivariant if it commutes with the actions $f(\mathfrak{g} \triangleright_\Omega u) = \mathfrak{g} \triangleright_{\Omega'} f(u) \quad \forall \mathfrak{g} \in \mathfrak{G}, \forall u \in \Omega$.

Example Using the vector $\mathbf{u}$, we can define a function h that shifts and scales all of the vertices that produce our N-gon

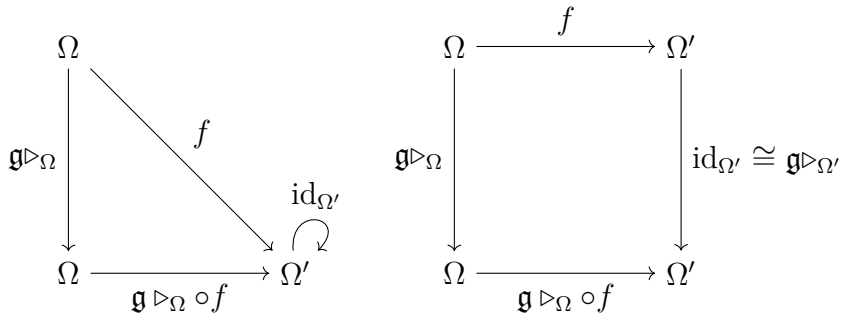
$$h(\mathbf{u}) = C \cdot \mathbf{u} + \mathbf{b} \tag{A.13}$$

where C is a scalar and $\mathbf{b}$ is the shift factor. This function is *equivariant* to the actions of $\mathfrak{D}_N$ since $f(\mathfrak{r} \triangleright_{\mathbb{R}^2} \mathbf{u}) = \mathfrak{r} \triangleright_{\mathbb{R}^2} f(\mathbf{u})$ and $f(\mathfrak{s} \triangleright_{\mathbb{R}^2} \mathbf{u}) = \mathfrak{s} \triangleright_{\mathbb{R}^2} f(\mathbf{u})$. In other words, applying these group actions before the function is calculated will yield the same output as applying them after the function is calculated.

Remark Whereas we related invariance to classification, equivariance can be also related to *regression* [Wei22]. Regression is a task in which a machine learning model is asked to predict a numerical value given some input. Indeed, this reflects the case above since the output domain Ω produces a linear numerical result. Similarly, we relate equivariance to Schaeffer's concept of *morphology* in Chapter 5.

Proposition $\mathfrak{G}$-invariance can be interpreted as a special case of $\mathfrak{G}$-equivariance.

Proof This is quite eloquently shown visually in [Wei22] using commutative diagrams. If we interpret the codomain of an invariant map as a trivial action $\mathrm{id}_{\Omega'}$, then the two diagrams become isomorphic.



(A.2.1)

For this reason, we will sometimes simply refer to invariance when talking about both invariance and equivariance.

□

A.3 Group Representations

While studying arbitrary group actions can be a promising way to learn more about a certain mathematical object, we can obtain more pertinent information by studying *group representations*. A group representation can be described as a *linearization* of a set Ω via the group action $\mathfrak{g}$ [Mal16]. This linear representation allows the group and its actions to easily be analyzed within the context of vector spaces and tensor algebras. In this section, we formally define the group representation and subrepresentation, while also reviewing two common vector space operations.

Definition (Representation) Consider $\mathrm{GL}(n, \mathbb{F})$ the general linear group of vector spaces over some field $\mathbb{F}$. This group consists of $n \times n$ invertible matrices with non-zero determinant (sometimes denoted $\mathrm{GL}(V)$). An n-dimensional real representation of a group $\mathfrak{G}$ is a map to the general linear group

$$\rho : \mathfrak{G} \to \mathrm{GL}(n, V) \tag{A.14}$$

assigning to each $\mathfrak{g} \in \mathfrak{G}$ a representation $\rho(\mathfrak{g})$, and satisfying the condition $\rho(\mathfrak{gh}) = \rho(\mathfrak{g})\rho(\mathfrak{h})$ for all $\mathfrak{g}, \mathfrak{h} \in \mathfrak{G}$.

A.3.1 Subrepresentations

Definition (Subrepresentation) Given a representation $\rho : \mathfrak{G} \to \mathrm{GL}(V)$, a subrepresentation is any vector subspace $W \leq V$ that is invariant under the action of $\mathfrak{G}$. That is, $\rho|_W(\mathfrak{g}) \in W \quad \forall \mathfrak{g} \in \mathfrak{G}, \forall w \in W$.

Remark Because subrepresentations are closely tied to their corresponding group actions, we will often denote a subrepresentation as a pair of both the representation and the vector space $\rho|_W$ following notation used in [Wei22] of a *restricted* representation. This notation means that the subspace W is invariant and closed under the action of $\mathfrak{G}$. Using a slight notational leap, this might also refer to the original representation $\rho|_V$ since trivially every representation is also a subrepresentation $\rho|_V \subseteq \rho|_V$

A.3.2 Direct Sum and Tensor Product

Definition (Direct Sum) The direct sum of vector spaces V and W over an arbitrary field $\mathbb{F}$, denoted $V \oplus W$, is the set of all ordered pairs (v, w), where $v \in V$ and $w \in W$, equipped with component-wise addition and scalar multiplication:

$$V \oplus W = \{(v, w) \mid v \in V, w \in W\}$$

with addition defined as

$$(v_1, w_1) + (v_2, w_2) = (v_1 + v_2, w_1 + w_2)$$

and scalar multiplication defined as

$$\alpha \cdot (v, w) = (\alpha \cdot v, \alpha \cdot w)$$

where α is a scalar.

Corollary We can incorporate the direct sum into our group representation notation by constructing, for instance, the sum of two group representations $(\rho_1 \oplus \rho_2) : \mathfrak{G} \to \mathrm{GL}(V \oplus W)$.

Definition (Tensor Product) The tensor product $V \otimes W$ of two vector spaces is the space based on elements $v \otimes w$, labelled by pairs of vectors $v \in V$ and $w \in W$ with the following distributive relations:

- *Addition*: $(v_1 + v_2) \otimes w = v_1 \otimes w + v_2 \otimes w$;
- *Multiplication*: $v \otimes (w_1 + w_2) = v \otimes w_1 + v \otimes w_2$
- *Scalar Multiplication*: $(k \cdot v) \otimes w = v \otimes (k \cdot w) = k \cdot (v \otimes w)$.

In other words, $V \otimes W$ is the quotient of the vector space with basis $\{v \otimes w\}$ by the subspace spanned by the differences of left- and right-hand sides in each identity above.

Corollary Similarly, we can incorporate the tensor product into our group representation notation by introducing the product of two group representations $(\rho_1 \otimes \rho_2) : \mathfrak{G} \to \mathrm{GL}(V \otimes W)$ [Tel05].

A.4 Representational Disentanglement

Finally, we turn to [HAP+18] for definitions concerning irreducible representations and representational disentanglement, which are integral to the study of invariance and equivariance in neural audio synthesis. Texts such as [HAP+18], [KPB+23], and [FWW21] are prefaced on the idea that the most desirable neural network architectures for representation learning are ones in which the layers *disentangle* input representations, since this allows for interpretability.

A.4.1 Irreducibility

So far we've examined group representations whose group transformations $\mathfrak{g} \in \mathfrak{G}$ act on an arbitrary vector spaces V. In this section, we consider vector spaces that *decompose* into some direct sum of the form $\rho|_{W_1} \oplus \rho|_{W_2} \oplus ... \oplus \rho|_{W_L}$. We continue to denote the composition of these subspaces as $\rho|_V$ such that:

$$\begin{aligned} \rho &: \mathfrak{G} \to \mathrm{GL}(V) \\ \rho|_V &:= \bigoplus_{l=1}^{L} \rho|_{W_l} \end{aligned} \tag{A.15}$$

Definition (Irreducible Representations) A representation $\rho|_V$ is said to be *irreducible* if its only subrepresentations $\rho|_W \subseteq \rho|_V$ are the two trivial subrepresentations $\rho|_W = \{0\}$ and $\rho|_W = \{\rho|_V\}$.

Corollary As a visual aid, the decomposition into irreducible representations can be shown as a block-diagonal matrix:

$$\rho|_V := \begin{bmatrix} \rho|_{W_1} & 0 & \cdots & 0 \\ 0 & \rho|_{W_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho|_{W_L} \end{bmatrix} \tag{A.16}$$

Under most conditions, the decomposition of $\rho|_V$ will always yield the same set of irreducible representations $\rho|_{W_L}$ up to isomorphism, order, and change of basis [HAP+18].

A.4.2 Disentanglement

Definition (Disentanglement) A group representation $\rho|_V : \mathfrak{G} \to \mathrm{GL}(V)$ is disentangled if there exists a decomposition of the following form:

$$\rho|_V = \bigoplus_{l=1}^{L} \bigotimes_{m=1}^{M} \rho^{(m)}|_{W_l} \tag{A.17}$$

where each factor $\rho^{(m)}$ is an irreducible representation of some $\mathfrak{G}_l$ with respect to the group decomposition $\mathfrak{G} = \mathfrak{G}_1 \times \mathfrak{G}_2 ... \mathfrak{G}_L$ and there is at most one non-trivial representation in each $\rho^{(m)}$.

Example Following [HAP+18], we can analyze a simpler example of a disentangled representation of the group $\mathfrak{G} = \mathfrak{G}_1 \times \mathfrak{G}_2$ where the representation contains two linearly independent vector spaces $V = W_1 \oplus W_2$ that each contain two irreducible factors $\rho^{(1)} \otimes \rho^{(2)}$.

$$\rho|_V = (\rho^{(1)}|_{W_1} \otimes \rho^{(2)}|_{W_1}) \oplus (\rho^{(1)}|_{W_2} \otimes \rho^{(2)}|_{W_2}) \tag{A.18}$$

Because we know that each subspace W_l is the minimal invariant subspace under the group actions of $\mathfrak{G}$, we can infer that at the very most only one $\rho^{(m)}$ is non-trivial, which allows each W_l to respectively represent only one $\mathfrak{G}_l$, hence the term disentanglement.

Caveat (Linear Disentanglement) It is worth mentioning that if we are dealing with *linear* group representations, our problem simplifies drastically. Namely, the factors of the tensor product $\bigotimes_{m=1}^{M} \rho^{(m)}$ disappear and we only have to deal with the linear group actions that decompose into a direct sum, such as those outlined previously in Eq. A.15. The problem of disentanglement then simply becomes a problem of associating linear subspaces W_l with their corresponding groups and group actions $\mathfrak{G}_l$.

Appendix B

Spectrotemporal Audio Descriptors

This appendix covers a number of spectrotemporal audio descriptors discussed in *The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals* [PGS+11]. We review a selection of audio descriptors chosen as parameters for our disentanglement experiments in Chapter 6, providing short descriptions and equations.

B.1 Statistical Descriptors

This section reviews the statistical audio descriptors used for our experiments. These include spectral centroid, spread, skewness, kurtosis, flatness, slope, decrease, roll-off, and crest. We denote the magnitude of the STFT in this section as $X(t, f_k)$, where f_k is the kth spectral bin of K bins and f_0 is the fundamental frequency.

Spectral Centroid: The "center of mass" of the spectrum, representing the perceived brightness of the sound.

$$c(t) = \frac{\sum_k f_k X(t, f_k)}{\sum_k X(t, f_k)}$$

Spectral Spread: The dispersion of the spectral energy around the spectral centroid, indicating the bandwidth.

$$\text{Spread}(t) = \sqrt{\frac{\sum_k (f_k - c(t))^2 X(t, f_k)}{\sum_k X(t, f_k)}}$$

Spectral Skewness: The asymmetry of the spectral shape, providing information about whether the direction energy favors lower or higher frequencies.

$$\text{Skew}(t) = \left(\frac{\sum_k \left(f_k - c(t)\right)^3 X(t, f_k)}{\sum_k X(t, f_k)} \right) \Big/ \text{Spread}(t)^3$$

Spectral Kurtosis: The peakedness of the spectral shape, representing how flat or peaked the spectrum is.

$$\text{Kurt}(t) = \left(\frac{\sum_k \left(f_k - c(t)\right)^4 X(t, f_k)}{\sum_k X(t, f_k)} \right) \Big/ \text{Spread}(t)^4$$

Spectral Flatness: Measures how flat or noise-like the spectrum is, calculated as the ratio of the geometric mean to the arithmetic mean of the power spectrum.

$$\text{Flatness}(t) = \frac{\left(\prod_k X(t, f_k)\right)^{\frac{1}{K}}}{\frac{1}{K} \sum_k X(t, f_k)}$$

Spectral Slope: The slope of the spectral envelope, computed using a linear regression over the spectral amplitude values. Here, $\overline{X}(t, f_k)$ denotes the mean amplitude of the STFT.

$$\text{Slope}(t) = \frac{\sum_k (f_k - c(t))(X(t, f_k) - \overline{X}(t, f_k))}{\sum_k (f_k - c(t))^2}$$

Spectral Decrease: Describes the rate of amplitude decrease, emphasizing the slopes of the lowest frequencies.

$$\text{Decrease}(t) = \frac{\sum_k \frac{X(t, f_{k+1}) - X(t, f_0)}{f_k}}{\sum_k X(t, f_k)}$$

Spectral Roll-off: The frequency below which a certain percentage (usually 95%) of the spectral energy is contained. Here, the denominator sum denotes the sum of frequencies up to the Nyquist limit $\frac{S_R}{2}$.

$$\text{Roll}(t) = f_c(t) \text{ s.t. } \left(\frac{\sum_{f \leq f_c} X(t, f)^2}{\sum_{f \leq f_{\frac{S_R}{2}}} X(t, f)^2} \geq 0.95 \right)$$

Spectral Crest: Measures the peakiness of the spectrum, calculated as the ratio of the maximum spectral value to the average spectral value.

$$\text{Crest}(t) = \frac{\max_k X(t, f_k)}{\frac{1}{K}\sum_k X(t, f_k)}$$

B.2 Harmonic Descriptors

This section reviews the harmonic audio descriptors used for our experiments. These include harmonic energy, noise energy, harmonic-to-noise ratio, inharmonicity, harmonic spectral variation, noisiness, and odd-to-even ratio. Here we denote harmonic partial frequencies as f_h in order to differentiate from spectral bins.

Harmonic Energy: The total energy of the harmonic components in the audio signal.

$$\text{HarmEn}(t) = \sum_{f \in \text{harm}} X(t, f)^2$$

Noise Energy: The total energy of the noise components in the audio signal.

$$\text{NoiseEn}(t) = \sum_{f \in \text{noise}} X(t, f)^2$$

Harmonic-to-Noise Ratio: The ratio of the energy of the harmonic components to the energy of the noise components, representing the degree of noisiness in the sound.

$$\text{HarmToNoise}(t) = \frac{E_{\text{harm}}(t)}{E_{\text{noise}}(t)}$$

Inharmonicity: Measures the deviation of the partials from the harmonic series, indicating the presence of inharmonic partials in the sound.

$$\text{Inharm}(t) = \frac{\sum_h \left(f_h - h f_0\right)^2 X(t, f_h)}{\sum_h X(t, f_h)} \cdot \frac{2}{f_0}$$

Harmonic Spectral Deviation: Measures the deviation of the harmonic amplitudes from a smooth spectral envelope (denoted $\hat{X}(t, f_h)$), indicating irregularities in the harmonic structure. Often $\hat{X}(t, f_h)$ is estimated by averaging the values of three adjacent

partials: $\hat{X}(t, f_h) = \frac{1}{3}[X(t, f_{h-1}) + X(t, f_h) + X(t, f_{h+1})]$ [PGS+11]. We denote H as the maximum number of harmonics in the analysis.

$$\text{HarmDev}(t) = \frac{1}{H} \frac{\sum_{h \in H} \left| X(t, f_h) - \hat{X}(t, f_h) \right|}{\sum_{h \in H} X(t, f_h)}$$

Odd-to-Even Harmonic Ratio: The ratio of the energy of the odd harmonics to the even harmonics, often related to the timbral characteristics of the sound.

$$\text{OddEven}(t) = \frac{\sum_{h \in \text{odd harmonics}} X(t, f_h)^2}{\sum_{h \in \text{even harmonics}} X(t, f_h)^2}$$

Noisiness: Quantifies the amount of noise present in the audio signal, often derived from the energy of non-harmonic components.

$$\text{Noisiness}(t) = \frac{E_{\text{noise}}(t)}{E_{\text{total}}(t)}$$

Bibliography

[AAE+22] Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Anden, Eugene Belilovsky, Joan Bruna, Vincent Lostanlen, Muawiz Chaudhary, Matthew J. Hirn, Edouard Oyallon, Sixin Zhang, Carmine Cella, and Michael Eickenberg. Kymatio: Scattering Transforms in Python, 2022.

[ABH+24] L.D. Abreu, P. Balazs, N. Holighaus, F. Luef, and M. Speckbacher. Time-Frequency Analysis on Flat Tori and Gabor Frames in Finite Dimensions. *Applied and Computational Harmonic Analysis*, 69:101622, 2024.

[AKZB11] Daniel Arfib, Florian Keiler, Udo Zölzer, and Jordi Bonada. Time-Frequency Processing. In Udo Zölzer, editor, *DAFX: Digital Audio Effects*, pages 219–278. Wiley Telecom, 2011.

[AKZV11] Daniel Arfib, Florian Keiler, Udo Zölzer, and Vincent Verfaille. Source-Filter Processing. In Udo Zölzer, editor, *DAFX: Digital Audio Effects*, pages 279–320. Wiley Telecom, 2011.

[ALM19] Joakim Anden, Vincent Lostanlen, and Stéphane Mallat. Joint Time-Frequency Scattering. *IEEE Transactions on Signal Processing*, 67(14):3704–3718, Jul 2019.

[AM14] Joakim Anden and Stéphane Mallat. Deep Scattering Spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, aug 2014.

[Art11] Michael Artin. *Algebra*. Pearson Education, 2011.

[Bat07] Marc Battier. What The GRM Brought to Music: From Musique Concrète to Acousmatic Music. *Organised Sound*, 12(3):189–202, 2007.

[BBCV21] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. `https://arxiv.org/pdf/2104.13478`, 2021.

[Ber10] Marcel Berger. *Geometry Revealed: A Jacob's Ladder to Higher Geometry*. Springer, Berlin, Heidelberg, 2010.

[Bou60] N. Bourbaki. *Eléments d'histoire des mathématiques*. Histoire de la pensée. Springer Berlin Heidelberg, 1960.

[BRC24] Adrián Barahona-Ríos and Tom Collins. NoiseBandNet: Controllable Time-Varying Neural Synthesis of Sound Effects Using Filterbanks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1573–1585, 2024.

[Bro21] Harley Brown. A Visual History of Spatial Sound. *Red Bull Music Academy Daily*, 2021. `https://daily.redbullmusicacademy.com/2018/09/a-visual-history-of-spatial-sound`.

[BSAL11] Jordi Bonada, Xavier Serra, Xavier Amatriain, and Alex Loscos. Spectral Processing. In Udo Zölzer, editor, *DAFX: Digital Audio Effects*, pages 393–445. Wiley Telecom, 2011.

[BSL13] Joan Bruna, Arthur Szlam, and Yann LeCun. Learning Stable Group Invariant Representations with Convolutional Networks. *CoRR*, abs/1301.3537, 2013.

[Cel17] Carmine-Emanuele Cella. Machine Listening Intelligence. *CoRR*, abs/1706.09557, Jun 2017. `http://arxiv.org/abs/1706.09557`.

[CG19] M. Chion and C. Gorbman. *Audio-Vision: Sound on Screen*. Columbia University Press, 2019.

[CHC+23] Vahidi Cyrus, Han Han, Wang Changhong, Lagrange Mathieu, Fazekas György, and Lostanlen Vincent. Mesostructures: Beyond Spectrogram Loss

in Differentiable Time–Frequency Analysis. *Journal of The Audio Engineering Society*, 71:577–585, September 2023.

[Chi83] Michel Chion. *Guide des objets sonores*. Éditions Buchet/Chastel, 21 bd Jules-Ferry, Paris, 1983.

[Dav16] Stephen Davismoon. Atomisation of Sound. *Contemporary Music Review*, 35(2):263–274, 2016.

[EHGR20] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differential Digital Signal Processing. *CoRR*, abs/2001.04643, Jan 2020. `https://arxiv.org/abs/2001.04643`.

[Fla99] Patrick Flandrin. *Time-Frequency/Time-Scale Analysis*, volume 10 of *Wavelet Analysis and Its Applications*. Academic Press, 1999.

[FWW21] Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3318–3328. PMLR, 18–24 Jul 2021.

[Gab46] Dennis Gabor. Theory of Communication. Part 1: The Analysis of Information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93:429–441(12), November 1946.

[Gay09] Évelyne Gayou. The GRM: Landmarks on a Historic Route. In *Proceedings of the 2009 International Computer Music Conference*, 2009. `https://music.arts.uci.edu/dobrian/CMC2009/OS12.3.Gayou.pdf`.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[GQ20] Jean Gallier and Jocelyn Quaintance. *Differential Geometry and Lie Groups: A Computational Perspective*. Springer, Cham, Switzerland, 2020.

[HAP+18] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations, 2018. https://arxiv.org/abs/1812.02230.

[HLL23] Han Han, Vincent Lostanlen, and Mathieu Lagrange. Perceptual–Neural–Physical Sound Matching. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[HSF+24] Ben Hayes, Jordie Shier, György Fazekas, Andrew McPherson, and Charalampos Saitis. A Review of Differentiable Digital Signal Processing for Music and Speech Synthesis. *Frontiers in Signal Processing*, 3, 2024. https://www.frontiersin.org/articles/10.3389/frsip.2023.1284100.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366, 1989.

[IL20] Christopher Ick and Vincent Lostanlen. Learning a Lie Algebra from Unlabeled Data Pairs. 2020. https://arxiv.org/abs/2009.09321.

[Jan98] Augustus J. E. M. Janssen. *The Duality Condition for Weyl-Heisenberg Frames*, pages 33–84. Birkhäuser Boston, Boston, MA, 1998.

[Kan14] Brian Kane. Pierre Schaeffer, the Sound Object, and the Acousmatic Reduction. In *Sound Unseen: Acousmatic Sound in Theory and Practice*. Oxford University Press, 07 2014.

[Kle72] Felix Klein. *Vergleichende Betrachtungen über neuere geometrische Forschungen*. University of Erlangen, Erlangen, Germany, 1872.

[KPB+23] Hamza Keurti, Hsiao-Ru Pan, Michel Besserve, Benjamin F Grewe, and Bernhard Schölkopf. Homomorphism AutoEncoder – Learning Group Structured Representations from Observed Transitions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on*

Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 16190–16215. PMLR, 23–29 Jul 2023.

[LBBH98] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied To Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LC04] Patrick J. Loughlin and Leon Cohen. The Uncertainty Principle: Global, Local, or Both? *IEEE Transactions on Signal Processing*, 52:1218–1227, 2004.

[LEHR+21] Vincent Lostanlen, Chady El-Hajj, Matthieu Rossignol, Gérard Lafay, Joakim Andén, and Mathieu Lagrange. Time-Frequency Scattering Accurately Models Auditory Similarities Between Instrumental Playing Techniques. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):3, 2021. Epub 2021 Jan 11.

[Lip16] Zachary C. Lipton. The Mythos of Model Interpretability. *Commun. ACM*, 61(10):36–43, Jun 2016. presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.

[Löh17] Clara Löh. *Geometric Group Theory: An Introduction*. Universitext. Springer International Publishing, 2017.

[LYY23] Vincent Lostanlen, Lingyao Yan, and Xianyi Yang. From HEAR to GEAR: Generative Evaluation of Audio Representations. *Proceedings of Machine Learning Research*, (166):48–64, February 2023.

[Mal89] Stéphane Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[Mal09] Stéphane Mallat. Chapter 2 - The Fourier Kingdom. In *A Wavelet Tour of Signal Processing (Third Edition)*, pages 33–57. Academic Press, Boston, third edition edition, 2009.

[Mal12a] Stéphane Mallat. Group Invariant Scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[Mal12b] Stéphane Mallat. Scattering Invariant Deep Networks for Classification, 2012. Graduate Summer School 2012: Deep Learning, Feature Learning, Institute for Pure and Applied Mathematics, UCLA, July 18, 2012.

[Mal16] Stéphane Mallat. Understanding Deep Convolutional Networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, Apr 2016.

[Mar78] Karl Marx. *Economic and Philosophic Manuscripts of 1844*. International Publishers, New York, 1978.

[MFSL19] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the Universality of Invariant Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4363–4371. PMLR, 09–15 Jun 2019.

[Mor91] Thomas Mormann. Husserl's Philosophy of Science and The Semantic Approach. *Philosophy of Science*, 58(1):61–83, 1991.

[MVW+22] John Muradeli, Cyrus Vahidi, Changhong Wang, Han Han, Vincent Lostanlen, Mathieu Lagrange, and George Fazekas. Differentiable Time-Frequency Scattering on GPU. In *Digital Audio Effects Conference (DAFx)*, Proceedings of the DAFx 2022 Conference, Vienna, Jul 2022.

[OC17] Chris Olah and Shan Carter. Research Debt. *Distill*, 2017. `https://distill.pub/2017/research-debt/`.

[Pal98] Carlos Palombini. Technology and Pierre Schaeffer: Pierre Schaeffer's Arts-Relais, Walter Benjamin's Technische Reproduzierbarkeit and Martin Heidegger's Ge-stell. *Organised Sound*, 3(1):35–43, 1998.

[PGS+11] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, Nov 2011.

[Por] Portraits GRM. Portraits GRM - Music on Bandcamp. `https://portraitsgrm.bandcamp.com/music`. Accessed: 2024-09-15.

[Rav20] Siamak Ravanbakhsh. Universal Equivariant Multilayer Perceptrons. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7996–8006. PMLR, 13–18 Jul 2020.

[Rey21] Simon Reynolds. Tapeheads: The History and Legacy of Musique Concrète. *TIDAL*, 2021. `https://tidal.com/magazine/article/musique-concrete/1-78792`.

[Roa01] Curtis Roads. *Microsound*. MIT Press, Cambridge, MA, 2001.

[Roa14] Curtis Roads. Rhythmic Processes in Electronic Music. In *Proceedings of the International Computer Music Conference (ICMC)*, Athens, Greece, 2014.

[Rou23] Michael Roubach. *Phenomenology and Mathematics*. Elements in The Philosophy of Mathematics. Cambridge University Press, 2023.

[Rud87] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.

[Sch66] Pierre Schaeffer. *Traité des objets musicaux*. Éditions du Seuil, 27 rue Jacob, Paris, 1966.

[SM23] Simon Schwär and Meinard Müller. Multi-Scale Spectral Loss Revisited. *IEEE Signal Processing Letters*, PP:1–5, 01 2023.

[SS90] Xavier Serra and Julius Orion Smith. Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic Plus Stochastic Decomposition. *Computer Music Journal*, 14:12–24, 1990.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*, 2014. `http://arxiv.org/abs/1312.6199`.

[Tel05] Constantin Teleman. Lecture Notes in Representation Theory. `https://math.berkeley.edu/~teleman/math/RepThry.pdf`, 2005. Last visited on 2024/06/01.

[Ter15] Daniel Teruggi. Musique Concrète Today: Its Reach, Evolution of Concepts and Role in Musical Thought. *Organised Sound*, 20:51–59, 04 2015.

[Tie05] Richard Tieszen. Free Variation and The Intuition of Geometric Essences: Some Reflections on Phenomenology and Modern Geometry. *Philosophy and Phenomenological Research*, 70(1):153–173, 2005.

[Tud91] David Tudor. COEFFICIENT, 1991. New York Public Library. https://www.nypl.org/research/research-catalog/bib/b21006280.

[VML23] Cyrus Vahidi, Christopher Mitcheltree, and Vincent Lostanlen. Ismir 2023 tutorial. `https://www.kymat.io/ismir23-tutorial/intro.html`, 2023. Presented at the International Society for Music Information Retrieval (ISMIR) 2023.

[VNWD14] Doug Van Nort, Marcelo M. Wanderley, and Philippe Depalle. Mapping Control Structures for Sound Synthesis: Functional and Topological Perspectives. *Computer Music Journal*, 38(3):6–22, 09 2014.

[Wal17] Irène Waldspurger. Exponential Decay of Scattering Coefficients. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 143–146, 2017.

[Wei22] Maurice Weiler. Groups, Representations, and Equivariant Maps. Slides, 2022. `https://www.sci.unich.it/geodeep2022/slides/Groups_Representations_and_Equivariance.pdf`.

[Wer90] Paul Werbos. Backpropagation Through Time: What It Does and How to Do It. *Proceedings of the IEEE*, 78:1550 – 1560, 11 1990.

[Won02] M. W. Wong. *The Weyl-Heisenberg Group*, pages 90–97. Birkhäuser Basel, Basel, 2002.

[Xen92] Iannis Xenakis. *Formalized Music: Thought and Mathematics in Composition.* Pendragon Press, Stuyvesant, NY, revised edition, 1992.